Unveiling Cellular Heterogeneity, Genetic Regulation, and Protein Trafficking Dynamics Via Novel Integrative Multi-Omics Approaches

by

Rekha Mudappathi

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved November 2023 by the
Graduate Supervisory Committee:
Li Liu, Chair
Valentin Dinu
Zhifu Sun

ARIZONA STATE UNIVERSITY
December 2023

ABSTRACT

Advancements in high-throughput biotechnologies have generated large-scale multi-omics datasets encompassing diverse dimensions such as genomics, epigenomics, transcriptomics, proteomics, metabolomics, metagenomics, and phenomics. Traditionally, statistical and machine learning-based approaches utilize single-omic data sources to uncover molecular signatures, dissect complicated cellular mechanisms, and predict clinical results. However, to capture the multifaceted pathological mechanisms, integrative multi-omics analysis is needed that can provide a comprehensive picture of the disease. Here, I present three novel approaches to multi-omics integrative analysis. I introduce a single-cell integrative clustering method, which leverages multi-omics to enhance the resolution of cell subpopulations. Applied to a Cellular Indexing of Transcriptomes and Epitopes (CITE-Seq) dataset from human Acute Myeloid Lymphoma (AML) and control samples, this approach unveiled nuanced cell populations that otherwise remain elusive. I then shift the focus to a computational framework to discover transcriptional regulatory trios in which a transcription factor binds to a regulatory element harboring a genetic variant and subsequently differentially regulates the transcription level of a target gene. Applied to whole-exome, whole-genome, and transcriptome data of multiple myeloma samples, this approach discovered synergetic cis-acting and trans-acting regulatory elements associated with tumorigenesis.

The next part of this work introduces a novel methodology that leverages the transcriptome and surface protein data at the single-cell level produced by CITE-Seq to model the intracellular protein trafficking process. Applied to COVID-19 samples, this approach revealed dysregulated protein trafficking associated with the severity of the infection.

DEDICATION

I dedicate this thesis to the most important people in my life, whose unwavering support, encouragement, and sacrifices have made this journey possible.

To my loving parents, Balakrishnan Mudappathi and Sreemathy Chakkaren, your belief in me has been my greatest motivation, your countless sacrifices and your guidance have been the driving force. I dedicate this work to you with profound gratitude and love. To my dedicated husband, Rahesh Chembakan, your steadfast support and profound understanding have been the bedrock of my academic journey. I dedicate this work to you with heartfelt appreciation and boundless love. To my precious children, Roshan Chembakan and Reyaan Chembakan, your smiles and laughter have been my daily inspiration. This thesis is dedicated to you with all my love.

To my loving mother-in-law and father-in-law, Radha Vayali and Hariharan Chembakan whose unwavering support and prayers have lightened my path. To my sisters and their husbands, Dr. Reena Roopesh and Dr. Roopesh Chakkaren, Nimmi Mudappathi and Babhith Arayullathil, Haripriya Chembakan and Sreejith Sreedhar, to my niece and nephews, Neha Roopesh, Niranjan Roopesh, Swarith Arayullathil, Sidharth Sreejith, Satheerth Sreejith, and Sidharth Arayullathil, your love and support has been a constant source of comfort in my life. I am profoundly grateful for the bond we share.

To my friends especially Verah Nyarige, Aparna Manu, and all those who have supported me in ways big and small, I dedicate it to you with heartfelt appreciation.

ACKNOWLEDGEMENTS

I extend my deepest gratitude to the respected members of my committee - Dr. Li Liu, Dr. Valentin Dinu, and Dr. Zhifu Sun for their exceptional guidance and unwavering support throughout my Ph.D. journey. Dr. Liu, serving as my chair, played a pivotal role in nurturing my academic growth. Her unwavering support and mentorship laid the foundation for my achievements and embodied the true essence of impactful research. Her guidance not only ensured the success of my academic pursuits but instilled in me a profound understanding of the core principles of research and academia. Dr. Dinu and Dr. Sun, equally instrumental in my journey, provided invaluable encouragement, feedback, and guidance, which significantly contributed to my progress.

I owe a debt of gratitude to Dr. Junwen Wang for the remarkable opportunity to work in his lab during his tenure at Mayo Clinic. Under his guidance, I had the privilege to collaborate on several projects. The depth of knowledge and experience I gained under his mentorship has been immeasurable. Even after his departure, I wished for his continued guidance and support as part of my committee throughout my Ph.D. journey. I am deeply appreciative of his mentorship and the financial support provided during my time in his lab.

I'm sincerely grateful to Dr. Dongwen Wang, the Director of Biomedical Informatics PhD program, for his consistent support and advocacy. His efforts in securing essential financial aid, encompassing diverse graduate fellowships and TAship opportunities, have been invaluable throughout my academic journey. I'm indebted to my Ph.D. cohort, especially Dr. George Karway, and my lab mates Verah Nyarige, Yanxi

Chen, Jingmin Shu, Hai Chen, and Tatiana Patton, for their unwavering support and collaboration. The mentors at Mayo Clinic, Dr. Ping Yang, Dr. Stephen Ansell, Dr. Mrinal S. Patnaik, and others, have been instrumental in my professional growth, providing opportunities to work on cutting-edge research projects. A special mention of gratitude to Dr. Panwen Wang, Dr. Vaishali Bhardwaj, Dr. Alejandro Ferrer, Dr. Alanna Maguire, and Dr. Isabella Zaniletti for their exceptional mentorship during our collaborative efforts.

I'd also like to extend my gratitude to the Mayo Clinic for the invaluable affiliation that significantly contributed to my academic and research journey. The rich experiences, skills, and resources from my affiliation have profoundly contributed to my academic growth and success. I want to specifically thank Dr. Ping Yang for sponsoring this affiliation. Finally, I'm deeply grateful for the consistent support from the College of Health Solutions, especially the guidance provided by Aaron Falvey, Lauren Madjidi, and Maria Hanlin, which significantly contributed to my academic journey.

TABLE OF CONTENTS

Page

CHAPTER Page

LIST OT TABLES

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Figure Page

CHAPTER 1

INTRODUCTION

In recent years, high-throughput technologies have enabled the generation of extensive datasets spanning various omics domains, including genomics, epigenomics, transcriptomics, proteomics, metabolomics, metagenomics, and phenomics. This "multi-omics revolution" has significantly expanded our understanding of biological systems, offering comprehensive insights into various molecular layers. Advanced computational methods for integrative multi-omics data analysis are instrumental to harnessing the power of the multi-omics data that helps uncover hidden connections and elucidate complex mechanisms. However, many computational models are developed based on black box algorithms, which hinders the efficient translation of computational results to biologically meaningful discoveries. Interpretable computational models are promising approaches to address this knowledge gap.

1.1 Overview of Omics Data Layers

Genomics, epigenomics, transcriptomics, and proteomics are the fundamental pillars of multi-omics studies, each offering unique insights into the intricate world of genetic and molecular processes (Reel et al., 2021). Genomics, dedicated to deciphering the complete DNA sequence of an organism, reveals the genetic blueprint of a cell, encompassing genetic codes, structural variations, and mutations that underlie diverse biological phenomena (*Primer of Human Genetics | Wageningen University and Research Library Catalog*, n.d.).

The rapid advancement of Next-generation sequencing (NGS) techniques has brought about a paradigm shift in DNA sequencing. It now permits the concurrent analysis of numerous genes and the detection of millions of genetic variants, all accomplished with remarkable efficiency (Pervez et al., 2022). This transformative technology has broadened the horizons of genomics research, with applications such as Whole-Genome Sequencing (WGS) offering a comprehensive view of an individual's entire DNA sequence, Whole-Exome Sequencing (WES) targeting protein-coding regions, and Targeted Sequencing focusing on specific gene regions. These applications find utility across diverse fields, from cancer research to population genetics and the discovery of novel genome assemblies (Satam et al., 2023).

While genomics examines the entirety of the genome, epigenomics delves into the alterations made to the chromatic regions through processes like DNA methylation and histone modification, all without inducing changes to the underlying DNA sequence (*Epigenomics | Learn Science at Scitable*, n.d.). Advanced epigenome sequencing techniques like Chromatin Immunoprecipitation Sequencing (ChIP-seq) and Assay for Transposase-Accessible Chromatin with High-Throughput Sequencing (ATAC-seq) have empowered researchers to explore chromatin states and associated transcription factors, shedding light on their critical roles in numerous diseases (S. Ma & Zhang, 2020).

In the realm of transcriptomics, the focus shifts to RNA, encompassing aspects of transcription, expression levels, functions, cellular locations, trafficking, and degradation (Skerrett-Byrne Anthony et al., 2023). Transcriptomics has seen remarkable evolution, ushering in various technologies designed to deduce and quantify the transcriptome.

Prominent among these are transcript profiling techniques using Microarrays and direct sequencing via RNA-Seq, which have become powerful methods, providing researchers with profound insights into the intricate landscape of RNA and its role in the cell (Zhao et al., 2014) (Z. Wang et al., 2009).

Proteomics, in contrast, represents a comprehensive evaluation of proteins, encompassing their structure, functions, interactions, and dynamic cellular activities. Given the dynamic nature of protein expression, influenced by time and environmental factors, proteomics presents an inherent complexity beyond that of genomics or transcriptomics. Proteomics methodologies have advanced from conventional techniques like immunohistochemistry (IHC) staining, western blotting, and enzyme-linked immunosorbent assay (ELISA) to high-throughput approaches such as tissue microarray (TMA), protein pathway arrays, and mass spectrometry (Cui et al., 2022). These evolving methods empower researchers to investigate and understand protein-related processes in diverse biological contexts, further enriching our comprehension of complex molecular mechanisms.

Although metabolomics, lipidomics, and glycomics are typically not part of the central dogma analysis, they provide valuable insights into the intermediate products, such as metabolites, lipids, and glycans, synthesized by the proteome via biosynthetic pathways, serving as excellent indicators of a cell's activity (Barh et al., 2016).

While omics studies on bulk cell populations offer a holistic perspective of the genetic and transcriptomic landscape across diverse cell types and tissues, they must be revised when scrutinizing less understood or rare cell populations. The advent of single-cell sequencing holds promises in mitigating these challenges. Single-cell RNA

sequencing (scRNA-seq), an advanced next-generation sequencing technology, sets itself apart from conventional bulk RNA sequencing by unveiling the distinct gene expression profiles, consequently exposing cellular composition and characteristics variations. This transformative approach is particularly adept at revealing rare cell populations, often obscured by bulk RNA-seq methods, as seen in cancer tissues (G. Deng et al., 2023). In a specific study harnessing the capabilities of single-cell RNA sequencing (scRNA-seq), researchers successfully identified five unique subgroups within gastric cancer, each distinguished by their exclusive expression profiles (M. Zhang et al., 2021).

Moreover, single-cell genome sequencing enables the discovery of new germline mutations and somatic mutations in healthy and cancerous cells (Sun et al., 2022). Additionally, single-cell technologies for studying epigenomics, such as single-cell ATAC-Seq (scATAC) and single-cell ChIP-Seq, contribute to an enhanced understanding of the epigenetic intricacies at a single-cell level. This multifaceted approach to single-cell epigenomics has uncovered cell-type-specific changes in chromatin accessibility within distinct heterochromatin domains in aging mice, elucidating the ramifications of heterochromatin loss in mammalian aging (Y. Zhang et al., 2022). In a distinct study, the combination of single-cell mass cytometry (CyTOF) and single-cell proteomics facilitated an extensive examination of white adipose tissue (WAT) during both homeostasis and dietary interventions, revealing a dynamic array of macrophage subpopulations with diverse developmental origins and functional roles. This illuminates the intricate interplay between environmental cues and the malleability of resident WAT macrophages (Félix et al., 2021).

While single-cell single-omic technologies have significantly advanced our understanding of cellular processes, multi-omics sequencing techniques open a new era of integrative analysis techniques like CITE-Seq (Cellular Indexing of Transcriptomes and Epitopes by Sequencing) (Stoeckius et al., 2017) and REAP-Seq (RNA Expression and Protein Sequencing) (Peterson et al., 2017) combine RNA-Seq with protein level information, offering comprehensive insights into both gene expression and protein profiles at the single-cell level. Techniques like Genome and Transcriptome sequencing (G&T-seq) (Macaulay et al., 2015) and gDNA–mRNA sequencing (DR-seq) (Dey et al., 2015) allow the concurrent exploration of both genome and transcriptome, shedding light on the transcriptional dynamics of our genetic material. Further expanding the horizon, scM&T-seq provides epigenome-transcriptome correlation for unraveling the diverse expression patterns originating from identical DNA sequences across distinct cells. These variations, driven by factors like DNA methylation, DNA accessibility, and histone modifications, are meticulously deciphered by scM&T-seq, shedding light on the intricate regulatory mechanisms that govern cellular diversity (Angermueller et al., 2016). Finally, the assay for transposase-accessible chromatin sequencing (ATAC-seq) is a single-cell multi-omics technology that identifies genomic regions displaying open chromatin states closely linked with transcriptional activity. This technology effectively couples epigenomic information with transcriptomics (Buenrostro et al., 2013). This expanding array of multi-omics sequencing technologies equips researchers with an extensive layer of data, enabling them to conduct integrative analyses for exploring orchestration of cellular processes in unprecedented depth and detail.

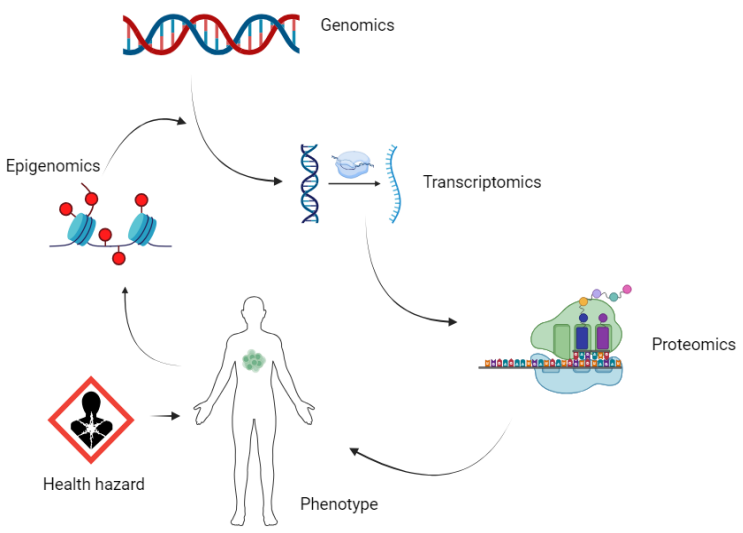

Figure 1.1: Layers of multi-omics integration

1.2 Integrative Analysis of Multi-Omics Data

The complexity of biological systems demands a comprehensive approach that combines different omes rather than just confining to single data types. Multi-omics integration has emerged as a crucial paradigm in data analysis in both bulk and single-cell data analysis. Single-level omics approaches often need more resolving power to establish transparent causal relationships between specific molecular alterations and the resulting phenotypic manifestations. Analyzing multi-omics data enhances the categorization of samples into biologically relevant groups, deepens our comprehension of prognostic and predictive traits, reveals cellular responses to treatments, and aids translational research through integrative models. For example, in a study conducted by Pradeep et al., the integration of transcriptomic and proteomic data unveiled molecular signatures associated with Alzheimer's disease (Kodam et al., 2023).

A deep learning-based model leveraging RNA sequencing, miRNA sequencing, and methylation data from The Cancer Genome Atlas (TCGA) effectively identifies robust survival subgroups of hepatocellular carcinoma (HCC) for improving HCC prognosis prediction (Chaudhary et al., 2018). Recent advances in single-cell multi-omics analysis, exemplified by innovative tools like Seurat's multimodal clustering, have demonstrated the power of integrating multiple omics layers to improve the resolution and accuracy of clustering biological samples, enabling the discovery of subtle yet biologically significant subpopulations.

1.2.1 Challenges in Multi-Omics Integration

The integration of multi-omics data, despite the expanding availability of omic datasets and analytical tools, continues to be challenging. Several factors contribute to this complexity. Firstly, the design of multi-omics studies can be intricate, as it requires careful consideration of which data types to combine, appropriate analytical methodologies, and the harmonization of data from diverse sources. Secondly, noise in the data, stemming from technical variability and experimental conditions, poses a significant hurdle. The challenge lies in distinguishing accurate biological signals from noise to extract meaningful insights. Furthermore, data interoperability is a crucial aspect of multi-omics integration. Different data types generated using various platforms and technologies may not readily align or be directly comparable. Data preprocessing and transformation are often necessary to ensure compatibility, which can be complex and time-consuming. Just as the curse of dimensionality presents a central challenge in single-omic studies, its impact is notably exacerbated in the realm of multi-omics

research. In multi-omics, the sheer quantity of measured variables significantly escalates, necessitating advanced dimensionality reduction techniques to navigate this complexity effectively (Wörheide et al., 2021). Lastly, there is the issue of varying definitions and scopes of what constitutes a multi-omics study. The field is evolving rapidly, and different researchers may have distinct criteria for what they consider multi-omics, further adding to the challenges of harmonizing and comparing studies in this space.

1.2.2 Approaches to Multi-Omics Integration

Integration methodologies are crucial in understanding the intricate relationships between various omic datasets. In recent years, a diverse array of multi-omics integration methods has emerged, employing various mathematical, statistical, and computational techniques. These strategies involve merging individual omic datasets, either sequentially or simultaneously, to unveil the complex interactions within the biological system (I. Subramanian et al., 2020) (Yan et al., 2018).

Sequential integration strategies adopt a step-wise approach, where omic datasets are initially analyzed individually or in specific combinations, followed by the integration of their findings in subsequent stages. This sequential integration method allows for comprehensive data analysis, even in cases where omic measurements for the same samples are unavailable. It is particularly suitable for bulk datasets, as demonstrated by a previous study that successfully integrated ATAC-seq and RNA-seq data to identify critical genes and regulatory pathways associated with the neuroprotection of S-adenosylmethionine (SAM) against perioperative neurocognitive disorder (PND) (Xu et al., 2023).

In this study, initial analysis focused on differential gene expression using RNA-seq, and later, chromatin accessibility and transcription factor binding information from ATAC-seq were incorporated to uncover epigenetic regulatory defects contributing to the aberrant expression of identified differential genes.

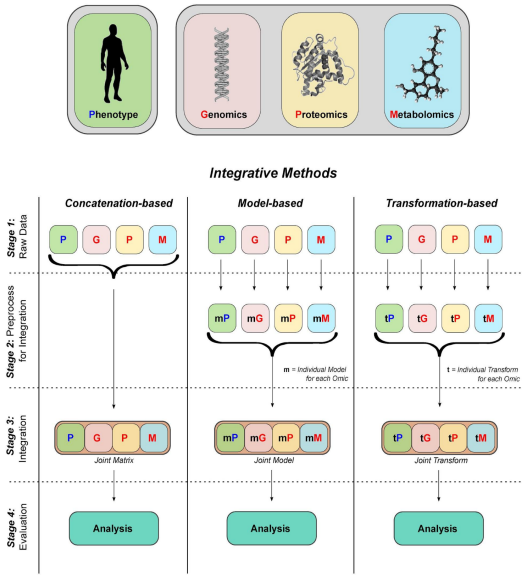

Figure 1. 2: Approaches for multi-omics integration (Reel et al., 2021)

In contrast to sequential integration, simultaneous integration involves the parallel integration of multiple omic layers, allowing the analysis of all available omic data at once within a single modeling step. While this approach requires that data originate from the same biological samples or individuals, it is a potent method for deriving valuable insights into cellular functions by capturing the interplay between different omic datasets. When appropriately applied, this versatile approach can provide a holistic understanding of complex biological processes. These integration methods can be broadly categorized

into concatenation, model-based, and transformation-based approaches (**Fig 1.2**) (Reel et al., 2021).

Concatenation-based integration methods involve the direct merging of data matrices from individual omics domains, such as genomics, proteomics, and metabolomics, creating a comprehensive multi-omics dataset without prior data preprocessing. This integrated multi-omics data is then harnessed as input for various machine-learning techniques for predictive modeling. For example, in a study focusing on ovarian cancer patients, multiple genomic data types (mRNA, DNA methylation, DNA copy-number alteration, and microRNA) from The Cancer Genome Atlas (TCGA) were harmoniously integrated using a multivariate Cox Lasso model. This integration led to the discovery of a robust prognostic signature for predicting progression-free survival (PFS) in these patients, underscoring the substantial potential of multi-omics integration in clinical prognosis for this challenging disease (Mankoo et al., 2011). Concatenation-based integration can also incorporate unsupervised methods, as demonstrated in a study on muscle-invasive bladder cancer (MIBC). Researchers employed iClusterBayes, a fully Bayesian latent variable method, as an example of unsupervised concatenation-based integration. This approach was pivotal in uncovering intrinsic MIBC subtypes and identifying biomarkers with significant prognostic value, ultimately enhancing patient stratification for frontline therapeutic strategies (Mo et al., 2020).

Model-based integration methods involve a multi-stage process where individual models are first developed for different omics data types, and then these models are

combined into a joint model, allowing for the integration of heterogeneous data sources from different patient groups with the same disease information.

Model-based integration methods can further be categorized as either supervised or unsupervised, depending on the nature of the analysis and the availability of labeled phenotypic or clinical information. As an illustrative example, MOSAE (Multi-omics Supervised Autoencoder) employs autoencoders designed for each omics data type to generate data-specific representations. These individual representations are subsequently merged into a consolidated representation, which is harnessed for predictive modeling. This method was applied to the TCGA Pan-Cancer dataset, effectively predicting four distinct clinical outcome endpoints (Tan et al., 2020).

Finally, transformation-based integration methods begin by converting individual omics datasets into graphs or kernel matrices, which are then combined to create a unified representation capturing the relationships between diverse omics data. The versatility of these methods is a significant advantage, allowing them to integrate various omics types effectively, making them valuable for comprehensive multi-omics analysis. Many supervised learning methods in the transformation-based category rely on kernel and graph-based algorithms. For instance, the fMKL-DR (fast multiple kernel learning for dimensionality reduction) method, as developed by Giang et al., employs kernel matrix transformation and a support vector machine (SVM) classifier to stratify samples effectively (Giang et al., 2020). The fMKL-DR method demonstrated significant success in classifying multiple cancer types, such as lung cancer, GBM, breast cancer, OV cancer, liver cancer, and kidney cancer. Additionally, its application to Alzheimer's

disease (AD) patients showed the potential to stratify patients based on different disease phases, promising early diagnosis and effective treatment monitoring, especially in the later stages of AD. Weighted-nearest neighbor (WNN) method is an example for transformation-based unsupervised algorithms that effectively integrate multiple data types within single cells. Modality weights based on the relative importance of each ome are learned and combined to construct a WNN graph, thus providing a single representation of multimodal datasets while preserving the richness of both data types (Hao et al., 2021). Applying WNN analysis to a CITE-Seq dataset of cord blood mononuclear cells, which involves the joint integration of RNA and protein modalities, effectively segregated CD4+ T cells from CD8+ T cells, while separate analyses of each modality failed to achieve this level of distinction.

1.3 Integrating Prior Knowledge

Knowledge-driven multi-omics integration leverages existing knowledge or prior biological insights to enhance the analysis of multi-omics datasets for gaining a deeper understanding of complex biological systems. In single-cell clustering analysis, dimensionality reduction is essential due to high dimensionality and noise. Traditional methods like Principal Component Analysis (PCA) and non-linear techniques like t-SNE and UMAP offer different approaches, but each has limitations in terms of complexity and interpretability. In single-cell data analysis, dimensionality reduction is essential due to high dimensionality and noise. Traditional methods like Principal Component Analysis (PCA) and non-linear techniques like t-SNE and UMAP offer different approaches, but each has limitations in terms of complexity and interpretability. A supervised neural

network algorithm incorporating prior biological knowledge from the protein-protein interaction (PPI) network for reducing dimensionality in single-cell RNA-seq data is being proposed (Gundogdu et al., 2022).

Correlation-based network construction, which relies on pairwise correlations and multiple testing corrections, faces notable challenges when dealing with larger sample sizes. As the sample size increases, weaker correlations become statistically significant, leading to denser networks. Additionally, the choice of multiple testing methods and significance levels can result in substantially different networks, even though they all meet statistical criteria, potentially obscuring the identification of biologically relevant relationships. A new approach to correlation-based network inference is introduced, shifting the focus from statistical cutoffs to the selection of a correlation threshold that maximizes agreement with a predefined biological reference (Benedetti et al., 2020). This method aims to identify networks with the highest concordance with known biological information, eliminating the need for arbitrary p-value cutoffs and incorporating prior knowledge as a guiding principle. The approach is shown to be applicable to metabolomics and transcriptomics data, even when only partial prior knowledge is accessible. While single-omic data analysis has seen notable strides in integrating prior biological knowledge to enhance the accuracy and interpretability of results, the application of a similar approach in the multi-omics domain remains an area with much untapped potential. While the benefits of incorporating prior knowledge are widely recognized, it is intriguing that this methodology has not been as extensively explored in the context of multi-omics data analysis.

1.4 Novel Multi-Omics Integration Methods

While multi-omics integration holds immense promise, it also presents unique challenges, including data harmonization, computational complexity, and the need for advanced statistical methodologies. To harness the full potential of these vast multi-omics datasets, it is imperative to develop innovative tools and methodologies that can effectively extract valuable insights, paving the way for new discoveries and advancements in the field.

In this research endeavor, I introduce three innovative approaches to multi-omics integrative analysis, each geared towards unraveling intricate biological processes and shedding light on unexplored frontiers of knowledge. First, I present a novel single-cell integrative clustering method, INCLOSE which harnesses the power of multi-omics data to significantly enhance the resolution of cell subpopulations. The second chapter covers in detail principles and methods of the INCLOSE algorithm with its practical application and comparison to well established single cell clustering algorithm, Seurat. When applied to a CITE-Seq dataset encompassing human Acute Myeloid Lymphoma (AML) and control samples, the method unveils previously hidden nuances within cell populations, offering a fresh perspective on this challenging disease. The method proves to be a good enhancement for use with Seurat clustering for improving the resolution of clustering.

In the third chapter, the focus shifts to a sophisticated computational framework designed to unearth transcriptional regulatory trios. A novel regression model is being designed and implemented aimed at exploring how transcription factors, regulatory elements, and genetic variants work in harmony to modulate gene expression, providing

essential insights into the dynamics of cellular regulation. These trios consist of a transcription factor binding to a regulatory element, housing a genetic variant, and subsequently exerting differential regulatory control over the transcription level of a target gene. Employing whole-exome, whole-genome, and transcriptome data from multiple myeloma samples, this innovative approach unveils the potential cis-acting and trans-acting regulatory elements associated with tumorigenesis, providing essential insights into the molecular mechanisms driving this disease.

Finally, in the fourth chapter, I introduce an innovative CITE-Traffick algorithm leveraging the transcriptome and surface protein data at the single-cell level, obtained through cellular indexing of transcriptomes and epitopes via sequencing assays (CITE-Seq). This approach dissects the intracellular trafficking of proteins, considering both cell type and disease context. The application of CITE-Traffick to a CITE-Seq dataset focused on COVID-19, reveals dysregulated protein trafficking pathways linked to the severity of COVID-19 infection. The markers discovered from this study consisting of both previously identified and novel, are being compared to benchmark markers and evaluated for performance, which provide better performance. These innovative approaches collectively mark a significant step forward in multi-omics research, offering new tools and perspectives to advance our understanding of complex biological phenomena.

CHAPTER 2

INTEGRATIVE CLUSTERING OF SINGLE CELLS: INCLOSE

2.1 Background

Single-cell sequencing, particularly single-cell RNA-sequencing (scRNA-seq), has transformed our understanding of cellular diversity by enabling the clustering of cells based on their individual gene expression profiles. ScRNA-seq technology provides a high-resolution examination of transcriptomes at the single-cell level, allowing the detection of transcriptional variations within cell populations. scRNA-seq has been instrumental in unraveling cell-type heterogeneity, providing valuable insights into functional distinctions within seemingly homogeneous cell populations. As demonstrated in (Mahata et al., 2014), scRNA-seq provided a high-resolution transcriptomic analysis of T helper 2 (Th2) cells, uncovering extensive heterogeneity and identifying the differential upregulation of Cyp11a1, highlighting the cells' capacity for de novo steroid synthesis.

Similarly, exploring gene co-expression patterns at the single-cell level has proven valuable in uncovering co-regulated gene modules and regulatory networks that distinguish cell types or samples (Haque et al., 2017). For instance, a study utilized a set of six co-expressed genes to categorize glioblastoma patients into distinct groups with significantly different survival outcomes, even though these genes were selected without prior knowledge of cancer biology, emphasizing the potential of single-cell gene co-expression analysis in patient stratification and the discovery of biologically relevant insights (J. Wang et al., 2016).

2.1.1 Single-Cell RNA-Seq Clustering Methods

With the rapid progress in sequencing techniques, diverse computational approaches based on data clustering have emerged to interpret and understand single-cell RNA-seq data. Two main popular techniques employed for clustering are K-means clustering and hierarchical clustering.

K-means clustering seeks to discover a set number of cluster centers, referred to as centroids. Its goal is to minimize the collective sum of squared Euclidean distances between data points and their corresponding centroids. This scalability with the number of data points makes it an efficient choice for handling large datasets. Multiple clustering tools, such as SAIC (L. Yang et al., 2017), and RaceID (Grün et al., 2015), employ K-means-based approaches to interpret single-cell RNA-seq data, aiming to identify distinctive gene subsets or rare cell types within the datasets. RaceID utilizes a K-means clustering method, employing a similarity matrix based on Pearson's correlation coefficients. It leverages the gap statistic to ascertain the ideal number of clusters, ultimately improving the separation and consistency of the clusters. SAIC, or Single-cell Analysis via Iterative Clustering, employs an iterative K-means clustering approach to systematically optimize a parameter space using the Davies-Bouldin index to select the most relevant signature genes for a given number of clusters and significance threshold.

K-means clustering, while widely used, has notable drawbacks when applied to single-cell clustering. This greedy algorithm may fail to find the global optimum and can be sensitive to outliers, making it less effective in identifying rare cell types. It also heavily depends on the predefined number of clusters, which can influence the clustering results significantly. Moreover, as recent advances in scRNA-seq technologies have led to increasingly large datasets containing thousands to millions of cells, K-means can become computationally intensive and slow due to its requirement to load the entire dataset into memory, potentially limiting its applicability for large-scale analyses. To mitigate these issues, researchers often use replicates with random starting points to improve the chances of finding a global minimum solution.

Hierarchical clustering employs agglomerative and divisive strategies for clustering by merging cells into clusters based on distance measures and recursively splitting clusters. This offers flexibility to identify rare cell types without requiring predetermined cluster numbers as in K-means. CIDR, introduced by (P. Lin et al., 2017), incorporates both dimension reduction and hierarchical clustering into single-cell RNA-seq analysis, utilizing an implicit imputation process to mitigate dropout effects and achieving a stable estimation of pairwise cell distances. pcaReduce is an agglomerative clustering method that combines PCA and hierarchical clustering, aiming to link the reduced principal component representations to the number of discernible cell types. The approach capitalizes on the expectation that broad cell type information is found in low-dimensional PCs, whereas more detailed cell type structures are represented in higher-dimensional PCs (žurauskienė & Yau, 2016).

However, classic hierarchical clustering algorithms are only suitable for small datasets due to their high computational complexity. Fast and memory-efficient scRNA-seq K-means clustering algorithms (Baker et al., 2021) and hierarchical clustering that efficiently handles large-scale single-cell data by constructing dendrograms based on shared nearest neighbor (SNN) graphs (Zou et al., 2021) are also recently developed.

Graph-based clustering in data analysis views a dataset's similarities as a weighted graph, treating its data points as nodes and their relationships as edges, their weights reflecting similarity (S. Zhang et al., 2023).This method identifies clusters, often referred to as communities, by pinpointing groups of highly connected nodes. In the intricate domain of single-cell RNA sequencing (scRNA-seq) analysis, several approaches like Spectral clustering, Louvain, and Leiden serve as common community detection algorithms (Z. Liu & Barahona, 2020). Spectral clustering applies eigenvalues to perform dimensionality reduction of the dataset before clustering, while the Secuer algorithm (Wei et al., 2022) stands out for its efficiency, leveraging an anchor-based graph construction and a refined similarity calculation between cells and anchors. Conversely, Louvain and Leiden techniques focus on breaking down graphs into connected subgroups. These methods are employed in popular analysis tools like Seurat (Satija et al., 2015) and SCANPY (Wolf et al., 2018).

Density-based techniques such as DBSCAN (Ester et al., n.d.) pinpoint dense regions within data space, separated by less dense regions, contributing a different perspective to the clustering process. Finally, the advent of deep learning has significantly influenced clustering approaches. For instance, scDeepCluster (Tian et al., 2019) is an example of a single-cell clustering method employing deep learning to learn feature representations and cluster cells based on this acquired insight, offering a novel approach to identify cellular groupings in scRNA-seq data.

2.1.2 Integrative Clustering of Single Cells

In recent years, various single-cell multi-omics sequencing technologies have evolved providing complementary data to explore cellular heterogeneity, identify novel cell types, and uncover intricate regulatory networks. This includes single-cell RNA-seq (scRNA-seq) provides gene expression profile, Single cell Assay for Transposase-Accessible Chromatin (scATAC-seq) provides chromatic accessibility data, Cytometry by Time-Of-Flight (CyTOF) employs mass spectrometry to measure multiple protein markers, and Cellular Indexing of Transcriptomes and Epitopes by sequencing (CITE-seq), allows simultaneous profiling of gene expression and surface protein abundance. The integration of these diverse omics data has further enriched the landscape of single-cell clustering methodologies.

Researchers have shown that manual curation of cell clusters inferred from CyTOF and scRNA-seq data enabled the identification of disease-specific immune cell subtypes and their functional states (Luo et al., 2022) (Kashima et al., 2021). Computational methods are also available that leverage these datasets to identify cell

populations. The commonly used Seurat package offers two ways to integrate single-cell multi-omics data for cell clustering. The first approach involves clustering the individual omics data separately and then integrating the resulting clusters (Stuart et al., 2019). The second approach integrates the multi-omics data while measuring cell similarities and then derives clusters based on these similarities (Hao et al., 2021). MAESTRO (C. Wang et al., 2020) , an open-source tool tailored for the integrative analysis of scRNA-seq and scATAC-seq datasets, employs Seurat for scRNA-seq clustering and either scABC (Zamanighomi et al., 2018) or latent semantic indexing (Cusanovich et al., 2018) for scATAC-seq clustering. It offers a specialized function for calculating gene regulatory potential scores to model gene activity, which is subsequently integrated with the scRNA gene expression profile for enhanced precision in cell type identification.

2.1.3 Discovering Context-Specific Cell Populations

Despite advancements in single-cell clustering and cell identification, identifying biologically meaningful subpopulations within complex samples remains challenging. It is well known that specific cell populations appear or proliferate in response to external stimuli or in specific disease conditions. For instance, in the context of trauma-induced heterotopic ossification (HO), research has elucidated how these cells react to stimuli, leading to the formation of abnormal bony growths and persistent chronic pain. Significantly, the conventional treatment approach, which involves surgical excision, often falls short in mitigating the enduring consequences of ectopic bone formation and frequently leads to recurrence (Agarwal et al., 2017). In another study involving bortezomib drug, it was found that extended exposure of CD4+ T cells to this drug led to

the emergence of a regulatory T-cell population capable of significantly suppressing the proliferation of effector T cells, reducing IFN-γ production, and downregulating CD40L expression in activated effector T cells (Blanco et al., 2009).

Conventional clustering techniques that examine gene markers aggregated over all samples may fail to detect these specialized populations as the broad cellular landscape may obscure the nuanced subpopulations. In such cases, the underrepresented or condition-specific cell populations will be merged with large or partially similar clusters. Studies show that condition-specific cell population expansion is distinguished by the upregulation of specific protein markers, and the widely accepted method for identifying these cell types involves employing a sequential gating strategy to assess protein markers. As an illustrative example, in a study focused on Waldenstrom macroglobulinemia (WM), the research team identified the expansion of a specific subset of myeloid-derived suppressor cells (MDSCs) distinguished by the expression of the CD66b marker, denoting them as CD66b+ MDSCs, within WM patients. Through advanced transcriptomic analysis, particularly using CITE-seq, the study unveiled an inflammatory immunosuppressive gene expression signature associated with this specific MDSC subset (Bhardwaj et al., 2022).

In a separate study, Waldenstrom macroglobulinemia (WM) patients requiring treatment displayed increased counts of a unique myeloid-derived suppressor cell (MDSC) population characterized by the expression of CD163 and CD138 markers, signifying a substantial expansion of this specific MDSC subtype potentially linked to WM progression.

Interestingly, traditional markers associated with monocyte-like (m-MDSC) and granulocyte-like (g-MDSC) MDSCs, like CD14 and CD15, were also prominently expressed within this cell population. This observation emphasizes the need for comprehensive phenotyping to distinguish MDSC subtypes (Jalali et al., 2019).

While traditional clustering techniques have made significant strides in cell classification, they occasionally struggle to identify condition-specific cell populations accurately, especially in the presence of contextual features such as various diseases, treatments, and diverse ethnic groups. These factors can trigger the emergence of specific cell populations, resulting in the amalgamation of these subtler cell groups within larger or partially similar clusters. Recent methodologies often overlook the critical consideration of these contextual features, leading to challenges in accurately capturing these distinct cell populations.

The SEAcells algorithm is a novel approach devised to address the limitations of traditional clustering techniques in identifying groups of cells in unique cell states called metacells. This method showcases superior performance in discerning metacells from diverse data types like RNA and ATAC, across datasets encompassing discrete cell types and continuous trajectories thus helping to reveal distinct cell states linked to specific diseases. CellSIUS is another recent method used for uncovering rare cell populations within scRNA-seq data.

This approach involves the use of cluster-specific candidate marker genes showcasing bimodal expression patterns within pre-clustered data. By leveraging a graph-based clustering algorithm, CellSIUS identifies correlated gene sets and then segregates cells into subgroups based on the collective expression of these identified gene sets.

Current multi-omics integration clustering techniques like SEAcells and specialized tools like CellSIUS have notably improved cell subpopulation identification. However, there's a distinct absence of clustering techniques that effectively integrate contextual information into multi-omics analyses. A crucial gap remains in incorporating cell labels that reflect their clinical sample sources in the clustering process to discover condition-specific cell subpopulations. This void in integration methodologies raises an intriguing hypothesis: By incorporating clinical features into the cell clustering process, we anticipate a substantial enhancement in the accuracy of cellular clustering. This enhancement could provide a more precise and nuanced identification of condition-specific cell subpopulations, thus addressing the current limitations in our cell clustering methodologies.

Previously, Maneck et. al introduced a semi-supervised clustering algorithm to discover coexpressed gene sets using two transcriptome data sets, one with cell perturbation labeled and one without (Maneck et al., 2011). In this algorithm, gene-gene similarities computed using the data set without labels are adjusted by differential expression patterns extracted from the data set with labels.

This helps segregate coexpressed genes with varying expression levels between different pathways. Enlightened by this approach, we have designed a novel single-cell clustering method called the Integrative Clustering Of Single Cells (INCLOSE) that integrates single-cell multi-omics sequencing data to discover condition-specific cell clusters.

Through the integration of single-cell transcriptomic data with proteomic or chromatin information, in conjunction with critical contextual factors, including exposure to perturbations or variations in ethnicity, INCLOSE enables the fine-tuning of cell clustering, resulting in the identification of more precise and biologically relevant subpopulations. INCLOSE is versatile, capable of integrating single-omic or multi-omics data at the single cell level with sample metadata that may contain a single variable, such as a disease group, or include additional covariates.

2.2 INCLOSE Algorithm

In the INCLOSE clustering method, multiple feature matrices representing different omic profiles and clinical data are utilized to compute cell-cell similarities. These similarities are determined using a correlation-based approach, with tuning parameters balancing information from various feature matrices. Clusters are identified by iteratively adding cells with high within-cluster similarity. By combining omics and clinical data, INCLOSE unveils condition-specific clusters, identifying new cell populations in a comprehensive and integrated manner.

In a schematic illustration (**Fig. 2.1**), we show that cells clustered on the omic profiles may not reflect their relationships in the clinical feature space (**Fig. 2.1.a**). By combining the omics and clinical data, INCLOSE reveals a condition-specific cluster (**Fig. 2.1.b**). This is achieved via fusing cell-cell similarity matrices calculated from individual omes into an overall similarity matrix, which helps detect the top-most densest modules of clusters that potentially represent newly emerged cell populations.

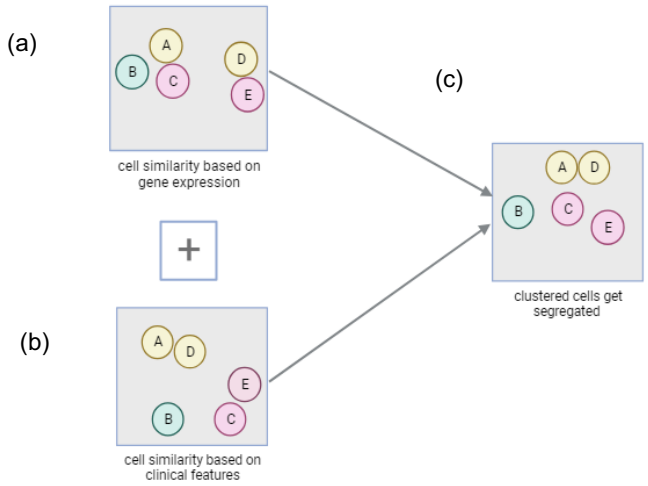

Figure 2. 1: Illustration of cell clusters using conventional clustering methods and after application of INCLOSE method. Cells with similar clinical features are colored same. (a) Clustering of cells based on gene expression similarity. Cells A, B, and C are clustered together, forming one group, while cells D and E are grouped together as another cluster. (b) Clustering of cells based on clinical features similarity. Cells A, B, and D, E clusters together since they have similar clinical feature**,** while cells D and E are placed in a separate cluster. (c) In the INCLOSE method, cells are reorganized based on a fused similarity matrix which integrates cell similarity information from both (gene expression as shown in (a)) and clinical features (as shown in (b)). As a result of this integration, cells C and B move apart**,** and cells A and D comes together to form a cluster.

2.2.1 Definitions and Notations

Let $\{T_1, T_2, \ldots, T_n\}$ be a set of feature matrices, each representing a specific omic profile of S single cells. For example, T_1 is a gene expression matrix obtained from scRNA-seq data, T_2 is a surface protein abundance matrix measured by Antibody Derived Tags (ADT), and T_3 is a matrix of phenotype features (e.g., disease group, treatment group, clinical characteristics, etc.). In these feature matrices, rows correspond to cells and columns correspond to features.

2.2.2 Estimation of Cell-Cell Similarities

For each feature matrix T_i, we compute a similarity matrix A_i containing pairwise cell similarities as:

$$A_i(g,h) = max(r(T_i^g, T_i^h),\ 0) \qquad [2.1]$$

where T_i^g and T_i^h are feature vectors of cells g and h, respectively; $d(T_i^g, T_i^h)$ is the distance calculated as $1 - max(Y(T_i^g, T_i^h), 0)$ where $Y(.,.)$ is the correlation coefficient (Spearman or Pearson) of two feature vectors; σ and ω_i are the tuning parameters applied to the distance matrix of a given T_i for balancing the information from different feature matrices.

INCLOSE supports Spearman and Pearson correlation - Spearman correlation is appropriate for studying nonlinear associations and Pearson correlation is suitable for capturing linear relationships (Hou et al., 2022).

We then fuse the correlation matrices in a cell-cell similarity matrix as:

$$W = \prod_{i=1}^{n} f(A_i) \qquad [2.3]$$

where $\prod_{i=1}^{n}$. is element-wise multiplication and $f(.)$ is a Gaussian smoothing function,

$$f(A_i) = exp\left(\frac{-\omega_i A_i{}^2}{2\sigma^2}\right)$$

The smoothing parameter σ is shared across all data modalities to reduce the noise caused by minor variations in correlation coefficients.

The weight ω_i is assigned to each data modality to balance the contribution to the combined similarity. W holds high values for pairs of cells that are similar across multiple data modalities. In **Fig. 2.2.A**, we present a schematic representation illustrating the process by which the combined similarity matrix is calculated.

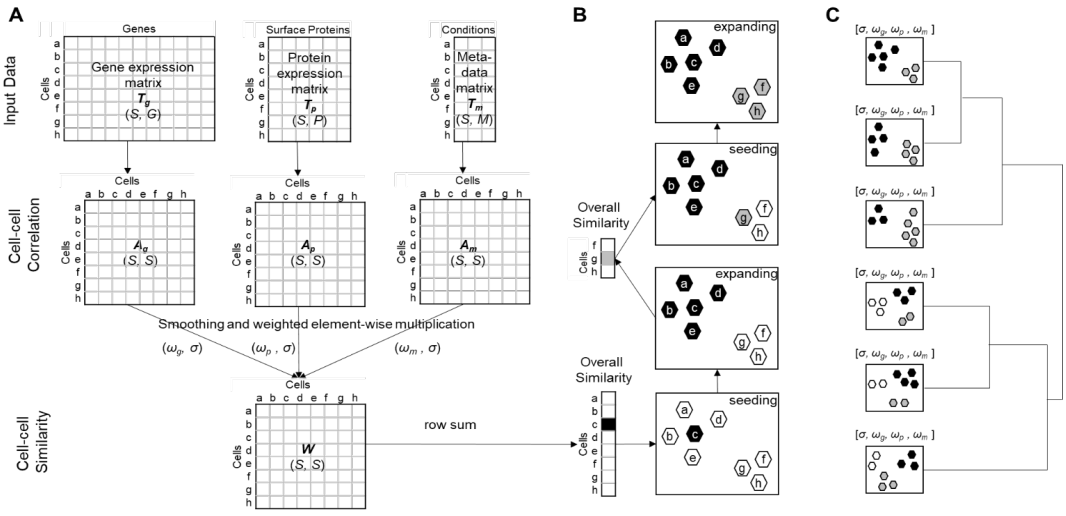

Figure 2. 2: Schematic illustration of the INCLOSE algorithm. (**A**) *S* cells are characterized by expression of *G* genes and *P* surface proteins and *M* meta features. Pairwise correlation between cells is calculated based on each data modality, which are subsequently combined into a cell-cell similarity matrix tuned by weights (ω) and smoothing parameter (σ). (**B**) The clustering process involves a series of seeding and expanding steps. At each iteration, the cell with the highest overall similarity serves as the seed. (**C**) Various combinations of tuning parameter values are tested. The clustering results are organized into a hierarchical tree based on J-score that quantifies alignments between two clustering results.

INCLOSE can accommodate metadata containing a single feature, such as the sample groups each cell belongs to. In this case, the corresponding correlation matrix A_i is a $S \times S$ square matrix where values of cells in the same group are set to 1 and values of cells in different groups are set to 0. Next, we smooth these values using a logistic sigmoid function,

$$f(A_i) = \frac{-\omega_i}{(1 + exp(-\sigma * (A_i - 0.5)))} \quad [2.4]$$

The overall similarity between a cell and all the other cells is measured as

$$H(g) = \sum_{j=1}^{S} W_{g,j} \qquad [2.5]$$

High H values indicate cells that share similar profiles with many other cells.

2.2.3 Identification of Clusters

The clustering process is like the one proposed by Maneck, et al. The clustering process starts by identifying the cell g_0 that has the highest overall similarity. Using g_0 as the seed, we grow this cluster C by progressively adding a cell g_k that maximizes the within-cluster similarity.

$$\gamma(g_0, g_1, \ldots, g_k) = \frac{\sum_{i,j \le k} W_{gi,gj}}{|C|+1} \qquad [2.6]$$

where $|C|$ is the number of cells in the cluster. The iteration is terminated if the within-cluster similarity stops increasing. Among the remaining cells, we repeat the steps of selecting the seed cell and growing the cluster until the user-specified maximum number of clusters is reached. Finally, we combine clusters containing fewer than 10 cells with the cells that have not been assigned to any clusters into a "trash cluster". Users must specify the maximum number of clusters the cells can be partitioned into. INCLOSE will produce K non-trash clusters and 1 trash cluster under the constraint the $K + 1$ does not exceed the user-specified upper limit.

2.2.4 Parameter Tuning and Optimization

INCLOSE uses the tuning parameter ω_i to balance the influence of feature matrix T_i on the clustering results. As the ω_i increases from 0 to 1, the influence of the feature matrix T_i increases. The sum of ω_i values is normalized to 1. The parameter σ, which governs the extent of smoothing, impacts the size of resultant clusters. As the σ value increases, the size of the clusters increases, and the number of clusters decreases. σ is constrained within the range [0,1].

INCLOSE optimizes the σ and ω_i values by performing a grid search with a series of ω_i values in the range of 0 and 1 and σ values in the range of 0.03 and 0.3. The goal is to find a combination of σ and ω values that maximizes the within-cluster similarities and between-cluster distances. Because clinical and phenotypic features have already contributed to the clustering steps, they are not used again in the parameter tuning step. As such, the parameters are tuned to fit the molecular profiles.

Given a specific combination of σ and ω values, INCLOSE groups the cells into a set of disjoint clusters excluding the trash cluster. Using an omic profile T_i, the within-cluster similarity metric ϕ_i quantifies the average similarity between pairs of cells in the same cluster,

$$\phi_i = 1/|C| \sum_k \sum_{g,h \epsilon C_k} max(r(T_i^g, T_j^h), 0) \qquad [2.7]$$

where g and h are two cells in the k^{th} cluster C_k, $r(.,.)$ is Pearson correlation coefficient, and $|C|$ is the number of cells not in the "trash cluster".

The overall within-cluster similarity across multiple omes is a weighted sum given by,

$$\phi = \sum_i \quad w_i \phi_i \qquad [2.8]$$

where w_i is the re-standardized weight computed from ω_i to ensure that $\Sigma w_i = 1$ after excluding the metadata modality.

To calculate between-cluster distances, we first find the centroid of each cluster. For cluster C_k, the centroid O_i^k based on an omic profile T_i is a vector in which each element is the mean value of a feature over all cells in this cluster,

$$O_{i,j}^k = \frac{1}{|C_k|} \sum_{g \in C_k} \quad T_{i,j}^g \qquad [2.9]$$

where $T_{i,j}^g$ is the abundance of j^{th} feature in cell g, i.e., value in row g and column j in matrix T_i, and $|C_k|$ is the number of cells in cluster C_k, i.e., the cluster size. The distance between two clusters C_u and C_v is

$$d_{u,v} = 1 - r(O_i^u, O_i^u) \qquad [2.10]$$

and $r(.,.)$ is the Pearson correlation coefficient. We then derive the mean between-cluster distance over all clusters as

$$\psi_i = \frac{1}{M-1} \sum_{v \neq u} \quad \sum_u \quad d_{u,v} \qquad [2.11]$$

where M is the number of non-trash clusters.

The overall between-cluster distance across multiple omes is a weighted sum given by,

$$\psi = \sum_i \quad w_i \psi_i \qquad [2.12]$$

The within-cluster similarity and between-cluster distance are combined to produce a segregation score,

$$Z = \phi + \psi \qquad [2.13]$$

We calculate Z for clustering results using each unique combination of σ and ω values. A high Z score indicates a good clustering result, in which cells within the same cluster are highly similar and cells in different clusters are highly dissimilar. Users can use the Z score to guide the selection of the final clustering results. **Fig. 2.3** illustrates schematically how within-cluster similarities and between-cluster distance are calculated for a given pair of clusters.

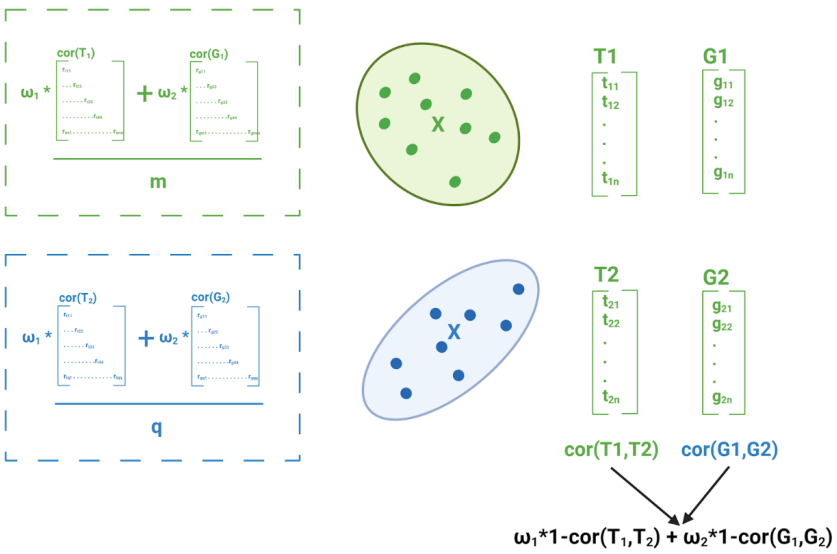

Figure 2. 3: Illustration of calculation of within cluster similarity and between cluster similarity for two clusters within a clustering set.

2.2.5 Identifying Optimal Clustering

Using different weights and smoothing parameters leads to different clustering results. The optimal combination of these tuning parameters was not known *in priori*. To address this issue, INCLOSE iterates through a series of parameter values and performs clustering analysis for each possible combination. The segregation Z score is a heuristic measure of the quality of clustering results. In practice, the best Z score does not necessarily indicate the most biologically meaningful clustering result. Furthermore, while some parameter combinations produce very different clustering patterns, other combinations may cause little changes in the clustering patterns. We, therefore, provide functions to visualize and align the large number of clustering results produced by INCLOSE.

To measure the similarity between two clustering results, we calculate the J-score (Ahmadinejad et al., 2023) that aligns clusters from two clustering based on cells that are consistently grouped together (i.e., mutual presence). A J-score of 1 indicates that the two clustering results are identical. A low J-score indicates that the two clustering results are highly discordant. After computing J-scores for each pair of clustering results, we perform hierarchical clustering analysis to group these results into a tree structure. A clade in this tree consists of a collection of clustering results that are similar to each other. Users can choose a node and display a UMAP graph to visualize the clustering result.

2.2.6 Relationship to the guided clustering algorithm

Guided clustering proposed by Maneck. et. al and INCLOSE clustering offer unsupervised clustering approaches for single-cell data analysis. Guided clustering strategy merges experimental and clinical high-throughput data of potentially distinct genomic types. It incorporates prior biological knowledge from experimental studies such as cell perturbation experiments to guide the gene clustering process. This helps in identifying gene sets that not only stand out in experimental data but also exhibit coherent expression patterns in clinical data. Guided clustering can accommodate various genomic platforms and provides the flexibility to adjust the balance between guiding and clinical data explicitly. INCLOSE is designed for the integration of single-cell multi-omics sequencing data and clinical data, enabling the fine-tuning of cell clustering. INCLOSE combines feature matrices representing various omics profiles to compute cell-cell similarities and subsequently identifies clusters.

While both methods involve data integration, Guided Clustering is tailored for diagnostic signature construction based on a biological focus and can predict pathway activation. In contrast, INCLOSE incorporates cell label information or other clinical data aiming to identify condition-specific cell clusters in single-cell data. INCLOSE excels in its adaptability to incorporate multiple omics layers as feature matrices, making it a versatile choice for studies involving diverse high-throughput data sources. Both methods share a common procedural framework, initiating the clustering process with selecting a seed element, followed by iterative additions of elements or genes to expand the cluster while maximizing within-cluster similarity.

In guided clustering, the user specifies the parameter σ, while ω is chosen automatically, and for each σ, clustering runs with various ω values. The best cluster is then selected based on the ω value that maximizes the sum of within-cluster strength, measured by average pairwise correlation and average gene activation. In contrast, INCLOSE allows users to specify both σ and ω values, and for each combination of these parameters, clusters are identified. The best clustering is determined based on the average pairwise similarities of all feature matrices provided, considering both within-cluster strength and between-cluster distance. Furthermore, INCLOSE offers the flexibility to choose from a range of clusters, providing more options than guided clustering.

2.3 Application of INCLOSE to CITE-Seq

2.3.1 Acute Myeloid Leukemia CITE-Seq Dataset

In this section, I discuss the application of the INCLOSE algorithm to a single-cell dataset. The clustering of samples was performed using Seurat and INCLOSE. INCLOSE clustering was compared to well-established clustering, Seurat elucidating the differences and unique contributions of INCLOSE clustering. The INCLOSE could successfully reveal novel clusters that were merged into larger clusters by Seurat enhancing the capability to understand the complexity of cellular subtypes which otherwise have been oversimplified by traditional clustering methods.

For this study, I utilized publicly available CITE-Seq data derived from an Acute Myeloid Leukemia (AML) study, accessible through the Gene Expression Omnibus (GEO) database, under the accession ID GSE220473. This study employed a CITE-seq panel of 131 oligo-tagged antibodies and sequenced eleven bone marrow samples, encompassing three age-matched normal donors and eight newly diagnosed acute myeloid leukemia (AML) patients. The primary focus of this study was to investigate the therapeutic potential of antibody therapies targeting the AML-specific cell surface marker U5 snRNP200. There was a total of 20987 cells from the AML group and an equivalent count of 20740 cells from the control group. This initial matching of cell numbers ensured a balanced starting point for the analysis of both groups.

A literature review on Myeloid-Derived Suppressor Cells (MDSCs) illuminated the critical role of this cell type in various cancer types including AML. The review underscored the complex nature of MDSCs, emphasizing its classification as a highly heterogeneous population of immature myeloid cells. Importantly, this heterogeneity is highly disease-specific, with distinct subsets of MDSCs identified in different types of cancers and diseases (Bhardwaj & Ansell, 2023). Studies have demonstrated the expansion of MDSCs in the tumor microenvironment. Their expansion has been strongly associated with immune suppression and the progression of diseases such as AML (Lv et al., 2019) (Hyun et al., 2020). This fundamental understanding of MDSCs and their pivotal role in disease progression, particularly in the context of AML, has established a robust foundation for the subsequent direction of this study. The primary objective, therefore, is the comprehensive exploration of MDSC heterogeneity to discern and delineate AML-specific MDSC subclusters through the application of our novel clustering algorithm, INCLOSE. By employing INCLOSE, we can capitalize on its unique ability to incorporate cell label information thus refining the clustering process tailored to identifying AML-specific or control-specific clusters.

The initial clustering and exploration of the dataset revealed a notable expansion of MDSC cells in the bone marrow of AML samples. Seurat clustering of MDSC cells uncovered AML-specific subclusters, shedding light on the heterogeneity of these immune suppressive cells. Considering the inherent heterogeneity of MDSC subtypes, it is hypothesized that more condition-specific subtypes could exist. To comprehensively explore and unveil additional subtypes within these cell groups, we employed the

INCLOSE algorithm. Notably, while there was some overlap with Seurat's clustering patterns, INCLOSE introduced distinctive clustering patterns. The subsequent comparative marker and enrichment analysis yielded intriguing findings, revealing that the clusters identified by INCLOSE featured unique immune-related markers, potentially holding considerable significance in AML research. By leveraging the J-score function, we could pinpoint subtypes of AML-specific MDSCs that were previously merged or dispersed within the Seurat clustering results, thus enhancing our understanding of MDSC heterogeneity in the context of AML.

2.3.1.1 Clustering by Seurat

The raw CITE-seq count matrices were loaded into R (v4.0.3) and processed using the Seurat R package (v4.1.2). Cells with less than 100 detected genes and genes detected in fewer than 5 cells were filtered out. Cells with mitochondrial gene expression greater than 5% of the total gene expression were also removed. A Seurat object was constructed for both the scRNA and protein data, and the two objects were integrated using the Seurat integration pipeline. The RNA expression levels were normalized using standard normalization to correct for batch effects and the top 2000 highly variable genes were identified for downstream analysis. The protein expression levels were normalized using centered log ratio normalization and scaling. Dimensional reduction using principal component analysis (PCA) was performed on the integrated scRNA and ADT data separately to compute 30 principal components (PC).

Clustering was performed on the integrated scRNA and ADT assays using the Seurat Weighted Nearest Neighbors (WNN) pipeline. At a resolution of 0.8 Seurat clustering identified 10 distinct clusters. Clusters with less than 5 cells were removed.

To identify marker genes for each cluster, we employed the FindAllMarkersMAESTRO function from the MAESTRO package in R. After identifying marker genes for each cluster, we annotated the clusters using RNAAnnotateCelltype function from the MAESTRO package based on canonical marker genes for immune cell types provided by Azimuth. The distinct cell types identified were - CD14+Monocyte, HematopoieticStemandProgenitorcell, Erythroidcell, CD4CentralMemoryT, CD4NaiveT, NaturalKiller, CD8EffectorMemoryT, NaiveBcell, CD16+Monocyte, and Plasmablast (**Fig. 2.4**).

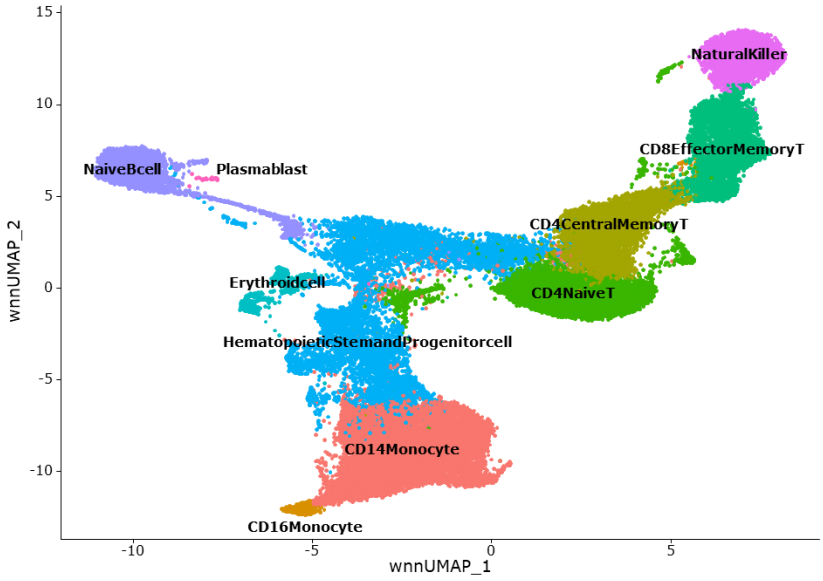

Figure 2.4: Seurat clustering of pooled AML and control dataset

After the clustering analysis, we leveraged the identified CD14+ and CD16+ monocyte populations to identify Myeloid Cells (MCs). This was achieved by employing a specific gating strategy that relied on the expression patterns of surface protein markers including CD11b+, CD33+, and HLA-DR low surface protein markers (Hyun et al., 2020) (**Fig. 2.5**). The AML samples exhibited a notable increase in the MC cell count, with a total of 4,234 MC cells, as compared to the control group, which had 2,543 MC cells. Statistical analysis revealed a t-statistic of 2.47 (p-value < 0.05) for the comparison of MDSC and non-MDSC proportions in samples between the AML and control groups. These results strongly indicate a significant expansion of MDSC populations within the AML samples, emphasizing the potential relevance of MDSCs in the context of AML.

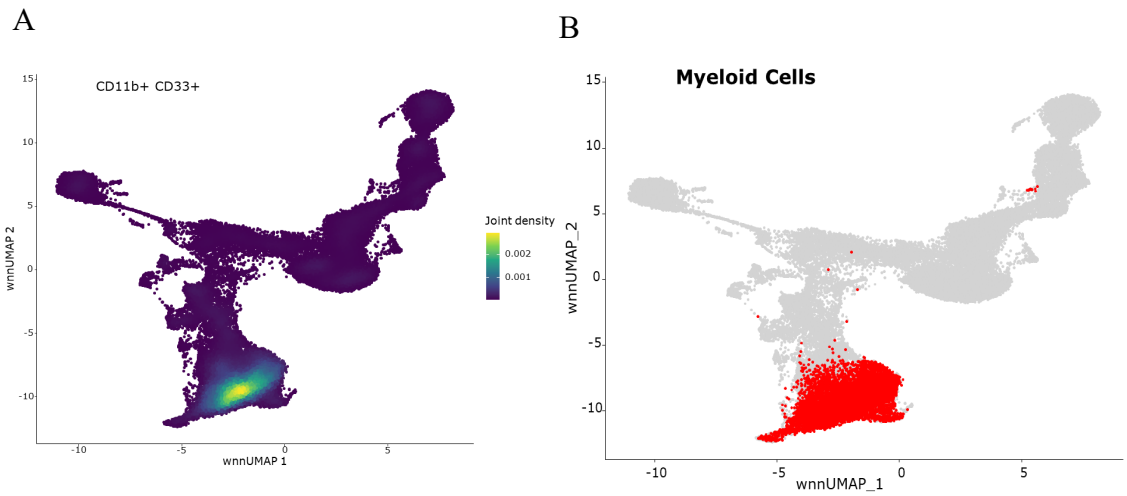

Figure 2.5: Gating of myeloid cells from monocytes. (A) Density plot of CD11b and CD33. (B) UMAP plot highlights the myeloid cells that will be further clustered by INCLOSE to discover subpopulations.

Subsequently, Seurat clustering was applied with a resolution parameter set to 0.15, revealing the presence of six distinct subclusters within the MC population. One of these subclusters emerged exclusively from the AML samples, signifying its unique presence within the context of (AML) (**Fig 2.6**).

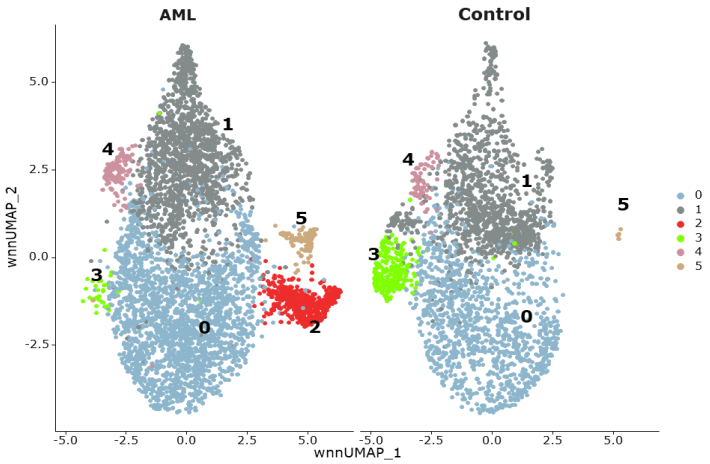

Figure 2.6: Seurat sub-clustering of MC. Six clusters identified by Seurat is displayed on the UMAP plots based on the top 10

2.3.1.2 Clustering by INCLOSE

The INCLOSE algorithm was applied to the MC cell population with the goal of identifying one or more sub-clusters of AML-specific MC cells. We applied the FindVariableFeatures Seurat function to both RNA and ADT assay to discern the top 2000 highly variable genes and the top 15 highly variable surface proteins within the MC cells of both AML and control samples. The resultant gene expression matrix surface protein abundance matrix of highly variable features and cell labels representing the phenotype of the samples were used as input feature matrices for INCLOSE. The cells were labeled 1 if from AML and 0 if from control.

We explored a series of tuning parameters including σ ranging from 1/3 to 0.1/3, weights assigned to the RNA and ADT feature matrix (ω_g and ω_p, respectively) ranging from 0.1 to 0.5, and weights assigned to the cell label matrix (ω_l) fixed at 0.03. A total of 177 unique combinations of these tuning parameters were explored, producing 177 clustering results. The segregation scores (Z scores) of these clustering results ranged from 0.0008 to 0.935 (**Fig. 2.7.A**). The 45 clustering results with segregation score >0.9 were distributed across four clades in the dendrogram (**Fig. 2.7.B**). Within each of these four clades, the clustering results exhibited a high degree of similarity (**Fig. 2.8 I-XII**), with the number of clusters ranging from six to eight, and J-scores ranging from 0.779 to 0.999. Notably, all these clustering results revealed condition-specific clusters.

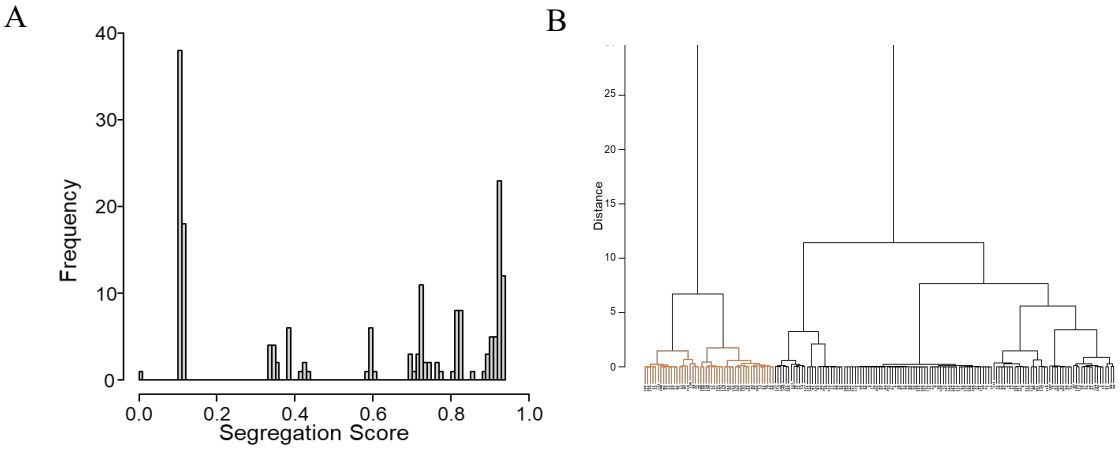

Figure 2.7: INCLOSE analysis of the MC data. (A) Dendrogram shows relationship between different clustering results. Orange color indicates clustering results with segregation score >0.9. (B) Histogram shows distribution of segregation scores of 177 clustering results using various combinations of tuning parameter values.

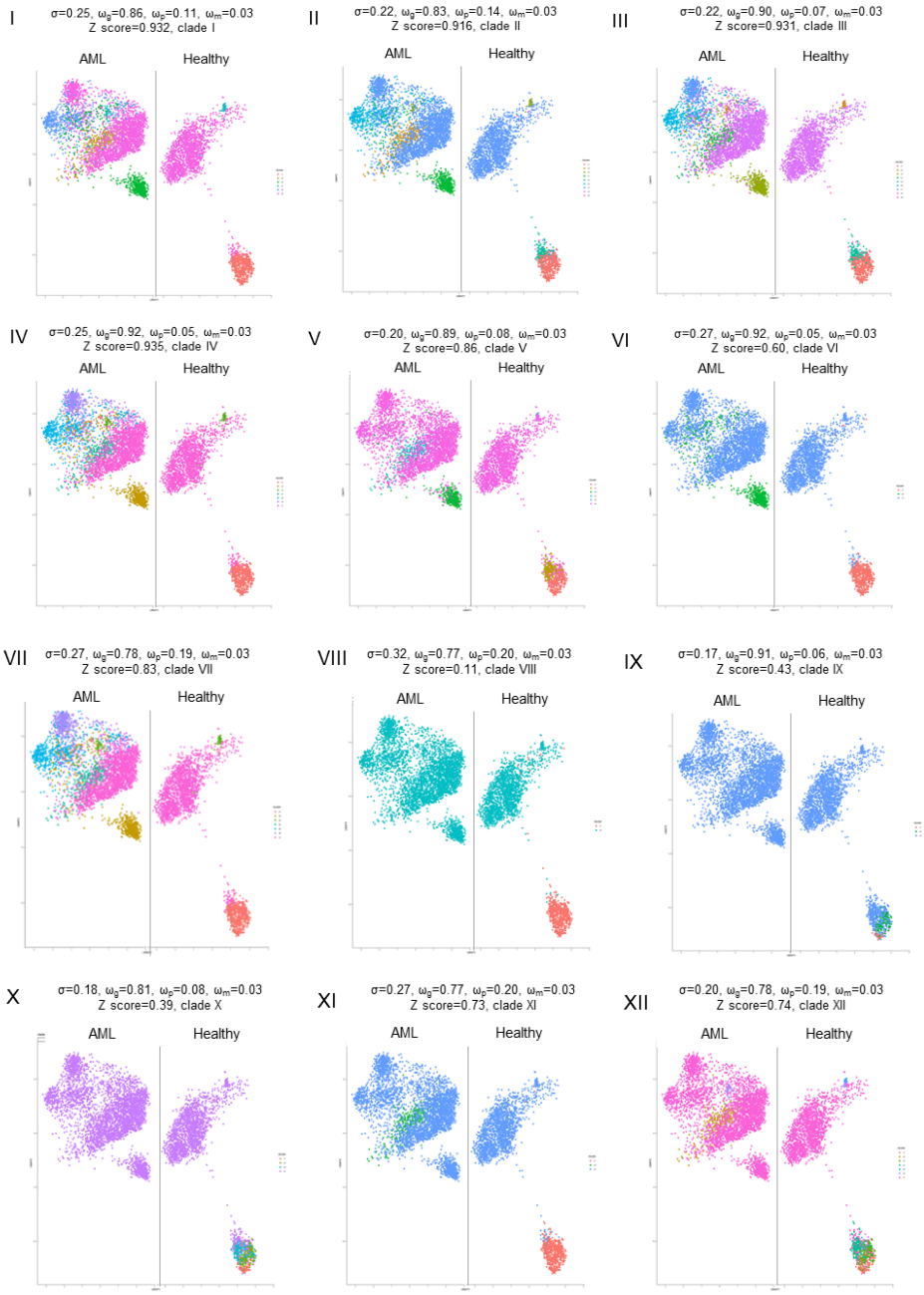

Figure 2. 8: INCLOSE clustering clades. In each clade, the clustering result with the highest segregation score was plotted in UMAP. Values of the tuning parameters are displayed, including smoothing parameter (σ) and weights for gene expression (ω_g), surface protein expression (ω_p), and metadata (ω_m). Different colors represent different cell clusters.

We closely examined one of the clustering results with parameters σ = 0.23 and wt = 0.11 where eight clusters were formed (**Fig. 2.9.A**). Some of these clusters consisted exclusively of cells from the healthy samples or the AML samples, while others were a mixture of both (**Fig. 2.9.B**). Notably, clusters 2, 4, 5, and 7 were exclusively present in the AML population, while clusters 1 and 6 were specific to the control samples.

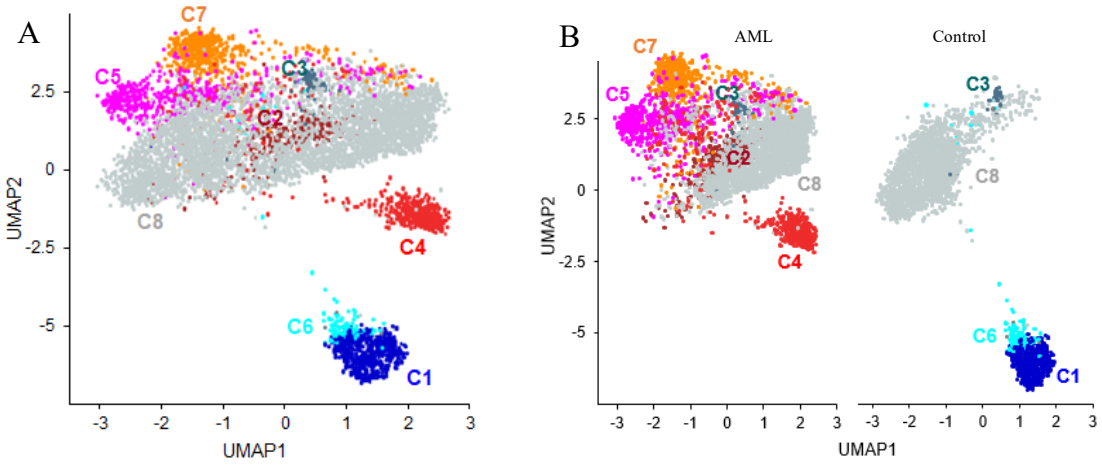

Figure 2.9: INCLOSE clustering of MDSC cells with tuning parameters σ=0.23, ω_p = 0.11 and ω_l = 0.3 revealed 9 total subclusters with a total weighted sum of 0.92. Out of the 9 clusters, 4 were unique in AML and 2 were unique in control.

2.3.1.3 Cluster Marker Analysis

After applying the INCLOSE clustering method, INCLOSE cluster numbers were mapped to cells within the Seurat object. This allowed us to perform subsequent differential expression gene enrichment analysis. Wilcoxon rank sum test, as implemented in the Seurat R package was used for identifying cluster-specific markers.

To account for multiple testing, we adjusted the p-values at a threshold of 0.05 using the Bonferroni method. Initially, we identified cluster-specific markers by comparing each cluster against the aggregate of all other clusters. This approach unveiled a discernible pattern of significant markers (RNA and ADT) unique to each cluster (**Fig 2.10**) highlighting the heterogeneity across the clusters.

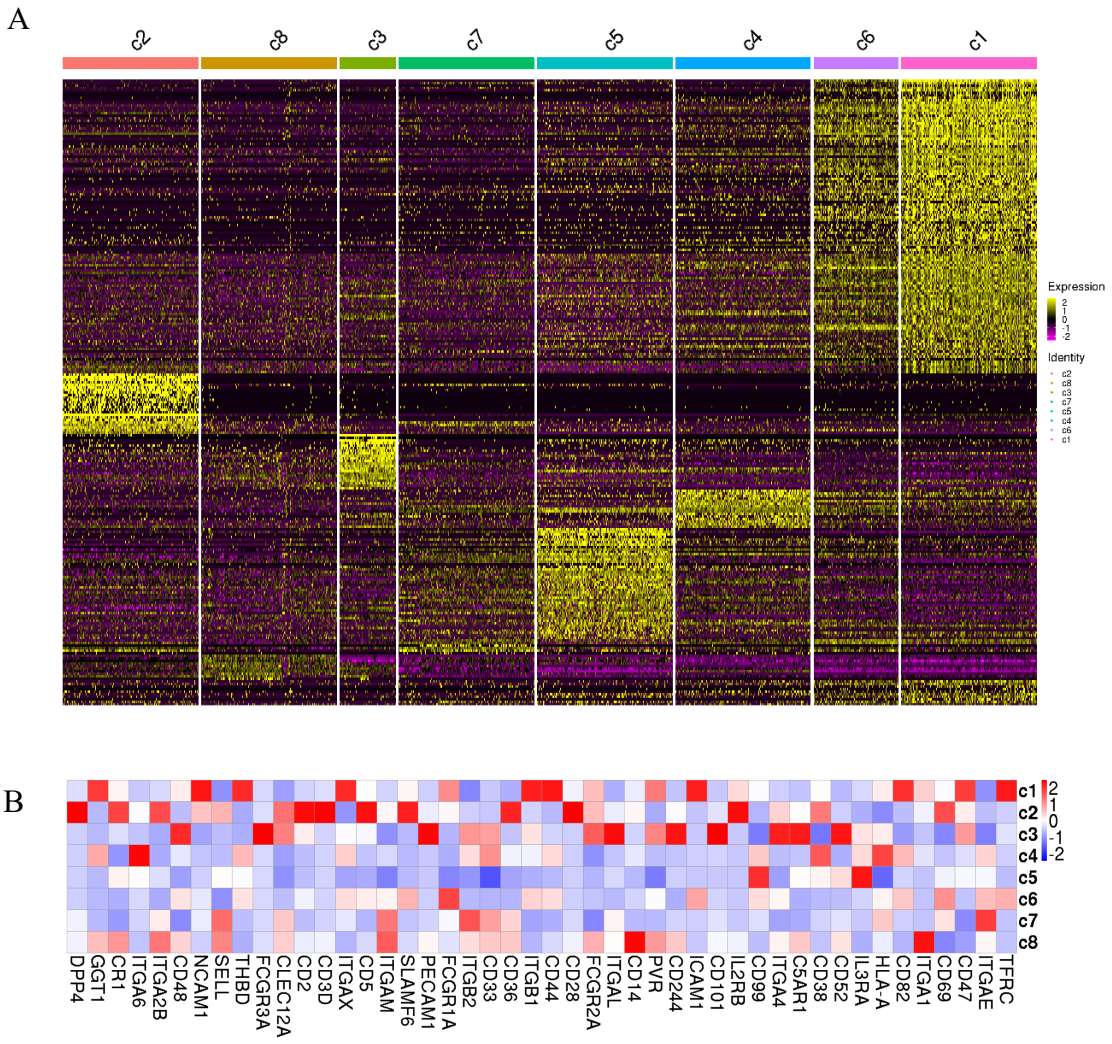

Figure 2. 10: INCLOSE Cluster Markers Heatmap. (A) RNA markers (B) ADT markers

Distinct clustering patterns between AML and control observed in our analysis are indicative of the underlying heterogeneity and differences in cell subtypes between the AML and control groups. This is because similar cell subtypes often display distinctive molecular profiles across conditions, reflecting the influence of different physiological states or pathological conditions. To elucidate the differences between similar cell types in AML and control groups, we identified markers specific to each condition's clusters, facilitating the identification of distinct MDSC subtypes within each sample set. Interestingly, marker analysis within AML and control revealed a distinct marker profile both at mRNA and ADT level (**Figures 2.11** and **2.12**).

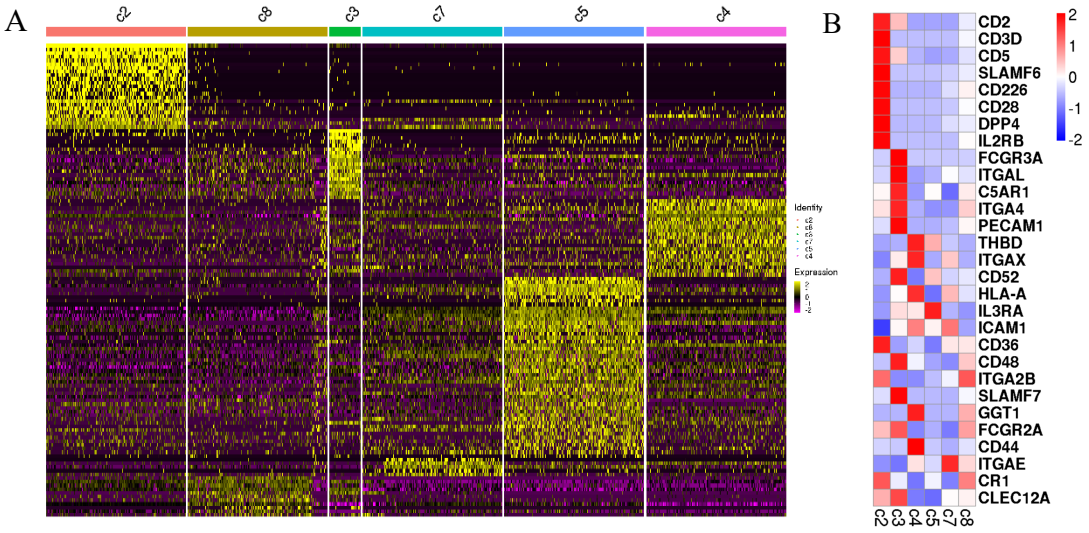

Figure 2. 11: AML Cluster Markers Heatmap. (A) RNA markers (B) ADT markers

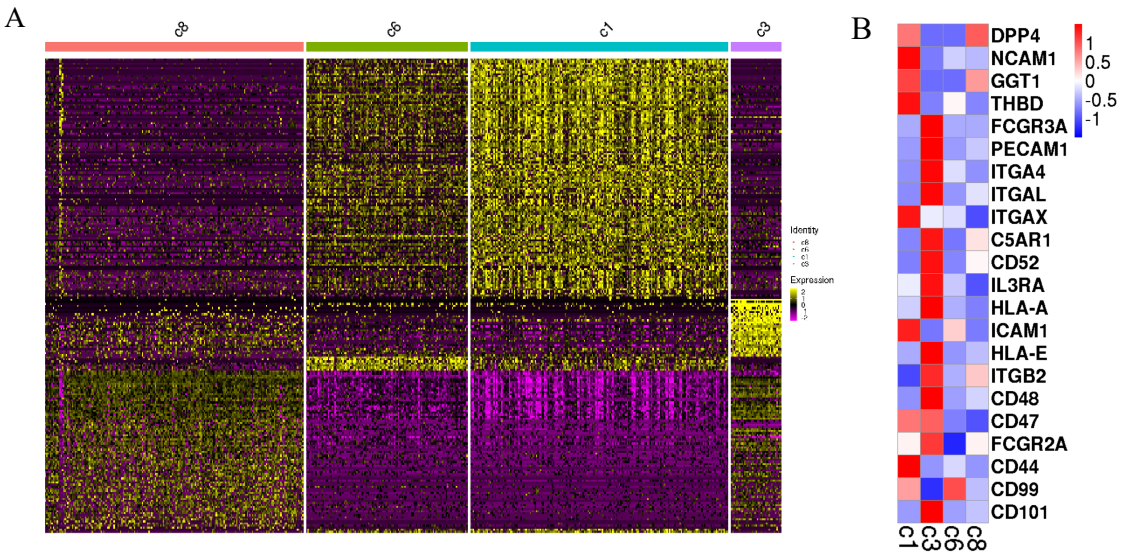

Figure 2.12: Control Cluster Markers Heatmap. (A) RNA markers (B) ADT

Furthermore, for clusters with a mix of cell types, we executed differential expression analysis between AML and healthy states, pooling cells from both groups within the same cluster. This approach provided insights into condition-specific variations across these mixed cell populations.

2.3.1.4 Identifying Cell Subtypes in AML and Control Samples

Analysis of RNA and protein markers within AML and control subclusters unveiled distinct expression patterns across all clusters for both modalities (**Fig 2.13**). Based on the markers defined by Gabrilovich (Gabrilovich, 2017), Myeloid-Derived Suppressor Cells (MDSCs), a subset of immature myeloid cells with immunosuppressive functions, can be categorized into two primary groups: Polymorphonuclear Myeloid-Derived Suppressor Cells (PMN-MDSCs) and Monocytic Myeloid-Derived Suppressor Cells (M-MDSCs).

PMN-MDSCs are characterized by markers CD11b+CD14−CD15+ or CD11b+CD14−CD66b+, while M-MDSCs exhibit markers CD11b+CD14+HLA-DR−/loCD15−. A third, minor population of MDSCs has also been identified, the early-stage MDSCs (e-MDSCs), which express neither CD15 nor CD14 (Bizymi et al., 2019).

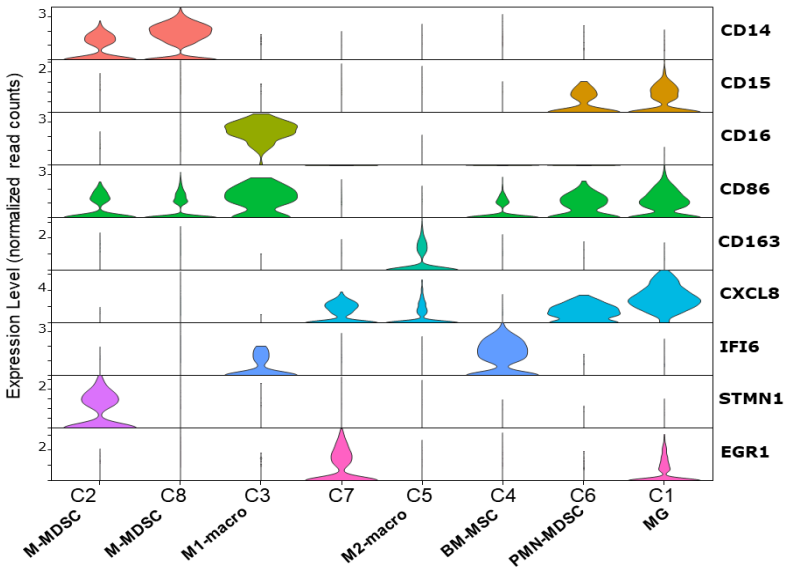

Figure 2.13: Distinct cluster markers in INCLOSE clustering.

Clusters 6 within the control group exhibit a CD14- CD15+ expression profile, aligning with the features of PMN-MDSC. While cluster 1 in the control also exhibited CD14- CD15+ profile. On differential gene expression analysis between clusters 1 and 6 revealed many inflammatory signature markers like CXCL8, CXCL2, and CCL2 in cluster 1 indicating the presence of mature granulocytes (**Fig 2.14.A**). Cluster 8 present in both AML and control subset exhibit an expression profile of CD14+ CD15-, a characteristic associated with Monocytic Myeloid-Derived Suppressor Cells (M-MDSC) subset within the AML and control subset.

Cluster 2 which is an AML only cluster exhibits a similar expression profile as cluster 8. DEG analysis (**Fig 2.14.B**) between clusters 2 and 8 revealed cell proliferation markers such as STMN1 (Vicari et al., 2022). The identification of M-MDSC and PMN-MDSC cell types highlights the robustness of INCLOSE clustering in delineating two major classifications of MDSCs.

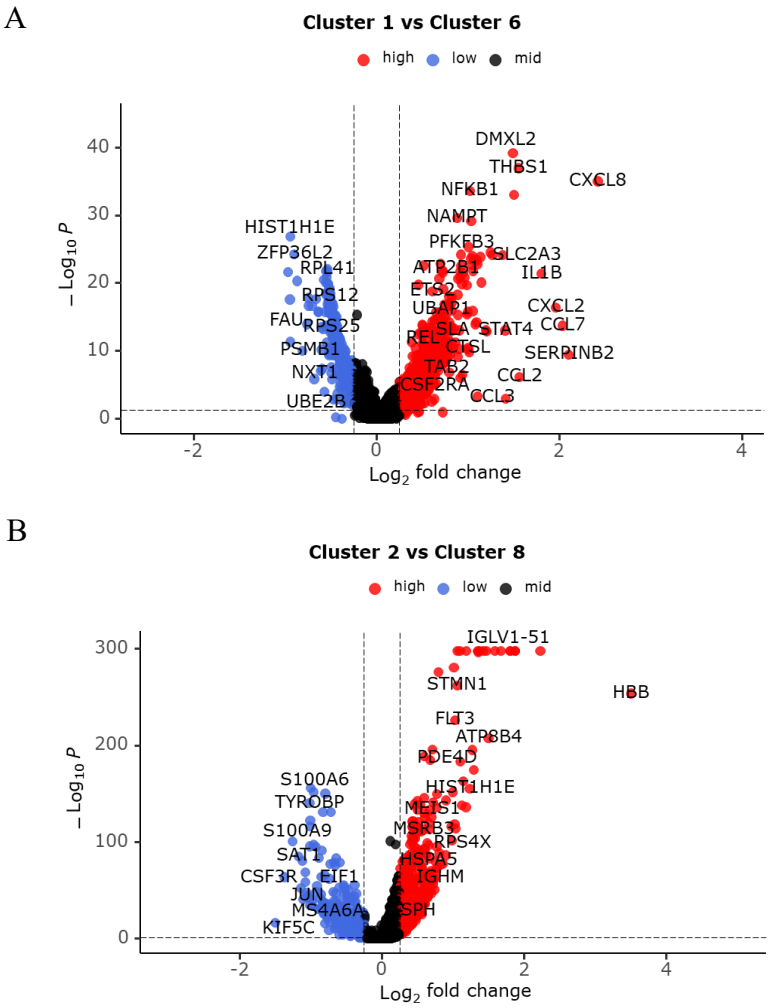

Figure 2.14: Volcano plots of Differentially Expressed Genes between mixed clusters. (A) Cluster 1 vs Cluster 6 (B) Cluster 2 vs Cluster 8

Cluster 3 which is a mixed cluster is positive for CD16 (*FCGR3A*) and *CD86* indicating M1-Macrophages (*Two Types of Macrophages*, n.d.) (Smithy & Luke, 2023). Cluster 5 which is an AML only cluster is positive for *CD68* and *CD163* (J. M. Hu et al., 2017) (Tremble et al., 2020) (Xie et al., 2022) revealing presence of M2-Macrophages in AML. Macrophages, as a heterogeneous group of myeloid cells, can be broadly categorized into two main types: M1-like, or classically activated macrophages, and M2-like, or alternatively activated macrophages.

While M1 macrophages are pro-inflammatory and thus participate in the positive immune response, function as an immune monitor, and act as tumoricidal, while M2-macrophages are anti-inflammatory and contribute to tumor progression and immune suppression (Ostrand-Rosenberg et al., 2012). In the tumor microenvironment, the close interaction with MDSCs significantly shapes macrophage phenotype, culminating in the emergence of M2-like macrophages, commonly known as "tumor-associated macrophages" (TAMs). These TAMs play a pivotal role in immune suppression and simultaneously contribute to the promotion of tumor progression, as documented in previous research (Veglia et al., 2021) (Mantovani et al., 2009).

Furthermore, within cluster 5, we identified upregulation of *MRC1* (CD206) signifying CD163+CD206+ M2-like macrophages in AML. A study examining the infiltration rate of CD163+CD206+ M2-like macrophages in the BM of AML patients and healthy volunteers demonstrated a significant increase in the frequency of CD163+CD206+ M2-like macrophages in the BM of AML patients compared to the healthy control group, confirming the expansion of these cells in the AML microenvironment (Al-Matary et al., 2016). The successful identification of TAMs within our clustering results represents a significant achievement, highlighting its ability to discern their presence.

Cluster 4 which is an AML only cluster is separated well from the rest of the cells. Their over-expression of marker *IFI6* and surface marker ITGA6 indicates presence of bone marrow mesenchymal stromal cells (BM-MSC) (C. Pan et al., 2023) (Nieto-Nicolau et al., 2020) (van Megen et al., 2019). These cells also express higher levels of IFNGR2, suggesting the activation of IFN-γ receptor. In a prior investigation, researchers investigated the implications of interferon-gamma (IFN-γ) in the context of mesenchymal stromal cells (MSCs) within AM (Goedhart et al., 2018). This study revealed how IFN-γ influences the intrinsic immunosuppressive functions of MSCs, particularly within the specific AML microenvironment. The discovery of clusters expressing IFNGR2 in AML samples is a compelling indicator of an immune-suppressive microenvironment. This finding amplifies the promise of the INCLOSE clustering method in identifying immune-suppressive MSCs within AML, offering valuable prospects for understanding and addressing this aspect of the disease.

Cluster 7, which is also an AML only cluster having an expression profile of CD14- CD15low is an unknown population of cells which we could not classify within the scope of already known markers of myeloid cells. Even though cluster 7 displays characteristics associated with PMN-MDSCs, yet they demonstrate distinct gene expression and ADT patterns, suggesting a specialized population of cells that must have emerged in AML and it needs to be further validated experimentally.

2.3.1.5 Differential Expression Analysis

Cluster 8, identified by the INCLOSE clustering method, comprises a combination of AML and control cells classified as M-MDSCs. It displayed distinct gene expression patterns between AML and control groups, uncovering 1435 genes with noteworthy differences (**Fig 2.15.A**). This underscores the algorithm's ability not only to segregate AML and control clusters but also to reveal a heterogeneous mix that, while of the same cell subtype, exhibits significant variations. This analysis brought attention to several critical genes that are over expressed in AML. IFITM3, known for its adverse prognostic influence on AML patients, was found to significantly impact their event-free and overall survival (Y. Liu et al., 2020). The gene MT2A which is also overexpressed in AML, has been studied before for its role in influencing AML cell proliferation and functions, notably modulating apoptosis, cell reproductive capacity, and impacting the NF-κB signaling pathway in HL60 cells (Y.-Q. Pan et al., 2021).

Additionally, in cluster 3, another hybrid of AML and control cells, a unique differential gene expression emerged (**Fig 2.15.B**). Within this cluster, a stark gene expression contrast was observed, particularly highlighting the gene FGL2. Recognized for its role in promoting tumor growth and elevating the population of MDSCs, FGL2's impact has been studied, particularly in the context of hepatocellular carcinoma (HCC) (B.-Q. Liu et al., 2021). Additionally, the gene ISG20, identified for its overexpression in AML, has been extensively studied. Its upregulation, notably triggered by interferons, has been linked to a poor prognosis in various malignant tumors, including AML (Peng et al., 2023), including AML(H. Xiong et al., 2021). The findings from these clusters bear immense potential for unraveling nuanced and critical aspects of AML and its distinct subtypes or cell populations.

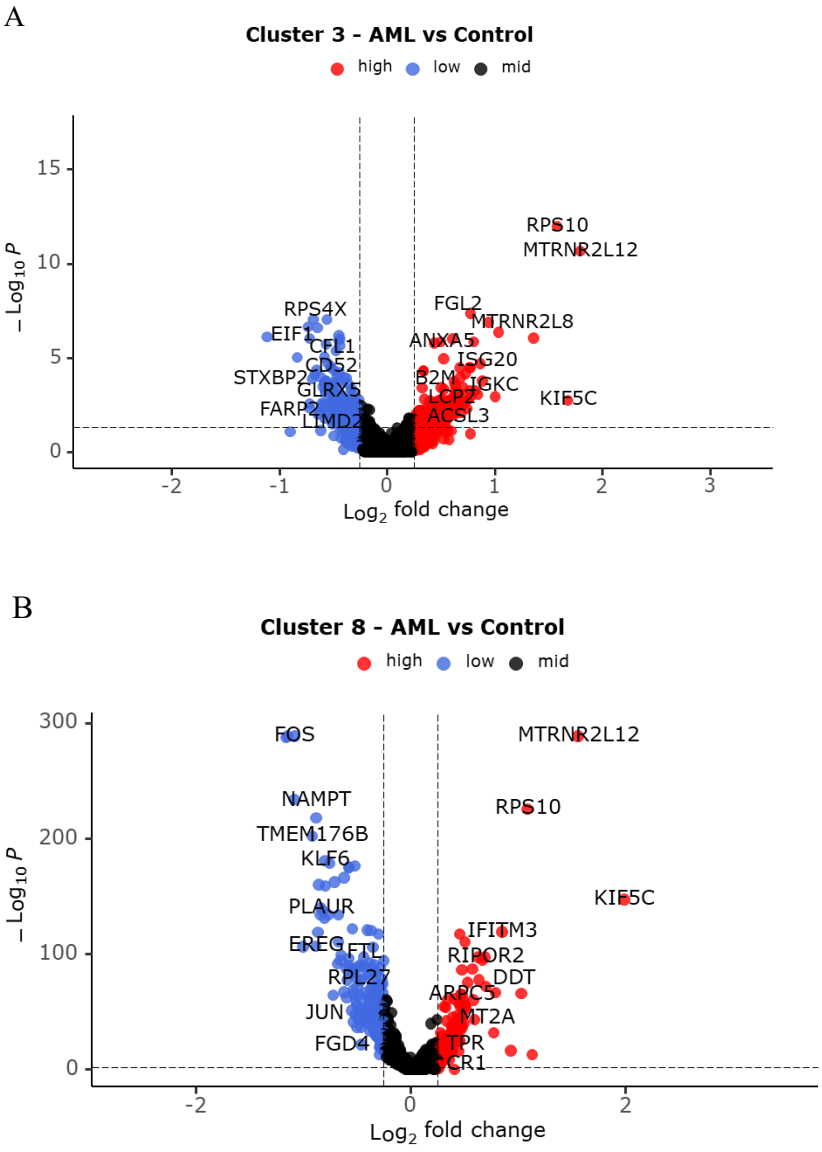

Figure 2.15: Volcano plots of Differentially Expressed Genes between AML and Control in (A) Cluster 3 and (B) Cluster 8

2.3.1.6 Comparing INCLOSE to Seurat Clustering

The J-score, a clustering accuracy metric, is employed to compare INCLOSE clusters to Seurat clusters, enabling the identification of correspondences between the two methods (**Fig 2.16**). This analysis provides insights into differences in cluster formation, such as whether larger Seurat clusters have been divided in INCLOSE, potentially revealing novel cell populations and highlighting variations in cell distribution between the two methods. We observed distinct clustering patterns when comparing Seurat and INCLOSE results. Specifically, what was a single, large cluster (Cluster 0 and 1) in Seurat has been partitioned into multiple clusters within INCLOSE.

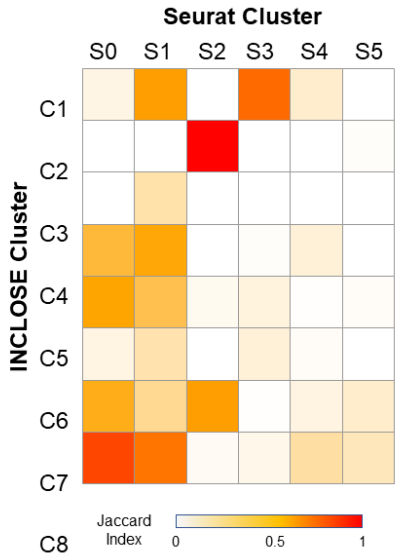

Figure 2.16: Bidirectional set matching based on Jaccard Index revealed correspondence between INCLOSE clusters and Seurat.

In Seurat, Cluster 2 exhibits a substantial correspondence with Cluster 2 in INCLOSE, reflecting a consistent cluster assignment. In contrast, Cluster 3 in Seurat, which primarily consisted of control samples, aligns notably with clusters 1 and 6 in INCLOSE, both of which are exclusively composed of control cells. In the INCLOSE analysis, Clusters 4, 5, and 7 are identified as AML-specific clusters, whereas Seurat splits cells from these INCLOSE clusters into various clusters. The most prominent cluster in INCLOSE, Cluster 8, corresponds strongly with one of the largest clusters in Seurat, Cluster 0. These findings highlight the differences and similarities in cluster assignments between the two methods.

We further conducted an extensive parameter exploration of the INCLOSE clustering method, aiming to identify the parameter combinations that best align with Seurat clusters. Our analysis involved systematically varying cell label weight (ω_l), ADT feature weight (ω_p), and sigma across predefined ranges. The ω_l was varied from 0 to 1 in increments of 0.03, ω_p ranged from 0.05 to 0.2 in increments of 0.01, and σ was explored between 1/3 and 0.1/3 in decrements of .05/3. Throughout this exploration, the highest J-score achieved was 0.6681886, signaling strong alignment with Seurat clusters. In **Fig. 2.17.A**, we showcase the outcomes featuring J-scores above 0.6, pinpointing those parameter combinations that yielded remarkably consistent clustering patterns closely mirroring Seurat's outcomes.

Each data point within the figure corresponds to a distinct parameter configuration. The size of these data points corresponds directly to the J-scores they represent, with larger data points indicating higher J-scores. Furthermore, the color-coded representation of the data points offers insights into the number of clusters generated under each parameter combination. Remarkably, our analysis consistently unveiled that the parameter sets leading to high correspondence predominantly formed patterns composed of 3 to 7 clusters. Strikingly, these parameter combinations consistently assigned significant weight to cell labels ($\omega_i > 0.5$), outweighing the cumulative weights of other feature matrices. Delving deeper into the parameter sets with elevated correspondence, we discerned an intriguing trend – mixed clusters were a rare occurrence (**Fig. 2.17 B** and **C**), indicating that these parameter combinations fostered distinct and homogenous cell populations.

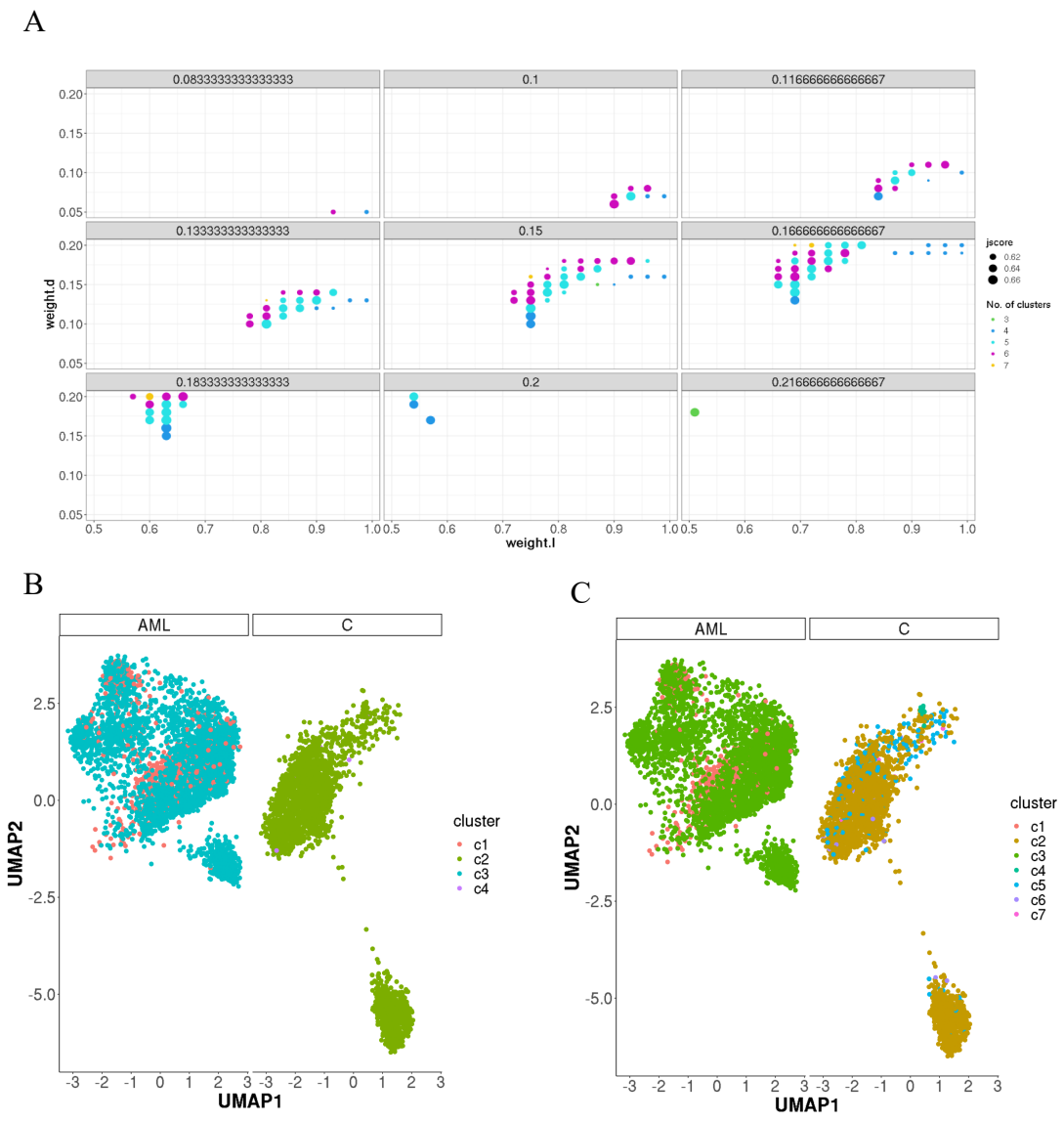

Figure 2. 17: Comparing INCLOSE to Seurat. (A) Scatterplot illustrating clustering patterns with top correspondence between INCLOSE and Seurat. Each data point represents a specific parameter set, with the size of the points indicating the J-scores (similarity scores) between the two clustering methods. Larger points correspond to higher J-scores. The color coding of the data points signifies the number of clusters produced under each parameter combination. (B) UMAP plots displaying INCLOSE clustering pattern with 4 clusters having high correspondence to Seurat clusters. (C) UMAP plots displaying INCLOSE clustering pattern with 4 clusters having high correspondence to Seurat clusters.

2.3.1.7 Effect of Phenotype Integration

To investigate the impact of integrating phenotypic data into the clustering process, we conducted a comprehensive analysis by systematically varying the cell label weight (ω_i) over a range from 0 to 1. In this analysis, we maintained a fixed ADT features weight at 0.11 and explored different sigma parameter values, specifically 0.0333, 0.1333, 0.2333, and 0.3333. For each distinct ω_i value, we assessed the number of mixed clusters, which are clusters containing cells from both the AML and Control conditions. As depicted in the **Fig 2.18**, the maximum number of mixed clusters attained in the AML dataset for all sigma values was consistently 2. Interestingly, as the ω_i value increased, the occurrence of mixed clusters decreased, ultimately leading to a lack of mixed clusters at higher ω_i values. This observation suggests that an appropriate selection of the cell label weight is essential for achieving mixed clusters that capture conditions with similar feature expression patterns.

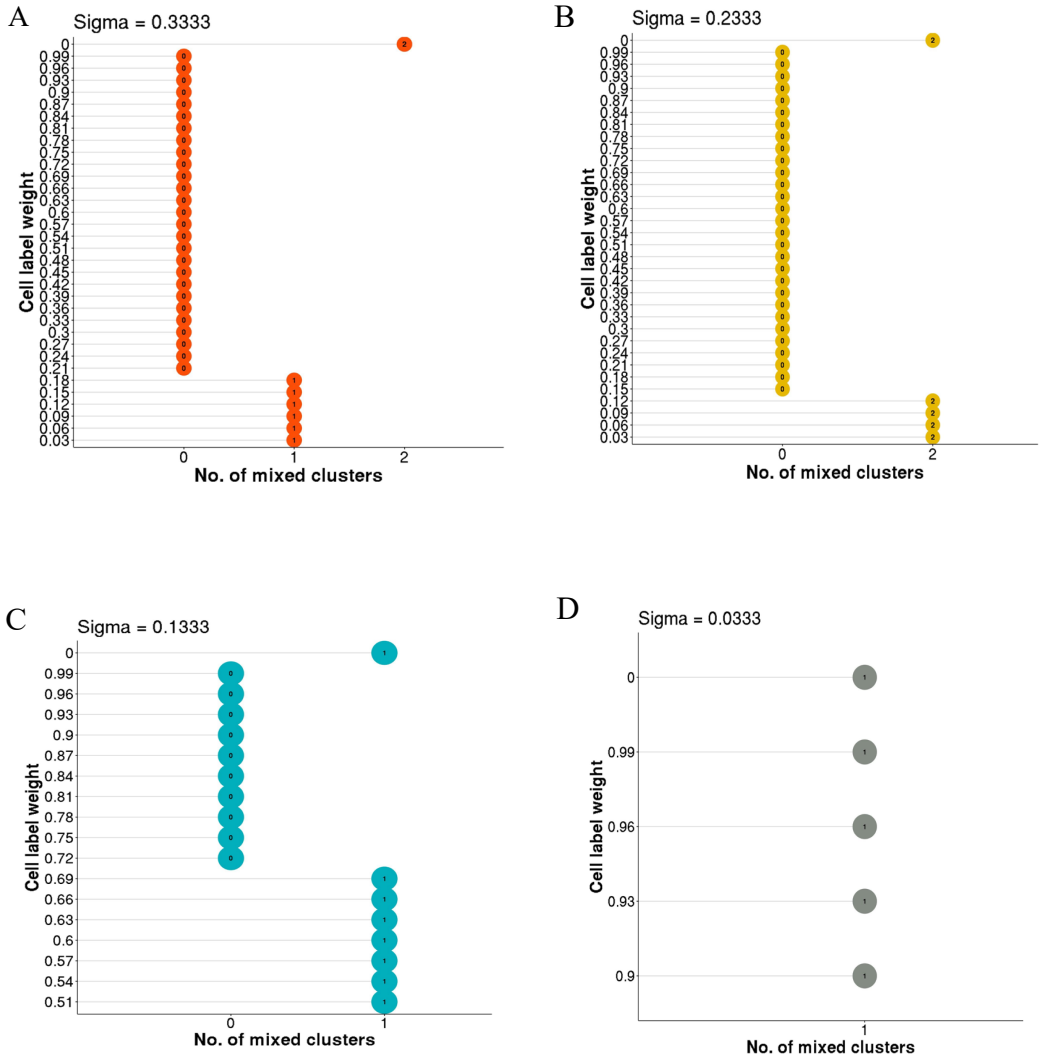

Figure 2. 18: Influence of cell label weight on INCLOSE clustering with ω_l varied between 0 and 1, ω_p set at 0.11 and σ set at (A) 0.3333, (B) 0.2333, (C) 0.1333, and (D) 0.0333.

2.4 Discussion

Traditional clustering techniques, rooted in the high-dimensional reduction of transcriptomic and proteomic profiles, sometimes fall short in capturing the full spectrum of MDSC subtypes, which often hinge on distinct surface markers (Cassetta et al., 2019) (Mandruzzato et al., 2016). Furthermore, the expansion of MDSC populations in tumor-infected tissues introduces unique clustering patterns not typically observed in healthy conditions. Even with the advent of advanced multimodal clustering tools such as Seurat, a significant challenge lies in their capacity to seamlessly integrate specific cell label information to pinpoint condition-specific clusters. The INCLOSE algorithm is designed to address these issues head-on. By incorporating cell label information directly into the clustering process, the INCLOSE method transcends conventional MDSC subcluster identification, which primarily relies on surface markers and transcriptomic profiles. This approach is meticulously tailored to the specific conditions under investigation, facilitating a comprehensive and context-aware exploration of MDSC diversity.

The INCLOSE method holds the potential to elucidate intricate cell subpopulations with application across diverse data types including scRNA-seq, CITE-seq, and scATAC-seq. With its application to scRNA-seq and CITE-Seq in contexts of radiation exposure and disease pathology, we were able to demonstrate its application to isolate the cell population that has emerged or expanded in tumors or treatment, which otherwise will get mixed with other cells that are similar based on a broader gene or protein expression profile.

INCLOSE is suited for clustering cells when dealing with a relatively smaller number of cells. However, as the number of cells increases, the computational requirements of the method become more demanding. This is primarily due to the initial construction of the similarity matrix, which relies on pairwise correlation analysis of the cells. Additionally, the larger number of cells will demand a greater number of clusters to capture the underlying complexity of cellular heterogeneity. Parameter tuning that governs the clustering process will become more complex since iterative adjustments and evaluations of tuning parameters will be required to ensure the best possible clustering outcome. So, it is suggested to employ clustering methods like Seurat, leveraging well-defined marker profiles such as CD4 T cells or CD14+ monocytes for the initial clustering of the cells. Subsequently, scGGC to uncover distinctive cell subpopulations within these prominent cell types, optimizing the identification of more nuanced cellular heterogeneity.

INCLOSE provides flexibility and adaptability for diverse research scenarios. It provides the option to obtain a predetermined number of clusters or to optimize clustering quality while keeping the number of clusters fixed. However, the current iterative approach utilizing the Z score heuristic for testing multiple parameter combinations in INCLOSE faces limitations in accurately identifying biologically relevant clustering outcomes. To refine this process, there is a pressing need for an enhanced heuristic that can discern the most biologically significant parameters, ensuring more precise and meaningful clustering results.

CHAPTER 3

IDENTIFYING CIS-TRANS REGULATORY TRIOS: Cis-Trans Trio

3.1 Background

3.1.1 Genetic Variants and Disease Susceptibility

Genetic variants are pivotal factors influencing an individual's susceptibility to diseases, treatment response, and clinical outcomes. These modifications can range from minute single nucleotide changes to significant structural alterations within the DNA sequence. Commonly encountered single-nucleotide polymorphisms (SNPs) or single-nucleotide variants (SNVs) represent the predominant type of genetic variants and have been extensively implicated in disease susceptibility (Eichler et al., 2007). Notably, they have been associated with a wide array of conditions such as diabetes, cardiovascular diseases, and various cancers, each demonstrating the genetic basis for predisposition (Shoily et al., 2021) (Zhu et al., 2023) (Bare et al., 2007) (N. Deng et al., 2017).

In breast cancer, studies analyzing SNP-related data within databases like TCGA have revealed key mutant genes significantly correlated with altered protein expression levels, pinpointing their involvement in cancer development pathways (Gao et al., 2019). Moreover, investigations linking specific SNPs to VEGF-A levels have highlighted their potential role in cardiovascular disease development and associated risk factors, offering vital insights for cardiovascular risk assessments (Meza-Alvarado et al., 2023).

One powerful tool for studying these associations is the Genome-wide Association Study (GWAS), where thousands of genetic variants across diverse individuals are scrutinized for links to various traits or diseases (Uffelmann et al., 2021). This comprehensive approach has been pivotal in identifying disease-associated SNPs, particularly notable in cancers such as breast cancer, colorectal cancer, and acute myeloid leukemia (AML). In AML, recent meta-analyses through GWAS have unearthed significant risk loci, such as KMT5B related to histone methylation and HLA associated with immune function, shedding substantial light on the disease's etiology (W.-Y. Lin et al., 2021). The findings from these studies underscore the fundamental impact of genetic variants on disease susceptibility and the essential role of comprehensive genomic analyses in unraveling their implications.

3.1.2 Genetic Variants and Gene Regulation

Single nucleotide variants (SNVs) encompass a broad spectrum of genetic changes occurring within protein-coding, non-coding, or intergenic regions between two genes with only less than 10% mapped in protein-coding regions. This emphasizes the vital role of investigating genetic variants in non-coding and intergenic domains, underscoring their impact on genetic regulation and functionality (Hindorff et al., 2009). Notably, the majority of variants discovered through Genome-Wide Association Studies (GWAS) reside in non-coding regions, predominantly in regulatory elements like promoters, enhancers, DNase hypersensitivity regions, and chromatin marks (Cano-Gamez & Trynka, 2020) (Trynka et al., 2013) (Maurano et al., 2012). While GWAS can identify associations between a single nucleotide polymorphism (SNP) and a phenotype,

it might not directly unveil the causal variants linked to a disease. Expression Quantitative Trait Loci (eQTLs) serve as a crucial bridge for understanding the sites in the genome likely to influence subsequent changes in gene expression (Varathan et al., 2022). Cis-eQTLs (**Fig. 3.1.A**) predominantly act on local genes, while trans-eQTLs **(Fig. 3.1.B**) affect distant genes or genes on different chromosomes (Shan et al., 2019). While cis-eQTLs exhibit potent effects on gene regulation, trans-eQTLs are also vital in regulating gene expression. The latter, however, demands larger sample sizes and innovative tools for the efficient detection of trans-eQTLs, as seen in the development of NetLIFT, a method that effectively addresses multiple-testing burdens (Weiser et al., 2014).

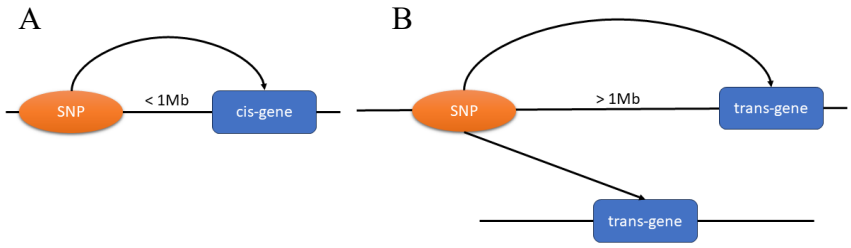

Figure 3. 1: Representation of eQTLs. (A) cis-eQTL and (B) trans-eQTL (Modified from (Weiser et al., 2014))

eQTLs help pinpoint genomic locations potentially influencing changes in gene expression profiles. For instance, an extensive meta-analysis on brain collections uncovered millions of significant eQTLs in cerebral and cerebellar regions, identifying potential implications for schizophrenia and reinforcing the value of brain eQTL findings (Sieberts et al., 2020).

Furthermore, studies on the impact of eQTLs within single-cell models reveal complex effects on cellular states and functions. For instance, a study on the impact of eQTLs on gene expression within single-cell models of memory T cells, revealed complex, context-dependent effects of these loci on cellular states and functions (Nathan et al., 2022).

3.1.3 Gene Regulation: Transcription Factors and Transcription Factor Binding Sites

Gene expression hinges on the coordinated interplay of numerous cis-regulatory elements, ranging from fundamental core promoters to proximal elements associated with promoters. In addition to these, several other modules exist, dispersed at varying distances from the transcription start sites (TSSs). These include enhancers, silencers, insulators, and tethering elements, each playing distinct roles in genetic regulation (Spitz & Furlong, 2012). Among these elements, enhancers are pivotal in initiating gene expression and have been a primary focus of extensive study. Enhancers, typically small DNA segments only a few hundred base pairs long, act as functional platforms for recruiting transcription factors (TF) facilitating the precise and intricate regulation of transcription. Enhancers (**Fig. 3.2.A**) orchestrate gene upregulation by engaging specific transcription factor binding sites in the promoter, facilitated by activator proteins. Conversely, silencers (**Fig. 3.2.B**) act as opposing agents, employing repressor proteins to bind to the promoter's TFBSs, leading to a reduction in gene expression. Meanwhile, insulators (**Fig 3.2.C**) execute a unique role, interfering with the binding between enhancers and promoters, thus restraining gene expression (Rojano et al., 2019a).

This ensemble of cis-regulatory elements collaboratively orchestrates the gene expression, from its inception to the finely tuned execution of genetic instructions. Any modifications or dysregulation in these elements could perturb gene expression, potentially contributing to various diseases.

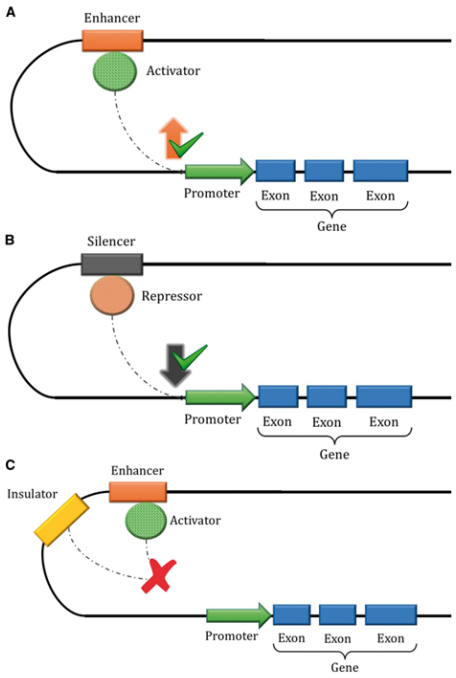

Figure 3. 2: Cis-regulatory elements, such as enhancers, silencers, and insulators, wield distinct effects on gene expression. Enhancers, depicted in (A), are regions that recruit activator proteins binding to specific transcription factor binding sites (TFBSs) in the promoter. This action upregulates the target gene. Silencers, illustrated in (B), work in an opposing manner by binding repressor proteins that also target TFBSs in the promoter, ultimately reducing gene expression. Meanwhile, insulators, as shown in (C), interact uniquely; they block the binding of the enhancer's activator protein to the promoter, curbing gene expression (Rojano et al., 2019)

Transcription factors (TFs) represent a class of proteins pivotal in orchestrating gene expression by regulating the transcription process. These proteins which include activators and repressors exert their influence by binding to specific DNA sequences, commonly found in gene promoter areas or distal regions known as enhancers. These binding sites, termed transcription factor binding sites (TFBS), play a crucial role in gene regulation. Notably, the proximity between a TFBS and the transcription start site (TSS) of the regulated gene can extend over vast genomic distances, even reaching several

megabases. The actual distance is intricately shaped by the chromatin structure and organization of the region, suggesting that TFs can impact gene expression across these considerable genomic spans. This ability to exert regulatory control over genes, both nearby and at a distance, highlights the intricate nature of TF-mediated gene expression regulation within the genome (Boeva, 2016). Technological advancements in genomics, including DNase-seq, FAIRE-seq, and ChIP-seq, have revolutionized the ability to locate and understand regulatory regions within the genome(Y. Wang et al., 2016). These high-throughput techniques allow for a comprehensive analysis of the genomic landscape, shedding light on elements that govern gene regulation and expression.

3.1.4 Regulatory Single Nucleotide Polymorphisms

Regulatory single nucleotide polymorphisms (rSNPs) are genetic variations situated within transcription factor binding sites (TFBSs), holding the potential to profoundly influence gene expression levels. These variations initiate intricate interactions between TFs and rSNPs, pivotal in shaping tissue-specific gene expression patterns (Degtyareva et al., 2021). Regulatory single nucleotide polymorphisms (rSNPs) significantly contribute to the diversity of complex traits, influencing the propensity for certain diseases to manifest. A notable instance of an rSNP is the substitution of G to A, positioned 376 base pairs from the TNF transcriptional start site. This genetic alteration has been shown to impact the interaction between the transcription factor OCT-1 and the genomic sequence. Studies have revealed a fourfold increase in the susceptibility to cerebral malaria in populations from both West and East Africa due to this specific rSNP (Degtyareva et al., 2021).

An earlier study revealed a crucial link between the transition from A to G within the Alu element just preceding the MPO gene and the onset of acute myelocytic leukemias. This specific alteration was found to create a robust binding site for the SP1 protein, subsequently initiating the transcription of the MPO gene and impacting its regulation in myeloid leukemias (Piedrafita et al., 1996).

The fusion of cutting-edge whole genome and exome sequencing, coupled with methodologies like GWAS and eQTL, has unveiled a vast array of genetic variants within these crucial regulatory domains. The fusion of experimental techniques and computational analysis of extensive omics data has offered valuable insights into understanding the significance of chromatin states, transcription factor binding regions, and the potential impact of genetic variations on these regulatory sites. For instance, through the application of self-transcribing active regulatory region sequencing (STARR-seq) on 10,000 cancer-associated SNPs, a study has revealed distal regulatory variants impacting gene expression positively and negatively. Investigating SNPs like rs11055880 (breast cancer) and rs12142375 (leukemia) revealed their distinct regulatory influences on genes ATF7IP and PDE4B, respectively, advancing our comprehension of how distal regulatory elements affect cancer risk in GWAS data (S. Liu et al., 2017).

3.1.5 Current Methods for Annotating Regulatory SNPs

The method of annotating regulatory variants involves precisely determining the specific regulatory components that are intersected by variants found within the genome. To accomplish this, a myriad of tools is employed, drawing information from various data sources like transcription factor binding sites (TFBSs), enhancers, promoters, DNA methylation sites, as well as introns and splicing sites. Multiple global collaborations and cross-disciplinary projects have been instituted to compile and organize these regulatory features found within the non-coding regions of the genome. Among the most notable initiatives are ENCODE (ENCyclopedia of DNA Elements) ("The ENCODE (ENCyclopedia Of DNA Elements) Project," 2004), FANTOM5 (Functional Annotation of the Mammalian Genome) (Abugessaisa et al., 2021), The Roadmap Epigenomics Project (Bernstein et al., 2010), and GTEx (Genotype-Tissue Expression) (Lonsdale et al., 2013). These endeavors employ a comprehensive spectrum of experimental methodologies such as Chromatin Immunoprecipitation Sequencing (ChIP-Seq), chromosome conformation capture methods, DNase I hypersensitivity assays, DNase Sequencing (DNase-Seq), and RNA Sequencing (RNA-Seq) to extensively chart and understand the regulatory elements entrenched within the non-coding domains of the human genome (Rojano et al., 2019b).

A multitude of computational tools leverages the wealth of data produced by these initiatives to annotate regulatory variants. Typically, these tools amalgamate genomic insights derived from multiple projects to ascertain the regulatory elements in proximity to a queried variant. RegulomeDB (Boyle et al., 2012) employs a scoring system that

considers the elements overlapped by a variant, referencing data from ENCODE and the Roadmap Epigenomics Project. HaploReg (Ward & Kellis, 2012) specializes in variants in linkage disequilibrium (LD) and those situated within or near regulatory elements, referencing data from ENCODE, GTEx, and the Roadmap Epigenomics Project. Meanwhile, FunciSNP (Coetzee et al., 2012a) prioritizes putative regulatory SNPs, drawing information from ENCODE and the Roadmap Epigenomics Project (Ward & Kellis, 2016).

3.1.6 Current Tools for Integrating Regulatory Elements with Regulatory SNPs

Several integrative tools have also been developed for linking regulatory variants with regulatory elements for dissecting the functional significance of genetic variations. FunciSNP is a bioinformatic tool that integrates data from whole-genome sequencing, GWAS SNPs, and chromatin maps to identify potentially functional genetic variants linked to specific phenotypes (Coetzee et al., 2012b). Scientists have also utilized eQTL and motif affinity analyzes to identify regulatory SNPs that map within canonical transcription factor binding motifs, potentially influencing transcription factor genomic occupancy (Jin et al., 2016).

Furthermore, modern research has also harnessed the power of machine learning techniques to understand the impact of human genetic variations in regulatory contexts. For instance, methods like Combined Annotation–Dependent Depletion (CADD) use machine learning, particularly support vector machines, to amalgamate varied annotations into a unified C score (Kircher et al., 2014). This empowers the prioritization of both functional and pathogenic variants across an array of genetic categories,

significantly exceeding the capabilities of single-annotation methods. In a parallel avenue, DeepSEA, a deep learning-based framework, deciphers regulatory sequences from extensive chromatin data. This in-depth analysis aids in precise predictions of how even single-nucleotide changes influence chromatin dynamics, improving the sorting of functional variants associated with gene expression (eQTLs) and diseases (Zhou & Troyanskaya, 2015). Sasquatch, a recent computational innovation, leans on DNase footprint data to evaluate the impact of non-coding variants on transcription factor binding (Schwessinger et al., 2017). Additionally, SEMpl, a novel pipeline, assesses the influence of single-nucleotide polymorphisms (SNPs) within functional transcription factor-binding sites (TFBSs) by scrutinizing changes in chromatin immunoprecipitation sequencing signal intensity. SEMpl's analysis provides insights into how SNPs might affect transcription factor binding, assisting in the recognition of potential disease-related regulatory regions within noncoding segments (Nishizaki et al., 2019).

While these methodologies meticulously explore the effects of regulatory SNPs or variants on transcription factor binding and/ or gene expression changes, none of these methodologies encompass the comprehensive interplay among SNPs within regulatory elements, the role of transcription factors in binding to these regulatory elements during the transcriptional regulatory processes, the influence of SNPs on transcriptional binding, and the resultant alterations in gene expression within a unified model. Building upon the context of the complex relationship between genetic variations, transcription factors, and gene expression, we introduce a novel computational methodology that combines information from SNPs within transcription factor binding sites (TFBS), the activity of specific transcription factors, and the expression of target genes into a comprehensive

regression model. This integration allows us to identify and characterize regulatory trios governing gene regulation, thus offering a robust framework to explore the regulatory mechanisms shaping complex phenotypic traits in diseases.

Fig.3.3 illustrates the effect of germline SNPs, transcription factors, and their interactions in transcription factor binding sites on target gene expression. This study focuses solely on germline SNPs, but by pinpointing tissue-specific somatic SNPs within regulatory elements, the scope can expand to encompass the impact of somatic SNPs, broadening the understanding of their effects.

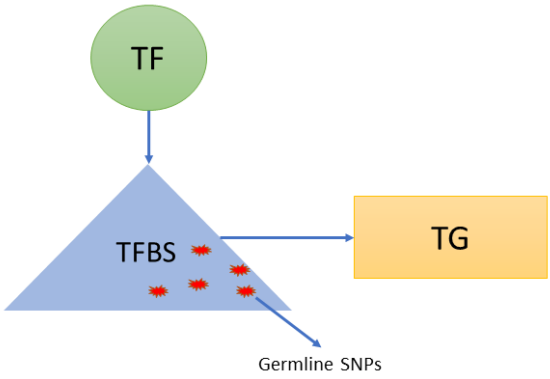

Figure 3. 3: Illustration of impact of germline SNP, TF, and their interaction in TFBS regions and their impact on the target gene expression

3.2 Methodology

This study aimed to investigate the complex regulatory landscape governing gene expression changes in disease contexts. Our method incorporates a multi-dimensional perspective, combining transcription factor (TF) expression, germline single nucleotide polymorphisms (SNPs) in the transcription factor binding sites (TFBS), and covariates

both genetic and clinical, to identify cis and trans-acting regulatory elements impacting the target gene (TG) expression. *A* SNP is considered *cis-acting* if it is within 1 Mb from the gene transcriptional start site (TSS), and *trans*-acting if the SNP is beyond that point (Jin et al., 2016). For sample *s*, we model the expression of target gene *g* as an outcome of the expression of a transcription factor *t* and a germline SNP *v,* both in the TFBS region of the target gene. Separate models *m* are run for different genotypes of the SNP, which are given by dominant (MM), recessive (mm), heterogenous (Mm), and additive models with an allele frequency of $\geq 10\%$ (J. Ma et al., 2018) (Gong et al., 2018). Given a sample *s*, we denote $X_{g,s}$ as the expression of a target gene *g*, $X_{t,s}$ as the expression of transcription factor *t,* and $S_{v,m,s}$ as the germline SNP genotype variable.

The regression model equation is as follows:

$$X_{g,s} = \alpha_0 + \alpha_1 X_{t,s} + \alpha_2 S_{v,m,s} + \alpha_3 X_{t,s} * S_{v,m,s} + \sum \beta G + \sum \gamma C + \in \qquad [3.1]$$

where α_0 is the intercept, α_1, α_2, and α_3 are coefficients representing the effect of $X_{t,s}$, $S_{v,m,s}$, and the interaction term between $X_{t,s}$ and $S_{v,m,s}$. *C* is a set of clinical covariates, each with a γ coefficient, and $\in$ is Gaussian-distributed errors. G is a set of genetic covariates, each with a β coefficient that includes the copy number variance of the target gene, the copy number variance of the transcription factor, and the diploidy class of the sample. By harnessing copy number variant information from the target gene and transcription factors (TFs), we employed a hierarchical clustering approach using the Ward D2 distance metric to categorize the samples into three distinct diploid classes:

diploid (D), partial hyperdiploid (P_HPD), and hyperdiploid (HPD). This allowed us to identify underlying patterns and relationships within the dataset, enabling the accurate classification of samples based on their genetic characteristics.

To correct for multiple comparisons, we adjusted the nominal p values using the Benjamini-Hochberg method and calculated the false discovery rate (FDR). Non-zero coefficients at FDR<0.05 indicate significant associations. A significant α_1 and α_2 indicates significant associations of the transcription factor, and germline SNP genotype with the target gene expression. Significant α_3 indicates the significant interaction between a transcription factor and the germline SNP in the TFBS region. The regression models fit on each trio are filtered based on the significance of these three coefficients, and we call these trios cis or trans-regulatory trios based on whether the SNP is cis or trans-acting.

Unlike traditional eQTL and ceQTL analyses (R. T. Wang et al., 2011), which primarily focus on genetic variations' direct impact on gene expression, and in contrast to methods based on binding affinities (Flynn et al., 2022) , our method combines multiple layers of regulatory information. The Integration of germline SNPs, transcription factor (TF) expression, and the interaction between TF expression and SNPs provides a more comprehensive, accurate, and biologically meaningful approach to understanding disease-associated gene expression changes.

Once the cis-trans regulatory trios were identified, we tested for transcription factors that are enriched in TFBS regions, indicating their potential roles in mediating

gene expression changes. To assess the enrichment of transcription factors (TFs) among significant and non-significant target genes, we employed the Fisher exact test. This statistical test allows us to determine whether the occurrence of TFs significantly differs between the two gene groups, shedding light on potential regulatory associations. For each TF, we examine its occurrence within the set of significant target genes identified in the regulatory trios. We then perform a Fisher exact test to determine whether the presence of the TF is significantly associated with these target genes, as compared to non-significant target genes. We hypothesize that these enriched transcription factors, as determined by Fisher's exact test, are pivotal nodes within the regulatory network, serving as master regulators capable of orchestrating coordinated responses across multiple genes and genetic variants. Ultimately, our investigation is extended to gene set over-representation analysis of the target genes of enriched TFs to ascertain their significance in regulating key biological processes or pathways associated with the disease under investigation and their functional implications.

3.3 Application

3.3.1 Cis and Trans-Acting Regulatory Trios in Multiple Myeloma

We leveraged the Multiple Myeloma Research Foundation (MMRF) dataset from the GDC portal, which encompasses mRNA expression matrix extracted from tumor bone marrow tissues offering insights into the gene expression patterns specific to multiple myeloma. Additionally, germline genotype data obtained from normal whole blood tissues were integrated into the analysis, providing a baseline genetic profile for each sample. Clinical covariates – age, gender, and stage are considered along with the

genetic covariates, including copy number variation of TG, TF, and diploidy class of the sample. **Fig 3.4** shows the diploidy classes identified from the MMRF dataset.

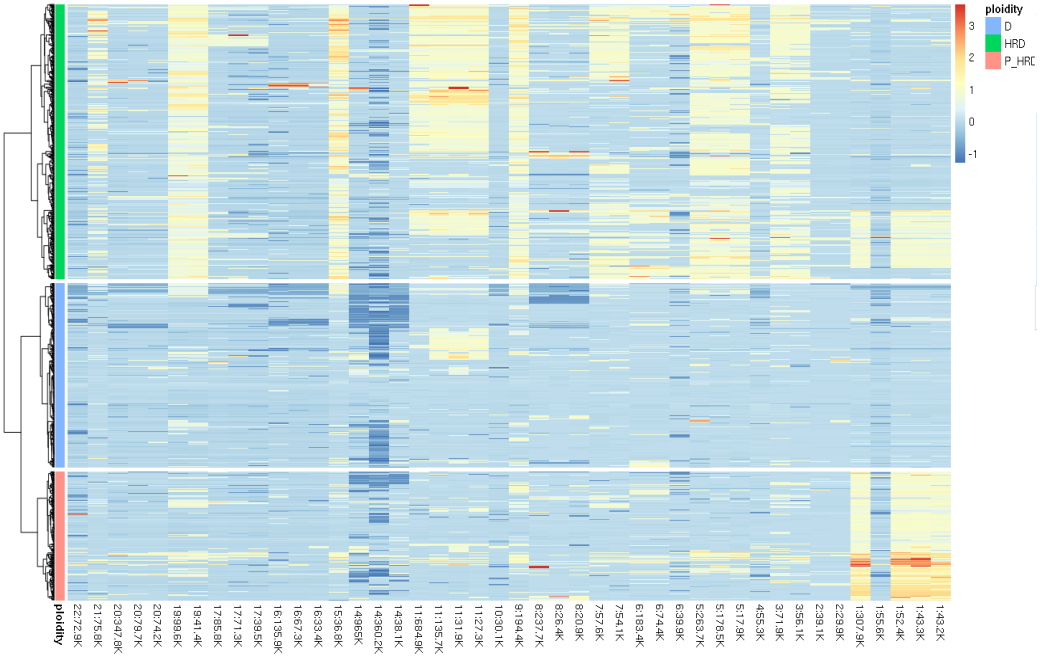

Figure 3. 4: Heatmap of MMRF Samples and Copy Number Variations Grouped by Chromosome Regions. Clustering pattern of samples along rows, indicates distinct classes related to diploidy (diploid), partial hyperdiploidy, and hyperdiploidy.

The regression analysis provided us with distinct sets of 43546 regulatory trios. Of these significant regulatory trios, there were 432 significant TFs, 4842 TGs, and 7129 significant germline SNPs at FDR < 0.05 for TF, SNP, and the interaction between TF and SNP. Multiple regression models were employed to examine the relationship between transcription factors (TFs) and target genes (TGs) across different genotypes of the single nucleotide polymorphism (SNP). This analysis identified a total of 43546 regulatory trios W6336 significant trios within additive models, 13,176 significant trios within homozygous dominant (AA) genotypes, 13,543 significant trios within heterozygous (Aa) genotypes, and 10,491 significant trios within homozygous recessive

(aa) genotypes. The **Fig 3.5** illustrates example scatterplots for each of the genotype models tested. Notably, the crossing of these regression lines indicates the presence of a significant interaction effect between the transcription factor and the genotype. This interaction suggests that the influence of the transcription factor on target gene expression varies across different genotypes of the SNP, adding depth to our understanding of the regulatory mechanisms at play.

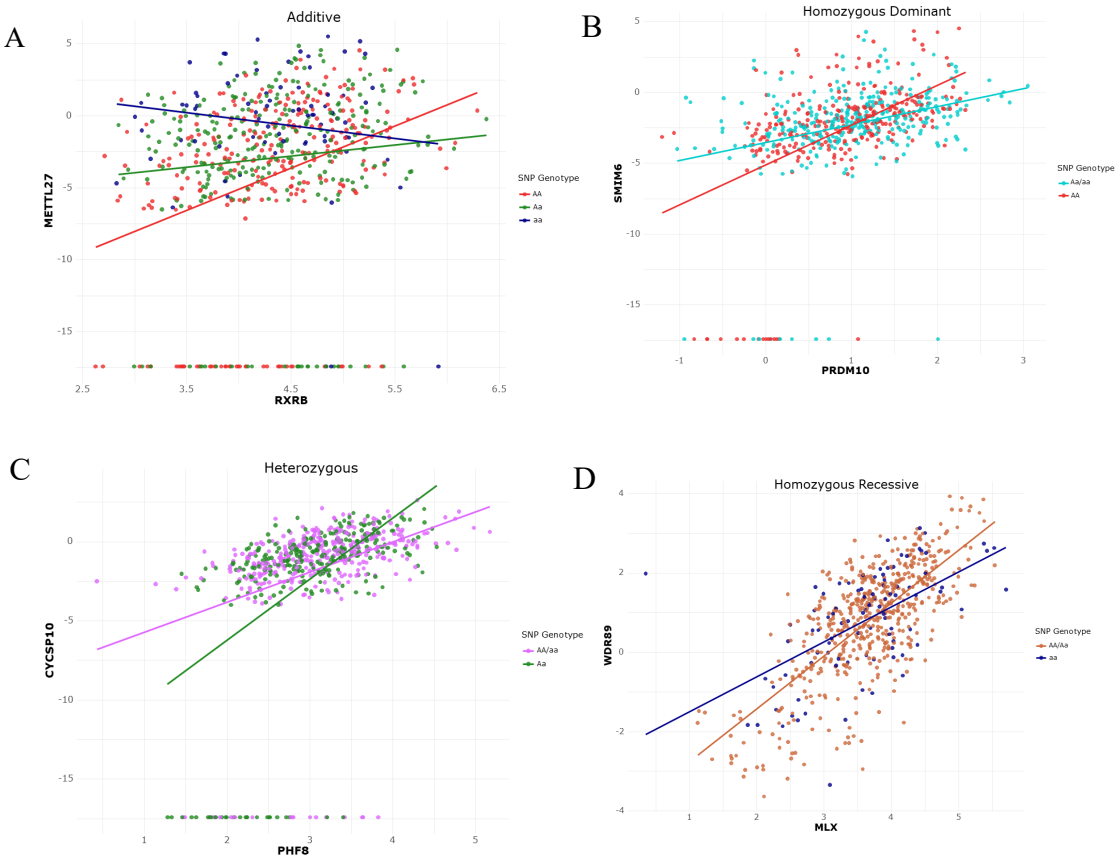

Figure 3. 5: Scatter plots demonstrating the relationship between transcription factors and target genes across different genotypes. (A) TF - METTL27 and TG - RXRB across additive genotype (B) TF - SMIM6 and TG - PRDM10 across homozygous dominant genotype (C) TF - CYCSP10 and TG - PHF8 across heterozygous genotype (D) TF - WDR89 and TG - MLX across homozygous recessive genotype

To unravel the nuanced relationships between transcription factors and the influence of regulatory SNPs on gene expression, a comprehensive scatter plot detailing the association between Transcription Factor (TF) and TF:SNP coefficients was devised (**Fig 3.6**). The x-axis represents TF coefficients, while the y-axis demonstrates the TF:SNP coefficients. The intricate interplay between these variables offers insightful categorizations of their regulatory effects on gene expression.

Instances where both the TF and the TF:SNP coefficients are positive implies that both the TF and the interaction term positively contribute to the target gene expression, which might suggest an additive effect in promoting gene expression. Conversely, when the TF coefficient is negative and the TF:SNP coefficient is positive, it suggests that the TF alone negatively affects expression, but in conjunction with the SNP, a compensatory or augmentative impact is observed. When a positive TF coefficient is accompanied by a negative TF:SNP coefficient, it hints at a counteractive role of the SNP, impeding the TF's regular function and subsequently decreasing gene expression. Furthermore, scenarios with both negative TF and negative TF:SNP coefficients represent a compounded inhibitory effect, indicating a combined reduction in gene expression. As shown in **Fig 3.6**, only a marginal fraction (0.01%) of TF-SNP pairs exhibit a compounded inhibitory effect, whereas a predominant number (55%) reveal that the SNP disrupts the positive effect of the TF. Additionally, 42% of TF-SNP pairs denote an additive effect, signifying their combined influence on gene expression. Notably, a mere 2% of the observed TF-SNP pairs demonstrate a compensatory effect by SNP.

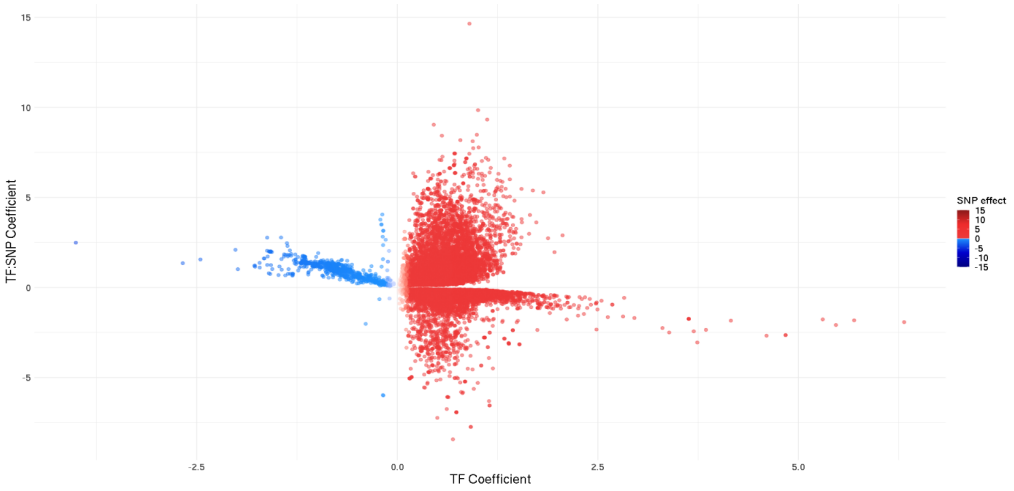

Figure 3. 6: Scatter plot depicting the interplay between Transcription Factor (TF) and SNP in the regulating gene expression depicting the additive, compensatory, or inhibitory effects in varying TF-SNP pair combinations.

Our method yielded a total of 43482 cis-acting regulatory trios, highlighting interactions occurring within localized genetic regions. Furthermore, we identified 64 trans-acting regulatory trios, indicative of interactions spanning distant genetic loci. The results of the Fisher exact test revealed that out of the 432 TFs analyzed, an impressive 384 TFs demonstrated significant enrichment within the set of target genes deemed significant by the regulatory trios ($p < 0.05$, Benjamini-Hochberg corrected for multiple testing). Each TF is ranked based on its enrichment p-value, highlighting the strongest associations with the regulatory trios. In **Table 3.1**, we present the top 10 TFs that demonstrated significant enrichment with the greatest number of target genes and in **Fig 3.7** we showcase the odds ratios associated with the top 100 transcription factors, providing a comprehensive view of their respective impacts.

Table 3. 1: Significant TG, TF, and SNPs from regression model

TF	TG Count	Fisher OR	Fisher Padj
MAX	2423	5.77335	0.00E+00
NCOR1	1275	3.985074	0.00E+00
RXRA	802	3.142674	0.00E+00
RUNX3	584	2.936895	0.00E+00
ZNF579	397	2.441118	7.66E-249
ZNF217	377	2.600185	0.00E+00
ZNF512B	365	3.533993	3.68E-259
ZNF362	307	2.246956	1.24E-214
ZNF140	281	3.570729	1.46E-134
RBFOX2	259	2.304046	0.00E+00

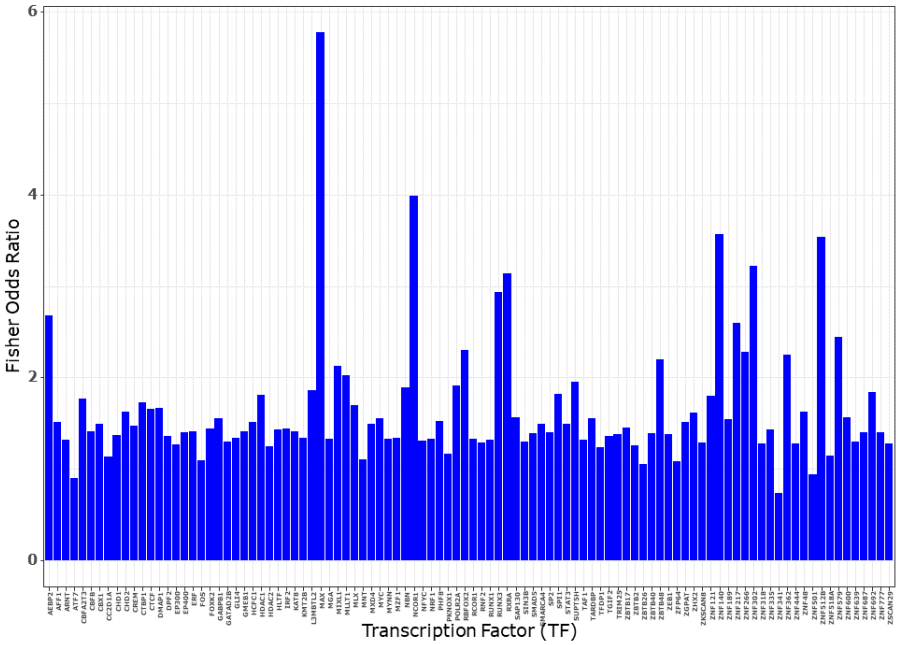

Figure 3. 7: Top 100 transcription factors along with their corresponding fisher exact test odds ratios

Over-representation analysis of all target genes of the enriched TFs unveiled several hallmark pathways that exhibited statistically significant associations with the target genes of the enriched TFs which include HALLMARK_DNA_REPAIR, HALLMARK_MYC_TARGETS_V1, HALLMARK_E2F_TARGETS, HALLMARK_MYC_TARGETS_V2, and HALLMARK_OXIDATIVE_PHOSPHORYLATION (**Fig 3.8)**.

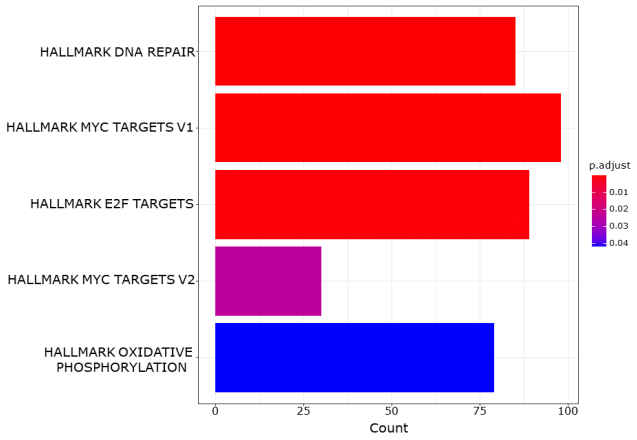

Figure 3. 8: Gene Enrichment bar plot illustrating immune-related pathways that are enriched among the target genes regulated by the enriched transcription factors identified in the MMRF regression model.

The dysregulation of pathways HALLMARK_MYC_TARGETS_V1 and HALLMARK_MYC_TARGETS_V2 indicates the potential role of enriched TFs on MYC target genes. MYC is a known oncogene that regulates cell growth, proliferation, and apoptosis. Dysregulation of MYC signaling is common in various cancers, including multiple myeloma (Holien et al., 2015). The enrichment of these pathways could suggest that the enriched TFs influence MYC-regulated genes, potentially impacting cell growth and survival pathways relevant to myeloma development (Wahlström & Arsenian

Henriksson, 2015) (Chng et al., 2011). Upon closer examination of our regression results, we observed that only 2 of the MYC target genes demonstrated a significant direct association with the MYC transcription factor. However, the remaining target genes enriched within this gene set appeared to be linked with other transcription factors. In our analysis, we identified a total of 105 transcription factors (TFs) that featured one or more MYC targets among their regulatory targets. Notably, among these 105 TFs, 45 TFs exhibited a particularly robust association, each having more than two MYC targets within their regulatory repertoire. Remarkably, MAX emerged as the transcription factor with the highest number of target genes (76) associated with MYC. **Table 3.2** gives the top 10 TFs arranged in the order of the greatest number of MYC target genes.

Table 3. 2: Top TFs with MYC targets

TF	TG_count
MAX	76
NCOR1	39
RXRA	20
ZNF217	18
ZNF512B	12
RUNX3	11
ZNF362	10
ZNF579	10
MLX	8
RBFOX2	7

This observation led us to delve further into the MYC-MAX association through a comprehensive literature review. Our investigation aimed to deepen our understanding of the intricate relationship between MYC and MAX and its implications for gene regulation in the context of our study. MYC-MAX interaction is a fundamental mechanism that governs gene expression programs crucial for cell growth and proliferation (Amati & Land, 1994). Its intricate regulatory network and potential implications in diseases, particularly cancer, continue to be an active research area (Madden et al., 2021). The enrichment analysis provides insights into potential mechanisms, but validation through experimental studies and clinical observations is necessary to confirm their significance in multiple myeloma.

3.4 Discussion

The analysis of the MMRF dataset in our study enabled a comprehensive exploration of the regulatory mechanisms in multiple myeloma. The integration of mRNA expression profiles derived from tumor bone marrow tissues with germline genotype data allowed for the identification of regulatory trios associated with the disease. Our approach produced a considerable number of significant regulatory trios, unveiling the influence of various transcription factors (TFs) and target genes (TGs). Employing regression models across different genotypes of the single nucleotide polymorphism (SNP) resulted in the identification of substantial trios, reflecting the interplay between TFs and TGs across distinct genotypes.

Notably, our approach identified both cis-acting and trans-acting regulatory trios, demonstrating the varied interactions occurring within localized and distant genetic regions. The enrichment analysis unveiled a significant association of multiple TFs with the set of target genes identified by the regulatory trios, particularly highlighting pathways associated with MYC targets.

Further investigation revealed the potential influence of enriched TFs on MYC-regulated genes, suggesting their involvement in pathways crucial for cell growth and survival, particularly relevant in the context of multiple myeloma. Although only a limited number of MYC target genes exhibited a direct association with the MYC transcription factor in our analysis, a comprehensive examination uncovered several TFs that displayed a robust link to MYC target genes, notably MAX emerging as a predominant regulator.

The examination of the MYC-MAX relationship, essential in governing gene expression programs crucial for cell growth and proliferation, suggests its relevance in the context of multiple myeloma. However, while our study has provided valuable insights into potential regulatory mechanisms, further experimental studies and clinical observations are warranted to validate and confirm their significance in the disease context.

CHAPTER 4

SINGLE CELL PROTEIN TRAFFICKING: CITE-Traffick

4.1 Background

4.1.1 Cell Surface Markers

Cell membranes exhibit a diverse array of proteins such as enzymes, transporters, ion channels, and receptors. Through changes in abundance and composition, these cell surface proteins actively contribute to a wide range of biological processes and play vital roles in shaping cell functions (Alberts et al., 2002). For instance, many cluster of differentiation (CD) markers are cell surface proteins that define cell types and cell differentiation stages ("CLUSTER OF DIFFERENTIATION (CD) ANTIGENS," 2004). Cell surface proteins serve a multifaceted role as receptors for cytokines, ligands associated with antigen presentation, signaling, and cell adhesion. Cytokines, soluble proteins secreted by various immune cell types, interact with these receptors, initiating intracellular signaling pathways that trigger and perpetuate a cascade of immune responses. An illustrative example of this is the crucial involvement of cytokine receptors in immune function. These receptors, upon binding specific ligands, induce conformational changes that activate JAKs, subsequently leading to tyrosine-based motif phosphorylation and the creation of docking sites for essential proteins like STATs. This interplay significantly contributes to the regulation of immune responses and immune-related disorders (Lee & Rhee, 2017). Furthermore, a specific category of cell surface proteins known as MHC proteins plays a pivotal role in antigen presentation. MHC

proteins are responsible for presenting antigens on the cell surface, facilitating recognition by the appropriate T cells. While all cells produce MHC class I molecules to present intracellular pathogens, antigen-presenting cells (APCs) like macrophages and dendritic cells produce MHC class II proteins for presenting extracellular pathogens to T cells (Neefjes et al., 2011). Tumor-specific antigens are mostly intracellular, and their recognition by T cell receptors (TCRs) expressed on the surface of T cells can trigger various effects, such as T cell proliferation and differentiation and cytokine or chemokine secretion (He et al., 2019).

Aberrant expression of cell surface proteins has been associated with numerous diseases and demonstrated to profoundly impact cellular function. For instance, abnormal expression of human leukocyte antigen (HLA), a critical cell surface protein involved in immune responses, has been implicated in various health conditions (Dendrou et al., 2018). Several drugs clinically used to treat cancers target integrins that are transmembrane adhesion receptors mediating cell-cell and cell-extracellular matrix interactions (Pang et al., 2023). Because of their distinct cellular locations and profound involvement in disease processes, cell surface proteins have emerged as promising targets for diagnostic biomarkers and therapeutic interventions (Yin & Flynn, 2016).

4.1.2 Intracellular Protein Trafficking

The expression of surface proteins is a complex process involving transcription, translation, post-translational modification, and intracellular protein transportation (ICT). Notably, ICT requires precise coordination of multiple organelles and genes to ensure proper protein trafficking and expression on cell membranes (**Fig 4.1**) (Tokarev et al.,

2013). Protein trafficking unfolds through two primary pathways: endocytosis and exocytosis. In the process of endocytosis, proteins journey from the cell's surface to the early endosomes. Subsequently, these internalized proteins are directed by the early endosome either to the lysosome for degradation or to the trans-Golgi network. Conversely, exocytosis entails the transfer of freshly synthesized proteins into the endoplasmic reticulum (ER), their passage through the cis-Golgi complex, and ultimate transportation via the trans-Golgi network (A. Kumar et al., 2020).

Proteins are synthesized in ribosomes by translating mRNA into peptides. The ribosome, along with the growing polypeptide chain, attaches to the endoplasmic membrane (ER) to facilitate the translocation of the nascent protein into the ER lumen. From the ER, membranous vesicles shuttle cargo to the Golgi apparatus. ER-derived cargo sequentially moves through the cis, medial, and trans cisternae regions of the Golgi apparatus, facilitating the processing, modification, and sorting of proteins. The proteins are then packed into secretory vesicles for transport to the plasma membrane, where they merge with the plasma membrane, releasing the proteins to their intended destinations. Golgi cargo is sorted not only to the plasma membrane for secretion but also to endosomes, lysosomes, and even back to the endoplasmic reticulum (ER). This process ensures that proteins or other cellular components that need to be broken down or recycled are properly targeted to these compartments for degradation and subsequent recycling of their constituent molecules. Cells can also internalize cell surface proteins by endocytosis. Endocytic vesicles generated from the plasma membrane (PM) fuse with recycling endosomes, from where they eventually move to lysosomes for degradation.

Proper protein trafficking from the endoplasmic reticulum (ER) to the plasma membrane (PM), various target organelles, or the extracellular (EC) space is required for cell survival Dysregulated ICT has been shown to alter cell surface protein expression and is linked to diseases such as obesity, diabetes, cancers (Sneeggen et al., 2020), and abnormal host responses to infections (Welstead et al., 2004) (Hassan et al., 2021), However, a comprehensive understanding of genome-wide patterns and interplays between ICT and surface protein expression in human diseases remains limited. In this study, we present a novel computational approach, CITE-trafficking, to investigate regulatory circuits of surface protein expression in the context of ICT processes.

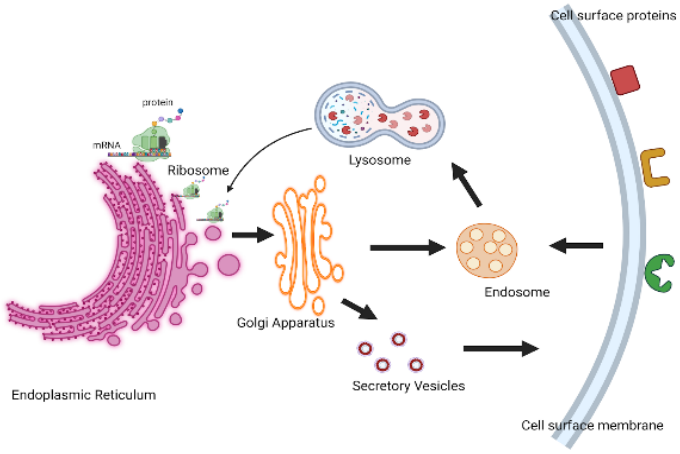

Figure 4. 1: Overview of intracellular protein trafficking

4.1.3 Cellular Indexing of Transcriptomes and Epitopes (CITE-Seq)

Single-cell transcriptomics provide valuable insights but may not provide a holistic understanding of protein trafficking dynamics. In contrast, proteins play immediate and crucial roles in maintaining cellular functions compared to transcripts.

Therefore, it is crucial to employ combined single-cell measurement techniques for both mRNA and proteins to achieve a more comprehensive understanding of intracellular protein trafficking and its impact on cellular states (Reimegård et al., 2021). The recent advent of cellular indexing of transcriptomes and epitopes by sequencing (CITE-Seq) introduces a high-throughput approach to simultaneously interrogate surface protein expression and gene transcript expression (Stoeckius et al., 2017). CITE-seq combines single-cell RNA sequencing (scRNA-seq) and quantification of antibody-derived tags (ADT) to profile the entire transcriptome and a selected panel of cell surface proteins at the single cell level. Including cell surface protein data has significantly improved the accuracy and resolution of cell type characterization compared to using transcriptomic data alone (Darden et al., 2021) (Bronte et al., 2016). However, once cell type clustering analysis is completed, these two modalities of data are analyzed independently (Tang et al., 2022) (Leblay et al., 2020) (Z.-Z. Yang et al., 2023), missing the opportunity to integrate the data and explore the interplays between cell surface protein expression and gene transcription. CITE-Traffick method leverages CITE-seq data to obtain information on the initial transcripts (i.e., mRNA abundance of cell surface proteins), end products (i.e., protein abundance on the cell membrane), and genes participating in the transportation (i.e., mRNA abundance of ICT genes). It examines the influence of ICT on differential expression of surface protein expression between disease groups in a mediation model, allowing multiple exposures and multiple mediators.

4.1.4 Regularized Mediation Analysis

Mediation analysis is used for studying the effect of an independent variable for e.g., a study exposure on an outcome through an intermediate variable, called mediator. Mediation analysis offers a valuable advantage in research by providing deeper biological insights into the potential causal mechanisms that underlie the observed associations between exposure and outcome variables. Three regression models can be used to describe the direct and indirect effects as shown in following equations:

$$E[Y_i] = \delta_0 + \delta_1 X_i \quad [4.1]$$

$$E[M_i] = \alpha_0 + \alpha_1 X_i \quad [4.2]$$

$$E[Y_i] = \beta_0 + \delta_2 X_i + \beta_1 M_i \quad [4.3]$$

In equation 1, δ_1 represents the total effect of the exposure X_i on the outcome Y_i. In equation 2, α_1 represents the association between the exposure X_i and a mediator M_i. In equation 3, β_1 represents the effect of surface marker, M_i on the outcome Y_i and, δ_2 represents the direct effect of the exposure X_i on the outcome Y_i after adjusting for the effect of the mediator, M_i.

In the case of high dimensional studies where we have multiple exposures and multiple mediators, the above equations can be extended by regressing Y_i simultaneously on all mediators and exposures and regressing each mediator on multiple exposures.

While separate tests can be conducted for each exposure-mediator, this procedure ignores the correlations between multiple exposures and multiple mediators. Hence regularized mediation analysis is developed to simultaneously estimate and select multiple exposures and mediators through use of penalized likelihood function (Schaid et al., 2022). Regularized mediation analysis methods offer a statistical framework to examine the indirect effects of the ICT genes on disease status with surface marker expression as the mediator.

4.1.5 Structural Equation Modeling

Mediation analysis is a special case of general structural equation modeling (SEM). SEM is a powerful statistical technique that can be used to model and analyze complex relationships among observed variables or indicators, latent variables, and outcome variables. In SEM, latent variables act as unobserved constructs representing underlying concepts or dimensions that cannot be directly measured but are inferred from multiple observed variables (Spearman, 1904) (Tarka, 2018). These latent variables serve as a bridge connecting the observed variables and mediators to control the outcome or dependent variable. They allow us to capture the common variance shared among multiple observed variables and provide a way to model complex relationships and interactions in a more parsimonious manner.

SEM offers a comprehensive framework that incorporates both a measurement model and a structural model (Fan et al., 2016).

- The measurement model facilitates the representation of latent variables, derived from observed variables, by establishing the connections between latent constructs and their corresponding observed indicators.
- Simultaneously, the structural model enables the establishment of relationships between latent variables themselves, providing a comprehensive view of how underlying constructs influence the observed outcomes.

Combining these two models offers a valuable advantage by allowing us to group related observed variables into latent variables and model their interactions. Instead of dealing with many individual observed variables, we can work with a smaller set of latent variables to explore their interrelationships. This approach is particularly beneficial when analyzing complex multivariate longitudinal data.

4.2. Unraveling Intracellular Trafficking Gene-Mediated Regulation of Surface Protein Expression in Disease

In this comprehensive study, we delve into the cellular mechanisms governing the expression of cell surface proteins and their relationship with disease phenotypes. It is well-established that the synthesis and transportation of surface proteins are finely orchestrated processes, guided by a network of genes responsible for intracellular trafficking (ICT).

In the context of disease, the alteration of surface protein expression, whether it be an increase or decrease, has been a focal point of investigation. Previous research, exemplified by studies in cancer (G. Chen et al., 2002) (Zerdes et al., 2021) (Reimegård et al., 2021), has unveiled a striking discordance between the levels of protein expression and the transcription of the corresponding coding genes. For instance, a study of lung adenocarcinomas found that a mere 17% (28 proteins) of the 165 examined displayed associations between protein abundance and mRNA levels, suggesting a post-translational mechanism primarily underlies the regulation of these proteins in such conditions. Our central hypothesis posits that the key to deciphering these complex alterations lies within the realm of intracellular trafficking genes. These genes are instrumental in the precise transportation of surface proteins, and we postulate that they hold the answers to understanding differential protein expression in disease conditions. Our research further uncovers a multifaceted landscape in which some intracellular trafficking genes not only directly connect to disease through specific signaling pathways but also intricately interact with other genes, thereby exerting influence over protein expression levels. Even when lacking direct associations with the disease, some ICT genes indirectly establish crucial links with the disease phenotype through their roles in regulating surface protein transport. Therefore, this study aims to disentangle the multifaceted roles of ICT genes, illuminating their potential in explaining the variations in surface protein expression during disease.

4.3 CITE-Traffick Methodology

The CITE-Traffick algorithm includes three modules that aim to (1) discover ICT genes associated with differential expression of surface proteins, (2) establish the mediation effect of surface protein expression on disease phenotype, and (3) identify dysregulated pathways, respectively (**Fig 4.2**). In the protein transportation network, an individual ICT gene may facilitate the transportation of multiple surface proteins, and the transportation of a single surface protein requires numerous ICT genes, which poses a challenging barrier for tractable computational modeling. To address this problem, CITE-Traffick dissects these complex networks into small trios, each comprising a specific surface protein expressed on the cell membrane, the transcript of its corresponding coding gene, and the transcript of an ICT gene.

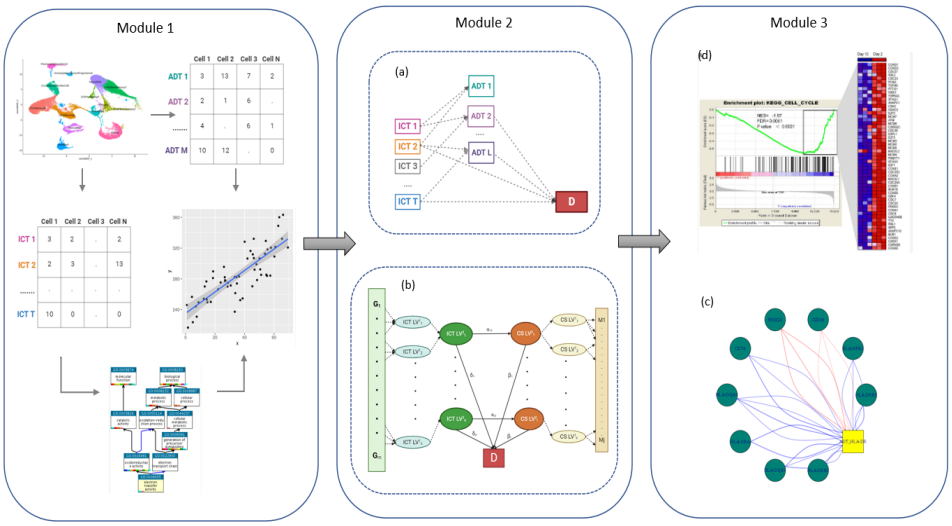

Figure 4. 2: Overview of CITE-Trafficking algorithm and modules. In Module 1, the mRNA and protein count matrices from CITE-Seq are used for running mixed effects regression models to create PTTs. The genes are pre-filtered based on GO terms annotations for biological processes, molecular functions and cellular components related to intracellular protein trafficking. In Module 2, the ICT genes and surface proteins from PTTs are tested for mediation effect through either () Regularized mediation analysis or (b) Structural equation modeling. In Module 3, the protein and ICT modules can be further analyzed for visualization as shown in (c) or functional annotation and interpretation through enrichment analysis as shown in (d).

4.3.1 Identifying Intracellular Trafficking (ICT) Genes

To identify the Gene Ontology (GO) terms associated with intracellular trafficking of proteins in cells, we utilized the Go.db R package (Ashburner et al., 2000). This package provides a convenient interface to access and analyze GO annotations and terms. For the biological processes (BP), we performed a search using the keywords 'intracellular protein transport,' 'intracellular transport,' and 'vesicle-mediated transport' within the GO database. We curated the parent terms related to these keywords and retrieved all their offspring terms. Similarly, for molecular functions (MF), we conducted a search using the keywords 'protein carrier activity,' 'folding chaperon,' and 'chaperone binding.' We selected the parent terms and obtained all their offspring terms.

Lastly, for cellular components (CC), we selected GO terms related to organelles participating in both endocytic and exocytic pathways of protein trafficking. These terms included vacuole, endoplasmic reticulum, lysosome, Golgi apparatus, endosome, vesicle, plasma membrane, intermediate compartment, cytoplasmic microtubule, autophagosome, and microtubule bundle. All genes from the mRNA count matrix that passed the quality control and were used for cell type clustering were included as input for the Gene Ontology (GO) analysis. By including all genes in the GO analysis, we aimed to capture a broad range of molecular functions, biological processes, and cellular components relevant to the intracellular trafficking observed in the dataset. We applied a filtering step to refine the selection of GO terms and genes by ensuring that the genes associated with biological processes and molecular functions were also linked to the cellular components involved in intracellular trafficking. This filtering step allowed us to focus specifically on the genes that were functionally relevant to both biological processes and molecular functions associated with intracellular trafficking, and that were localized within the appropriate cellular components. **Table 4.1**, **4.2**, and **4.3** give a selected list of GO terms for MF, BP, and CC

Table 4. 1: GO terms for MF related to intracellular transport.

GO	ONTOLOGY	TERM
GO:0044183	MF	protein folding chaperone
GO:0051087	MF	chaperone binding
GO:0140597	MF	protein carrier activity

Table 4. 2: GO terms for BP related to intracellular protein transport.

GO	ONTOLOGY	TERM
GO:0006886	BP	intracellular protein transport
GO:0006888	BP	endoplasmic reticulum to Golgi vesicle-mediated transport
GO:0006890	BP	retrograde vesicle-mediated transport, Golgi to endoplasmic reticulum
GO:0006891	BP	intra-Golgi vesicle-mediated transport
GO:0006892	BP	post-Golgi vesicle-mediated transport
GO:0016192	BP	vesicle-mediated transport
GO:0019060	BP	intracellular transport of viral protein in host cell
GO:0030705	BP	cytoskeleton-dependent intracellular transport
GO:0032386	BP	regulation of intracellular transport
GO:0032387	BP	negative regulation of intracellular transport
GO:0032388	BP	positive regulation of intracellular transport
GO:0033157	BP	regulation of intracellular protein transport
GO:0046907	BP	intracellular transport
GO:0048219	BP	inter-Golgi cisterna vesicle-mediated transport
GO:0060627	BP	regulation of vesicle-mediated transport
GO:0060628	BP	regulation of ER to Golgi vesicle-mediated transport
GO:0075521	BP	microtubule-dependent intracellular transport of viral material towards nucleus
GO:0075733	BP	intracellular transport of virus
GO:0090316	BP	positive regulation of intracellular protein transport
GO:0099003	BP	vesicle-mediated transport in synapse
GO:1901253	BP	negative regulation of intracellular transport of viral material
GO:1901254	BP	positive regulation of intracellular transport of viral material
GO:1902953	BP	positive regulation of ER to Golgi vesicle-mediated transport
GO:2000156	BP	regulation of retrograde vesicle-mediated transport, Golgi to ER

Table 4. 3: GO terms for CC related to intracellular transport

GO	ONTOLOGY	TERM
GO:0000137	CC	Golgi cis cisterna
GO:0000138	CC	Golgi trans cisterna
GO:0000139	CC	Golgi membrane
GO:0005783	CC	endoplasmic reticulum
GO:0005786	CC	signal recognition particle, endoplasmic reticulum targeting
GO:0005788	CC	endoplasmic reticulum lumen
GO:0005789	CC	endoplasmic reticulum membrane
GO:0005790	CC	smooth endoplasmic reticulum
GO:0005791	CC	rough endoplasmic reticulum
GO:0005793	CC	endoplasmic reticulum-Golgi intermediate compartment
GO:0005794	CC	Golgi apparatus
GO:0005795	CC	Golgi stack
GO:0005796	CC	Golgi lumen
GO:0005797	CC	Golgi medial cisterna
GO:0005798	CC	Golgi-associated vesicle
GO:0005801	CC	cis-Golgi network
GO:0005802	CC	trans-Golgi network
GO:0012507	CC	ER to Golgi transport vesicle membrane
GO:0012510	CC	trans-Golgi network transport vesicle membrane
GO:0017119	CC	Golgi transport complex
GO:0030130	CC	clathrin coat of trans-Golgi network vesicle
GO:0030134	CC	COPII-coated ER to Golgi transport vesicle
GO:0030140	CC	trans-Golgi network transport vesicle
GO:0030142	CC	COPI-coated Golgi to ER transport vesicle
GO:0030173	CC	integral component of Golgi membrane
GO:0030176	CC	integral component of endoplasmic reticulum membrane
GO:0030660	CC	Golgi-associated vesicle membrane
GO:0030867	CC	rough endoplasmic reticulum membrane
GO:0030868	CC	smooth endoplasmic reticulum membrane
GO:0031205	CC	endoplasmic reticulum Sec complex

4.3.2 Formation of ICT Trios

Given a surface marker M, we identified its coding gene R from the HUGO Gene Nomenclature Committee (HGNC) database. For surface markers with different isoforms, splice variants, or multiple subunits, we retrieved all the corresponding coding genes. For example, the CD3 surface protein has 3 subunits coded by the CD3E, CD3D, and CD3G genes. For each coding gene, we constructed unique trios that included the gene itself, the corresponding protein, and one of the ICT genes meticulously curated using the GO database. These trios, named ICT Trios, were designed to address the complex nature of cellular trafficking, where multiple ICT genes collaborate in transporting diverse proteins. Recognizing that the roles and cellular locations of ICT genes can vary significantly based on the specific protein being transported, our approach systematically explores potential interactions across a spectrum of ICT genes. By subjecting each trio to individual testing, we ensure a comprehensive examination that leaves no potential interaction unexplored, enriching our understanding of this complex system.

4.3.3 Module I: Inferring Putative Transportation Trios

The algorithm starts by establishing the regulatory relationship within each trio. Considering the cell-type specificity of surface protein expression, CITE-Traffick analyzes different cell populations separately. For cells belonging to the same cell type and from the same disease group, we model the expression of a cell surface protein *m* as an outcome of the transcription of its coding gene *g* and an ICT gene *t*.

Given a cell c from the sample s, we denote $E_{m,c}$ as the protein abundance measured by ADT, $X_{g,c}$ and $X_{t,c}$ as the transcript abundance of its coding gene and an ICT gene, respectively, measured by scRNA-seq. For cells of the same type from the same phenotypic group, we build a mixed-effect linear regression model given by,

$$E_{m,c} = \alpha_0 + \alpha_1 X_{g,c} + \alpha_2 X_{t,c} + \sum \beta V + (s_c) + \epsilon \qquad [4.4]$$

where $\alpha_0$is the intercept, $\alpha_1$and α_2 are coefficients representing the fixed main effect of $X_{g,c}$ and $X_{t,c}$, respectively, V is a set of covariates each with a β coefficient, $1|s_c$ represents the random effect accounting for multiple cells from the same sample s, and ϵ is Gaussian-distributed errors. To correct for multiple comparisons, we adjust the nominal p values using the Benjamini-Hochberg method and calculate the false discovery rate (FDR). Non-zero coefficients at FDR<0.05 indicate significant associations. Because the cell surface expression of a protein is closely connected to the transcription of its coding gene, we expect α_1 to be significant. Furthermore, significant α_2 indicates the transcription level of the ICT gene is associated with the expression level of the surface protein. We fit this model to each trio, and those with significant non-zero α_1 and $\alpha_2$values are putative transportation trios (PTTs).

4.3.4 Module II: ICT-Protein-Disease Mediation Network

In this module, we aggregate PTTs across different disease groups to build regulatory networks. CITE-Traffick offers two different approaches for this purpose - one based on regularized mediation analysis and the other based on SEM.

4.3.4.1 Modeling the network via regularized mediation analysis

Given a set of surface proteins M and a set of ICT genes T involved in these PTTs, we hypothesize that differential transcriptions of T may be directly associated with the disease status D, or indirectly associated with D by regulating cell surface expressions of M. We model these relationships as a mediation network, in which transcription levels of T are exposures, cell surface expression levels of M are mediators, and D is the outcome (**Fig. 4.3**). The directional arrows show the effects of T on M, effects of M on D, and effects of T on D. Specifically, $\alpha_{i,j}$ is the coefficient of an arrow connecting exposure *i* to mediator *j*, β_j is the coefficient of an arrow connecting mediator *j* to the outcome, and δ_i is the coefficient of an arrow connecting exposure *i* to the outcome.

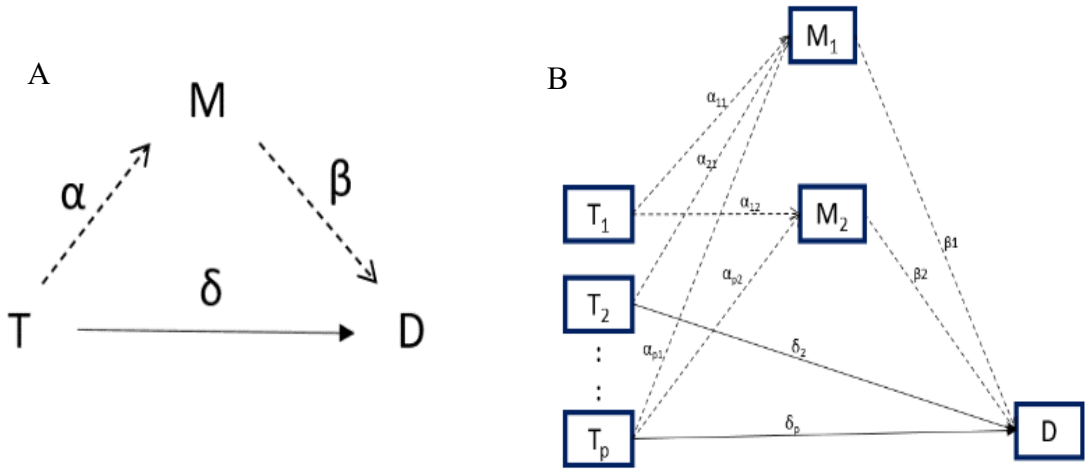

Figure 4. 3: Mediation analysis framework. (A) Single mediator single exposure mediation network with disease status, D as outcome and a single ICT, T as exposure and a single surface protein, M as mediator, (B) Mediation analysis framework with disease status, D as outcome and multiple ICTs T_1, T_2, T_p as exposures and multiple surface proteins M_1 and M_2 as mediators. An arrow connecting an ICT gene to the disease status indicate direct association. An arrow connecting an ICT gene to a cell surface protein that is in turn connected to the disease status indicates indirect association mediated by cell surface protein. $\alpha_{i,j}$ is the coefficient of an arrow connecting exposure *i* to mediator *j*, β_j is the coefficient of an arrow connecting mediator *j* to the outcome, and δ_i is the coefficient of an arrow connecting exposure *i* to the outcome

Not all ICT genes transport all cell surface proteins and not all differential expressions cause diseases. Arrows with coefficients α_{ij}, β_j, or $\delta_i = 0$ indicate the corresponding effects are not detected. We formulate this task as a regularized multivariate mediation analysis (Schaid et al., 2022). Regmed package in R implements penalized model for mediation analysis with multiple exposures, multiple mediators, and multiple outcomes using lasso (L1) type penalty for all coefficients α, β, and δ. Specifically, our task is to find a set of edges with non-zero coefficients that maximize the log-likelihood lnL of the model with L1 penalties P,

$$ln\ ln\ L\ + P \qquad [4.5]$$

The $ln\ L$ is calculated as:

$$ln\ L\ =\ ln\ (det\ (\sum)\ )\ +tr(\sum^{-1}\Lambda) \qquad [4.6]$$

where $det\ (\sum)$ is the determinant of the joint covariance matrix $\sum$ of exposures, mediators, and outcomes, *tr()* is the trace of a matrix, and Λ is the sample covariance matrix. The L1 penalties P is calculated as:

$$P(\alpha,\beta,\delta;\ \lambda) = \lambda\left(w(q,r)\sum_{i=1}^{q}\ \ \sum_{j=1}^{r}\ \ |\alpha_{i,j}|\ + w(r,1)\sum_{j=1}^{r}\ \ |\beta_j|\ + w(q,1)\sum_{i=1}^{q}\ \ |\delta_i|\right)$$

[4.7]

where λ is the regularization parameter, *w(q,r)*, *w(r,1)*, and *w(q,1)* are weight functions $w(d_1,d_2) = (d_1d_2)^4$ to balance the number of exposures *q* and the number of mediators *r*. The λ value controls the sparsity of the solutions - a large λ leads to strong penalization and subsequently few non-zero coefficients. Instead of choosing one λ value, we perform stability selection in which a series of values between 0.01 4 and 0.4 is tested and the edges received non-zero coefficients in at least 10% of tests are retained.

Various types of associations are represented in this model. Full mediation is characterized by a path where an arrow from T_i to M_j, followed by an arrow from *Mj* to D. In this scenario, the association between the ICT gene and disease status is completely explained by the mediating presence of the cell surface protein. On the other hand, in partial mediation, the ICT gene exhibits a direct effect on disease status, as indicated by the arrow connecting T_i to D, in addition to the mediation path involving T_i to M_j and M_j to D.

Conversely, in cases of no mediation effect, the ICT gene solely exerts a direct impact on the disease status, represented by an arrow from T_i to D, while lacking any arrows in the mediation path encompassing T_i to M_j and M_j to D. This absence signifies that the association between the ICT gene and disease status does not involve dysregulated transportation but rather results from other pathways or mechanisms.

The mediation network described above includes multiple ICT genes and multiple cell surface proteins, allowing to model many-to-many relationships. However, if the focus is to examine the effects of a single ICT gene or the transportations of a single surface protein, it can be easily reduced to a single-exposure-multiple-mediator model or multiple-exposure-single-mediator model, respectively. In these cases, only PTTs involving the specific ICT gene, or the specific surface protein will be aggregated.

Two step residual regression to test for the influence of ICT on protein expression

To investigate the extent to which the effect on the abundance of surface protein could be explained by the expression profile of selected ICT genes, we also conducted a two-step regression analysis. In the first step, we regressed the protein expression on coding gene expression and all other relevant control variables to estimate the total effect.

The regression equation for the first step is given by:

$$S_{m,c} = \alpha_0 + \alpha_1 X_{g,c} + \alpha_2 D_c + \sum \beta V + r \qquad [4.7]$$

where $S_{m,c}$ is the protein abundance measured by ADT, $\alpha_0$is the intercept, α_1 is the coefficient representing the main effect of coding gene $X_{g,c}$, α_2 is the coefficient representing the effect of disease status D_c, V is a set of covariates each with a β coefficient.

The residuals (r) in the regression represent the unexplained portion of the dependent variable that cannot be accounted for by $X_{g,c}$ and V. The regression equation for the second step is as follows:

$$r_c = \beta_0 + \beta_1 X_{t,c} + \beta_2 X_{t,c} * D_c + \epsilon \qquad [4.7]$$

The coefficient β_1 and β_2 associated with ICT gene $X_{t,c}$ and the interaction of $X_{t,c}$ with disease $D_c$represents the additional effect of $X_{t,c}$ and its interaction withD_c on the protein expression after accounting for the control variables. If all ICT genes tested have a significant effect on the residuals (significant β_1 and β_2 with p-value < 0.05), this will mean that the ICT genes could explain a significant portion of the variance in the protein expression that was not already accounted for by its coding gene expression and other control variables like age, sex etc.

4.3.4.2 Modeling the network via SEM

The transport and expression of cell surface markers is not solely governed by isolated gene activities, but rather by the intricate web of interactions among ICT genes within the complex framework of exocytic and endocytic pathways. These pathways collaboratively shape the expression and dynamics of surface markers present on the cell's plasma membrane. It is crucial to note that numerous surface proteins participate in complex interplays, yielding a significant influence on both the cellular landscape and the ultimate disease outcome. Importantly, these interplays remain unobservable and lie beyond the realm of direct measurement. While regularized mediation analysis offers valuable insights into the relationships between intracellular trafficking genes and proteins, it has a limitation that it does not consider their intricate interconnectedness. Moreover, as the number of exposures and mediators increases within the framework of regularized mediation analysis, the inference of appropriate regularization parameters becomes increasingly computationally intensive. The Lasso method used in regularization forces the coefficients of less important factors to shrink to zero, potentially eliminating connections between exposures and mediators. Consequently, surface proteins that exhibit a mediation effect within a single mediator model may be overlooked in a multiple mediator model, owing to the dominance of other mediators within the model. Acknowledging these limitations and recognizing the significance of complex interconnections within the cellular landscape, we introduce a novel Structural Equation Modeling (SEM) framework for ICT mediation analysis.

The CITE-Traffick SEM framework is designed to address the unique challenges posed by the interplay of numerous exposures and mediators, offering a comprehensive and interpretable solution. The methodological approach undertaken in this research is driven by two major goals.

- Uncovering the latent structure of interactions among the ICT genes and surface proteins. CITE-Traffick uses a SEM framework featuring two measurement models to achieve this goal. The measurement model 1 reveals the latent framework of ICT gene interconnections within the endocytic and exocytic pathways, wherein endocytic and exocytic pathways function as latent variables.

  The measurement model 2 encapsulates a latent variable that signifies the collective impact of interplaying surface proteins on disease outcomes, encompassing both positive and negative effects on the disease (refer to **Fig 4.4**).
- Modeling a structural relationship between ICT genes and disease, mediated through surface protein expression. The structural model of the SEM framework is designed for conducting mediation analysis with the latent variables representing ICT gene interactions acting as exposures and the protein latent variables acting as mediators. This connection allows us to infer the role of the latent structure of ICT genes, operating in concert, in influencing the dynamics of surface proteins and investigating the regulatory circuits underlying disease outcome (refer to **Fig 4.4**).

The SEM model in CITE-Traffick establishes a sophisticated yet interpretable network to examine the relationships between ICT genes, cell surface proteins, and disease outcomes.

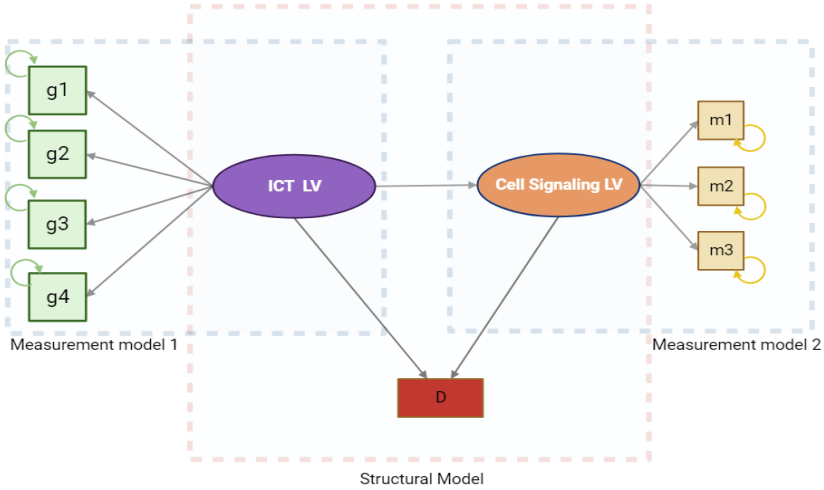

Figure 4. 4: Simplified illustration of the structural equation modeling (SEM) framework for ICT. The framework includes the ICT genes (g1, g2, g3, and g4) as indicators in measurement model 1, surface markers (m1, m2, m3) as indicators in measurement model 2., ICT LV is the latent variable in the measurement model 1, and Cell Signaling LV is the latent variable in measurement model 2. Disease D is the outcome predicted from the ICT LV and Cell Signaling LV. The self-pointing arrows on the indicators represent variances.

SEM Structural Model

The construction of latent variables is a pivotal yet challenging component of the SEM model in multi-omics studies. The measurement models shall aggregate hundreds to thousands of ICT genes and surface proteins into latent variables, all while balancing the intricate complexity of the relationships between intracellular trafficking pathways and disease outcomes. An intuitive solution is to use all ICT genes and surface proteins as inputs to construct all latent variables. However, not all ICT genes participate in the transportation of all surface proteins.

We therefore design a strategy to group ICT genes and surface proteins into subsets such that molecules within each subset likely participate in a common pathway. This strategy aims to strike an equilibrium between comprehensiveness and practicality, ensuring that our model remains informative and manageable.

The PTTs identified in the first module of the CITE-Traffick algorithm provide a foundation for selecting ICT genes and surface proteins of interest. While clustering genes and proteins based on their pairwise correlation could help us create partitions, the large number of cells with gene expression near the average level has a large influence on the clustering results. Thus, it falls short in representing the variances across the samples in the dataset. To address this issue, we employ the Principal Component Analysis (PCA) to unveil ICT genes and surface proteins that contribute to a substantial amount of expression variances across cells.

Specifically, we apply PCA to the expression matrix of significant surface proteins indicated in the PTTs. We then extract eigenvectors and eigenvalues of each surface protein which quantifies the variances attributed to each component. To determine the optimal number of principle components (PCs) to retain, a scree plot of the eigenvalues associated with each PC is utilized. The eigen values corresponding to the selected PCs are organized into a matrix with surface proteins in rows and eigen values from each PC in columns. We then use this matrix as input for hierarchical clustering analysis to group surface proteins into a tree structure. Next, the determination of clusters is a crucial step in the analysis. The hierarchical clustering dendrogram proves instrumental in identifying the optimal number of protein modules for including in the SEM model. Visual examination of the dendrogram is conducted to identify the

segmentation. A level is chosen to result in clusters containing more than three proteins, which aids in ensuring meaningful and statistically robust clusters. In cases where clusters consist of fewer than three proteins, they are merged with the nearest neighboring cluster. Clustering defines distinct groups of proteins, which we term "protein modules." Within each protein module, we identify ICT genes significantly associated with the transportation of the proteins, as elucidated by the PTTs. We then apply the same PCA analysis and hierarchical clustering analysis to these ICT genes, yielding protein-specific gene modules. **Fig 4.5** demonstrates the steps involved in the formation of latent variables through PCA and hierarchical clustering with respect to ICT genes. Note that this analysis yields disjoint protein modules, i.e., no shared surface proteins between modules. However, the gene modules allow overlaps, i.e., the same ICT gene can help transport multiple surface proteins.

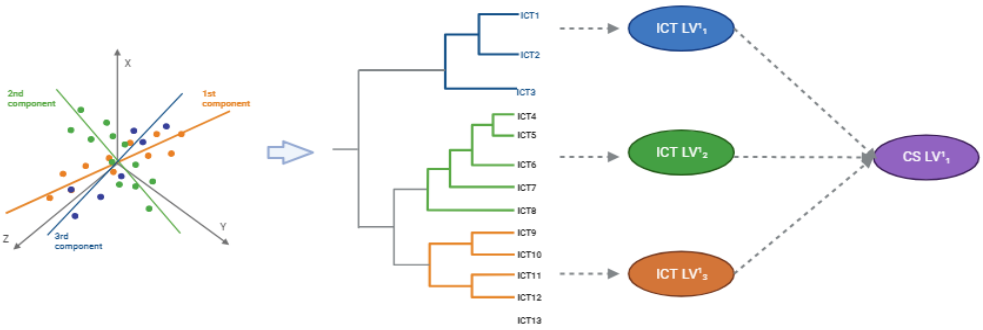

Figure 4. 5: PCA eigen vector loadings used for hierarchical clustering of proteins and ICT genes

For each module composed of at least three proteins or ICT genes, we build a SEM model to translate them into a latent variable (**Fig 4.6**). In addition to constructing latent variables from individual protein and ICT gene modules, we create higher-order latent variables. Higher-order latent variables are formed from other latent variables, and they represent a higher level of abstraction in molding complex relationships. With these higher-order constructs, we not only gain a more comprehensive understanding of the underlying biological processes but also reduce the computational complexity of our model. We represent latent variables created from ICT genes as ICT LV with first order latent variables written as CIT LV^1 and second order latent variables as ICT LV^2, similarly for latent variables created from proteins we represent using symbols CS LV^1 (Cell signaling LV) for first order and CS LV^2 as 2nd order.

Using these latent variables (first order and second order) and their contributing ICT genes and surface proteins, we construct the structural model for mediation analysis in the SEM framework. The latent variables associated with ICT genes, ICT LV function as exposure variables, while those representing proteins, CS LV serve as mediators. Within the SEM framework, we calculate direct and indirect effects of ICT modules on disease outcome. Each path in the SEM model is accompanied by coefficients and p-values, which offer statistical significance assessments for the relationships examined. The significance of direct and indirect paths enables us to infer the nature of mediation, distinguishing between partial and full mediation.

Fig 4.6. provides an illustrative representation of the final SEM framework tailored for CITE-Traffic. Within this framework, ICT genes G_1, G_2, ..., G_m serve as indicators in Measurement Model 1, while surface markers M_1, M_2, …, M_j act as indicators in Measurement Model 2. The latent variables ICT LV^1_1, ICT LV^1_2, … ICT LV^1_n emerges from the ICT gene modules, where some of these latent variables further combine to form a higher-order latent variable, for example ICT LV^2_1. Similarly, protein modules are derived from the protein latent variables CS LV^1_1, CS LV^1_2, … CS LV^1_n. The arrows in the diagram pointing from the gene latents (ICT LV^2_1 and ICT LV^1_n) towards the protein latents (CS LV^1_1, … CS LV^1_n), represent the associations between ICT gene latents and protein latents. The direct effects of ICT latents on disease outcome, D are indicated by arrows with coefficients $\boldsymbol{\delta}_1$,..,$\boldsymbol{\delta}_n$. The effects of the ICT latents on D are represented by $\boldsymbol{\alpha}_1$,..,$\boldsymbol{\alpha}_n$ and $\boldsymbol{\beta}_1$,..., $\boldsymbol{\beta}_n$ represents the effects of CS LVs on D. The combined pathway involving both "$\boldsymbol{\alpha}$" and "$\boldsymbol{\beta}$" coefficients illustrates the indirect effects of the ICT latents on disease outcomes. In the SEM structural model, a rigorous testing process is undertaken to evaluate the significance of direct and indirect effects. This analytical approach provides insights into whether the mediation is partial or full, shedding light on the complex interplay between ICT genes, surface markers, and disease outcomes.

The SEM framework is implemented using the Lavaan package in R (*Structural Equation Modeling with Lavaan | Wiley*, n.d.), a well-regarded approach for Structural Equation Modeling (SEM). Lavaan provides a versatile platform for specifying and estimating complex relationships among observed and latent variables.

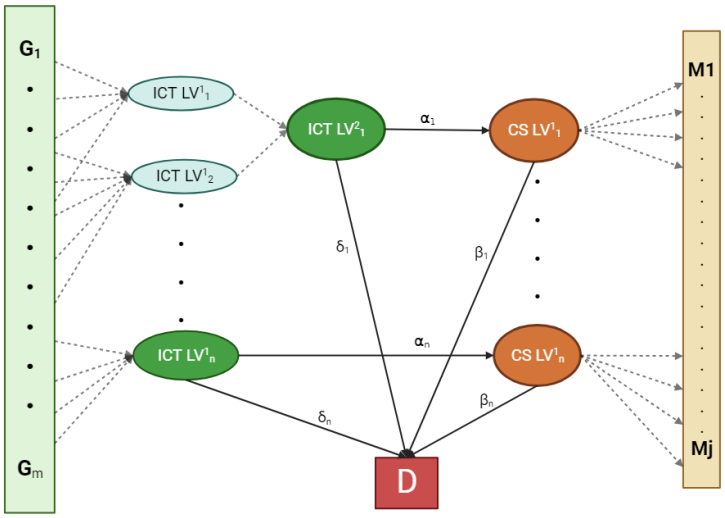

Figure 4. 6: First order and second order latent variable construction in CITE-Traffick SEM model. Modules of ICT genes form first order latent variables $ICT\ LV^{1}_{1}$, $ICT\ LV^{1}_{2}$, ... $ICT\ LV^{1}_{n}$. Some of these latent variables are combined to form second order latent variables for example, $ICT\ LV^{2}_{1}$. The surface marker modules are derived from the first order protein latent variables $CS\ LV^{1}_{1}$, ..., $CS\ LV^{1}_{n}$. The arrows pointing from the gene latents $ICT\ LV^{2}_{1}$ and $ICT\ LV^{1}_{n}$ towards the protein latents $CS\ LV^{1}_{1}$ and $CS\ LV^{1}_{n}$ give the association between gene latents and protein latents. The direct effects of ICT latents on disease outcome, *D* are indicated by arrows with coefficients $\boldsymbol{\delta}_1,\ldots\boldsymbol{\delta}_n$. The effects of the ICT latents on D are represented by $\boldsymbol{\alpha}_1,\ldots\boldsymbol{\alpha}_n$ and $\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_n$ represents the effects of CS LVs on D. The combined pathway involving both "$\boldsymbol{\alpha}$" and "$\boldsymbol{\beta}$" coefficients illustrates the indirect effects of the ICT latents on disease outcomes. Disease D is the outcome predicted from the latents.

Because of the increased complexity in modeling all the ICT genes and proteins in the same model, each protein-ICT module can be tested in separate models. For example, a model can be formed with the ICT latent variable $ICT\ LV^{2}_{1}$ as exposure and protein latent variable $CS\ LV^{1}_{1}$ as mediator and D as outcome. The structural part of the SEM will include the computation of the first order latent variables $ICT\ LV^{1}_{1}$, $ICT\ LV^{1}_{2}$ from their corresponding ICT genes, the computation of the second order ICT latent variable $ICT\ LV^{2}_{1}$ from $ICT\ LV^{1}_{1}$, $ICT\ LV^{1}_{2}$ and the computation of the ADT latent variable $CS\ LV^{1}_{1}$ from its corresponding ADTs.

Evaluation of SEM model fit and restructuring of SEM

To assess the fit of an SEM model, multiple measures and indices come into play. For instance, the chi-square test serves to examine the null hypothesis that the anticipated model and the observed data are equivalent. An insignificant result from this test is indicative of a well-fitting model. However, it's worth noting that the chi-square test has a drawback in that it is highly sensitive to sample size. As the sample size increases, the likelihood of obtaining a statistically significant chi-square result also grows. This sensitivity becomes especially pronounced in our case, where we are working with single-cell data, causing the chi-square test to produce significant outcomes even when using higher significance cutoffs, such as .01 or .001. As a result, we turn to a range of other widely accepted fit measures, including RMSEA, CFI, SRMR, IFI, NFI, GFI, PNFI, and RFI, (Schumacker & Lomax, 2010) (L. Hu & Bentler, 1999) to offer a more comprehensive and robust evaluation of model fit and performance. These metrics collectively provide a more nuanced and sample-size-independent assessment of the model's adequacy.

In our pursuit of the optimal SEM structure, we employ a systematic forward selection process. Initially, all first-order latent variables are individually integrated into models, and the resulting model fit is meticulously evaluated. Subsequently, we explore various combinations of these latent variables, each time assessing the model fit. If a combination of latent variables enhances the model fit, we retain the current SEM structure; otherwise, we proceed to test the next combination.

This iterative approach allows us to derive models that may consist exclusively of first-order latent variables or incorporate specific combinations of latent variables to construct second-order latent variables, depending on which model yields the most favorable fit measures. It's important to note that, in this study, we restrict our focus to first-order and second-order latent variables and do not consider higher-order structures.

4.3.5 Module III: Integration of Differential Gene Expression and Gene Set Enrichment Analysis

Once dysregulated ICT genes are identified, it becomes crucial to understand their involvement in other immune-related pathways and how their dysregulation impacts the overall disease phenotype. Differential gene expression analysis allows for the identification of genes that exhibit significant changes in expression levels between disease and control groups. Subsequently, GSEA enhances our understanding of the dysregulated genes by assessing their enrichment within pre-defined sets of genes representing specific biological pathways, functions, or disease signatures.

In module three, we conduct DEG and GSEA analyses to identify whether the ICT genes are significantly overrepresented or underrepresented immune-related gene sets, thus providing insights into the broader immune pathways affected by these genes. This knowledge provides insights into the broader molecular mechanisms underlying the disease and aids in comprehending the specific roles of these genes in disease progression, immune dysregulation, and potential therapeutic interventions.

4.3.6 Performance evaluation and comparisons

In the conclusive phase of our methodology, we conducted an exhaustive evaluation of the CITE-Trafficking approach's ability to predict disease outcomes. This comprehensive assessment involved a meticulous comparison between ICT genes, ADTs, and latent variables extracted from both ICT genes and proteins, as identified by the CITE-Traffick algorithm, and established markers. To establish a solid benchmark, we identified genes and proteins that exhibited differential expression between comparison groups using the Wilcoxon rank sum test.

Following this, we divided our dataset into training and test sets, adhering to an 80:20 ratio. Within the training set, we performed rigorous stability testing for feature selection, exploring a range of lambda values from 0.3 to 0.0003. This initial feature selection was limited to differentially expressed genes and ADTs. For each lambda value, we executed a ten-fold cross-validation, further enhancing the feature set using LASSO logistic regression. The top 50 features with the highest frequency across iterations formed the foundation for our multi-variable Generalized Linear Model (GLM) on the untouched test set.

In the subsequent phase, our approach extended to encompass various sets of features identified through the CITE-Traffick framework: (i) the complete list of significant ICT genes and ADTs, as selected by CITE-Traffick Module I, (ii) the compilation of ICT genes and ADTs pinpointed by the multiple-mediator-multiple-exposure regularized mediation model in CITE-Traffick Module II, (iii) the latent variables originating from ICT genes and ADTs, as generated within CITE-Traffick SEM in Module II, and (iv) an enriched feature set achieved by incorporating latent variables produced by SEM in CITE-Traffick Module II, in conjunction with the differentially expressed genes and ADTs. This comprehensive approach enabled us to meticulously assess CITE-Traffick's feature selection capabilities, using these findings to make detailed comparisons with DEG features.

Consistently, we applied the same meticulous process of stability testing and Generalized Linear Model (GLM) modeling to all feature sets. The performance of all GLM models was meticulously assessed on the independent test dataset, using a set of robust metrics, including accuracy, F1-score, and AUC-ROC values.

4.4 Advantages of using CITE-Traffick against traditional analysis

Differential expression analysis is limited in its ability to fully elucidate the association between genes, surface proteins, and intracellular trafficking. While it can identify surface proteins and genes that are significantly different between comparison groups, it does not provide a comprehensive understanding of their involvement in protein trafficking. To address this limitation and identify genes that significantly participate in the pathway of protein trafficking, we employed regression analysis.

Correlation analysis could be used to explore the association between surface protein expression and intracellular trafficking genes. However, to gain a more comprehensive understanding of this association, it is essential to employ regression analysis, which allows to incorporate additional predictors and covariates including corresponding mRNA expression, and other covariates such as age, gender, stage of disease, or disease subtype. This comprehensive approach allowed us to assess the independent contributions of each predictor to the association with intracellular trafficking, providing a more nuanced understanding of the molecular mechanisms underlying protein trafficking. Correlation analysis assesses the strength and direction of the linear relationship between two variables but does not account for potential confounding factors. Regression analysis provides a framework to incorporate covariates that can influence the relationship between the surface protein expression and intracellular trafficking genes. In studies involving biological data, it is crucial to consider and control for confounding variables that may influence the observed correlations.

Moreover, regression analysis allows for the estimation of regression coefficients, which can provide insights into the direction and magnitude of the associations between predictors and the outcome variable. These coefficients can help identify which predictors are significantly associated with surface protein expression and quantify the strength of these associations, aiding in the interpretation of the results.

4.5 Application to COVID-19 CITE-Seq Data

We analyzed a CITE-Seq data set (GSE155673) from a previously published study of immunity in COVID-19 patients (Arunachalam et al., 2020). This dataset contains single-cell transcriptome profiles and abundances of 36 cell surface proteins in peripheral blood leukocytes from twelve age-matched individuals, including five healthy controls, three mild COVID-19 cases, and four severe COVID-19 cases.

4.5.1 CITE-Seq Data Clustering and Annotation

The raw CITE-seq count matrices were loaded into R (v4.0.3) and processed using the Seurat R package (v4.1.2). Cells with less than 100 detected genes and genes detected in fewer than 5 cells were filtered out. Cells with mitochondrial gene expression greater than 5% of the total gene expression were also removed. A Seurat object was constructed for both the scRNA and protein data, and the two objects were integrated using the Seurat integration pipeline. The RNA expression levels were normalized using standard normalization to correct for batch effects and the top 2000 highly variable genes were identified for downstream analysis. The protein expression levels were normalized using centered log ratio normalization and scaling. Dimensional reduction using principal component analysis (PCA) was performed on the integrated scRNA and ADT data separately to compute 30 principal components (PC). Clustering was performed on the integrated scRNA and protein assays using Seurat Weighted Nearest Neighbors (WNN) pipeline at a resolution of 0.8, which yielded a total of 15 clusters, each composed of cells originating from healthy, mild, and severe samples. Clusters with less than 5 cells were removed.

To identify marker genes for each cluster, we employed the FindAllMarkersMAESTRO function from the MAESTRO package in R. After identifying marker genes for each cluster, we annotated the clusters using RNAAnnotateCelltype function from the MAESTRO package based on canonical marker genes for immune cell types. Cell type annotation identified 13 cell populations, CD16+ Monocyte, CD4 Naïve T cell, CD8 Effector Memory T cell, Naïve B cell, Erythroid cell, CD14+ Monocyte, Platelet, Hematopoietic Stem and Progenitor cell, CD4 Proliferating T cell, Natural Killer cell, classical Dendritic Cell, plasmacytoid Dendritic Cell, and Plasmablast (**Fig 4.7**).

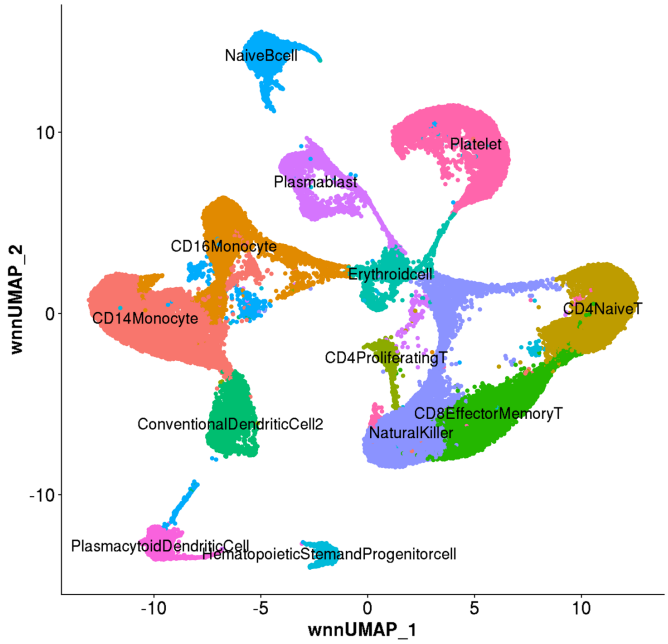

Figure 4. 7: UMAP visualization of CITE-seq cell clustering using Seurat and Azimuth PBMC Reference Annotation with cells from Healthy, Mild, and Severe COVID-19 samples

4.5.2 Identifying PTTs

In the CITE-Traffick first module, we compiled a total of 1,582 genes that are related to intracellular protein trafficking using GO analysis. Out of the 36 ADTs within the CITE-Seq data, we selected 30 ADTs considering the availability of coding genes in the mRNA assay. Recognizing multiple subunits for certain surface proteins, we thoughtfully incorporated the coding genes for all relevant subunits.
This encompassed, CD3D, CD3G, and CD3E for ADT CD3; HLA-DRA, HLA-DRB1, and HLA-DRB5 for ADT HLA-DR; and FCGR1A, FCGR2A, FCGR3A, FCGR1B, FCGR2B, and FCGR3B for ADT CD16. A total of 39 ADT-coding gene pairs, in conjunction with 1582 ICT genes, was subsequently employed to generate a set of 61,698 unique ICT trios.

Independent mixed effects regression models were run for all the ICT trios in all the three comparison groups – healthy, mild, and severe. Based on the significant coefficients in equation [4.1] that represent the association between cell surface protein expression and ICT gene expression, CITE-Trafficking identified 4754, 732, and 2243 PTTs in healthy controls, mild COVID-19 cases, and severe COVID-19 cases, respectively, at FDR<0.05. A total of 1034 unique significant ICTs associated with 25 unique ADTs in Healthy, a total of 637 unique ICTs associated with 19 unique ADTs in Severe and 365 unique ICTs associated with 18 unique ADTs in Mild were identified.

We performed two-way hierarchical clustering of all the PTTs which exhibited clustering patterns showing the association of cell surface protein expression with the transcription of multiple ICT genes, and vice versa, confirming the many-to-many

relationships (**Fig 4.8)**. The cell surface protein associated with the largest number of ICT genes was CD11c (551) and CD16 (489) in healthy controls, CD16 (168) and CD613 (63) in mild cases, and CD16 (301) and HLA-DR (274) in severe cases. Clusters observed in the heatmap indicated ICT genes associated with a common set of proteins. For example, GAS6, DAB2, and RAB3B are associated with surface proteins CD16, CD123, CD163, CD33, etc. in severe, mild, and healthy samples. Transcription of most ICT genes facilitated transportation, as implied by the positive associations (in red color). For a few ICT genes, the transcription level was negatively associated with cell surface protein expression (blue color in heatmap). These genes, such as *STAB1* and *DYSF,* which are negatively associated with CD16 and CD123 surface markers, have GO annotations related to endosome and endocytosis, plausibly participated in the process that internalizes, breaks downs, or recycles cell surface proteins (Kzhyshkowska et al., 2006). There are clusters of ICT genes showing a positive association with certain surface proteins and a negative association with some other surface protein. For example, all ICT genes *MARCO* and *CLEC10A* is positively associated with HLA-DR, CD33, CD38, CD163, and CD14, and it has a negative association with CD16, and CD123.

The two-way clustering results also show that the associations between ICT genes and cell surface protein expression varies by disease status. For example, gene *ACTB* is shown to have a negative association with HLA-DR in healthy (-0.13) and a positive association in severe (0.11). To systematically investigate these changes, CITE-Trafficking was used to build mediation networks comparing ICT processes between disease groups.

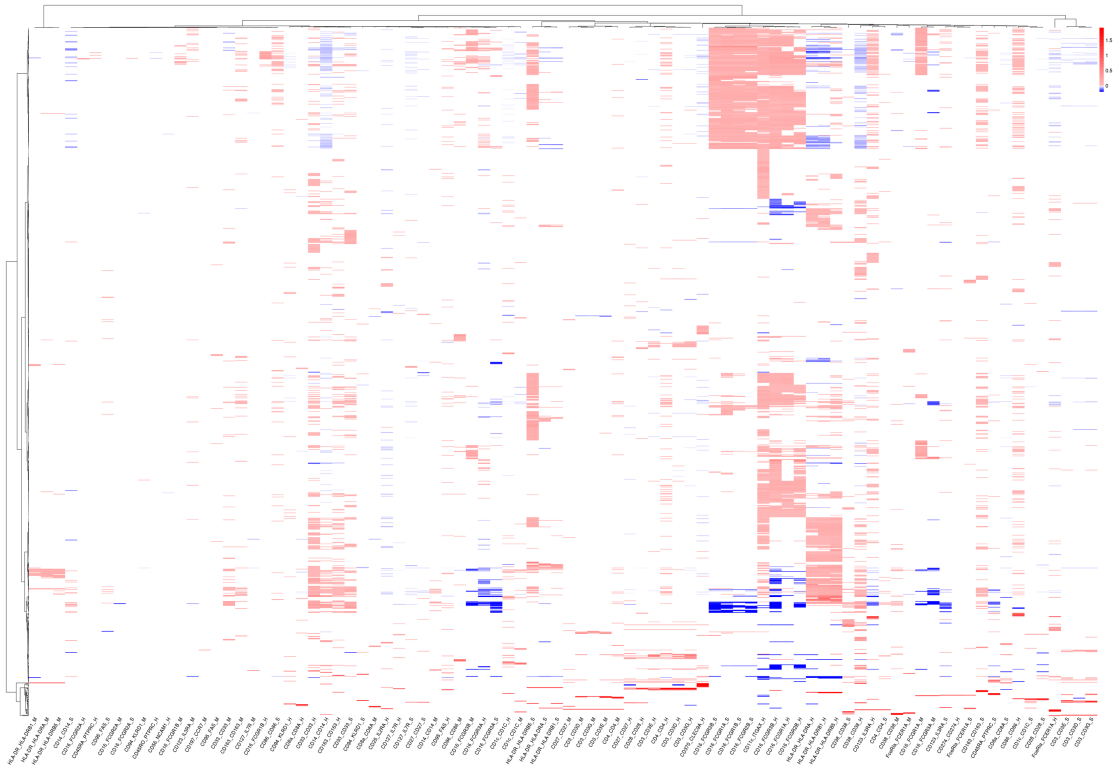

Figure 4. 8: UMAP visualization of CITE-seq cell clustering using Seurat and Azimuth PBMC Reference Annotation with cells from Healthy, Mild, and Severe COVID-19 samples

4.5.3 CITE-Traffick Reveals Dysregulated ICTs Associated with HLA-DR Expression in CD16+ Monocyte

The original study reported underexpression of human leukocyte antigen class DR (HLA-DR) on the surface of monocytes in COVID-19 patients, with the lowest expression level observed in severe cases. HLA-DR is a major histocompatibility complex (MHC) class II molecule that is expressed on antigen-presenting cells (APCs), including monocytes, dendritic cells, macrophages, and B cells. It presents antigen peptides to T cells and activates them. Interestingly, differential expression analysis of

the surface proteins revealed significant under expression of HLA-DR in CD16+ Monocytes of severe compared to healthy and mild (**Fig 4.9**). Therefore, CITE-Traffick analysis was centered around CD16+ Monocytes identified through our cell clustering analysis. To investigate the dysregulation of intracellular trafficking (ICT) genes in the context of low HLA-DR expression in COVID-19 monocytes, we conducted separate comparisons between severe COVID-19 samples and both healthy samples and mild COVID-19 samples. Through this comparative analysis, we aimed to identify ICT genes that exhibited consistent and distinct dysregulation across different severity groups.

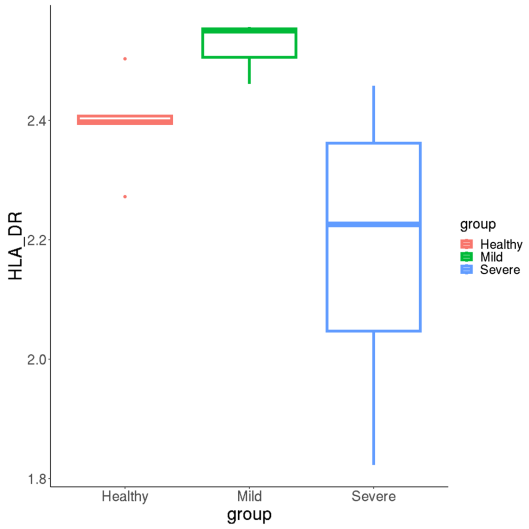

Figure 4. 9: Boxplot showing significant under expression of HLA-DR surface protein in severe compared to healthy controls and mild patients.

4.5.3.1 Severe - Healthy Comparison

In our initial assessment of PTTs for HLA-DR from the mixed effects regression model, we identified 359 ICT genes that exhibited a statistically significant association with HLA-DR expression in either healthy controls or severe COVID-19 cases or both with thresholds set at adjusted p-value < 0.05 and absolute value of regression coefficient > 0.1. Out of the 359, 131(36%) of the genes were uniquely identified in the healthy group, 177(49%) were unique to the severe group and 51(14%) were common to both severe and healthy. We examined the coefficients from the regression model to estimate the effect of ICT genes on surface marker expression.

Some ICT genes showed a consistently positive association with the HLA-DR expression in both severe and healthy. It is interesting to note that some of these ICT genes were also widely reported in COVID related studies. For example, *CD74* from our analysis is shown to be significantly positively correlated with the HLA-DR expression. The HLA-DR surface marker has two subunits, the alpha chain encoded by the *HLA-DRA* gene and the beta chain encoded by multiple *HLA-DRB* genes (*Nomenclature for Factors of the HLA System, 2010 - Marsh - 2010 - Tissue Antigens - Wiley Online Library*, n.d.). Two of these genes (*HLA-DRA* and *HLA-DRB1*) were expressed in a sufficient number of cells (>5) to be tested using the regression model in both severe and healthy.

Trafficking trios identified by the regression model showed that the expression of HLA-DR on cell surface was consistently positively correlated with the transcription of these two coding genes (0.1 - 0.2) as well as the ICT gene *CD74* (0.18 - 0.23) *in* healthy and severe (**Fig. 4.10 A**). These concordant patterns imply that *CD74* is a key ICT gene in trafficking both subunits of HLA-DR. The **Fig 4.10 B** further demonstrates a positive correlation between ICT gene *CD74* and HLA-DR surface protein in Healthy and Severe.

Based on GO annotations, *CD74 is* an integral component of the luminal side of ER membrane (GO:0071556) and part of ER to Golgi transport vesicle membrane (GO:0012507); CD74 participates in intracellular protein transport (GO:0006886) and is a protein folding chaperone (GO:0044183). There are multiple studies indicating the role of CD74 as a crucial protein-binding chaperone for HLA-DR, facilitating the proper assembly and presentation of antigens by major histocompatibility complex class II (MHC-II) molecules (Schröder, 2016). In the study conducted by Kvedaraitea et. al, using high-dimensional flow cytometry analysis on mononuclear phagocyte (MNP) lineages in SARS-CoV-2-infected patients with moderate and severe COVID-19, researchers have found lower expression levels of MHC class II and CD74 in severe COVID-19 patients (Kvedaraite et al., 2021).

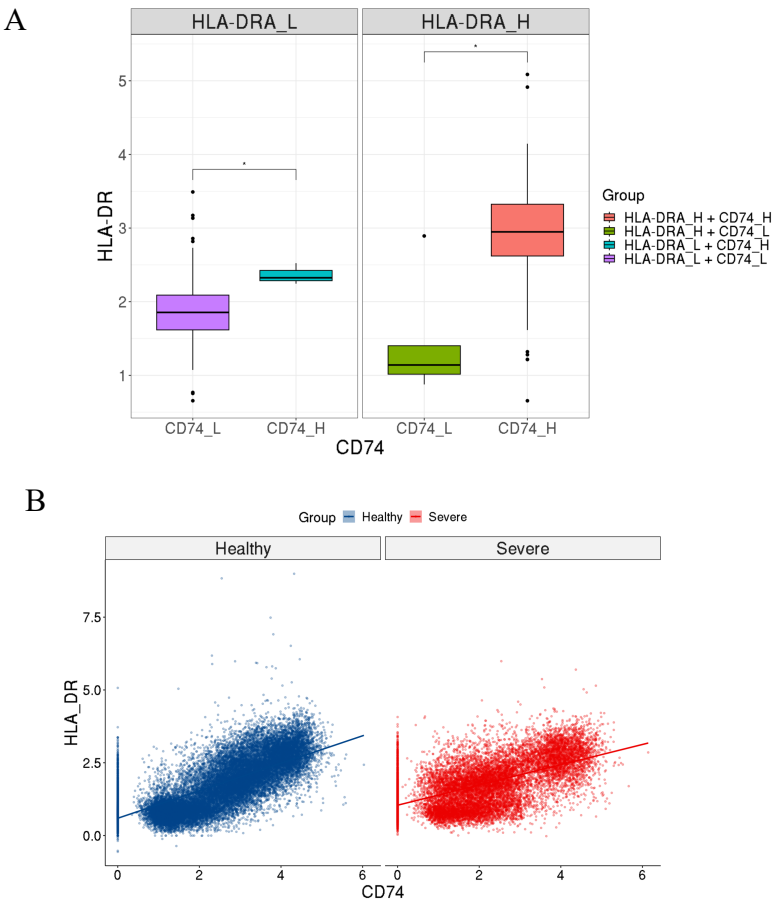

Figure 4. 10: HLA-DR and CD74. (A) Boxplots showing the protein expression of HLA-DR surface protein at varying levels of its coding gene *HLA-DRA* and ICT gene *CD74* with expression level above top 80% and bottom 20% quantiles. (B) Scatter plot showing positive correlation between ICT gene *CD74* and surface protein HLA-DR in both Severe and Healthy

Another example, CLU exhibits a negative association with HLA-DR expression with coefficients ranging between -0.26 and -0.11 in severe and healthy. Trafficking trios identified by the regression model showed that the expression of HLA-DR on cell surface was consistently positively correlated with the transcription of its two coding genes *HLA-DRA and* HLA-*DRB1* (0.1 - 0.2) and negatively correlated with the ICT gene *CLU* in healthy and severe (**Fig. 4.11**).

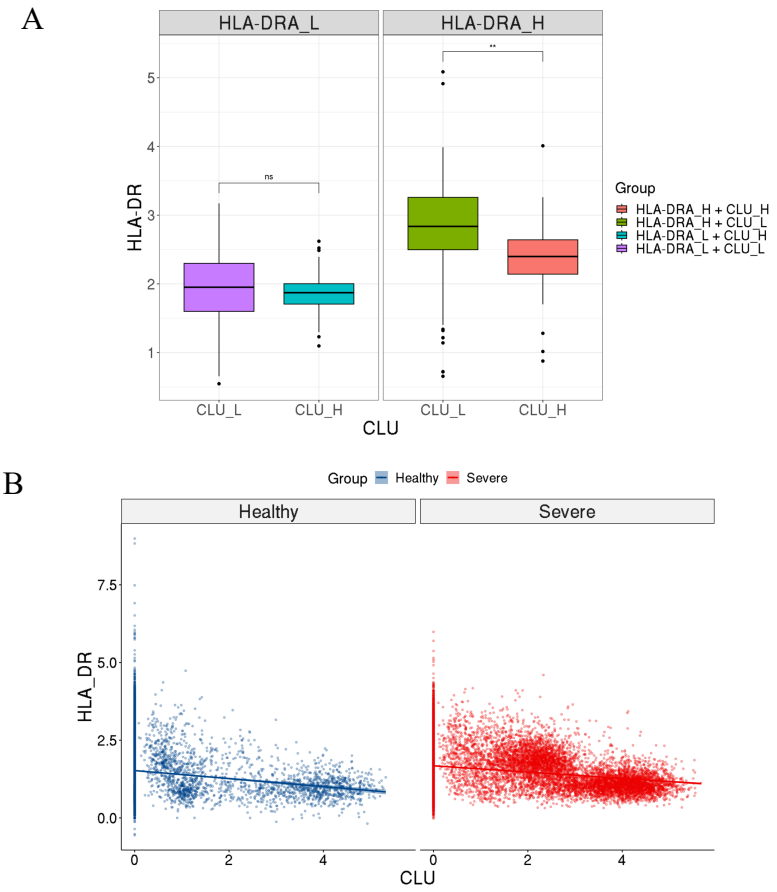

Figure 4. 11: *HLA-DRA* and *CLU.* A) Boxplots showing the protein expression of HLA-DR surface protein at varying levels of its coding gene *HLA-DRA* and ICT gene *CLU* with expression level above top 80% and bottom 20% quantiles. While there is no significant difference in expression level at 20%, the change in expression level at 80% is significant. (B) Scatter plot showing negative correlation between ICT gene *CLU* and surface protein HLA-DR in both Severe and Healthy

On examining the GO terms for CLU, we saw that it is associated with endocytosis and autophagy as it participates in positive regulation of receptor-mediated endocytosis (GO:0048260) and protein targeting to lysosome involved in chaperone-mediated autophagy (GO:0061740). A study has revealed the crucial role of endocytosis in the transport of HLA-DR proteins to the cell surface, as it facilitates the delivery of newly synthesized major histocompatibility complex (MHC) class II molecules from the trans-Golgi network (TGN) to early endosomes, enabling peptide loading and subsequent antigen presentation to CD4+ T lymphocytes (Brachet et al., 1999).

A study on autophagy in the pathology of COVID and its potential therapeutic implications, reviews how viruses, including coronavirus, exploit this cellular process for their replication and, therefore, medications that have modulatory effects on autophagy could be potential treatments against this virus. Several of the ICT genes identified through our method are consistent with the aforementioned process, highlighting the potential relevance of endocytosis and autophagy in the transport of HLA-DR proteins to the cell surface. The regression results for the ICT genes CD74 and CLU and the GO terms associated with them are displayed on **Tables 4.4** and **4.5**.

Table 4. 4: CITE-Traffick regression results for PTTs with ICT genes CD74 and CLU

ICT	mRNA	ICT.coef	ICT.Padj	mRNA.coef	mRNA.Padj
CD74	HLA-DRB1	0.21	0.00	0.10	0.00
CD74	HLA-DRA	0.19	0.00	0.13	0.00
CD74	HLA-DRB1	0.23	0.00	0.15	0.00
CD74	HLA-DRA	0.18	0.00	0.20	0.00
CLU	HLA-DRB1	-0.11	0.00	0.22	0.00
CLU	HLA-DRA	-0.14	0.00	0.24	0.00
CLU	HLA-DRB1	-0.23	0.00	0.28	0.00
CLU	HLA-DRA	-0.26	0.00	0.31	0.00

Table 4. 5: GO terms associated with ICT genes CD74 and CLU

SYMBOL	GO	ONTOLOGY	TERM
CD74	GO:0000139	CC	Golgi membrane
CD74	GO:0005771	CC	multivesicular body
CD74	GO:0005765	CC	lysosomal membrane
CD74	GO:0005773	CC	Vacuole
CD74	GO:0005886	CC	plasma membrane
CD74	GO:0006886	BP	intracellular protein transport
CD74	GO:0009897	CC	external side of plasma membrane
CD74	GO:0012507	CC	ER to Golgi transport vesicle membrane
CD74	GO:0032588	CC	trans-Golgi network membrane
CD74	GO:0042613	CC	MHC class II protein complex
CD74	GO:0043202	CC	lysosomal lumen
CD74	GO:0044183	MF	protein folding chaperone
CD74	GO:0071556	CC	integral component of lumenal side of endoplasmic reticulum membrane
CLU	GO:0005794	CC	Golgi apparatus
CLU	GO:0048260	BP	positive regulation of receptor-mediated endocytosis
CLU	GO:0051087	MF	chaperone binding
CLU	GO:0061740	BP	protein targeting to lysosome involved in chaperone-mediated autophagy
CLU	GO:0097440	CC	apical dendrite
CLU	GO:0099020	CC	perinuclear endoplasmic reticulum lumen
CLU	GO:0140597	MF	protein carrier activity

Regularized Mediation Analysis

To test whether the ICT genes identified are associated with severity of COVID outcome and to test whether this association is mediated by the expression of HLA-DR on cell surface, we ran a single-mediator-multiple-exposure regularized mediation model. All the 359 ICT genes identified from the first module were used as exposures; HLA-DR protein was designed as mediator and severe COVID disease as outcome (severe vs healthy). The results, illustrated in **Fig 4.12** revealed negative association of HLA-DR with disease severity, with a **β** coefficient of -0.05. Only a total of 124 (35%) ICT genes were selected by the model with HLA-DR showing full mediation with 38 (31%) ICT genes with $\boldsymbol{\alpha}$ coefficients ranging from -0.05 to 0.06; partial mediation effects for 20 (16%) ICT genes, where $\boldsymbol{\alpha}$ coefficients ranged from -0.12 to 0.28 and **δ** coefficients ranged from -0.13 to 0.14; and 66 (53%) ICT genes without an indirect effects on disease through HLA-DR.

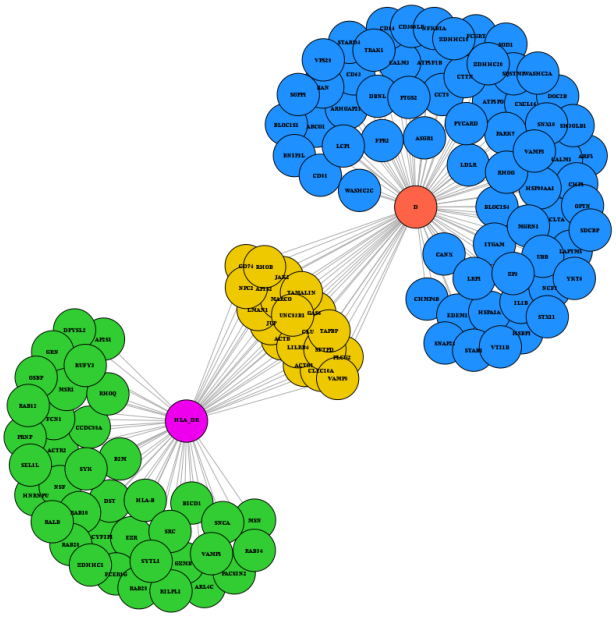

Figure 4. 12: Mediation network with HLA-DR as mediator (in pink), ICT genes as exposures (blue, green, and yellow) and D as outcome (red). The figure highlights the ICT genes with different mediation levels - full mediation (yellow), partial mediation (green) or without any mediation connection (blue)exposure as green circles), the HLA-DR surface marker (mediator as blue circle), and disease outcome (red circle).

On comparing the $\boldsymbol{\alpha}$ coefficients from the regularized mediation model with the mixed effects regression coefficients for the 58 ICT genes, we noticed that while most of the genes shows association in the same direction (positive or negative) some of them were showing association in opposite direction. The ICT genes *CD74* and *CLU* discussed in the previous section consistently exhibit positive and negative association with HLA-DR. The ICT gene *CD74* exhibits partial mediation through HLA-DR, with $\boldsymbol{\alpha}$ = 0.28, and concurrently displays a direct negative association with the disease with $\boldsymbol{\delta}$ = -0.12. This strongly suggests that CD74 plays a critical role in regulating HLA-DR expression, implying that reduced CD74 expression could contribute to the lower levels of HLA-DR expression observed in the context of the disease. The ICT gene *CLU* also exhibits partial mediation through HLA-DR wth $\boldsymbol{\alpha}$ = -0.12 and $\boldsymbol{\delta}$ = 0.192 re confirming negative association of *CLU* gene on HLA-DR expression and also positive association with the disease. The **Table 4.6** provides a comprehensive overview of ICT genes derived from the regularized mediation model with HLA-DR as the mediator.

Table 4. 6: Overview of ICT genes derived from regularized mediation model for HLA-DR

Mediation	ICT	α	Δ
Full	ACTR2 AP2S1 ARL4C B2M BICD1 CCDC88A CYFIP1 DPYSL2 DST EZR FCER1G FCN1 GRN GZMB HLA-B HNRNPU MSN MSR1 NSF OSBP PACSIN2 PRNP RAB10 RAB12 RAB20 RAB28 RAB34 RALB RHOQ RILPL1 RUFY3 SEL1L SNCA SRC SYK SYTL1 VAMP5 ZDHHC1	-0.05 - 0.06	
Nil	ABCG1 ARF5 ARHGAP21 ASGR1 ATP5F1B ATP5PO BLOC1S1 BLOC1S4 BNIP3L CALM1 CALM3 CANX CCT8 CD14 CD300LF CD63 CD81 CHMP4B CHP1 CLTA CTTN CXCL16 DBNL DOC2B EDEM1 FCGRT FPR2 HSBP1 HSP90AA1 HSPA1A IL1B ITGAM LAPTM5 LCP1 LDLR LRP1 MGRN1 NCF1 NFKBIA OPTN PARK7 PTGS2 PYCARD RAN RHOG SDCBP SGPP1 SH3GLB1 SNAP23 SNX10 SOD1 SQSTM1 STAB1 STARD4 STX11 TRAK1 UBB VAMP3 VPS29 VTI1B WASHC2A WASHC2C YKT6 ZDHHC18 ZDHHC20 ZP3		-0.07 - 0.26
Partial	ACTB ACTG1 AP1S2 CD74 CLEC10A CLU GAS6 JAK2 JUP LILRB4 LMAN1 MARCO NPC2 PLCG2 RHOB SFTPD TAMALIN TAPBP UNC93B1 VAMP8	-0.12 - 0.28	-0.13 - 0.15

The **Fig 4.13** visually illustrates the distribution of various ICT genes, highlighting both positive and negative associations with HLA-DR transport, across diverse cellular components within the realm of intracellular trafficking. It is evident that ICT genes participate in both exocytic and endocytic pathways within the cellular compartment. For example, ICT genes CLU and MARCO are seen participating in endocytosis by forming part of the early endosome and late endosome.

The ICT genes like GRN and TAPBP which are part of the ER and Golgi, RAB10 and NSF are part of exocytosis. We can also find ICT genes that are helping with movement of protein from Golgi to the plasma membrane where they get embedded. Also, ICT genes part of lysosome are seen helping with recycling of proteins back to ER.

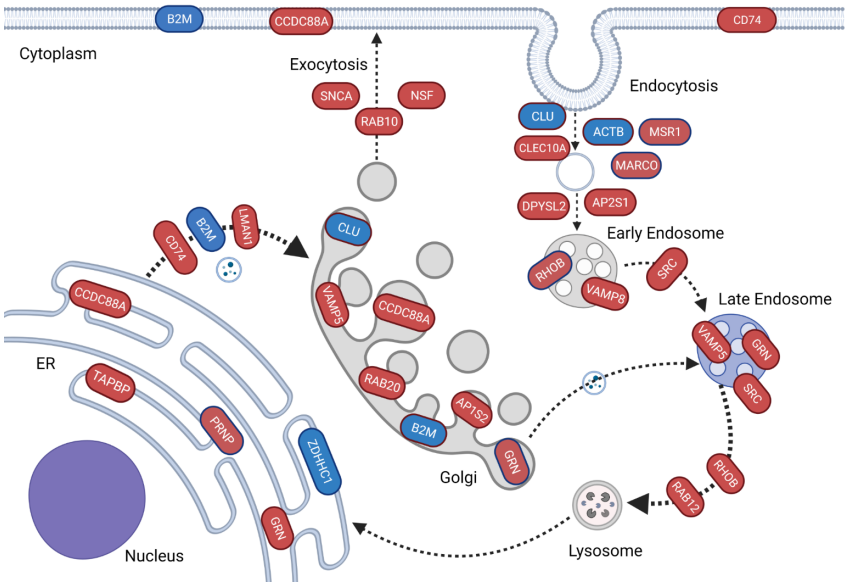

Figure 4. 13: Schematic diagram of intracellular trafficking pathways. Dysregulated ICTs that are positively associated with the transport of HLA-DR surface protein are shown in red and the dysregulated ICT genes with negative association are shown in red and are positioned in the place where they are believed to function.

To explore the enrichment of ICT genes within biological processes and cellular components related to intracellular protein trafficking, we conducted an enrichment analysis utilizing GOBP (Gene Ontology Biological Processes) and GOCC (Gene Ontology Cellular Component) gene sets obtained from the MsiGDB database.

As displayed in the **Fig 4.14** several of the ICT genes are seen to be enriched in GOCC_ENDOSOME geneset with a total count of 39, GOCC_VACUOLE with a count of 33, and GOCC_ENDOSOME_MEMBRANE with a count of 21 genes. In the GOBP category, the genesets GOBP_PROTEIN_LOCALIZATION_TO_PLASMA_MEMBRANE (count = 22) and GOBP_REGULATION_OF_INTRACELLULAR_TRANSPORT (count = 22) are the top gene sets with most genes enriched.

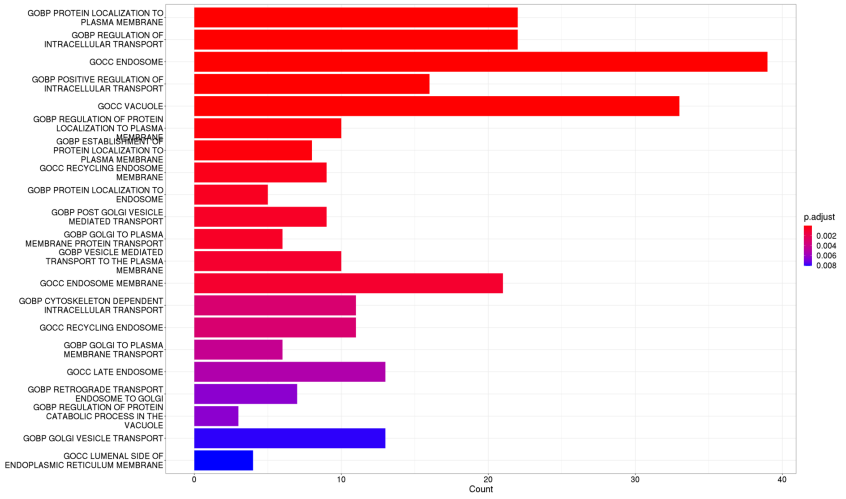

Figure 4. 14: Gene sets enriched in Gene Ontology Biological Processes (GOBP) for the ICT genes from CITE-Traffick

To comprehensively assess the extent to which significant ICT genes explain the variability in HLA-DR protein abundance, we conducted a two-step regression analysis. In the initial step, our aim was to establish that the expression of HLA-DR protein could be accounted for by the expression of its constituent coding genes, namely HLA-DRA and HLA-DRB1, while controlling for relevant covariates like age and sex.

Subsequently, in the second step, we regressed the residuals obtained from the first regression analysis on the ICT genes to examine whether these genes could clarify the remaining variability, even after considering coding gene expression and covariates. Notably, in this second step, the analysis revealed that all ICT genes obtained from the regularized mediation process exhibited significant associations with the residuals (significance indicated by α with $p < 0.05$). This compelling outcome underscores the substantial influence of ICT genes in elucidating a noteworthy portion of the variance in HLA-DR protein expression. Importantly, their impact is observed beyond the effects of coding gene expression and other control variables.

Functional enrichment analysis revealed dysregulated inflammation-related pathways

To identify differentially expressed genes (DEG) between severe and healthy samples, we compared the expression level of each of the genes within the CD16 monocyte population. There were 1669 significant DEGs at a p-value threshold of < 0.05 and absolute value of average log2 fold change > 0.1. The DEGs were further filtered for ICT genes which resulted in 231 significant ICT DEGs. The **Fig 4.15** shows the overlapping of ICT genes identified through the CITE-Traffic penalized regression model with significant ICT DEGs. Out of the 124 ICT genes identified in the regularized mediation model, 72(58%) are DEGs while 52 (42%) were uniquely identified by CITE-Traffick. Among the 52 ICTs, HLA-DR was showing full-mediation effect for 30 (58%) ICTs, partial mediation effect for 8 (15%) ICTs, and no mediation effect in the case of 14 (27%) ICTs. Among the 72 ICTs that are also DEGs, 8 (11%) have full mediation, 12 (17%) have partial mediation, and 52 (72%) did not have mediation effect.

Out of 231 DEGs, 159 (69%) were not identified by the mediation model as significantly associated with HLA-DR protein or disease outcome. When examining the regression model results, this number came down to 136, so 23 genes while they exhibited significant association with HLA-DR, they did not pass the mediation test. The rest of the 136 genes even though they are differentially expressed between Severe and Healthy they are not significantly associated with the transport of HLA-DR.

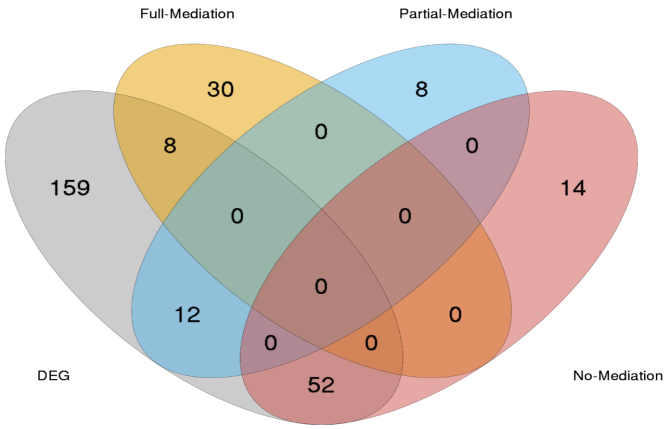

Figure 4. 15: Overlap of DEG and ICT genes with full, partial or nil mediation through HLA-DR

To understand the common biological functions and pathways these DEGs share, we conducted a GSEA analysis using the differentially expressed genes between severe and healthy cases, employing both Hallmark and Curated gene sets from MSigDB (A. Subramanian et al., 2005) (Liberzon et al., 2011). Following this initial analysis, we further refined our findings by filtering the results with CITE-Traffick ICT genes.

This approach aimed to delve into the enrichment of these genes within immune pathways pertinent to the disease phenotype, providing deeper insights into the underlying mechanisms. The **Fig 4.16** showcases the Hallmark pathways and Reactome pathways enriched with ICT genes.

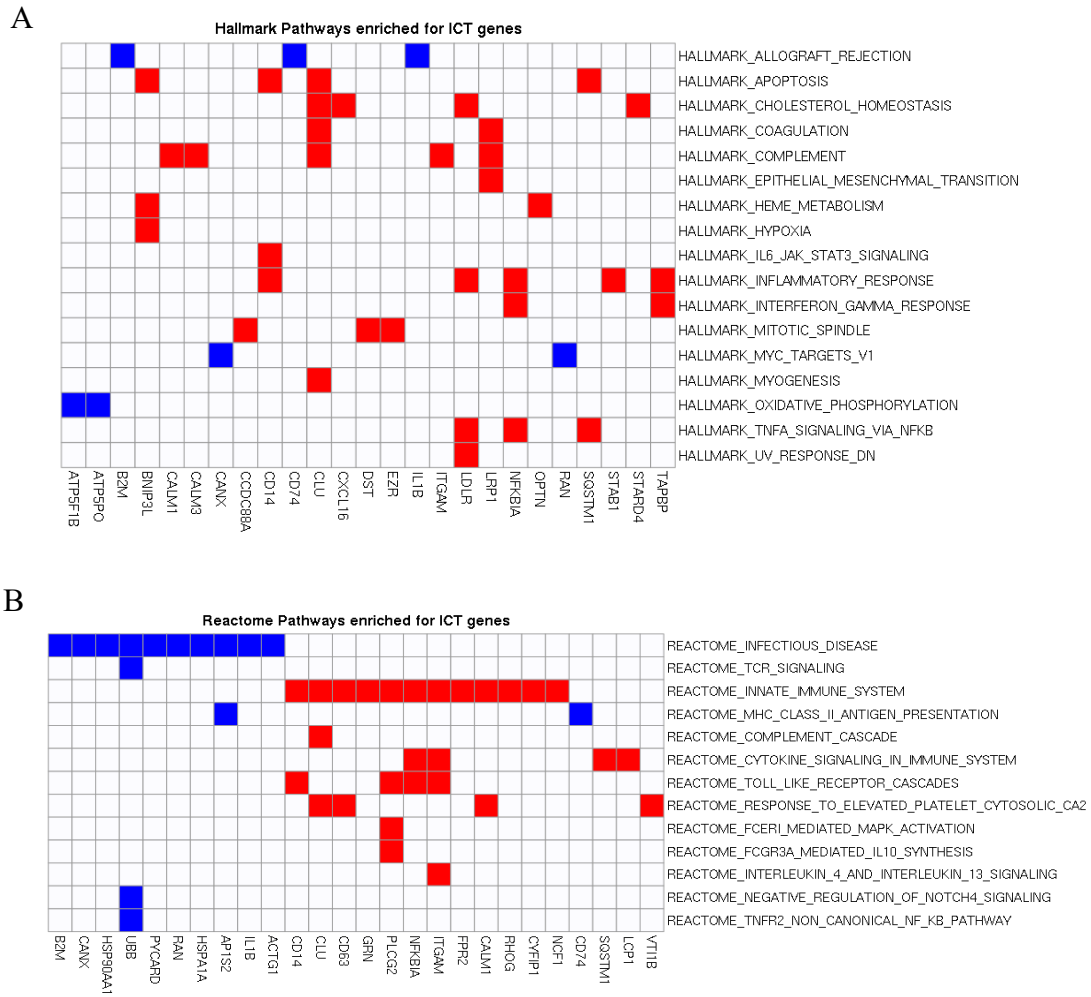

Figure 4. 16: Pathway enrichment analysis of ICT genes. (A) Hallmark pathways (B) Reactome pathways

From the enriched pathways it is noted that CD74, IL1B, and B2M which are downregulated in severe compared to healthy are all part of HALLMARK_ALLOGRAFT_REJECTION. A study on sepsis gene expression signatures in severely afflicted COVID-19 patients underscores the importance of organ dysfunction and its potential implications for allograft rejection in severe COVID-19 patients (Baghela et al., 2023). ICT genes *CALM1, CALM2, CLU, ITGAM, and LRP1 CLU, CALM3, and FCN1,* which are upregulated in severe compared to healthy, are part of the HALLMARK_COMPLEMENT pathway. Several studies (Mazzoni et al., 2021) (Rovito et al., 2022) (Kircheis et al., 2020) provide compelling evidence that the inflammatory response is a hallmark of severe COVID-19. Their investigation, which involved analyzing immune profiles in patients with mild, moderate, and severe COVID-19, demonstrated distinct patterns of immune activation and cytokine dysregulation in individuals with severe disease. *NFKBIA, TAPBP, STAB1, LDLR* all of which are upregulated in severe compared to healthy are the ICT genes that show part of inflammatory related pathways - HALLMARK_INFLAMMATORY_RESPONSE, HALLMARK_INTERFERON_GAMMA_RESPONSE, and HALLMARK_TNFA_SIGNALING_VIA_NFKB. The ICT genes *STAB1*, *LDLR*, *CLU,* and *CXCL16* were found to be significantly enriched in the HALLMARK_CHOLESTEROL_HOMEOSTASIS pathway.

Studies (Dai et al., 2022) (Bakillah et al., 2022), have investigated the dysregulation of cholesterol balance, which has emerged as a significant factor in COVID-19, with implications for both viral replication and the host immune response. ICT genes such as *STAB1*, *LDLR*, and *CLU* displayed notable upregulation in severe patients, whereas *CXCL16* did not attain significance in the DEG list.

A recent study has indicated increased plasma concentration of *CXCL16* in COVID-19 hospitalized patients and demonstrated its association with disease severity (Smieszek et al., 2022). Previous studies have emphasized the critical role of the WASF regulatory complex, a 5-subunit protein complex associated with invasion and metastasis phenotypes, with *CYF1P1* being one of its coding genes (Y. Xiong et al., 2019). Notably, *CYF1P1* which is a significant ICT gene in severe-healthy analysis and is implicated as part of the REACTOME_INNATE_IMMUNITY_SYSTEM pathway. On the other hand, DEG analysis did not identify *CYF1P1* as a significant gene. Although not identified as a significant DEG in our study, the ICT genes *CYF1P1* and *CXCL16* in the CITE-Traffick analysis underscores their potential as valuable marker discovery tools, illuminating their importance beyond traditional differential expression analysis. Similarly, *CLU, CALM3, and FCN1,* which were also upregulated in severe compared to healthy, is part of the HALLMARK_COMPLEMENT pathway. Studies conducted by (Jarlhelt et al., 2021) and (L. Ma et al., 2021) have extensively investigated the role of complement immune system activation as a prominent feature of severe COVID-19.

4.5.3.2 Severe - Mild Comparison

In the severe to mild (SM) comparison analysis, 365 ICT genes were selected by the mixed effects regression model for the Mild group out of which 248 ICTs (68%) were common between Severe and Mild. Only 46 (13%) ICTs passed the regularized mediation test with HLA-DR having full mediation effect for 21 ICTs (46%), partial mediation with 11 ICTs (24%) and no mediation for 14 ICTs (30%). In the severe to mild comparison also, HLA-DR protein revealed negative association with covid severity (β =0.15). To investigate the dysregulation of intracellular trafficking (ICT) genes in the context of low HLA-DR expression in severe and mild COVID-19 samples, we compared SM and SH regularized mediation analysis results (**Fig 4.17.A**). Intriguingly, a set of 96 Intracellular Trafficking Genes (ICTs) exhibited specific dysregulation in the Healthy (SH) group, while 18 ICTs uniquely displayed dysregulation in the Severe (SM) group. The ICT genes that exhibited dysregulation unique to the SH may be genes that are specifically associated with the pathological processes and immune response observed in severe cases of COVID-19 and dysregulated ICT genes unique to SM may be involved in modulating the transition from mild to severe COVID-19.

Interestingly, it was revealed that 28 ICTs (61%) were found to be dysregulated in severe COVID-19 samples compared to both healthy (SH) and mild (SM) COVID-19 comparisons. While some of the ICTs for example, CLU, CD74, CXCL16 had similar mediation effects in both SH and SM analysis, other ICTs were showing different mediation effects in SH and SM comparisons (**Fig. 4.17 B**). For e.g. CLU and CD74 are having partial mediation in both SH and SM, EZR and RAB34 are having full mediation

in both, and CXCL16 and HSPA1A display no meditation in both. The ICT GAS6 shows partial mediation in SH, while it shows full mediation in SM. Another example, RHOQ which shows full mediation in SH, but has no mediation effect in SM. Next, we compared the delta coefficients associated with these ICT genes against the DEG (**Fig. 4.17.C**). Interestingly all the 28 ICTs were found to be DEGs. Except for the genes: NSF, RAB34, RALB, RUFY3, UNC93B1, and VAMP8, the direction of differential expression for all other ICTs was consistent across both the SM and SH comparisons.

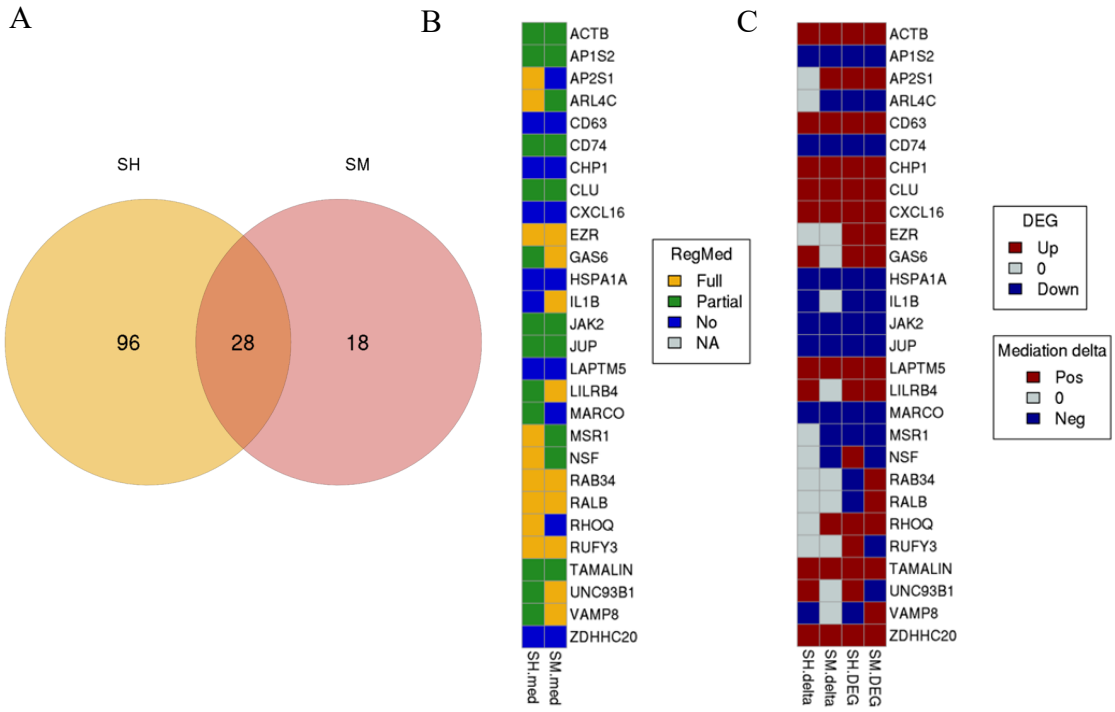

Figure 4. 17: Comparative Analysis of ICT Dysregulation in Severe- Healthy (SH) and Severe - Mild (SM) Comparisons. (a) Comparison of regularized mediation analysis reveals unique and overlapping ICT genes in SM and SH groups. (b) Mediation effects of overlapping ICT genes in both SM and SH comparisons show variability in associations, with different genes displaying varying mediation effects across the groups. (c) Comparison of Differential Expression (DEG) and mediation coefficients of overlapping ICT genes across SH and SM.

4.5.4 CITE-Traffick Identifies Dysregulated ADT-ICT Mediation Network in CD16+ Monocyte

CITE-Traffick has been successful in identifying the dysregulated ICT genes associated with the transport of HLA-DR protein in CD16+ Monocyte and also was able to identify the direct and indirect effects of these genes on the disease outcome, it is worthwhile to see its application on a wide scale of ADTs and ICT genes. Hence with the aim to explore a wider network of dysregulated ADT-ICT networks we shifted our focus to a multi-mediation network which involves multiple exposures and multiple mediators. CITE-Traffick offers two ways to explore the multi-mediation mediation network: (i) Regularized mediation and (ii) Structural Equation modeling. In the following sections we will apply both these methods to CD16+ Monocyte cell population in Severe COVID and Healthy groups from the COVID dataset.

4.5.4.1 Regularized Mediation Model

Regularized mediation analysis using Regmed was run using all the PTTs identified in Module I at an adjusted p-value threshold of 0.05 and absolute correlation coefficient threshold of 0.05. At this threshold, a total of 2015 PTTs were identified in the Severe group and 3778 PTTs were identified in the Healthy group, which included a total of 23 ADTs and1123 ICT genes. With ADTs as mediators and ICT genes as exposures, regularized mediation analysis. At a lambda of 0.01, a total of 10 ADTs were retained by the regularization which includes HLA-DR, CD14, CD16, CD33, CD38, CD163, CD4, CD123, CD3, and CD1c out of which CD33 did not form mediation connection with any of the ICT genes. Out of 1123 ICT genes only 159 (14%) genes were retained with non-

zero coefficients in the model. Out of the 159 ICT genes, only 52 (33%) had mediator connections, the rest 107 (67%) of the ICT genes had no mediators connected**. Table 4.7** gives all the ADTs and the corresponding exposure ICTs which are divided into partial, full or nil based on the type of mediation the ADT forms with the ICT genes. The **Fig 4.18** illustrates the mediation network of ICT genes and proteins. For the HLA-DR protein, a total of 517 ICT genes were tested. In the multiple mediator model, It is noticed that for the HLA-DR protein, only 103 ICT genes (20%) were retained. Out of these 103 ICT genes, fewer ICTs with mediation effects were identified in the multiple-mediator model (18%) compared to the single-mediator models. Conversely, more ICTs (82%) without mediation effects were identified.

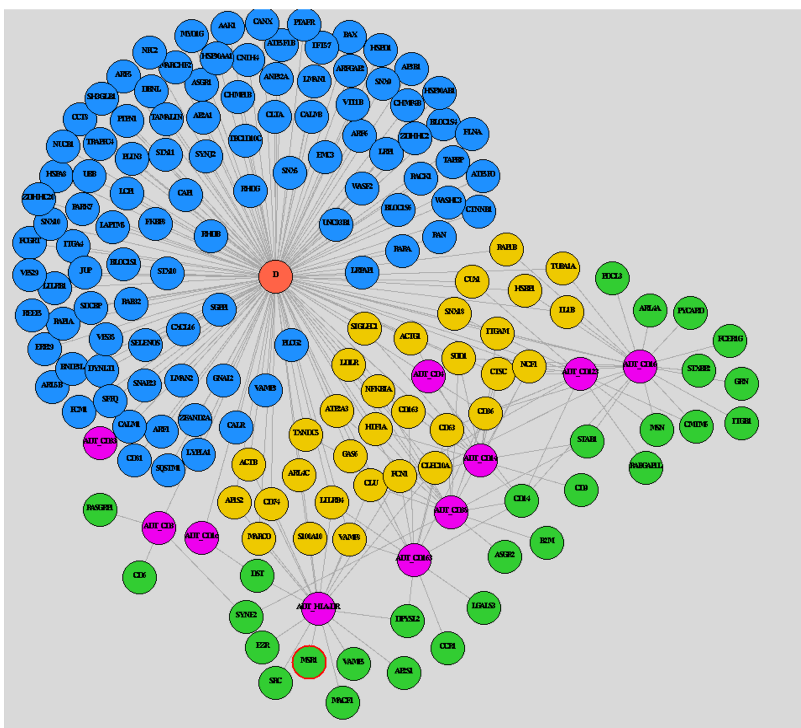

Figure 4. 18: Multiple-Exposure-Multiple-Mediator regularized mediation model in Severe-Healthy

Table 4. 7: Results from multiple-mediator-multiple-exposure regularized mediation model in Severe-Healthy

ADT	Mediation	ICT	α	Δ
CD123	full	CD14 CMTM6 MSN RABGAP1L	-0.08 - 0.03	
	partial	CD36 CD63 CTSC CUX1 FCN1 NCF1 SNX18	-0.01 - 0.03	-0.04 - 0.24
CD14	full	CD14 CD9 STAB1	0.05 - 0.33	
	partial	ACTG1 CD163 CD36 CLU CTSC HIF1A ITGAM NCF1 NFKBIA SIGLEC1 SOD1	-0.02 - 0.08	-0.07 - 0.09
CD16	full	ARL4A CD14 CMTM6 FCER1G GRN ITGB1 MSN PDCL3 PYCARD STAB1 STXBP2	-0.15 - 0.09	
	partial	CD36 CLEC10A CTSC HSBP1 IL1B ITGAM NCF1 RAP1B SOD1 TUBA1A	-0.02 - 0.07	-0.08 - 0.03
CD163	full	CCR1 CD14 DPYSL2 LGALS3	0 - 0.11	
	partial	CD163 CD36 CD63 CLEC10A FCN1 GAS6 LILRB4 VAMP8	0.01 - 0.06	-0.02 - 0.24
CD1c	full	DST	0.01	
CD3	full	CD6 RASGRP1 SYNE2	0.04 - 0.06	
CD33	nil	ASGR1 CD14 CD36 CD63 FCN1 ITGAM NCF1 NFKBIA STAB1	0.01 - 0.14	
CD38	full	ASGR2 B2M CD14 STAB1 SYNE2	-0.02 - 0.13	
	partial	ATP2A3 CD36 CD63 CLU CTSC HIF1A LDLR NCF1 NFKBIA SOD1 TXNDC5	-0.05 - 0.07	-0.04 - 0.24
CD4	full	CD14	0.02 - 0.02	
HLA-DR	full	AP2S1 DPYSL2 DST EZR MACF1 MSR1 SRC VAMP5	0.01 - 0.03	
	partial	ACTB AP1S2 ARL4C CD74 CLEC10A CLU FCN1 LILRB4 MARCO S100A10 VAMP8	-0.03 - 0.27	-0.09 - 0.09

4.5.4.2 Structural Equation Modeling

The PTTs for the SEM model are filtered for adjusted p-value threshold < 0.05, absolute correlation coefficient > 0.1 and the ICT genes were filtered based on total number of cells with non-zero expression > 200. Using these criteria a total of 23 ADTs and 415 ICT genes were selected. Hierarchical clustering of the ADTs revealed the formation of 1 to 4 distinct clusters (see **Fig 4.19**). As illustrated in the figure, we specifically examined the clustering within the first protein module, which intriguingly placed the HLA-DR protein alongside the ADTs CD86 and CD11c. This arrangement aligns with previous findings from the original paper, which reported a noteworthy underexpression of CD86, coupled with HLA-DR, in monocytes of COVID patients, as confirmed by CyTOF analysis (Arunachalam et al., 2020). While our analysis of the differential ADT expression in the CITE-Seq dataset did not explicitly validate this observation, the co-clustering of CD86 and HLA-DR proteins offers a compelling perspective on this intriguing relationship.

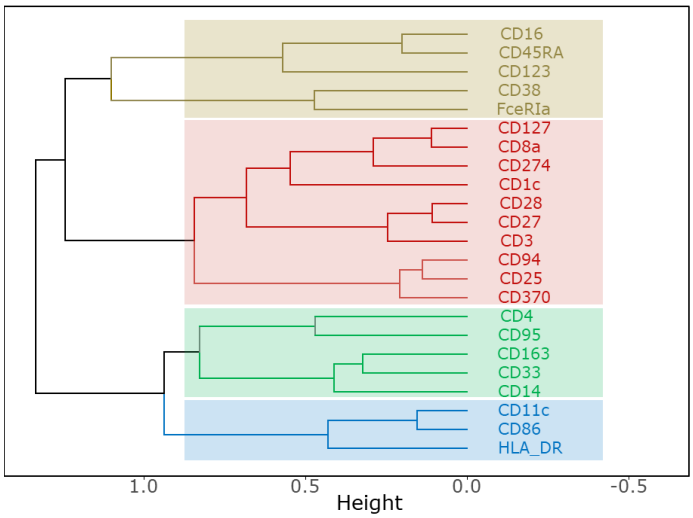

Figure 4. 19: Hierarchical clustering of proteins in Severe-Healthy based on top 5 PCs

The hierarchical clustering of the ICT genes for each protein module resulted in 3 to 4 ICT clusters. The **Fig 4.20** demonstrates the clustering of ICT genes in the first protein module.

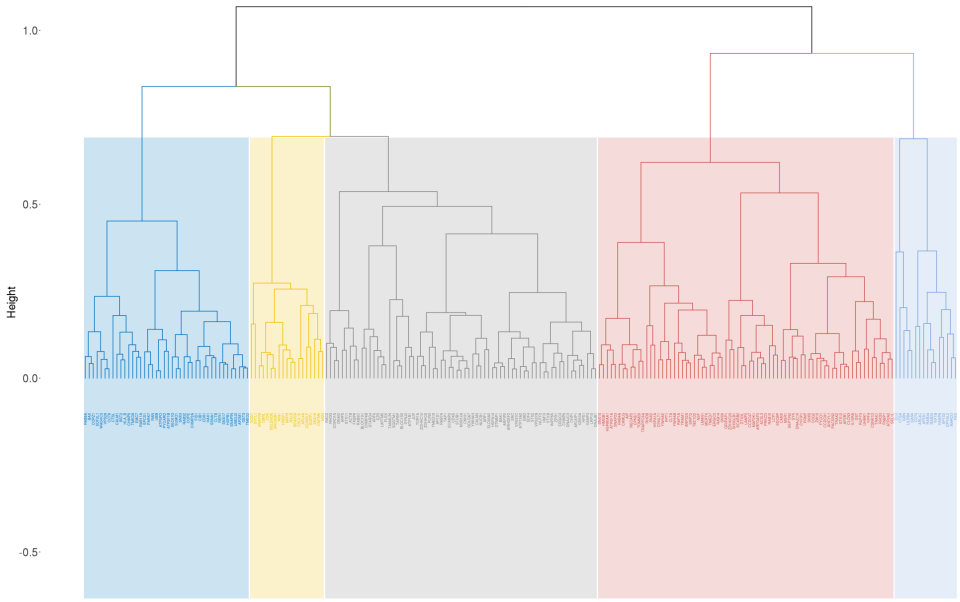

Figure 4. 20: Hierarchical clustering of ICT genes in Severe-Healthy based on top 5 PCs

In the context of our SEM mediation analysis, we constructed distinct mediation models, each comprising of two structural components. The first component was dedicated to the computation of latent variables derived from ICT genes, while the second component was responsible for the computation of latent variables from ADTs. The measurement component of our model was designed to facilitate a comprehensive mediation analysis.

Upon thorough evaluation of the model fit measures for various combinations and structures, a deliberate decision was made to confine our models to include only first-order latent variables. As a result, we subsequently developed separate mediation models, each corresponding to an ICT module along with its respective protein module (**Fig 4.21**). This refined approach was chosen to ensure the most effective and interpretable representation of our data.

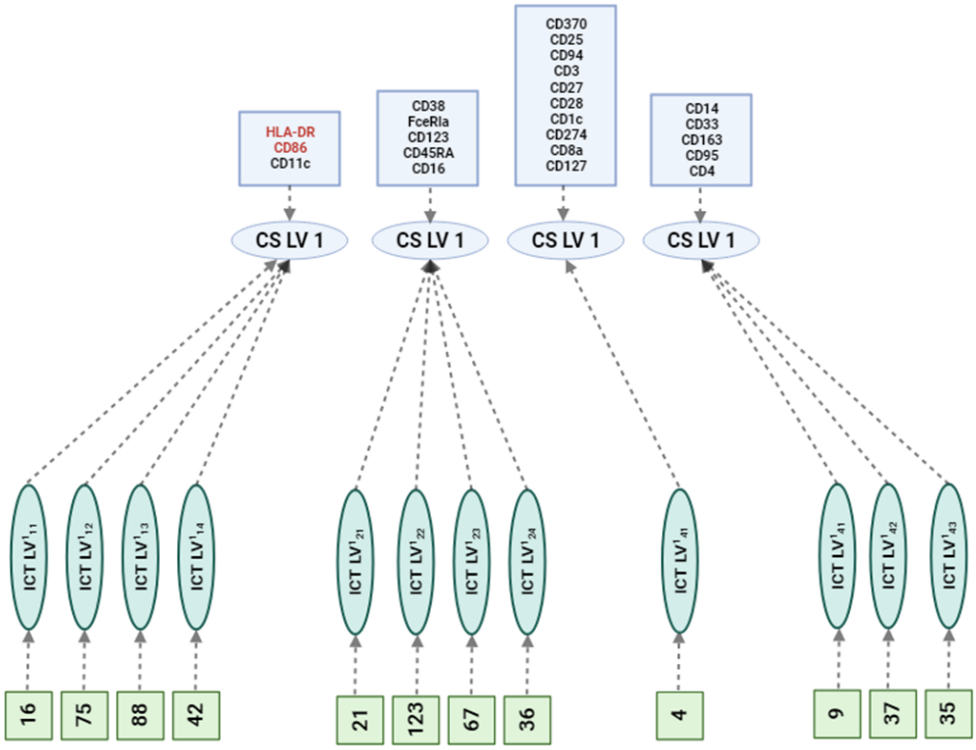

Figure 4. 21: SEM framework for significant ICT genes and ADTs from Severe to Healthy comparison

A total of 11 SEM models were tested with only 2 models having non-significant paths connecting the disease outcome to either ICT latent variable (D ~ ICT_LV) or ADT latent variable (D ~ CS_LV). All the paths in the mediation network connecting ICT and ADT latent variables (CS_LV ~ ICT_LV) were significant. Out of 358 ICTs and 23 ADTs tested in SEM models, 344 (96%) ICTs had significant association with the ICT latent variable ICT_LV. Also, all of the 23 ADTs have significant association with the ADT latent variable CS_LV.

The **Table 4.8** shows the performance metrics for each of the models tested for Severe -Healthy comparison. The chi-square test of all the models was statistically significant. The other fit measures attained the recommended target values for some of the models. The Standardized Root Mean Square Residual (SRMR) represents the square-root of the difference between the residuals of the sample covariance matrix and the hypothesized model. The recommended values for RMSEA and SRMR are < 0.08. The RMSEA and SRMR for all four models are below this cut-off.

The Normed Fit Index (NFI) indicates the proportion by which the model of interest improves the fit and it is recommended to be > 0.90. The Incremental Fit Index (IFI) adjusts the Normed Fit Index (NFI) for sample size and degrees of freedom and it is recommended that > 0.90 is a good fit. In our case IFI for all models is between 0.74 and 0.96. The Goodness of Fit (GFI) is the proportion of variance accounted for by the estimated population covariance and its cut-off is > 0.9. All our models are above this cut-off. Finally, the Parsimony-Adjusted Measures Index (PNFI) is recommended to have a value > 0.50. All four models have achieved this threshold. The Relative Fit Index (RFI) close to 1 indicates a good fit.

Table 4.8. SEM models in Severe-Healthy with their fit measures

Model	p_Chi2	GFI	NFI	RFI	IFI	PNFI	RMSEA	SRMR
cluster_1_g1	0.00	0.89	0.63	0.58	0.64	0.56	.07	.07
cluster_1_g2	0.00	0.96	0.76	0.75	0.85	0.74	.02	.02
cluster_1_g3	0.00	0.95	0.78	0.77	0.86	0.76	.02	.02
cluster_1_g4	0.00	0.98	0.93	0.93	0.96	0.89	.02	.02
cluster_2_g1	0.00	0.92	0.82	0.80	0.83	0.75	.05	.05
cluster_2_g2	0.00	0.95	0.65	0.65	0.80	0.64	.01	.02
cluster_2_g3	0.00	0.94	0.88	0.88	0.91	0.86	.03	.03
cluster_2_g4	0.00	0.95	0.84	0.83	0.86	0.80	.03	.03
cluster_3	0.00	0.99	0.82	0.79	0.87	0.69	.02	.02
cluster_4_g1	0.00	0.97	0.92	0.91	0.93	0.77	.05	.03
cluster_4_g2	0.00	0.97	0.76	0.75	0.83	0.73	.02	.02
cluster_4_g3	0.00	0.96	0.73	0.72	0.78	0.69	.03	.03

As evident in **Table 4.8** the models cluster_1_g4 and cluster_4_g1 consistently outperformed other models across all fit measures, establishing them as the optimal fit models. Upon examining the ICT genes and proteins within these models, we discovered a high and predominantly positive correlation (**Fig 4.22.A** and **B**) among all the genes and proteins.

Conversely, clusters cluster_1_g1 and cluster_2_g2 exhibited the lowest fit scores, and their correlation plot revealed a weaker overall correlation, with some instances of negative correlation (**Fige 4.22.C** and **D**). This observation underscores the significance of feature cohesiveness in determining the model's overall fit, with no discernible relationship to the size of the feature set.

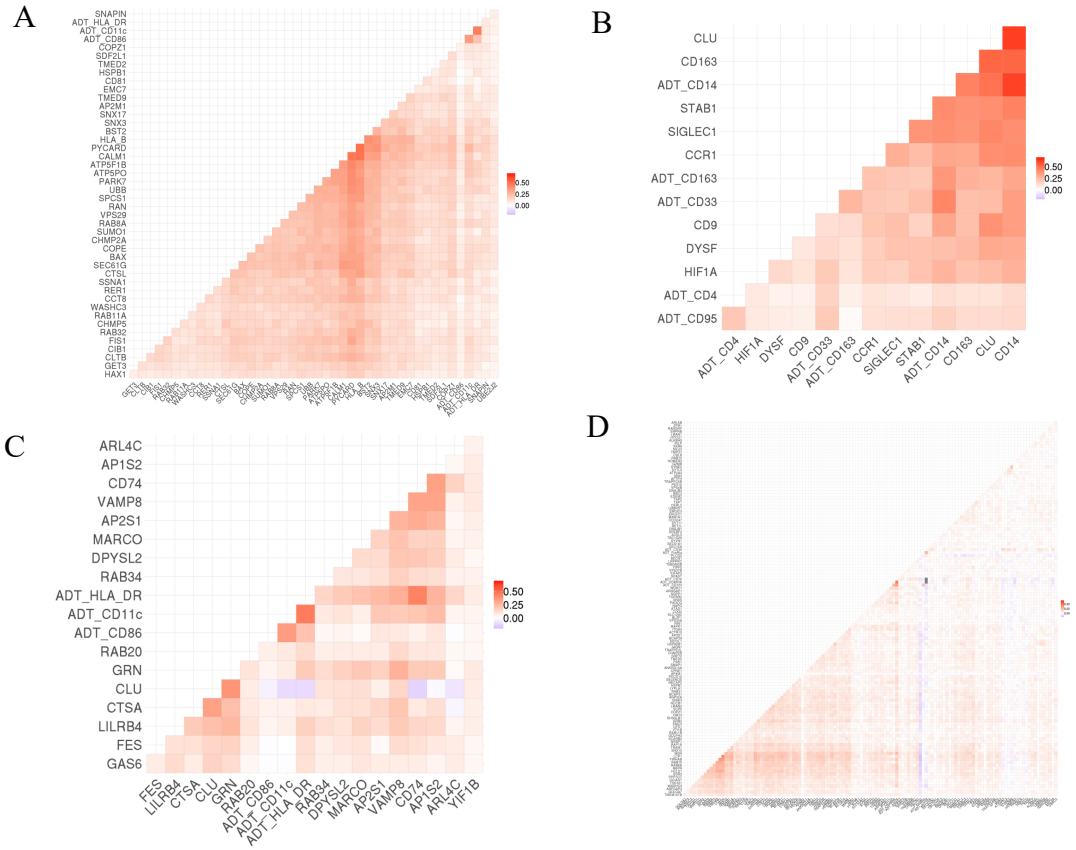

Figure 4. 22: Correlation plots for ICT genes and ADTs in the SEM models (A) cluster_1_g4, (B) cluster_4_g1, (C) cluster_1_g1, and (D) cluster_2_g2

The **Fig 4.23** illustrates the schematic of the cluster_1_4 SEM model that has all fitness quality scores above the thresholds and the **Table A 1** (Appendix A) provides a comprehensive overview of the SEM coefficients with the p-values and confidence intervals for each component in this model. Within this framework, the regression formula D ~ G refers to the direct effect of the ICT latent variable G on the disease outcome D. Additionally, the formulas M ~ G and D ~ M define the indirect effect of G on D through the mediator M. The cluster_1_4 SEM model includes 42 ICTs and 3 ADTs - CD11c, HLA-DR, and CD86. As given in the table all the paths in the given model are significant with p-value < 0.001. The path coefficients reveal essential insights into the relationships between these variables. Notably, the path from G to D and that from M to D exhibit coefficients of -0.09 and -0.15, signifying a negative association between the ICT latent variable and the protein latent variable with the disease. Furthermore, the path from G to M is characterized by a coefficient of 0.36, indicating a positive association between the ICT latent variable and the protein latent variable. It is also important to note that the paths from all 42 ICTs to G have positive coefficients ranging 0.17 - 0.73. In summary, the presence of significant direct and indirect paths in this model underscores the existence of a partial mediation effect, signifying a complex interplay between the ICT latent variable, the protein latent variable, and their collective influence on the disease outcome.

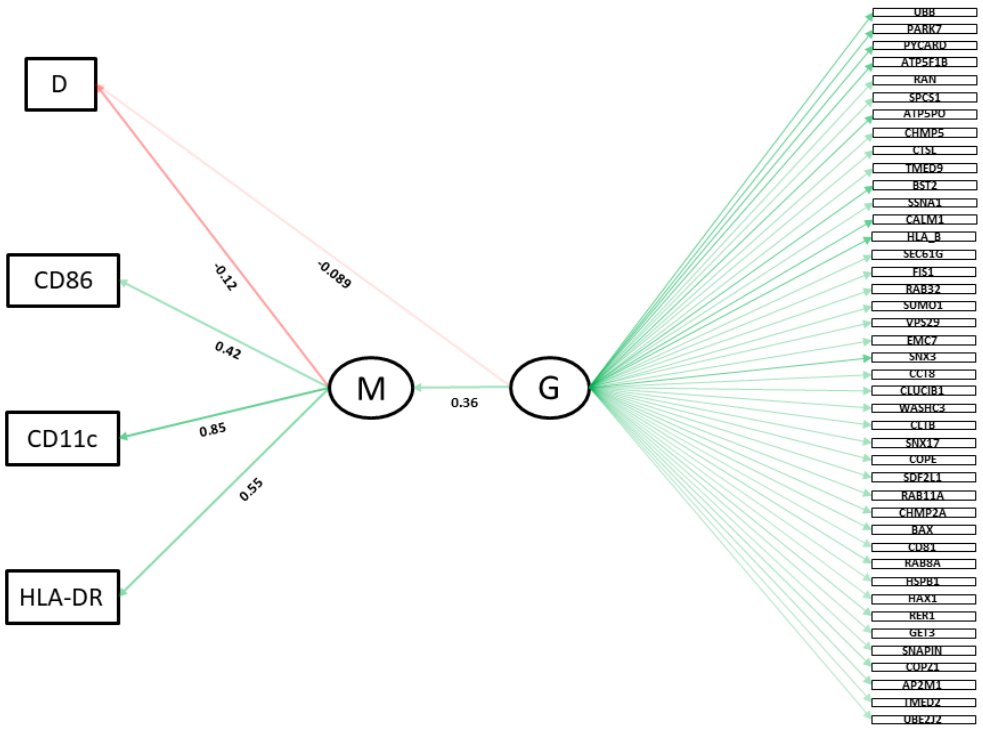

Figure 4. 23: Illustration of cluster1_g4 SEM model with G as the ICT latent variable, M as the ADT latent variable, and disease D as the outcome.

On examining the coefficients of the paths from ICTs to G, the top 10 ICT genes are *PYCARD (0.73), HLA_B (0.69), CALM1 (0.65), ATP5F1B (0.56), PARK7 (0.55), ATP5PO (0.53), BST2 (0.53), SNX3 (0.53), UBB (0.52)*, and *SPCS1 (0.50)*. The *CALM1* gene, and its significance has been studied before. Specifically, the connection between CALM1 and ACE2 is grounded in a study conducted by Lambert et al. (*Calmodulin Interacts with Angiotensin-converting Enzyme-2 (ACE2) and Inhibits Shedding of Its Ectodomain - Lambert - 2008 - FEBS Letters - Wiley Online Library*, n.d.) (Wruck & Adjaye, 2020), where they provide evidence of CALM1's interaction with the coronavirus receptor ACE2.

Notably, their research highlights that CALM1 plays a role in inhibiting the shedding of ACE2's ectodomain. This process is of particular importance in the context of COVID-19, as inhibiting ACE2 shedding can have implications for viral entry and pathogenesis.

To gain deeper insights into the potential roles of these 42 ICT genes in immune-related functions, we conducted a comprehensive gene enrichment analysis focused on the cluster_1_g4 model. The results of this analysis unveiled a multitude of pathways, some of which have significant implications in COVID-19 studies (**Fig 4.24**). For instance, we observed pathways such as "REACTOME_SARS_COV_2_INFECTION" and "REACTOME_INFECTIOUS_DISEASE," both of which have been extensively associated with COVID-19 research. These findings underscore the relevance of these genes in the context of viral infections and immune responses. Furthermore, our analysis also highlighted the presence of pathways related to metabolic processes, including "REACTOME_GLYCOGEN_METABOLISM." This aligns with previous research that has pointed to dysregulated glucose metabolism as a noteworthy aspect of COVID-19 pathogenesis (R. Kumar et al., 2022) (P. Chen et al., 2023). Most importantly, we noted the enrichment of multiple pathways related to protein trafficking, such as "REACTOME_GOLGI_TO_ER_RETROGRADE_TRANSPORT," "REACTOME_VESICLE_MEDIATED_TRANSPORT," and "REACTOME_LATE_ENDOSOMAL_MICROAUTOPHAGY." However, it's notable that we did not observe significant enrichment of pathways directly linked to inflammation. This suggests that these specific ICT genes may not be directly involved in the inflammatory responses in the context of COVID-19.

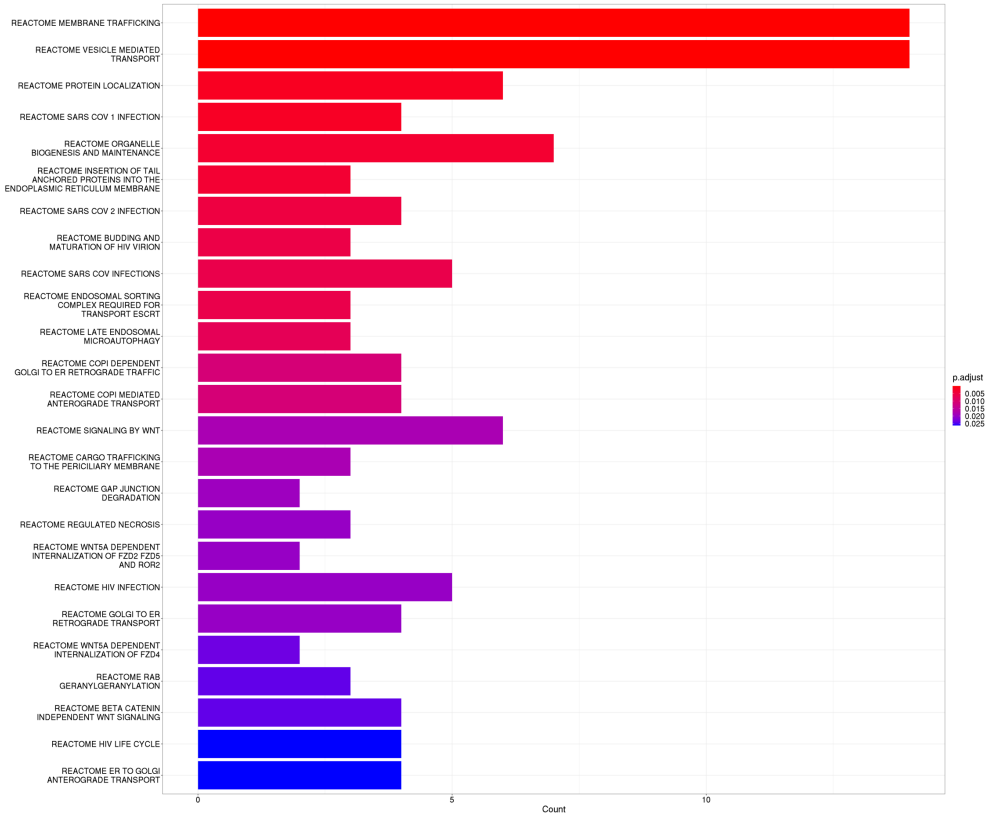

Figure 4. 24: GSEA analysis on ICT genes from cluster_1_g4 model

4.5.5 CITE-Traffick identifies CLU as a marker for targeting multiple ADTs

Within the domain of the CITE-Traffick algorithm's modules, an intriguing discovery emerged regarding the Clusterin (CLU) gene. Clusterin (CLU), a versatile glycoprotein acting as a stress-activated chaperone is involved in the suppression of protein aggregation during under stress conditions. CLU exhibits heightened expression in Alzheimer's (Foster et al., 2019) and cancers, exerting anti-apoptotic effects and contributing to treatment resistance in cancer (Yom et al., 2009). CLU is also involved in autophagy. Under stress conditions excessive autophagy can lead to type II programmed cell death (Gozuacik & Kimchi, 2007).

Hence regulated expression of CLU is critical for protein homeostasis within the cell. CLU inhibitors (Custirsen) as therapeutic targets have demonstrated promising outcomes, revealing increased overall survival and a significant reduction in mortality rates when combined with docetaxel based on a phase II study (Chi et al., 2010).

Differential gene expression revealed over-expression of CLU in severe compared to healthy and mild, further indicating its potential role in COVID. CLU revealed a significant negative association with HLA-DR transport in the regression model for severe (α = -0.08) and healthy (α = -0.21). Also, in the single mediator regularized mediation model with HLA-DR as partial mediator, CLU revealed negative association with HLA-DR (α = -0.12) and positive association with severe COVID outcome (δ = 0.15). Finally, in the multiple mediator regularized mediation model, CLU revealed negative association with HLA-DR once again acting as partial mediator (α = -0.03) and revealed a positive association with severe COVID outcome (δ= 0.09). Also CLU was identified as a marker associated with multiple other ADT targets (CD38 and CD14) in the multiple mediator model (**Fig 4.25**). This finding shed light on CLU's role as a potential marker for multiple ADT targets within the context of intracellular protein trafficking. In CITE-Traffick SEM models, CLU had significant associations in two of the SEM models (cluster_1_g1 and cluster_1_g2) within which ADTs tested were - HLA-DR, CD86, CD14, CD33, CD163, CD95, and CD4. On examining, 20 out of 23 ICT genes within these two SEM models were DEGs.

On conducting enrichment analysis of all the 20 ICT genes including CLU revealed multiple immune inflammatory related pathways. This could indicate that these genes may collectively form a module orchestrating HLA-DR expression transport, and directly correlated with the disease through immune and inflammatory pathways in the context of severe COVID conditions.

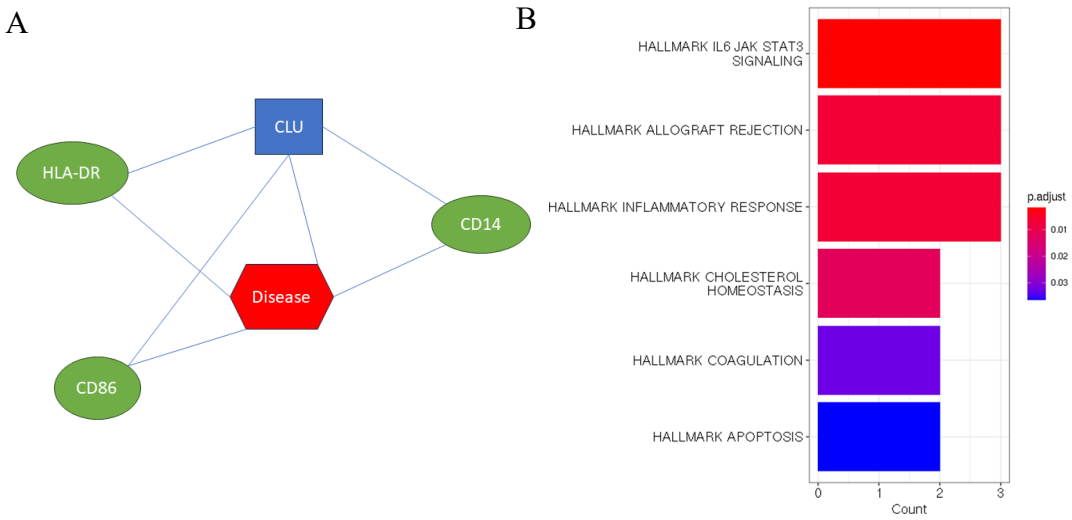

Figure 4. 25: Role of CLU in targeting multiple ADTs and immune functions.
(A) Representation of CLU's association with multiple ADT targets as observed in the CITE-Traffick Structural Equation Models (SEMs) for CLU with ADTs. (B) Pathway enrichment analysis of CLU and other ICTs in SEM model revealed Immune-inflammatory pathways

4.5.6 Integrating Insights from Original Studies and CITE-Traffick Analysis

The original study reports impaired type I IFN production and enhanced expression of interferon-stimulated genes (ISGs) like IFI27, IFITM3, and ISG15 in COVID patients. Also, it is reported a significant decrease in the expression of genes such as CD74 involved in antigen-presentation pathways in myeloid cells and reduced expression of CD86 and HLA-DR on monocytes and mDCs in COVID-19 patients, particularly in severe cases. Increased levels of proinflammatory mediators in plasma and suppressed immunity response in PBMC monocytes and DCs are also being reported in the original study.

Previous research has studied the crucial role of IFN-gamma in regulating HLA class II expression and has found that IFN-gamma is responsible for a temporary increase in HLA antigen expression during influenza A virus infection. The reduction in HLA class I and II mRNA at late times of infection implicated dynamic regulation of these genes over the course of the infection (Keskinen et al., 1997). These important findings connecting IFN-gamma, ISG genes and HLA class II have important implications in COVID as well.

The CITE-Traffick method applied to the COVID-19 CITE-seq dataset revealed a comprehensive landscape of intracellular trafficking (ICT) genes enriched in key pathways such as HALLMARK_INTERFERON_GAMMA_RESPONSE, HALLMARK_INFLAMMATORY_RESPONSE, and HALLMARK_COMPLEMENT. Notably, several identified genes, including TAPBP, TAP1, NFKBIA, ARL4A, CD74, VAMP5, VAMP8, BST2, and IL10RA, exhibited associations with interferon-stimulated

genes (ISGs) and inflammatory responses. While some findings aligned with the original study, such as the upregulation of known ISGs, the CITE-Traffick method uncovered genes like ARL4A that were not detected by traditional differential expression analysis. The downregulation of CD74 which is also a reported ISG gene and its positive correlation with HLA-DR, exemplifies relationship between IFN-γ, ISGs, and MHC expression underscoring the method's ability to capture complex immune dynamics. The ICT genes IL10RA, NFKBIA, TAPBP identified by your method are not only linked to ICT but are also seen associated with the IFN-gamma response and inflammatory pathways.

It is intriguing to observe that certain genes, such as CTSL, FCN1, GNAI2, PRKCD, and SRC, identified as positively correlated with HLA-DR expression and enriched in the hallmark complement pathway, were not detected in the differential expression analysis (DEG). Notably, HSPA1A, situated within the HLA class III region, is recognized for its roles in transporting antigenic peptides from tumor and virus-infected cells, acting as a protein chaperone (Ucisik-Akkaya et al., 2010). Its downregulation in COVID, despite its involvement in immune pathways, is a noteworthy discovery. Another significant immune marker is CLU. The consistent upregulation of CLU reported in several COVID studies (Singh et al., 2021), its identification in the CITE-Traffick analysis, and its negative correlation with HLA-DR highlight its potential as a crucial target for further exploration in the context of immune functions.

The **Fig 4.26** illustrates a comprehensive heatmap showcasing Intracellular Trafficking (ICT) genes with significant correlations to HLA-DR transport, categorizing them based on their involvement in IFN-gamma, inflammatory, and immune pathways, as well as their differential expression status. Additionally, it incorporates information on whether they are DEGs, the positive or negative correlation reported in the CITE-Traffick regression model and indicates their significance in mediation analysis, depicting whether mediation with HLA-DR is partial, full, or nil.

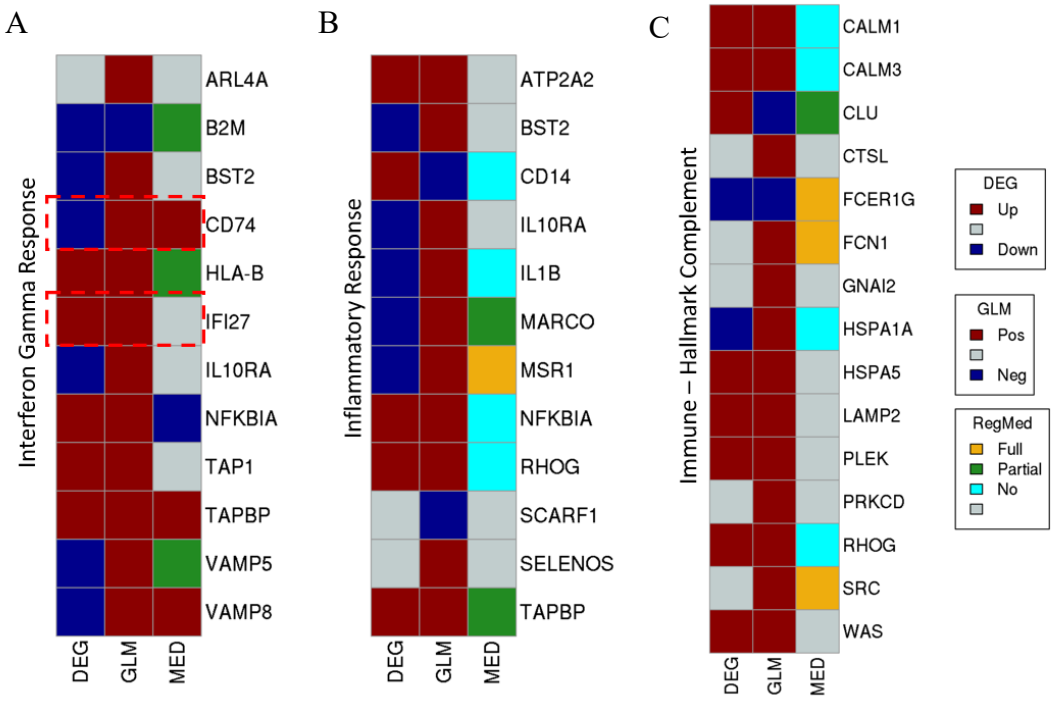

Figure 4. 26: The heatmap of Intracellular Trafficking (ICT) genes with significant correlations with HLA-DR transport. Genes are categorized based on their participation in IFN-gamma, inflammatory, and immune pathways, along with details on their differential expression, correlation status in the CITE-Traffick model, and their roles in mediation analysis with HLA-DR.

4.5.7 Comparison of CITE-Traffick with other methods

To assess the feature selection capabilities of the CITE-Traffick algorithm in predicting the severity of COVID-19, we conducted a comprehensive evaluation. Our analysis involved a comparative study of different feature sets, including ICT genes, Antibody-Derived Tags (ADTs), and latent variables identified by CITE-Traffick, against differentially expressed genes and ADTs derived from the comparison between severe and healthy groups.

The evaluation process encompassed several key steps. Firstly, we employed LASSO models that were complemented by stability testing and cross-validation to identify the most relevant features within each feature set. Subsequently, a multiple Generalized Linear Model (GLM) was applied utilizing these selected top features on an independent test dataset. The feature sets considered for evaluation were: (i) Differentially expressed genes and ADTs identified between the severe and healthy groups. (ii) Significant ICT genes and ADTs identified in either severe or healthy groups, using the mixed effects regression model within CITE-Traffick Module I. (iii) ICT genes and ADTs specific to the severe-healthy comparison, as pinpointed by the multiple-mediator-multiple-exposure regularized mediation model in CITE-Traffick Module II. (iv) A combined feature set consisting of differentially expressed genes and ADTs (from feature set i) complemented by latent variables produced by SEM in CITE-Traffick Module II. This meticulous evaluation process allowed us to gain insights into the performance of the CITE-Traffick algorithm and its ability to select features that can effectively predict the severity of COVID-19.

Upon evaluating the Generalized Linear Model (GLM) on the test set, we noticed that the combined feature set which included latent variables alongside differentially expressed genes and ADTs achieved an impressive accuracy of 0.96, while the feature set that solely incorporated differentially expressed genes resulted in an accuracy of 0.94. Furthermore, our analysis revealed that among the top 50 selected features in the DEG_plus_Latent feature set, the latent variables "cluster_4_g1_G", "cluster_1_g1_g2_M", "cluster_4_g3_M", and “cluster_4_g3_G" were included. These results strongly indicate that CITE-Traffick's latent variables significantly enhance the predictive power of the model when combined with DEGs, reinforcing their role in predicting disease outcomes. For a comprehensive overview of our findings, please refer to the performance metrics provided in the **Table 4.9**. Additionally, the accompanying **Fig 4.27.A**, including the ROC curves and the Variable Importance plot (**Fig 4.27.B** and **Fig 4.27.C**), demonstrate the variables selected for the top two performing models.

Table 4.9. Performance metrics for all feature sets

Metric	DEG	DEG_plus_Latents	GLMM	RegMed
Accuracy	0.94	0.96	0.92	0.91
Precision	0.93	0.96	0.92	0.90
Sensitivity	0.87	0.91	0.79	0.74
F1_Score	0.90	0.94	0.85	0.82
AUC_ROC	0.98	0.99	0.97	0.96

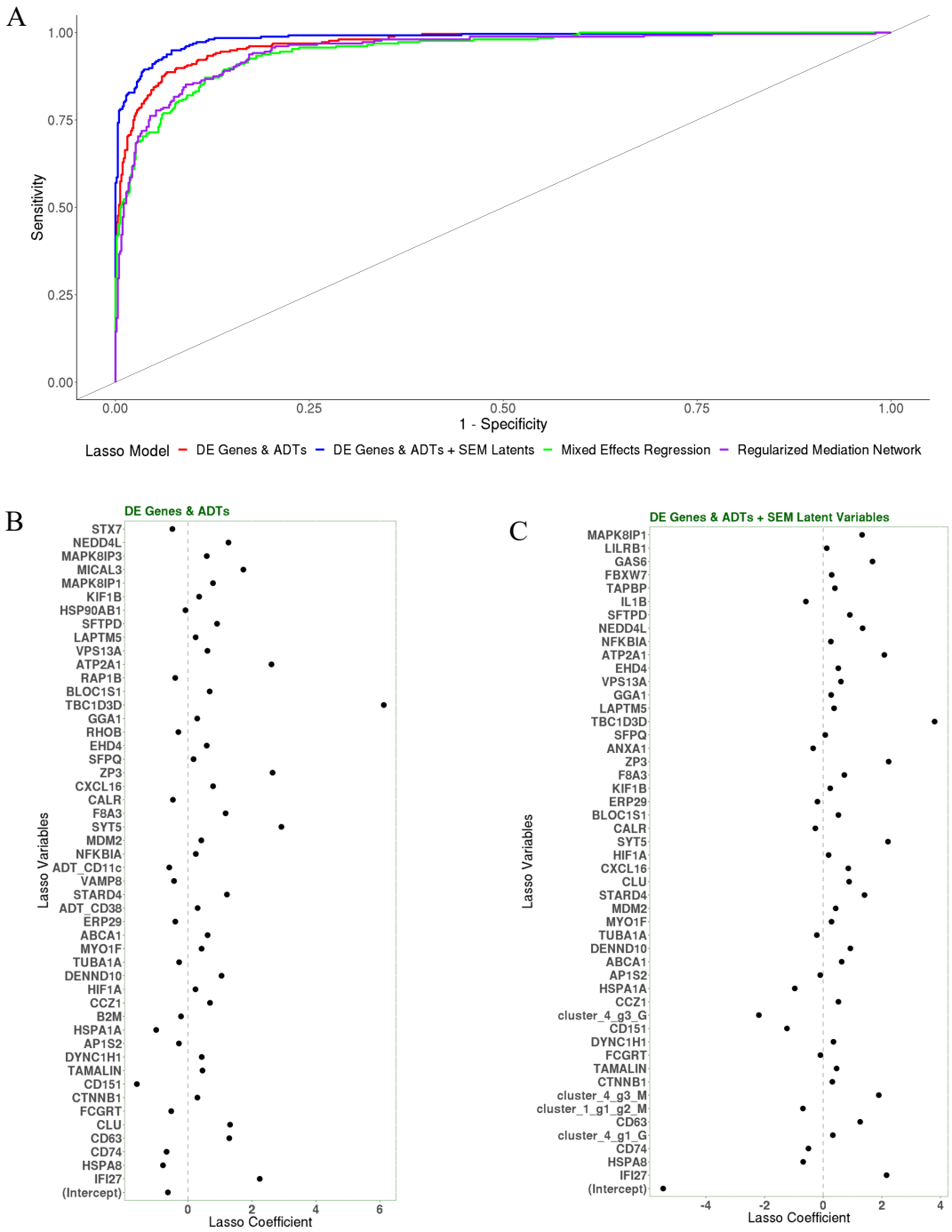

Figure 4. 27: SEM latent variables for feature selection. (A) ROC curves of all feature set models compared. (B) Variable Importance plot showing the top 100 differentially expressed genes and ADTs selected by Lasso stability with their relative importance in the GLM model (C) Variable importance plot showing the CITE-Traffick SEM latent variables in the top 50 selected features

4.5.8 Impact of Sample Size Variability in ICT Analysis

In analyzing the PTTs associated with healthy and varying severities of COVID-19, a noticeable trend emerged wherein fewer significant PTTs were observed in both severe and mild COVID groups compared to the healthy group, with the mild cases demonstrating the fewest significant markers .The disparities in sample sizes across these groups — 5 healthy, 4 severe, and only 3 mild samples may have notably influenced the derived p-values within the mixed-effects regression model. Of the total 5167 CD16+ monocyte cells analyzed, 63% belonged to the healthy group, 25% to the severe cases, and only 12% to the mild cases. This variation in sample distribution has led to considerable variability in the p-values, particularly considering the increased complexity of mixed-effects models requiring larger sample sizes to characterize random effects effectively (*Power Analysis and Effect Size in Mixed Effects Models: A Tutorial - Journal of Cognition*, n.d.) (Jenkins & Quintana-Ascencio, 2020). An alternative approach could involve pooling the samples, disregarding the case-control status, and considering the actual sample as the random effect to identify significant PTTs (Fonseka et al., 2018).

4.6 Discussion

CITE-Traffick algorithm, which employs a three-module approach, enabled us to identify ICT genes associated with the surface protein profiles, establish mediation effects on the disease phenotype, and uncover dysregulated pathways. In the application of CITE-Traffick to COVID-19, we harnessed the power of CITE-Traffick to scrutinize the dysregulation of intracellular trafficking (ICT) genes, in the context of the observed low HLA-DR expression in monocytes of severe COVID-19 patients.

Notably, our investigation unearthed a subset of ICT genes exhibiting consistent dysregulation across both the severe COVID-19 versus healthy and severe COVID-19 versus mild COVID-19 comparisons. These commonly dysregulated ICT genes offer valuable insights into potential factors contributing to disease progression and severity in COVID-19.

In addition to these shared findings, our analysis further unveiled a subset of ICT genes with unique dysregulation specific to the severe and healthy COVID-19 comparisons. These genes may hold intricate associations with the pathological processes and immune responses characteristic of severe COVID-19 cases, potentially playing a pivotal role in driving disease progression and severity. Similarly, in the comparison between severe COVID-19 samples and their mild counterparts, we identified yet another distinct set of ICT genes exhibiting significant associations. These genes might play a crucial role in modulating the transition from mild to severe COVID-19, thus contributing to the observed differences in clinical presentation and disease outcomes.

Even though the regularized mediation models within the CITE-Traffick methodology provided significant insights, it forces less critical factor coefficients to zero, potentially resulting in the elimination of certain connections between exposures and mediators. However, the connections that are retained demonstrate robust and substantial strength, emphasizing their importance within the intricate network of intracellular trafficking (ICT) genes. Additionally, with the growing number of exposures and mediators, the computational intensity required to infer the regularization parameters in regularized mediation analysis increases.

Structural Equation Modeling (SEM) offers a robust means of exploring latent relationships within ICT genes and proteins, providing a more comprehensive view of the intricate network of ICT genes. In essence, the SEM results shed light on the intricate and multifaceted nature of ICT genes. It became apparent that while some ICT genes exhibit distinctive impacts on protein trafficking and disease individually, while others wield substantial influence within the context of a network. This underscores the imperative for future research to develop a comprehensive model that accommodates both individual gene effects and their interplay within the network, essential for a deeper understanding of their roles in cellular processes and diseases. This holistic approach could unveil a more comprehensive understanding of ICT functionalities and implications in various disease contexts.

These promising results underscore the potential of CITE-Traffick as a valuable tool for unraveling the dynamics of intracellular trafficking and its connection to disease phenotypes. Furthermore, the ability of CITE-Traffick to dissect complex protein transportation networks into smaller trios and latent structures has proven instrumental in capturing essential molecular insights.

CHAPTER 5

CONCLUSIONS & FUTURE DIRECTIONS

In this dissertation, three innovative approaches to multi-omics integrative analysis were introduced, each shedding light on complex biological processes and expanding the horizons of knowledge. INCLOSE, a novel single-cell integrative clustering method, excels in enhancing the resolution of cell subpopulations by harnessing the power of multi-omics data. The algorithm demonstrated its prowess when applied to challenging datasets, offering a fresh perspective on conditions like Acute Myeloid Leukemia (AML) by revealing hidden nuances within cell populations.

The exploration extended further to uncover transcriptional regulatory trios in the third chapter, where a unique regression model was implemented. By dissecting the interactions between transcription factors, regulatory elements, and genetic variants, this approach provided essential insights into the dynamics of cellular regulation, particularly in the context of multiple myeloma. These findings brought to light the potential regulatory elements associated with tumorigenesis.

The journey culminated with the introduction of the CITE-Traffick algorithm in the fourth chapter, dedicated to unraveling intracellular protein trafficking. With a focus on COVID-19, this approach identified dysregulated protein trafficking pathways linked to the severity of infection, thus offering a unique perspective on the disease. Notably, this analysis highlighted shared and unique dysregulated ICT genes in severe COVID-19 cases, enhancing our understanding of the factors contributing to disease progression and severity.

Collectively, these innovative approaches mark a substantial advancement in multi-omics research, presenting promising results. However, there is ample room for further refinement and development. In the case of INCLOSE, one crucial avenue involves the incorporation of additional variables, such as sex and batch information, into the clustering analysis. This expansion aims to empower INCLOSE with a more comprehensive understanding of the intricacies within single-cell omic data, enabling it to discern nuanced patterns associated with a broader spectrum of experimental conditions. Concurrently, an intriguing exploration lies in a comparative analysis with CellSIUS, a recent method tailored for the identification of rare cell populations in scRNA-seq data. This comparative scrutiny seeks to unravel the performance nuances between the two methodologies, focusing particularly on CellSIUS's effectiveness in delineating rare cell populations and assessing whether INCLOSE can surpass or augment its capabilities. Furthermore, the algorithm's mettle will be rigorously tested through its application to well-established datasets like the Seurat reference or simulated datasets, providing a stringent assessment of its reliability and generalizability. While

INCLOSE excels at identifying condition-specific clusters, addressing mixed clusters composed of cells with notable differential gene expression between conditions presents another exciting avenue for improvement. On the technical front, future work entails the optimization of the correlation algorithm within INCLOSE for speed and efficiency. As the number of cells increases, it imposes heavier computational demands and necessitates more complex parameter tuning. Future work should focus on streamlining the computation of cell-cell affinities and devising more efficient approaches for parameter optimization. Consideration will be given to integrating fast correlation methods for rapid correlation computation, a pivotal enhancement for accommodating larger cell populations and ensuring the scalability of the algorithm.

One limitation inherent in the current approach of iteratively testing multiple parameter combinations in INCLOSE is the challenge in identifying the most biologically meaningful clustering result solely based on the segregation Z score. The Z score serves as a heuristic measure, offering insights into clustering quality, but it may not always align with the most biologically relevant outcomes. To address this limitation, there is a crucial need to develop a more sophisticated heuristic or optimization strategy that can better discern biologically significant clustering outcomes from the multitude of potential parameter combinations. These forward-looking initiatives collectively aim to fortify INCLOSE's adaptability, performance, and reliability, positioning it as a robust tool for unraveling intricate biological insights from diverse single-cell omic datasets.

In the roadmap for advancing the CITE-Traffick algorithm, an essential future step involves the development and application of simulated datasets tailored to represent

the nuanced dynamics and intricacies inherent in ICT processes, offering a robust testing ground for the algorithm's capabilities. By simulating datasets that mirror the multifaceted nature of ICT, the algorithm's performance, adaptability, and capacity to decode the complexities within such cellular processes can be rigorously evaluated and refined, ensuring its effectiveness in handling real-world complexities. Within the CITE-Traffick SEM method, we currently employ hierarchical clustering, followed by manual cluster selection through visual examination. Future research should concentrate on the development of clustering methods capable of autonomously deriving clusters without the need for user intervention. Additionally, computationally testing clusters for latent variable computation and refining clusters to enhance the SEM model's performance are areas ripe for further enhancement.

In the trajectory of advancing the CITE-Traffick algorithm within the domain of COVID datasets, a pivotal future task involves its application in discerning molecular deregulations underlying disease progression and severity. This will encompass deploying the method to conduct comparisons between mild vs. healthy and severe vs. healthy conditions. A critical component of this future endeavor involves synthesizing the findings into a comprehensive working hypothesis illustrated through a diagram that servers as a visual narrative that encapsulates the intricacies of deregulation in various molecular pathways, contributing to the progression of COVID from a molecular standpoint.

The goal is to uncover distinct molecular profiles characterizing disease severity and progression, thus refining our understanding of the pathophysiological mechanisms at play to unveil potential therapeutic targets and deeper comprehension of disease dynamics.

The observed discrepancy in the number of significant PTTs among the different COVID severity groups, notably fewer in the mild group, could be attributed to the limited sample sizes across these categories, influencing the statistical power and p-values. This showcases the importance of robust sample representation, particularly for mixed-effects models. Future studies could benefit from addressing these limitations by exploring larger and more balanced sample sets by pooling samples from different COVID-19 severity groups to enhance the model's sensitivity and reliability in identifying potential ICTs for various COVID-19 severities.

These future endeavors will serve to fortify the effectiveness and applicability of these innovative approaches in multi-omics research. Overall, this dissertation contributes to the growing body of knowledge in multi-omics integrative analysis, and the results and methods presented herein pave the way for further research and discovery in the realm of complex biological processes and diseases.

REFERENCES

Abugessaisa, I., Ramilowski, J. A., Lizio, M., Severin, J., Hasegawa, A., Harshbarger, J., Kondo, A., Noguchi, S., Yip, C. W., Ooi, J. L. C., Tagami, M., Hori, F., Agrawal, S., Hon, C. C., Cardon, M., Ikeda, S., Ono, H., Bono, H., Kato, M., … Kasukawa, T. (2021). FANTOM enters 20th year: Expansion of transcriptomic atlases and functional annotation of non-coding RNAs. *Nucleic Acids Research*, *49*(D1), D892–D898. https://doi.org/10.1093/nar/gkaa1054

Agarwal, S., Loder, S., Cholok, D., Li, J., Breuler, C., Drake, J., Brownley, C., Peterson, J., Li, S., & Levi, B. (2017). Surgical Excision of Heterotopic Ossification Leads to Re-Emergence of Mesenchymal Stem Cell Populations Responsible for Recurrence. *Stem Cells Translational Medicine*, *6*(3), 799–806. https://doi.org/10.5966/sctm.2015-0365

Ahmadinejad, N., Chung, Y., & Liu, L. (2023). J-score: A robust measure of clustering accuracy. *PeerJ Computer Science*, *9*, e1545. https://doi.org/10.7717/peerj-cs.1545

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). Membrane Proteins. In *Molecular Biology of the Cell. 4th edition*. Garland Science. https://www.ncbi.nlm.nih.gov/books/NBK26878/

Al-Matary, Y. S., Botezatu, L., Opalka, B., Hönes, J. M., Lams, R. F., Thivakaran, A., Schütte, J., Köster, R., Lennartz, K., Schroeder, T., Haas, R., Dührsen, U., & Khandanpour, C. (2016). Acute myeloid leukemia cells polarize macrophages towards a leukemia supporting state in a Growth factor independence 1 dependent manner. *Haematologica*, *101*(10), 1216–1227. https://doi.org/10.3324/haematol.2016.143180

Amati, B., & Land, H. (1994). Myc-Max-Mad: A transcription factor network controlling cell cycle progression, differentiation and death. *Current Opinion in Genetics & Development*, *4*(1), 102–108. https://doi.org/10.1016/0959-437x(94)90098-1

Angermueller, C., Clark, S. J., Lee, H. J., Macaulay, I. C., Teng, M. J., Hu, T. X., Krueger, F., Smallwood, S. A., Ponting, C. P., Voet, T., Kelsey, G., Stegle, O., & Reik, W. (2016). Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nature Methods*, *13*(3), Article 3. https://doi.org/10.1038/nmeth.3728

Arunachalam, P. S., Wimmers, F., Mok, C. K. P., Perera, R. A. P. M., Scott, M., Hagan, T., Sigal, N., Feng, Y., Bristow, L., Tak-Yin Tsang, O., Wagh, D., Coller, J., Pellegrini, K. L., Kazmin, D., Alaaeddine, G., Leung, W. S., Chan, J. M. C., Chik, T. S. H., Choi, C. Y. C., … Pulendran, B. (2020). Systems biological assessment of immunity to mild versus severe COVID-19 infection in humans. *Science*, *369*(6508), 1210–1220. https://doi.org/10.1126/science.abc6261

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene Ontology: Tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29. https://doi.org/10.1038/75556

Baghela, A., An, A., Zhang, P., Acton, E., Gauthier, J., Brunet-Ratnasingham, E., Blimkie, T., Freue, G. C., Kaufmann, D., Lee, A. H. Y., Levesque, R. C., & Hancock, R. E. W. (2023). Predicting severity in COVID-19 disease using sepsis blood gene expression signatures. *Scientific Reports*, *13*(1), Article 1. https://doi.org/10.1038/s41598-023-28259-y

Baker, D. N., Dyjack, N., Braverman, V., Hicks, S. C., & Langmead, B. (2021). Fast and memory-efficient scRNA-seq k-means clustering with various distances. *ACM-BCB ... ... : The ... ACM Conference on Bioinformatics, Computational Biology and Biomedicine. ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, *2021*, 24. https://doi.org/10.1145/3459930.3469523

Bakillah, A., Hejji, F. A., Almasaud, A., Jami, H. A., Hawwari, A., Qarni, A. A., Iqbal, J., & Alharbi, N. K. (2022). Lipid Raft Integrity and Cellular Cholesterol Homeostasis Are Critical for SARS-CoV-2 Entry into Cells. *Nutrients*, *14*(16), 3417. https://doi.org/10.3390/nu14163417

Bare, L. A., Morrison, A. C., Rowland, C. M., Shiffman, D., Luke, M. M., Iakoubova, O. A., Kane, J. P., Malloy, M. J., Ellis, S. G., Pankow, J. S., Willerson, J. T., Devlin, J. J., & Boerwinkle, E. (2007). Five common gene variants identify elevated genetic risk for coronary heart disease. *Genetics in Medicine*, *9*(10), Article 10. https://doi.org/10.1097/GIM.0b013e318156fb62

Barh, D., Blum, K., & Madigan, M. A. (2016). *OMICS: Biomedical Perspectives and Applications*. CRC Press.

Benedetti, E., Pučić-Baković, M., Keser, T., Gerstner, N., Büyüközkan, M., Štambuk, T., Selman, M. H. J., Rudan, I., Polašek, O., Hayward, C., Al-Amin, H., Suhre, K., Kastenmüller, G., Lauc, G., & Krumsiek, J. (2020). A strategy to incorporate prior knowledge into correlation network cutoff selection. *Nature Communications*, *11*(1), Article 1. https://doi.org/10.1038/s41467-020-18675-3

Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M. A., Beaudet, A. L., Ecker, J. R., Farnham, P. J., Hirst, M., Lander, E. S., Mikkelsen, T. S., & Thomson, J. A. (2010). The NIH Roadmap Epigenomics Mapping Consortium. *Nature Biotechnology*, *28*(10), Article 10. https://doi.org/10.1038/nbt1010-1045

Bhardwaj, V., & Ansell, S. M. (2023). Modulation of T-cell function by myeloid-derived suppressor cells in hematological malignancies. *Frontiers in Cell and Developmental Biology*, *11*, 1129343. https://doi.org/10.3389/fcell.2023.1129343

Bhardwaj, V., Jalali, S., Villasboas, J. C., Yang, Z.-Z., Tang, X., Mukherjee, P., Mondello, P., Kim, H., Mudappathi, R., Wang, J., Krull, J. E., Wenzl, K., Novak, A. J., & Ansell, S. M. (2022). Increased Tumor-Associated CD66b+ Myeloid-Derived Suppressor Cells in Waldenstrom Macroglobulinemia Inhibit T-Cell Immune Function. *Blood*, *140*(Supplement 1), 2904–2905. https://doi.org/10.1182/blood-2022-166450

Bizymi, N., Bjelica, S., Kittang, A. O., Mojsilovic, S., Velegraki, M., Pontikoglou, C., Roussel, M., Ersvær, E., Santibañez, J. F., Lipoldová, M., & Papadaki, H. A. (2019). Myeloid-Derived Suppressor Cells in Hematologic Diseases: Promising Biomarkers and Treatment Targets. *HemaSphere*, *3*(1), e168. https://doi.org/10.1097/HS9.0000000000000168

Blanco, B., Pérez-Simón, J. A., Sánchez-Abarca, L. I., Caballero-Velazquez, T., Gutierrez-Cossío, S., Hernández-Campo, P., Díez-Campelo, M., Herrero-Sanchez, C., Rodriguez-Serrano, C., Santamaría, C., Sánchez-Guijo, F. M., del Cañizo, C., & San Miguel, J. F. (2009). Treatment with bortezomib of human CD4+ T cells preserves natural regulatory T cells and allows the emergence of a distinct suppressor T-cell population. *Haematologica*, *94*(7), 975–983. https://doi.org/10.3324/haematol.2008.005017

Boeva, V. (2016). Analysis of Genomic Sequence Motifs for Deciphering Transcription Factor Binding and Transcriptional Regulation in Eukaryotic Cells. *Frontiers in Genetics*, *7*. https://www.frontiersin.org/articles/10.3389/fgene.2016.00024

Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., Karczewski, K. J., Park, J., Hitz, B. C., Weng, S., Cherry, J. M., & Snyder, M. (2012). Annotation of functional variation in personal genomes using RegulomeDB. *Genome Research*, *22*(9), 1790–1797. https://doi.org/10.1101/gr.137323.112

Brachet, V., Péhau-Arnaudet, G., Desaymard, C., Raposo, G., & Amigorena, S. (1999). Early Endosomes Are Required for Major Histocompatiblity Complex Class II Transport to Peptide-loading Compartments. *Molecular Biology of the Cell*, *10*(9), 2891–2904. https://doi.org/10.1091/mbc.10.9.2891

Bronte, V., Brandau, S., Chen, S.-H., Colombo, M. P., Frey, A. B., Greten, T. F., Mandruzzato, S., Murray, P. J., Ochoa, A., Ostrand-Rosenberg, S., Rodriguez, P. C., Sica, A., Umansky, V., Vonderheide, R. H., & Gabrilovich, D. I. (2016). Recommendations for myeloid-derived suppressor cell nomenclature and characterization standards. *Nature Communications*, *7*, 12150. https://doi.org/10.1038/ncomms12150

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, *10*(12), Article 12. https://doi.org/10.1038/nmeth.2688

Calmodulin interacts with angiotensin-converting enzyme-2 (ACE2) and inhibits shedding of its ectodomain—Lambert—2008—FEBS Letters—Wiley Online Library. (n.d.). Retrieved October 26, 2023, from https://febs.onlinelibrary.wiley.com/doi/10.1016/j.febslet.2007.11.085

Cano-Gamez, E., & Trynka, G. (2020). From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Frontiers in Genetics*, *11*. https://www.frontiersin.org/articles/10.3389/fgene.2020.00424

Cassetta, L., Baekkevold, E. S., Brandau, S., Bujko, A., Cassatella, M. A., Dorhoi, A., Krieg, C., Lin, A., Loré, K., Marini, O., Pollard, J. W., Roussel, M., Scapini, P., Umansky, V., & Adema, G. J. (2019). Deciphering myeloid-derived suppressor cells: Isolation and markers in humans, mice and non-human primates. *Cancer Immunology, Immunotherapy*, *68*(4), 687–697. https://doi.org/10.1007/s00262-019-02302-2

Chaudhary, K., Poirion, O. B., Lu, L., & Garmire, L. X. (2018). Deep Learning–Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. *Clinical Cancer Research*, *24*(6), 1248–1259. https://doi.org/10.1158/1078-0432.CCR-17-0853

Chen, G., Gharib, T. G., Huang, C.-C., Taylor, J. M. G., Misek, D. E., Kardia, S. L. R., Giordano, T. J., Iannettoni, M. D., Orringer, M. B., Hanash, S. M., & Beer, D. G. (2002). Discordant Protein and mRNA Expression in Lung Adenocarcinomas *. *Molecular & Cellular Proteomics*, *1*(4), 304–313. https://doi.org/10.1074/mcp.M200008-MCP200

Chen, P., Wu, M., He, Y., Jiang, B., & He, M.-L. (2023). Metabolic alterations upon SARS-CoV-2 infection and potential therapeutic targets against coronavirus infection. *Signal Transduction and Targeted Therapy*, *8*, 237. https://doi.org/10.1038/s41392-023-01510-8

Chi, K. N., Hotte, S. J., Yu, E. Y., Tu, D., Eigl, B. J., Tannock, I., Saad, F., North, S., Powers, J., Gleave, M. E., & Eisenhauer, E. A. (2010). Randomized Phase II Study of Docetaxel and Prednisone With or Without OGX-011 in Patients With Metastatic Castration-Resistant Prostate Cancer. *Journal of Clinical Oncology*, *28*(27), 4247–4254. https://doi.org/10.1200/JCO.2009.26.8771

Chng, W.-J., Huang, G. F., Chung, T. H., Ng, S. B., Gonzalez-Paz, N., Troska-Price, T., Mulligan, G., Chesi, M., Bergsagel, P. L., & Fonseca, R. (2011). Clinical and biological implications of MYC activation: A common difference between MGUS and newly diagnosed multiple myeloma. *Leukemia*, *25*(6), Article 6. https://doi.org/10.1038/leu.2011.53

Coetzee, S. G., Rhie, S. K., Berman, B. P., Coetzee, G. A., & Noushmehr, H. (2012a). FunciSNP: An R/bioconductor tool integrating functional non-coding data sets with genetic association studies to identify candidate regulatory SNPs. *Nucleic Acids Research*, *40*(18), e139. https://doi.org/10.1093/nar/gks542

Coetzee, S. G., Rhie, S. K., Berman, B. P., Coetzee, G. A., & Noushmehr, H. (2012b). FunciSNP: An R/bioconductor tool integrating functional non-coding data sets with genetic association studies to identify candidate regulatory SNPs. *Nucleic Acids Research*, *40*(18), e139. https://doi.org/10.1093/nar/gks542

Cui, M., Cheng, C., & Zhang, L. (2022). High-throughput proteomics: A methodological mini-review. *Laboratory Investigation*, *102*(11), Article 11. https://doi.org/10.1038/s41374-022-00830-7

Cusanovich, D. A., Hill, A. J., Aghamirzaie, D., Daza, R. M., Pliner, H. A., Berletch, J. B., Filippova, G. N., Huang, X., Christiansen, L., DeWitt, W. S., Lee, C., Regalado, S. G., Read, D. F., Steemers, F. J., Disteche, C. M., Trapnell, C., & Shendure, J. (2018). A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell*, *174*(5), 1309-1324.e18. https://doi.org/10.1016/j.cell.2018.06.052

Dai, J., Wang, H., Liao, Y., Tan, L., Sun, Y., Song, C., Liu, W., Qiu, X., & Ding, C. (2022). Coronavirus Infection and Cholesterol Metabolism. *Frontiers in Immunology*, *13*. https://www.frontiersin.org/articles/10.3389/fimmu.2022.791267

Darden, D. B., Bacher, R., Brusko, M. A., Knight, P., Hawkins, R. B., Cox, M. C., Dirain, M. L., Ungaro, R., Nacionales, D. C., Rincon, J. C., Gauthier, M.-P. L., Kladde, M., Bihorac, A., Brusko, T. M., Moore, F. A., Brakenridge, S. C., Mohr, A. M., Moldawer, L. L., & Efron, P. A. (2021). Single Cell RNA-seq of Human Myeloid Derived Suppressor Cells in Late Sepsis Reveals Multiple Subsets with Unique Transcriptional Responses: A Pilot Study. *Shock (Augusta, Ga.)*, *55*(5), 587–595. https://doi.org/10.1097/SHK.0000000000001671

Degtyareva, A. O., Antontseva, E. V., & Merkulova, T. I. (2021). Regulatory SNPs: Altered Transcription Factor Binding Sites Implicated in Complex Traits and Diseases. *International Journal of Molecular Sciences*, *22*(12), 6454. https://doi.org/10.3390/ijms22126454

Dendrou, C. A., Petersen, J., Rossjohn, J., & Fugger, L. (2018). HLA variation and disease. *Nature Reviews Immunology*, *18*(5), Article 5. https://doi.org/10.1038/nri.2017.143

Deng, G., Zhang, X., Chen, Y., Liang, S., Liu, S., Yu, Z., & Lü, M. (2023). Single-cell transcriptome sequencing reveals heterogeneity of gastric cancer: Progress and prospects. *Frontiers in Oncology*, *13*, 1074268. https://doi.org/10.3389/fonc.2023.1074268

Deng, N., Zhou, H., Fan, H., & Yuan, Y. (2017). Single nucleotide polymorphisms and cancer susceptibility. *Oncotarget*, *8*(66), 110635–110649. https://doi.org/10.18632/oncotarget.22372

Dey, S. S., Kester, L., Spanjaard, B., Bienko, M., & van Oudenaarden, A. (2015). Integrated genome and transcriptome sequencing from the same cell. *Nature Biotechnology*, *33*(3), 285–289. https://doi.org/10.1038/nbt.3129

Eichler, E. E., Nickerson, D. A., Altshuler, D., Bowcock, A. M., Brooks, L. D., Carter, N. P., Church, D. M., Felsenfeld, A., Guyer, M., Lee, C., Lupski, J. R., Mullikin, J. C., Pritchard, J. K., Sebat, J., Sherry, S. T., Smith, D., Valle, D., Waterston, R. H., & The Human Genome Structural Variation Working Group. (2007). Completing the map of human genetic variation. *Nature*, *447*(7141), Article 7141. https://doi.org/10.1038/447161a

Epigenomics | Learn Science at Scitable. (n.d.). Retrieved October 17, 2023, from http://www.nature.com/scitable/topicpage/epigenomics-the-new-tool-in-studying-complex-694

Ester, M., Kriegel, H.-P., & Xu, X. (n.d.). *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*.

Fan, Y., Chen, J., Shirkey, G., John, R., Wu, S. R., Park, H., & Shao, C. (2016). Applications of structural equation modeling (SEM) in ecological studies: An updated review. *Ecological Processes*, *5*(1), 19. https://doi.org/10.1186/s13717-016-0063-3

Félix, I., Jokela, H., Karhula, J., Kotaja, N., Savontaus, E., Salmi, M., & Rantakari, P. (2021). Single-Cell Proteomics Reveals the Defined Heterogeneity of Resident Macrophages in White Adipose Tissue. *Frontiers in Immunology*, *12*. https://www.frontiersin.org/articles/10.3389/fimmu.2021.719979

Flynn, E. D., Tsu, A. L., Kasela, S., Kim-Hellmuth, S., Aguet, F., Ardlie, K. G., Bussemaker, H. J., Mohammadi, P., & Lappalainen, T. (2022). Transcription factor regulation of eQTL activity across individuals and tissues. *PLoS Genetics*, *18*(1), e1009719. https://doi.org/10.1371/journal.pgen.1009719

Fonseka, C. Y., Rao, D. A., Teslovich, N. C., Korsunsky, I., Hannes, S. K., Slowikowsi, K., Gurish, M. F., Donlin, L. T., Lederer, J. A., Weinblatt, M. E., Massarotti, E. M., Coblyn, J. S., Helfgott, S. M., Todd, D. J., Bykerk, V. P., Karlson, E. W., Ermann, J., Lee, Y. C., Brenner, M. B., & Raychaudhuri, S. (2018). Mixed-Effects Association of Single Cells Identifies an Expanded Effector CD4+ T Cell Subset in Rheumatoid Arthritis. *Science Translational Medicine*, *10*(463), eaaq0305. https://doi.org/10.1126/scitranslmed.aaq0305

Foster, E. M., Dangla-Valls, A., Lovestone, S., Ribe, E. M., & Buckley, N. J. (2019). Clusterin in Alzheimer's Disease: Mechanisms, Genetics, and Lessons From Other Pathologies. *Frontiers in Neuroscience*, *13*. https://www.frontiersin.org/articles/10.3389/fnins.2019.00164

Gabrilovich, D. I. (2017). Myeloid-Derived Suppressor Cells. *Cancer Immunology Research*, *5*(1), 3–8. https://doi.org/10.1158/2326-6066.CIR-16-0297

Gao, C., Zhuang, J., Zhou, C., Li, H., Liu, C., Liu, L., Feng, F., Liu, R., & Sun, C. (2019). SNP mutation-related genes in breast cancer for monitoring and prognosis of patients: A study based on the TCGA database. *Cancer Medicine*, *8*(5), 2303–2312. https://doi.org/10.1002/cam4.2065

Giang, T.-T., Nguyen, T.-P., & Tran, D.-H. (2020). Stratifying patients using fast multiple kernel learning framework: Case studies of Alzheimer's disease and cancers. *BMC Medical Informatics and Decision Making*, *20*(1), 108. https://doi.org/10.1186/s12911-020-01140-y

Goedhart, M., Cornelissen, A. S., Kuijk, C., Geerman, S., Kleijer, M., van Buul, J. D., Huveneers, S., Raaijmakers, M. H. G. P., Young, H. A., Wolkers, M. C., Voermans, C., & Nolte, M. A. (2018). Interferon-Gamma Impairs Maintenance and Alters Hematopoietic Support of Bone Marrow Mesenchymal Stromal Cells. *Stem Cells and Development*, *27*(9), 579–589. https://doi.org/10.1089/scd.2017.0196

Gong, J., Mei, S., Liu, C., Xiang, Y., Ye, Y., Zhang, Z., Feng, J., Liu, R., Diao, L., Guo, A.-Y., Miao, X., & Han, L. (2018). PancanQTL: Systematic identification of cis-eQTLs and trans-eQTLs in 33 cancer types. *Nucleic Acids Research*, *46*(D1), D971–D976. https://doi.org/10.1093/nar/gkx861

Gozuacik, D., & Kimchi, A. (2007). Autophagy and Cell Death. In *Current Topics in Developmental Biology* (Vol. 78, pp. 217–245). Academic Press. https://doi.org/10.1016/S0070-2153(06)78006-1

Grün, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., Clevers, H., & van Oudenaarden, A. (2015). Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*, *525*(7568), Article 7568. https://doi.org/10.1038/nature14966

Gundogdu, P., Loucera, C., Alamo-Alvarez, I., Dopazo, J., & Nepomuceno, I. (2022). Integrating pathway knowledge with deep neural networks to reduce the dimensionality in single-cell RNA-seq data. *BioData Mining*, *15*(1), 1. https://doi.org/10.1186/s13040-021-00285-4

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E. P., Jain, J., Srivastava, A., Stuart, T., Fleming, L. M., Yeung, B., … Satija, R. (2021). Integrated analysis of multimodal single-cell data. *Cell*, *184*(13), 3573-3587.e29. https://doi.org/10.1016/j.cell.2021.04.048

Haque, A., Engel, J., Teichmann, S. A., & Lönnberg, T. (2017). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Medicine*, *9*(1), 75. https://doi.org/10.1186/s13073-017-0467-4

Hassan, Z., Kumar, N. D., Reggiori, F., & Khan, G. (2021). How Viruses Hijack and Modify the Secretory Transport Pathway. *Cells*, *10*(10), 2535. https://doi.org/10.3390/cells10102535

He, Q., Liu, Z., Liu, Z., Lai, Y., Zhou, X., & Weng, J. (2019). TCR-like antibodies in cancer immunotherapy. *Journal of Hematology & Oncology*, *12*(1), 99. https://doi.org/10.1186/s13045-019-0788-4

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., & Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, *106*(23), 9362–9367. https://doi.org/10.1073/pnas.0903103106

Holien, T., Misund, K., Olsen, O. E., Baranowska, K. A., Buene, G., Børset, M., Waage, A., & Sundan, A. (2015). MYC amplifications in myeloma cell lines: Correlation with MYC-inhibitor efficacy. *Oncotarget*, *6*(26), 22698–22705. https://doi.org/10.18632/oncotarget.4245

Hou, J., Ye, X., Feng, W., Zhang, Q., Han, Y., Liu, Y., Li, Y., & Wei, Y. (2022). Distance correlation application to gene co-expression network analysis. *BMC Bioinformatics*, *23*, 81. https://doi.org/10.1186/s12859-022-04609-x

Hu, J. M., Liu, K., Liu, J. H., Jiang, X. L., Wang, X. L., Chen, Y. Z., Li, S. G., Zou, H., Pang, L. J., Liu, C. X., Cui, X. B., Yang, L., Zhao, J., Shen, X. H., Jiang, J. F., Liang, W. H., Yuan, X. L., & Li, F. (2017). CD163 as a marker of M2 macrophage, contribute to predict aggressiveness and prognosis of Kazakh esophageal squamous cell carcinoma. *Oncotarget*, *8*(13), 21526–21538. https://doi.org/10.18632/oncotarget.15630

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hyun, S. Y., Na, E. J., Jang, J. E., Chung, H., Kim, S. J., Kim, J. S., Kong, J. H., Shim, K. Y., Lee, J. I., Min, Y. H., & Cheong, J. (2020). Immunosuppressive role of CD11b+CD33+HLA-DR− myeloid-derived suppressor cells-like blast subpopulation in acute myeloid leukemia. *Cancer Medicine*, *9*(19), 7007–7017. https://doi.org/10.1002/cam4.3360

Jalali, S., Villasboas, J., Shi, J., Bothun, C., Kim, H., Yang, Z.-Z., & Ansell, S. M. (2019). Mass Cytometry Identifies a Novel Signature for Myeloid-Derived Suppressor-Cells in Waldenstrom's Macroglobulinemia. *Blood*, *134*(Supplement_1), 3976. https://doi.org/10.1182/blood-2019-124850

Jarlhelt, I., Nielsen, S. K., Jahn, C. X. H., Hansen, C. B., Pérez-Alós, L., Rosbjerg, A., Bayarri-Olmos, R., Skjoedt, M.-O., & Garred, P. (2021). SARS-CoV-2 Antibodies Mediate Complement and Cellular Driven Inflammation. *Frontiers in Immunology*, *12*. https://www.frontiersin.org/articles/10.3389/fimmu.2021.767981

Jenkins, D. G., & Quintana-Ascencio, P. F. (2020). A solution to minimum sample size for regressions. *PLOS ONE*, *15*(2), e0229345. https://doi.org/10.1371/journal.pone.0229345

Jin, H.-J., Jung, S., DebRoy, A. R., & Davuluri, R. V. (2016). Identification and validation of regulatory SNPs that modulate transcription factor chromatin binding and gene expression in prostate cancer. *Oncotarget*, *7*(34), 54616–54626. https://doi.org/10.18632/oncotarget.10520

Kashima, Y., Togashi, Y., Fukuoka, S., Kamada, T., Irie, T., Suzuki, A., Nakamura, Y., Shitara, K., Minamide, T., Yoshida, T., Taoka, N., Kawase, T., Wada, T., Inaki, K., Chihara, M., Ebisuno, Y., Tsukamoto, S., Fujii, R., Ohashi, A., … Doi, T. (2021). Potentiality of multiple modalities for single-cell analyses to evaluate the tumor microenvironment in clinical specimens. *Scientific Reports*, *11*(1), Article 1. https://doi.org/10.1038/s41598-020-79385-w

Keskinen, P., Ronni, T., Matikainen, S., Lehtonen, A., & Julkunen, I. (1997). Regulation of HLA class I and II expression by interferons and influenza A virus in human peripheral blood mononuclear cells. *Immunology*, *91*(3), 421–429. https://doi.org/10.1046/j.1365-2567.1997.00258.x

Kircheis, R., Haasbach, E., Lueftenegger, D., Heyken, W. T., Ocker, M., & Planz, O. (2020). NF-κB Pathway as a Potential Target for Treatment of Critical Stage COVID-19 Patients. *Frontiers in Immunology*, *11*. https://www.frontiersin.org/articles/10.3389/fimmu.2020.598444

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, *46*(3), Article 3. https://doi.org/10.1038/ng.2892

Kodam, P., Sai Swaroop, R., Pradhan, S. S., Sivaramakrishnan, V., & Vadrevu, R. (2023). Integrated multi-omics analysis of Alzheimer's disease shows molecular signatures associated with disease progression and potential therapeutic targets. *Scientific Reports*, *13*(1), Article 1. https://doi.org/10.1038/s41598-023-30892-6

Kumar, A., Ahmad, A., Vyawahare, A., & Khan, R. (2020). Membrane Trafficking and Subcellular Drug Targeting Pathways. *Frontiers in Pharmacology*, *11*. https://www.frontiersin.org/articles/10.3389/fphar.2020.00629

Kumar, R., Kumar, V., Arya, R., Anand, U., & Priyadarshi, R. N. (2022). Association of COVID-19 with hepatic metabolic dysfunction. *World Journal of Virology*, *11*(5), 237–251. https://doi.org/10.5501/wjv.v11.i5.237

Kvedaraite, E., Hertwig, L., Sinha, I., Ponzetta, A., Hed Myrberg, I., Lourda, M., Dzidic, M., Akber, M., Klingström, J., Folkesson, E., Muvva, J. R., Chen, P., Gredmark-Russ, S., Brighenti, S., Norrby-Teglund, A., Eriksson, L. I., Rooyackers, O., Aleman, S., Strålin, K., … Unge, C. (2021). Major alterations in the mononuclear phagocyte landscape associated with COVID-19 severity. *Proceedings of the National Academy of Sciences*, *118*(6), e2018587118. https://doi.org/10.1073/pnas.2018587118

Kzhyshkowska, J., Gratchev, A., & Goerdt, S. (2006). Stabilin-1, a homeostatic scavenger receptor with multiple functions. *Journal of Cellular and Molecular Medicine*, *10*(3), 635–649. https://doi.org/10.1111/j.1582-4934.2006.tb00425.x

Leblay, N., Maity, R., Barakat, E., McCulloch, S., Duggan, P., Jimenez-Zepeda, V., Bahlis, N. J., & Neri, P. (2020). Cite-Seq Profiling of T Cells in Multiple Myeloma Patients Undergoing BCMA Targeting CAR-T or Bites Immunotherapy. *Blood*, *136*, 11–12. https://doi.org/10.1182/blood-2020-137650

Lee, M., & Rhee, I. (2017). Cytokine Signaling in Tumor Progression. *Immune Network*, *17*(4), 214–227. https://doi.org/10.4110/in.2017.17.4.214

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., & Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, *27*(12), 1739–1740. https://doi.org/10.1093/bioinformatics/btr260

Lin, P., Troup, M., & Ho, J. W. K. (2017). CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biology*, *18*(1), 59. https://doi.org/10.1186/s13059-017-1188-0

Lin, W.-Y., Fordham, S. E., Hungate, E., Sunter, N. J., Elstob, C., Xu, Y., Park, C., Quante, A., Strauch, K., Gieger, C., Skol, A., Rahman, T., Sucheston-Campbell, L., Wang, J., Hahn, T., Clay-Gilmour, A. I., Jones, G. L., Marr, H. J., Jackson, G. H., … Allan, J. M. (2021). Genome-wide association study identifies susceptibility loci for acute myeloid leukemia. *Nature Communications*, *12*(1), Article 1. https://doi.org/10.1038/s41467-021-26551-x

Liu, B.-Q., Bao, Z.-Y., Zhu, J.-Y., & Liu, H. (2021). Fibrinogen-like protein 2 promotes the accumulation of myeloid-derived suppressor cells in the hepatocellular carcinoma tumor microenvironment. *Oncology Letters*, *21*(1), 47. https://doi.org/10.3892/ol.2020.12308

Liu, S., Liu, Y., Zhang, Q., Wu, J., Liang, J., Yu, S., Wei, G.-H., White, K. P., & Wang, X. (2017). Systematic identification of regulatory variants associated with cancer risk. *Genome Biology*, *18*(1), 194. https://doi.org/10.1186/s13059-017-1322-z

Liu, Y., Lu, R., Cui, W., Pang, Y., Liu, C., Cui, L., Qian, T., Quan, L., Dai, Y., Jiao, Y., Pan, Y., Ye, X., Shi, J., Cheng, Z., & Fu, L. (2020). High IFITM3 expression predicts adverse prognosis in acute myeloid leukemia. *Cancer Gene Therapy*, *27*(1), Article 1. https://doi.org/10.1038/s41417-019-0093-y

Liu, Z., & Barahona, M. (2020). Graph-based data clustering via multiscale community detection. *Applied Network Science*, *5*(1), Article 1. https://doi.org/10.1007/s41109-019-0248-7

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., Foster, B., Moser, M., Karasik, E., Gillard, B., Ramsey, K., Sullivan, S., Bridge, J., Magazine, H., Syron, J., … Moore, H. F. (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, *45*(6), Article 6. https://doi.org/10.1038/ng.2653

Luo, Y., Liu, S., Li, H., Hou, J., Lin, W., Xu, Z., Lu, T., Li, Y., Peng, B., Zhang, S., Han, X., Kuang, Z., Wen, Y., Cai, J., Liu, F., & Chen, X.-L. (2022). Mass Cytometry and Single-Cell Transcriptome Analyses Reveal the Immune Cell Characteristics of Ulcerative Colitis. *Frontiers in Molecular Biosciences*, *9*. https://www.frontiersin.org/articles/10.3389/fmolb.2022.859645

Lv, M., Wang, K., & Huang, X. (2019). Myeloid-derived suppressor cells in hematological malignancies: Friends or foes. *Journal of Hematology & Oncology*, *12*(1), 105. https://doi.org/10.1186/s13045-019-0797-3

Ma, J., Fu, Y., Tu, Y., Liu, Y., Tan, Y., Ju, W., Pickering, C. R., Myers, J. N., Zhang, Z., & Zhong, L. (2018). Mutation allele frequency threshold does not affect prognostic analysis using next-generation sequencing in oral squamous cell carcinoma. *BMC Cancer*, *18*(1), 758. https://doi.org/10.1186/s12885-018-4481-8

Ma, L., Sahu, S. K., Cano, M., Kuppuswamy, V., Bajwa, J., McPhatter, J., Pine, A., Meizlish, M. L., Goshua, G., Chang, C. H., Zhang, H., Price, C., Bahel, P., Rinder, H., Lei, T., Day, A., Reynolds, D., Wu, X., Schriefer, R., … Kulkarni, H. S. (2021). Increased complement activation is a distinctive feature of severe SARS-CoV-2 infection. *Science Immunology*, *6*(59), eabh2259. https://doi.org/10.1126/sciimmunol.abh2259

Ma, S., & Zhang, Y. (2020). Profiling chromatin regulatory landscape: Insights into the development of ChIP-seq and ATAC-seq. *Molecular Biomedicine*, *1*(1), 9. https://doi.org/10.1186/s43556-020-00009-w

Macaulay, I. C., Haerty, W., Kumar, P., Li, Y. I., Hu, T. X., Teng, M. J., Goolam, M., Saurat, N., Coupland, P., Shirley, L. M., Smith, M., Van der Aa, N., Banerjee, R., Ellis, P. D., Quail, M. A., Swerdlow, H. P., Zernicka-Goetz, M., Livesey, F. J., Ponting, C. P., & Voet, T. (2015). G&T-seq: Parallel sequencing of single-cell genomes and transcriptomes. *Nature Methods*, *12*(6), Article 6. https://doi.org/10.1038/nmeth.3370

Madden, S. K., de Araujo, A. D., Gerhardt, M., Fairlie, D. P., & Mason, J. M. (2021). Taking the Myc out of cancer: Toward therapeutic strategies to directly inhibit c-Myc. *Molecular Cancer*, *20*(1), 3. https://doi.org/10.1186/s12943-020-01291-6

Mahata, B., Zhang, X., Kolodziejczyk, A. A., Proserpio, V., Haim-Vilmovsky, L., Taylor, A. E., Hebenstreit, D., Dingler, F. A., Moignard, V., Göttgens, B., Arlt, W., McKenzie, A. N. J., & Teichmann, S. A. (2014). Single-Cell RNA Sequencing Reveals T Helper Cells Synthesizing Steroids De Novo to Contribute to Immune Homeostasis. *Cell Reports*, *7*(4), 1130–1142. https://doi.org/10.1016/j.celrep.2014.04.011

Mandruzzato, S., Brandau, S., Britten, C. M., Bronte, V., Damuzzo, V., Gouttefangeas, C., Maurer, D., Ottensmeier, C., van der Burg, S. H., Welters, M. J. P., & Walter, S. (2016). Toward harmonized phenotyping of human myeloid-derived suppressor cells by flow cytometry: Results from an interim study. *Cancer Immunology, Immunotherapy*, *65*(2), 161–169. https://doi.org/10.1007/s00262-015-1782-5

Maneck, M., Schrader, A., Kube, D., & Spang, R. (2011). Genomic data integration using guided clustering. *Bioinformatics*, *27*(16), 2231–2238. https://doi.org/10.1093/bioinformatics/btr363

Mankoo, P. K., Shen, R., Schultz, N., Levine, D. A., & Sander, C. (2011). Time to Recurrence and Survival in Serous Ovarian Tumors Predicted from Integrated Genomic Profiles. *PLOS ONE*, *6*(11), e24709. https://doi.org/10.1371/journal.pone.0024709

Mantovani, A., Sica, A., Allavena, P., Garlanda, C., & Locati, M. (2009). Tumor-associated macrophages and the related myeloid-derived suppressor cells as a paradigm of the diversity of macrophage activation. *Human Immunology*, *70*(5), 325–330. https://doi.org/10.1016/j.humimm.2009.02.008

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutyavin, T., Stehling-Sun, S., Johnson, A. K., Canfield, T. K., Giste, E., Diegel, M., … Stamatoyannopoulos, J. A. (2012). Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science (New York, N.Y.)*, *337*(6099), 1190–1195. https://doi.org/10.1126/science.1222794

Mazzoni, A., Salvati, L., Maggi, L., Annunziato, F., & Cosmi, L. (2021). Hallmarks of immune response in COVID-19: Exploring dysregulation and exhaustion. *Seminars in Immunology*, *55*, 101508. https://doi.org/10.1016/j.smim.2021.101508

Meza-Alvarado, J. C., Page, R. A., Mallard, B., Bromhead, C., & Palmer, B. R. (2023). VEGF-A related SNPs: A cardiovascular context. *Frontiers in Cardiovascular Medicine*, *10*, 1190513. https://doi.org/10.3389/fcvm.2023.1190513

Mo, Q., Li, R., Adeegbe, D. O., Peng, G., & Chan, K. S. (2020). Integrative multi-omics analysis of muscle-invasive bladder cancer identifies prognostic biomarkers for frontline chemotherapy and immunotherapy. *Communications Biology*, *3*(1), Article 1. https://doi.org/10.1038/s42003-020-01491-2

Nathan, A., Asgari, S., Ishigaki, K., Valencia, C., Amariuta, T., Luo, Y., Beynor, J. I., Baglaenko, Y., Suliman, S., Price, A. L., Lecca, L., Murray, M. B., Moody, D. B., & Raychaudhuri, S. (2022). Single-cell eQTL models reveal dynamic T cell state dependence of disease loci. *Nature*, *606*(7912), 120–128. https://doi.org/10.1038/s41586-022-04713-1

Neefjes, J., Jongsma, M. L. M., Paul, P., & Bakke, O. (2011). Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nature Reviews Immunology*, *11*(12), Article 12. https://doi.org/10.1038/nri3084

Nieto-Nicolau, N., de la Torre, R. M., Fariñas, O., Savio, A., Vilarrodona, A., & Casaroli-Marano, R. P. (2020). Extrinsic modulation of integrin α6 and progenitor cell behavior in mesenchymal stem cells. *Stem Cell Research*, *47*, 101899. https://doi.org/10.1016/j.scr.2020.101899

Nishizaki, S. S., Ng, N., Dong, S., Porter, R. S., Morterud, C., Williams, C., Asman, C., Switzenberg, J. A., & Boyle, A. P. (2019). Predicting the effects of SNPs on transcription factor binding affinity. *Bioinformatics*, *36*(2), 364–372. https://doi.org/10.1093/bioinformatics/btz612

Nomenclature for factors of the HLA system, 2010—Marsh—2010—Tissue Antigens—Wiley Online Library. (n.d.). Retrieved October 26, 2023, from https://onlinelibrary.wiley.com/doi/10.1111/j.1399-0039.2010.01466.x

Ostrand-Rosenberg, S., Sinha, P., Beury, D. W., & Clements, V. K. (2012). Cross-talk between myeloid-derived suppressor cells (MDSC), macrophages, and dendritic cells enhances tumor-induced immune suppression. *Seminars in Cancer Biology*, *22*(4), 275–281. https://doi.org/10.1016/j.semcancer.2012.01.011

Pan, C., Hu, T., Liu, P., Ma, D., Cao, S., Shang, Q., Zhang, L., Chen, Q., Fang, Q., & Wang, J. (2023). BM-MSCs display altered gene expression profiles in B-cell acute lymphoblastic leukemia niches and exert pro-proliferative effects via overexpression of IFI6. *Journal of Translational Medicine*, *21*(1), 593. https://doi.org/10.1186/s12967-023-04464-1

Pan, Y.-Q., Niu, M., Liu, S., Bao, Y.-X., Yang, K., Ma, X.-B., He, L., Li, Y.-X., Cao, J.-X., Zhang, X., & Du, Y. (2021). Effect of MT2A on apoptosis and proliferation in HL60 cells. *International Journal of Medical Sciences*, *18*(13), 2910–2919. https://doi.org/10.7150/ijms.57821

Pang, X., He, X., Qiu, Z., Zhang, H., Xie, R., Liu, Z., Gu, Y., Zhao, N., Xiang, Q., & Cui, Y. (2023). Targeting integrin pathways: Mechanisms and advances in therapy. *Signal Transduction and Targeted Therapy*, *8*, 1. https://doi.org/10.1038/s41392-022-01259-6

Peng, Y., Liu, H., Wu, Q., Wang, L., Yu, Y., Yin, F., Feng, C., Ren, X., Liu, T., Chen, L., & Zhu, H. (2023). Integrated bioinformatics analysis and experimental validation reveal ISG20 as a novel prognostic indicator expressed on M2 macrophage in glioma. *BMC Cancer*, *23*(1), 596. https://doi.org/10.1186/s12885-023-11057-0

Pervez, M. T., Hasnain, M. J. ul, Abbas, S. H., Moustafa, M. F., Aslam, N., & Shah, S. S. M. (2022). A Comprehensive Review of Performance of Next-Generation Sequencing Platforms. *BioMed Research International*, *2022*, 3457806. https://doi.org/10.1155/2022/3457806

Peterson, V. M., Zhang, K. X., Kumar, N., Wong, J., Li, L., Wilson, D. C., Moore, R., McClanahan, T. K., Sadekova, S., & Klappenbach, J. A. (2017). Multiplexed quantification of proteins and transcripts in single cells. *Nature Biotechnology*, *35*(10), Article 10. https://doi.org/10.1038/nbt.3973

Piedrafita, F. J., Molander, R. B., Vansant, G., Orlova, E. A., Pfahl, M., & Reynolds, W. F. (1996). An Alu Element in the Myeloperoxidase Promoter Contains a Composite SP1-Thyroid Hormone-Retinoic Acid Response Element*. *Journal of Biological Chemistry*, *271*(24), 14412–14420. https://doi.org/10.1074/jbc.271.24.14412

Power Analysis and Effect Size in Mixed Effects Models: A Tutorial—Journal of Cognition. (n.d.). Retrieved November 12, 2023, from https://journalofcognition.org/articles/10.5334/joc.10#B25

Primer of human genetics | Wageningen University and Research Library catalog. (n.d.). Retrieved October 17, 2023, from https://library.wur.nl/WebQuery/titel/2219407

Reel, P. S., Reel, S., Pearson, E., Trucco, E., & Jefferson, E. (2021). Using machine learning approaches for multi-omics data analysis: A review. *Biotechnology Advances*, *49*, 107739. https://doi.org/10.1016/j.biotechadv.2021.107739

Reimegård, J., Tarbier, M., Danielsson, M., Schuster, J., Baskaran, S., Panagiotou, S., Dahl, N., Friedländer, M. R., & Gallant, C. J. (2021). A combined approach for single-cell mRNA and intracellular protein expression analysis. *Communications Biology*, *4*, 624. https://doi.org/10.1038/s42003-021-02142-w

Rojano, E., Seoane, P., Ranea, J. A. G., & Perkins, J. R. (2019a). Regulatory variants: From detection to predicting impact. *Briefings in Bioinformatics*, *20*(5), 1639–1654. https://doi.org/10.1093/bib/bby039

Rojano, E., Seoane, P., Ranea, J. A. G., & Perkins, J. R. (2019b). Regulatory variants: From detection to predicting impact. *Briefings in Bioinformatics*, *20*(5), 1639–1654. https://doi.org/10.1093/bib/bby039

Rovito, R., Augello, M., Ben-Haim, A., Bono, V., d'Arminio Monforte, A., & Marchetti, G. (2022). Hallmarks of Severe COVID-19 Pathogenesis: A Pas de Deux Between Viral and Host Factors. *Frontiers in Immunology*, *13*. https://www.frontiersin.org/articles/10.3389/fimmu.2022.912336

Satam, H., Joshi, K., Mangrolia, U., Waghoo, S., Zaidi, G., Rawool, S., Thakare, R. P., Banday, S., Mishra, A. K., Das, G., & Malonia, S. K. (2023). Next-Generation Sequencing Technology: Current Trends and Advancements. *Biology*, *12*(7), Article 7. https://doi.org/10.3390/biology12070997

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., & Regev, A. (2015). Spatial reconstruction of single-cell gene expression. *Nature Biotechnology*, *33*(5), 495–502. https://doi.org/10.1038/nbt.3192

Schaid, D. J., Dikilitas, O., Sinnwell, J. P., & Kullo, I. J. (2022). Penalized mediation models for multivariate data. *Genetic Epidemiology*, *46*(1), 32–50. https://doi.org/10.1002/gepi.22433

Schumacker, R. E., & Lomax, R. G. (2010). *A beginner's guide to structural equation modeling, 3rd ed* (pp. xx, 510). Routledge/Taylor & Francis Group.

Schwessinger, R., Suciu, M. C., McGowan, S. J., Telenius, J., Taylor, S., Higgs, D. R., & Hughes, J. R. (2017). Sasquatch: Predicting the impact of regulatory SNPs on transcription factor binding from cell- and tissue-specific DNase footprints. *Genome Research*, *27*(10), 1730–1742. https://doi.org/10.1101/gr.220202.117

Shan, N., Wang, Z., & Hou, L. (2019). Identification of trans-eQTLs using mediation analysis with multiple mediators. *BMC Bioinformatics*, *20*(3), 126. https://doi.org/10.1186/s12859-019-2651-6

Shoily, S. S., Ahsan, T., Fatema, K., & Sajib, A. A. (2021). Common genetic variants and pathways in diabetes and associated complications and vulnerability of populations with different ethnic origins. *Scientific Reports*, *11*(1), Article 1. https://doi.org/10.1038/s41598-021-86801-2

Sieberts, S. K., Perumal, T. M., Carrasquillo, M. M., Allen, M., Reddy, J. S., Hoffman, G. E., Dang, K. K., Calley, J., Ebert, P. J., Eddy, J., Wang, X., Greenwood, A. K., Mostafavi, S., Omberg, L., Peters, M. A., Logsdon, B. A., De Jager, P. L., Ertekin-Taner, N., & Mangravite, L. M. (2020). Large eQTL meta-analysis reveals differing patterns between cerebral cortical and cerebellar brain regions. *Scientific Data*, *7*(1), Article 1. https://doi.org/10.1038/s41597-020-00642-8

Singh, M. K., Mobeen, A., Chandra, A., Joshi, S., & Ramachandran, S. (2021). A meta-analysis of comorbidities in COVID-19: Which diseases increase the susceptibility of SARS-CoV-2 infection? *Computers in Biology and Medicine*, *130*, 104219. https://doi.org/10.1016/j.compbiomed.2021.104219

Skerrett-Byrne Anthony, D., Jiang Chen, C., Nixon, B., & Hondermarck, H. (2023). Transcriptomics. In R. A. Bradshaw, G. W. Hart, & P. D. Stahl (Eds.), *Encyclopedia of Cell Biology (Second Edition)* (pp. 363–371). Academic Press. https://doi.org/10.1016/B978-0-12-821618-7.00157-7

Smieszek, S. P., Polymeropoulos, V. M., Polymeropoulos, C. M., Przychodzen, B. P., Birznieks, G., & Polymeropoulos, M. H. (2022). Elevated plasma levels of CXCL16 in severe COVID-19 patients. *Cytokine*, *152*, 155810. https://doi.org/10.1016/j.cyto.2022.155810

Smithy, J. W., & Luke, J. J. (2023). CD16+ Macrophages: An Emerging Biomarker for Combined CTLA-4 and PD-1 Blockade. *Clinical Cancer Research*, *29*(13), 2345–2347. https://doi.org/10.1158/1078-0432.CCR-23-0490

Sneeggen, M., Guadagno, N. A., & Progida, C. (2020). Intracellular Transport in Cancer Metabolic Reprogramming. *Frontiers in Cell and Developmental Biology*, *8*, 597608. https://doi.org/10.3389/fcell.2020.597608

Spearman, C. (1904). *General intelligence,'objectively determined and measured. First published in American Journal of Psychology, 15, 201-293*.

Spitz, F., & Furlong, E. E. M. (2012). Transcription factors: From enhancer binding to developmental control. *Nature Reviews Genetics*, *13*(9), Article 9. https://doi.org/10.1038/nrg3207

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P. K., Swerdlow, H., Satija, R., & Smibert, P. (2017). Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, *14*(9), Article 9. https://doi.org/10.1038/nmeth.4380

Structural Equation Modeling with lavaan | Wiley. (n.d.). Wiley.Com. Retrieved August 22, 2023, from https://www.wiley.com/en-us/Structural+Equation+Modeling+with+lavaan-p-9781786303691

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., Hao, Y., Stoeckius, M., Smibert, P., & Satija, R. (2019). Comprehensive Integration of Single-Cell Data. *Cell*, *177*(7), 1888-1902.e21. https://doi.org/10.1016/j.cell.2019.05.031

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, *102*(43), 15545–15550. https://doi.org/10.1073/pnas.0506580102

Subramanian, I., Verma, S., Kumar, S., Jere, A., & Anamika, K. (2020). Multi-omics Data Integration, Interpretation, and Its Application. *Bioinformatics and Biology Insights*, *14*, 1177932219899051. https://doi.org/10.1177/1177932219899051

Sun, K., Xu, R., Ma, F., Yang, N., Li, Y., Sun, X., Jin, P., Kang, W., Jia, L., Xiong, J., Hu, H., Tian, Y., & Lan, X. (2022). scRNA-seq of gastric tumor shows complex intercellular interaction with an alternative T cell exhaustion trajectory. *Nature Communications*, *13*(1), Article 1. https://doi.org/10.1038/s41467-022-32627-z

Tan, K., Huang, W., Hu, J., & Dong, S. (2020). A multi-omics supervised autoencoder for pan-cancer clinical outcome endpoints prediction. *BMC Medical Informatics and Decision Making*, *20*(3), 129. https://doi.org/10.1186/s12911-020-1114-3

Tang, X., Yang, Z.-Z., Kim, H. J., Anagnostou, T., Yu, Y., Wu, X., Chen, J., Krull, J. E., Wenzl, K., Mondello, P., Bhardwaj, V., Wang, J., Novak, A. J., & Ansell, S. M. (2022). Phenotype, Function, and Clinical Significance of CD26+ and CD161+Tregs in Splenic Marginal Zone Lymphoma. *Clinical Cancer Research*, *28*(19), 4322–4335. https://doi.org/10.1158/1078-0432.CCR-22-0977

Tarka, P. (2018). An overview of structural equation modeling: Its beginnings, historical development, usefulness and controversies in the social sciences. *Quality & Quantity*, *52*(1), 313–354. https://doi.org/10.1007/s11135-017-0469-8

The ENCODE (ENCyclopedia Of DNA Elements) Project. (2004). *Science*, *306*(5696), 636–640. https://doi.org/10.1126/science.1105136

Tian, T., Wan, J., Song, Q., & Wei, Z. (2019). Clustering single-cell RNA-seq data with a model-based deep learning approach. *Nature Machine Intelligence*, *1*(4), Article 4. https://doi.org/10.1038/s42256-019-0037-0

Tokarev, A. A., Alfonso, A., & Segev, N. (2013). Overview of Intracellular Compartments and Trafficking Pathways. In *Madame Curie Bioscience Database [Internet]*. Landes Bioscience. https://www.ncbi.nlm.nih.gov/books/NBK7286/

Tremble, L. F., McCabe, M., Walker, S. P., McCarthy, S., Tynan, R. F., Beecher, S., Werner, R., Clover, A. J. P., Power, X. D. G., Forde, P. F., & Heffron, C. C. B. B. (2020). Differential association of CD68+ and CD163+ macrophages with macrophage enzymes, whole tumour gene expression and overall survival in advanced melanoma. *British Journal of Cancer*, *123*(10), Article 10. https://doi.org/10.1038/s41416-020-01037-7

Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B. E., Liu, X. S., & Raychaudhuri, S. (2013). Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature Genetics*, *45*(2), 10.1038/ng.2504. https://doi.org/10.1038/ng.2504

Two Types of Macrophages: M1 and M2 Macrophages. (n.d.). CUSABIO. Retrieved October 26, 2023, from https://www.cusabio.com/c-20938.html

Ucisik-Akkaya, E., Davis, C. F., Gorodezky, C., Alaez, C., & Dorak, M. T. (2010). HLA complex-linked heat shock protein genes and childhood acute lymphoblastic leukemia susceptibility. *Cell Stress & Chaperones*, *15*(5), 475–485. https://doi.org/10.1007/s12192-009-0161-6

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., & Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, *1*(1), Article 1. https://doi.org/10.1038/s43586-021-00056-9

van Megen, K. M., van 't Wout, E.-J. T., Lages Motta, J., Dekker, B., Nikolic, T., & Roep, B. O. (2019). Activated Mesenchymal Stromal Cells Process and Present Antigens Regulating Adaptive Immunity. *Frontiers in Immunology*, *10*. https://www.frontiersin.org/articles/10.3389/fimmu.2019.00694

Varathan, P., Gorijala, P., Jacobson, T., Chasioti, D., Nho, K., Risacher, S. L., Saykin, A. J., & Yan, J. (2022). Integrative analysis of eQTL and GWAS summary statistics reveals transcriptomic alteration in Alzheimer brains. *BMC Medical Genomics*, *15*(2), 93. https://doi.org/10.1186/s12920-022-01245-5

Veglia, F., Sanseviero, E., & Gabrilovich, D. I. (2021). Myeloid-derived suppressor cells in the era of increasing myeloid cell diversity. *Nature Reviews Immunology*, *21*(8), Article 8. https://doi.org/10.1038/s41577-020-00490-y

Vicari, H. P., Coelho-Silva, J. L., Pereira-Martins, D. A., Lucena-Araujo, A. R., Lima, K., Lipreri da Silva, J. C., Scheucher, P. S., Koury, L. C., de Melo, R. A., Bittencourt, R., Pagnano, K., Nunes, E., Fagundes, E. M., Kerbauy, F., de Figueiredo-Pontes, L. L., Costa-Lotufo, L. V., Rego, E. M., Traina, F., & Machado-Neto, J. A. (2022). STMN1 is highly expressed and contributes to clonogenicity in acute promyelocytic leukemia cells. *Investigational New Drugs*, *40*(2), 438–452. https://doi.org/10.1007/s10637-021-01197-0

Wahlström, T., & Arsenian Henriksson, M. (2015). Impact of MYC in regulation of tumor cell metabolism. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, *1849*(5), 563–569. https://doi.org/10.1016/j.bbagrm.2014.07.004

Wang, C., Sun, D., Huang, X., Wan, C., Li, Z., Han, Y., Qin, Q., Fan, J., Qiu, X., Xie, Y., Meyer, C. A., Brown, M., Tang, M., Long, H., Liu, T., & Liu, X. S. (2020). Integrative analyses of single-cell transcriptome and regulome using MAESTRO. *Genome Biology*, *21*(1), 198. https://doi.org/10.1186/s13059-020-02116-x

Wang, J., Xia, S., Arand, B., Zhu, H., Machiraju, R., Huang, K., Ji, H., & Qian, J. (2016). Single-Cell Co-expression Analysis Reveals Distinct Functional Modules, Co-regulation Mechanisms and Clinical Outcomes. *PLOS Computational Biology*, *12*(4), e1004892. https://doi.org/10.1371/journal.pcbi.1004892

Wang, R. T., Ahn, S., Park, C. C., Khan, A. H., Lange, K., & Smith, D. J. (2011). Effects of genome-wide copy number variation on expression in mammalian cells. *BMC Genomics*, *12*(1), 562. https://doi.org/10.1186/1471-2164-12-562

Wang, Y., Jiang, R., & Wong, W. H. (2016). Modeling the causal regulatory network by integrating chromatin accessibility and transcriptome data. *National Science Review*, *3*(2), 240–251. https://doi.org/10.1093/nsr/nww025

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*(1), Article 1. https://doi.org/10.1038/nrg2484

Ward, L. D., & Kellis, M. (2012). HaploReg: A resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Research*, *40*(Database issue), D930–D934. https://doi.org/10.1093/nar/gkr917

Ward, L. D., & Kellis, M. (2016). HaploReg v4: Systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Research*, *44*(Database issue), D877–D881. https://doi.org/10.1093/nar/gkv1340

Wei, N., Nie, Y., Liu, L., Zheng, X., & Wu, H.-J. (2022). Secuer: Ultrafast, scalable and accurate clustering of single-cell RNA-seq data. *PLOS Computational Biology*, *18*(12), e1010753. https://doi.org/10.1371/journal.pcbi.1010753

Weiser, M., Mukherjee, S., & Furey, T. S. (2014). Novel Distal eQTL Analysis Demonstrates Effect of Population Genetic Architecture on Detecting and Interpreting Associations. *Genetics*, *198*(3), 879–893. https://doi.org/10.1534/genetics.114.167791

Welstead, G. G., Hsu, E. C., Iorio, C., Bolotin, S., & Richardson, C. D. (2004). Mechanism of CD150 (SLAM) Down Regulation from the Host Cell Surface by Measles Virus Hemagglutinin Protein. *Journal of Virology*, *78*(18), 9666–9674. https://doi.org/10.1128/JVI.78.18.9666-9674.2004

Wolf, F. A., Angerer, P., & Theis, F. J. (2018). SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biology*, *19*, 15. https://doi.org/10.1186/s13059-017-1382-0

Wörheide, M. A., Krumsiek, J., Kastenmüller, G., & Arnold, M. (2021). Multi-omics integration in biomedical research – A metabolomics-centric review. *Analytica Chimica Acta*, *1141*, 144–162. https://doi.org/10.1016/j.aca.2020.10.038

Wruck, W., & Adjaye, J. (2020). SARS-CoV-2 receptor ACE2 is co-expressed with genes related to transmembrane serine proteases, viral entry, immunity and cellular stress. *Scientific Reports*, *10*, 21415. https://doi.org/10.1038/s41598-020-78402-2

Xie, Y., Yang, H., Yang, C., He, L., Zhang, X., Peng, L., Zhu, H., & Gao, L. (2022). Role and Mechanisms of Tumor-Associated Macrophages in Hematological Malignancies. *Frontiers in Oncology*, *12*. https://www.frontiersin.org/articles/10.3389/fonc.2022.933666

Xiong, H., Zhang, X., Chen, X., Liu, Y., Duan, J., & Huang, C. (2021). High expression of ISG20 predicts a poor prognosis in acute myeloid leukemia. *Cancer Biomarkers: Section A of Disease Markers*, *31*(3), 255–261. https://doi.org/10.3233/CBM-210061

Xiong, Y., He, L., Shay, C., Lang, L., Loveless, J., Yu, J., Chemmalakuzhy, R., Jiang, H., Liu, M., & Teng, Y. (2019). Nck-associated protein 1 associates with HSP90 to drive metastasis in human non-small-cell lung cancer. *Journal of Experimental & Clinical Cancer Research*, *38*(1), 122. https://doi.org/10.1186/s13046-019-1124-0

Xu, F., Cong, P., Lu, Z., Shi, L., Xiong, L., & Zhao, G. (2023). Integration of ATAC-Seq and RNA-Seq identifies key genes and pathways involved in the neuroprotection of S-adenosylmethionine against perioperative neurocognitive disorder. *Computational and Structural Biotechnology Journal*, *21*, 1942–1954. https://doi.org/10.1016/j.csbj.2023.03.001

Yan, J., Risacher, S. L., Shen, L., & Saykin, A. J. (2018). Network approaches to systems biology analysis of complex disease: Integrative methods for multi-omics data. *Briefings in Bioinformatics*, *19*(6), 1370–1381. https://doi.org/10.1093/bib/bbx066

Yang, L., Liu, J., Lu, Q., Riggs, A. D., & Wu, X. (2017). SAIC: An iterative clustering approach for analysis of single cell RNA-seq data. *BMC Genomics*, *18*(6), 689. https://doi.org/10.1186/s12864-017-4019-5

Yang, Z.-Z., Kim, H. J., Wu, H., Tang, X., Yu, Y., Krull, J., Larson, D. P., Moore, R. M., Maurer, M. J., Pavelko, K. D., Jalali, S., Pritchett, J. C., Mudappathi, R., Wang, J., Villasboas, J. C., Mondello, P., Novak, A. J., & Ansell, S. M. (2023). T-cell phenotype including CD57+ T follicular helper cells in the tumor microenvironment correlate with a poor outcome in follicular lymphoma. *Blood Cancer Journal*, *13*(1), Article 1. https://doi.org/10.1038/s41408-023-00899-3

Yin, H., & Flynn, A. D. (2016). Drugging Membrane Protein Interactions. *Annual Review of Biomedical Engineering*, *18*(1), 51–76. https://doi.org/10.1146/annurev-bioeng-092115-025322

Yom, C. K., Woo, H.-Y., Min, S. Y., Kang, S. Y., & Kim, H. S. (2009). Clusterin overexpression and relapse-free survival in breast cancer. *Anticancer Research*, *29*(10), 3909–3912.

Zamanighomi, M., Lin, Z., Daley, T., Chen, X., Duren, Z., Schep, A., Greenleaf, W. J., & Wong, W. H. (2018). Unsupervised clustering and epigenetic classification of single cells. *Nature Communications*, *9*(1), 2410. https://doi.org/10.1038/s41467-018-04629-3

Zerdes, I., Karafousia, V., Mezheyeuski, A., Stogiannitsi, M., Kuiper, R., Moreno Ruiz, P., Rassidakis, G., Bergh, J., Hatschek, T., Foukakis, T., & Matikas, A. (2021). Discordance of PD-L1 Expression at the Protein and RNA Levels in Early Breast Cancer. *Cancers*, *13*(18), 4655. https://doi.org/10.3390/cancers13184655

Zhang, M., Hu, S., Min, M., Ni, Y., Lu, Z., Sun, X., Wu, J., Liu, B., Ying, X., & Liu, Y. (2021). Dissecting transcriptional heterogeneity in primary gastric adenocarcinoma by single cell RNA sequencing. *Gut*, *70*(3), 464–475. https://doi.org/10.1136/gutjnl-2019-320368

Zhang, S., Li, X., Lin, J., Lin, Q., & Wong, K.-C. (2023). Review of single-cell RNA-seq data clustering for cell-type identification and characterization. *RNA*, *29*(5), 517–530. https://doi.org/10.1261/rna.078965.121

Zhang, Y., Amaral, M. L., Zhu, C., Grieco, S. F., Hou, X., Lin, L., Buchanan, J., Tong, L., Preissl, S., Xu, X., & Ren, B. (2022). Single-cell epigenome analysis reveals age-associated decay of heterochromatin domains in excitatory neurons in the mouse brain. *Cell Research*, *32*(11), Article 11. https://doi.org/10.1038/s41422-022-00719-6

Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K., & Liu, X. (2014). Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. *PLOS ONE*, *9*(1), e78644. https://doi.org/10.1371/journal.pone.0078644

Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, *12*(10), Article 10. https://doi.org/10.1038/nmeth.3547

Zhu, C., Baumgarten, N., Wu, M., Wang, Y., Das, A. P., Kaur, J., Ardakani, F. B., Duong, T. T., Pham, M. D., Duda, M., Dimmeler, S., Yuan, T., Schulz, M. H., & Krishnan, J. (2023). CVD-associated SNPs with regulatory potential reveal novel non-coding disease genes. *Human Genomics*, *17*(1), 69. https://doi.org/10.1186/s40246-023-00513-4

Zou, Z., Hua, K., & Zhang, X. (2021). HGC: Fast hierarchical clustering for large-scale single-cell data. *Bioinformatics*, *37*(21), 3964–3965. https://doi.org/10.1093/bioinformatics/btab420

žurauskienė, J., & Yau, C. (2016). pcaReduce: Hierarchical clustering of single cell transcriptional profiles. *BMC Bioinformatics*, *17*(1), 140. https://doi.org/10.1186/s12859-016-0984-y

APPENDIX A

SUPPLEMENTARY TABLES FOR CHAPTER 4

Table A 1: SEM model path coefficients and p-value for cluster_1_g4 model

Term	std.est	*z*	*p*	CI.Lo	CI.Up
G =~ UBB	0.52	35.92	< .001***	0.35	0.39
G =~ PARK7	0.55	38.46	< .001***	0.33	0.37
G =~ PYCARD	0.73	55.90	< .001***	0.55	0.59
G =~ ATP5F1B	0.56	39.52	< .001***	0.35	0.39
G =~ RAN	0.46	31.19	< .001***	0.26	0.30
G =~ SPCS1	0.50	34.36	< .001***	0.28	0.32
G =~ ATP5PO	0.53	37.35	< .001***	0.31	0.35
G =~ CHMP5	0.32	21.38	< .001***	0.15	0.18
G =~ CTSL	0.44	29.61	< .001***	0.27	0.30
G =~ TMED9	0.40	26.75	< .001***	0.18	0.21
G =~ BST2	0.53	37.28	< .001***	0.34	0.38
G =~ SSNA1	0.36	23.91	< .001***	0.18	0.21
G =~ CALM1	0.65	47.37	< .001***	0.49	0.53
G =~ HLA_B	0.69	51.80	< .001***	0.52	0.56
G =~ SEC61G	0.48	33.01	< .001***	0.27	0.31
G =~ FIS1	0.39	25.95	< .001***	0.20	0.24
G =~ RAB32	0.36	23.70	< .001***	0.19	0.22
G =~ SUMO1	0.43	29.39	< .001***	0.22	0.25

G =~ VPS29	0.46	31.26	< .001***	0.27	0.31
G =~ EMC7	0.35	23.61	< .001***	0.15	0.18
G =~ SNX3	0.53	36.89	< .001***	0.32	0.36
G =~ CCT8	0.38	25.52	< .001***	0.20	0.24
G =~ CIB1	0.34	22.27	< .001***	0.20	0.23
G =~ WASHC3	0.30	19.95	< .001***	0.12	0.15
G =~ CLTB	0.37	24.86	< .001***	0.18	0.21
G =~ SNX17	0.39	26.28	< .001***	0.20	0.23
G =~ COPE	0.47	32.19	< .001***	0.27	0.31
G =~ SDF2L1	0.28	18.18	< .001***	0.10	0.12
G =~ RAB11A	0.30	20.03	< .001***	0.14	0.16
G =~ CHMP2A	0.43	29.38	< .001***	0.24	0.28
G =~ BAX	0.47	32.07	< .001***	0.27	0.30
G =~ CD81	0.26	17.34	< .001***	0.13	0.16
G =~ RAB8A	0.47	31.85	< .001***	0.26	0.29
G =~ HSPB1	0.26	16.93	< .001***	0.11	0.13
G =~ HAX1	0.25	16.31	< .001***	0.11	0.14
G =~ RER1	0.36	24.03	< .001***	0.18	0.21
G =~ GET3	0.29	19.20	< .001***	0.11	0.14
G =~ SNAPIN	0.23	15.33	< .001***	0.07	0.09
G =~ COPZ1	0.33	22.05	< .001***	0.14	0.16

G =~ AP2M1	0.38	25.29	< .001***	0.20	0.23
G =~ TMED2	0.27	17.84	< .001***	0.11	0.14
G =~ UBE2J2	0.17	10.97	< .001***	0.05	0.08
M =~ ADT_CD86	0.42	24.67	< .001***	0.15	0.18
M =~ ADT_CD11c	0.85	38.09	< .001***	0.45	0.50
M =~ ADT_HLA_DR	0.55	31.49	< .001***	0.32	0.36
D ~ G	-0.09	-5.17	< .001***	-0.05	-0.02
M ~ G	0.36	19.00	< .001***	0.35	0.43
D ~ M	-0.15	-7.91	< .001***	-0.08	-0.05