PhD-FSTM-2025-027

The Faculty of Science, Technology and Medicine

DISSERTATION

Defence held on 06/03/2025 in Esch-sur-Alzette
to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG EN BIOLOGIE

by

Mathias GALATI
Born on 7 May 1995 in Colmar (France)

DATA MINING IN INTEGRATING IMMUNE STRATEGIES FOR TRANSLATIONAL RESEARCH IN ONCOLOGY, ALLERGOLOGY AND NUTRITIONAL MEDICINE

Dissertation defence committee
Prof. Reinhard Schneider, dissertation supervisor
Luxembourg Centre for Systems Biomedicine

Dr. Venkata Satagopam
Luxembourg Centre for Systems Biomedicine

Dr. Enrico Glaab, Chairman
Luxembourg Centre for Systems Biomedicine

Dr. Feng He
Luxembourg Institute of Health

Prof. Miguel Andrade
Johannes Gutenberg University Mainz

A dissertation by

Mathias GALATI

submitted to the University of Luxembourg

Approved by the dissertation defence committee:

Committee members:	Dr. Enrico Glaab
	Dr. Feng He
	Prof. Maria Secrier
	Prof. Miguel Andrade
Supervisor:	Prof. Reinhard Schneider
Co-supervisor:	Dr. Venkata Satagopam

Affidavit

I hereby confirm that the PhD thesis entitled "DATA MINING IN INTEGRATING IMMUNE STRATEGIES FOR TRANSLATIONAL RESEARCH IN ONCOLOGY, ALLERGOLOGY AND NUTRITIONAL MEDICINE" has been written independently and without any other sources than cited.

In ..... date ....

Author's signature

I extend my profound gratitude to my co-supervisors — Reinhard Schneider and Venkata Satagopam —, whose wise guidance and mentorship throughout my doctoral journey have been invaluable. Their commitment to my development as a scientist and support for international training and conference opportunities have significantly enhanced my research capabilities and scientific outreach skills. I am deeply grateful to the Biocore team for cultivating an intellectually stimulating environment that fosters scientific growth and innovation.

To my family, I express my deepest appreciation for their never ending support throughout my higher education journey. Their unwavering backing has been an invaluable blessing, enabling me to pursue my passions and follow my scientific interests with complete dedication.

My academic journey has been enriched by many meaningful friendships. To my friends Steve, Jean-Baptiste, Ludovic, Lucas, Victor from the Master's program in Strasbourg, who shared the challenges of computer science coursework and many more adventures. To those who supported me through the challenges of beginning doctoral research during the COVID-19 pandemic, my friends from my Master's internship in Montpellier, Alberto, Damien, Pierre, Mathis, Thibault — your companionship provided essential balance. To my Strasbourg friends Nicolas, Robin, Frederic, Julien, Victor, Paul, Guillaume and Bruno — our weekend gatherings, despite the distance, have been invaluable.

Special recognition goes to the exceptional scientists I've had the opportunity to collaborate with at Biocore and in the i2TRON DTU. To Ahmad, for our parallel growth and deep memories; to Mirek for your wise mentorship and our memorable late-night scientific discussions; to Guilherme where we can always translate our meeting to late philo-scientific discussions; to Cyril my Thionville buddy and our never ending debates; and Mohammed and Gleb, for making our shared office space conductive to productive scientific fussballish environment.

I extend my gratitude to the growing young bioinformatics communities, particularly the RSG France Jebif, where I had the opportunity to engage in meaningful outreach and event organization for fellow bioinformaticians. A special thanks to Jeff for his dedication to RSG Luxembourg and his continuous efforts in co-cultivating the growth of Luxembourg's young bioinformatics community.

To the Faluche communities, which provided invaluable respite during challenging times — my godparents Joel and Morgane in Strasbourg, Alberto and Antoine in Montpellier, and the communities in Nancy and Metz — your fellowship has maintained balance all along this journey.

To my partner, Anna, who has been always source of support during the final stages of this doctoral journey, sharing emotional highs and lows, our travels, and the joy of connecting with so many friends near and far. Our beloved cat and dog — Zola and Vayana — whose boundless enthusiasm in greeting us after long workdays has been a constant source of happiness.

Finally, I acknowledge the importance of maintaining authentic connections in an increasingly digital world. The genuine interactions with colleagues and friends, often accompanied by yeasted products — of course good cheese —, have proved far more enriching than any algorithm-driven entertainment.

Contents

Acronyms

CAF Cancer-associated fibroblasts. 5, 56, 83, 85–88, 90, 91, 93, 94, 99, 103, 108, 114, 145

CCC Cell-Cell Communication. 88, 90, 91, 106, 108, 113, 145, 147

CRC Colorectral cancer. 30, 45, 46, 51, 55, 56, 83–88, 94, 96, 99, 103, 106, 110, 111, 114, 135, 145, 147, 155

DEG Differentially expressed genes. 17, 19, 36, 90, 91, 93, 94, 96

DF Dietary fibers. 11, 47, 51, 117–124, 130, 131, 145, 156

DNA Deoxyribonucleic acid. 10, 12, 13, 16–18, 26, 45, 84

GO Gene Ontology. 94–106

GWAS Genome-Wide Association Study. 15, 24, 26–28, 125

IgE Immunoglobulin E. 29, 46, 47, 57–60

LCSB Luxembourg Center for Systems Biomedicine. 58, 81, 115, 131

LIH Luxembourg Institute of Health. 58, 81, 131

ML Machine Learning. 20, 21, 33, 34, 124, 126, 127, 136, 142, 147

MO Multi-Omics. 9, 20, 23, 24, 31–36, 41, 49, 56, 129, 135, 138, 142–145, 147–152

NCD Noncommunicable diseases. 5, 9–11, 55

NGS Next generation sequencing. 13

OFC Oral Food challenge. 57, 58, 60, 68, 70, 145

PA Peanut Allergy. 46, 57–59, 136, 145

PCR Polymerase Chain Reaction. 13, 15, 18

PM Personalized medicine. 13, 29, 32, 40, 44, 147, 151

PPI Protein-protein interaction networks. 23

PRS Polygenic risk score. 27, 28, 121, 123, 125–129

QC Quality Control. 20, 26, 28, 124, 149

RNA Ribonucleic acid. 10, 17–20, 25, 28, 36

SCFA Short-Chain Fatty Acids. 46–48, 117–120, 122–125, 130, 131

scRNA-seq Single-cell RNA sequencing. 5, 18, 28, 88, 112, 113, 115, 135, 136, 145, 147, 149, 152, 155

SNP Single-nucleotide polymorphism. 15, 26, 27, 56, 118, 121, 123–127, 148, 156

SPT Skin Prick Test. 57–59

TGS Third generation sequencing. 13

TME Tumor microenvironment. 83–86, 88, 96, 99, 103, 106, 109, 111, 114

WGS Whole Genome Sequencing. 12, 56, 123, 124

Wnt Wingless and Int-1. 87, 110, 111, 118

Abstract

The role of inflammation and immune reactions in Noncommunicable diseases (NCD) remains a critical area of research. Most NCDs are complex and often share common early clinical phenotypes despite the underlying molecular behavior. We aim to i) identify biomarkers that drive the clinical phenotype development, ii) characterize immune interaction profiles, and iii) model systemic inflammatory and immune-related responses.

Noteworthy, we focus on three independent modalities of disease complexity. Our project integrates heterogeneous types of datasets (omics / non-omics datasets) to investigate inflammatory and immune-related pathways. We use Single-cell RNA sequencing (scRNA-seq) to understand pathways driven by fibroblast stratification in cancer. Further, we implement a network-based modeling approach to understand allergy reaction profiles. We analyze established genetic predispositions to explore their impacts on the associated inflammatory disruptions in the gut.

Our results help to advance the understanding of inflammatory and immune-related mechanisms. In the first modality, we identify iCAF's WNT5A-related receptors to specific epithelial stem cells. As for the second modality, we propose correlated immune cell and cytokines signatures that could drive the allergy responses. The third modality foresees characterizing genetic and food-related immune profiles tailored to the food habits of the population.

Part I

Introduction

Chapter 1

Biological complexity: Insights from multi-omics integration

This introduction aims to assemble essential concepts for building a comprehensive understanding of the bioinformatic approaches developed to explore and analyze organisms and the complexity of their molecular pathways and interactions. It poses basis for understanding how inflammation, a complex biological response, connects diverse molecular and cellular factors to Noncommunicable diseases (NCD), as outlined in figure 1.1.

The background section introduces the foundational concepts of biological information. It begins with a discussion on how biological signatures rule physiological complexity in section 1.1. Building upon this, the narrative transitions into describing how omics technologies enable us to study individual biological dimensions in section 1.2. Finally, the concept of Multi-Omics (MO) is presented as a holistic framework that captures the interplay of various biological layers in section 1.3.

The figure 1.1 serves as a visual summary of the key biological factors and parameters influencing inflammation. Fixed factors, such as genetics and epigenetics, combine with variable environmental and lifestyle factors, leading to inflammatory responses. These responses, characterized at the cellular and systemic levels, have significant implications in the context of acute and chronic conditions. The progression from inflammation to NCD (*e.g.* cardiovascular diseases, cancer, diabetes) emphasizes the translational relevance of the thesis work.

The second part of the introduction connects the theory of omics technologies to their practical applications in modern medicine. The section 2.1 outlines the role of translational research in bridging scientific discoveries with clinical applications, highlighting the importance

of converting molecular insights into actionable patient care strategies.

The section 2.2 introduces the i2TRON translational research project, which investigates inflammation in NCD context, as shown in figure 1.1. The project underscores the importance of integrating fixed and variable factors in understanding the systemic effects of inflammation. The exploration of cytokines, immune cell pathways, and systemic responses is the backbone of this thesis, aligning with the broader research goal of identifying actionable bio-signatures for patient stratification.

1.1 Biological signatures reveal complexity in Life sciences

Recent advances in omics technologies have fundamentally transformed the field of molecular biology, providing unprecedented insights into cellular processes and complex biological systems. While breakthrough technologies, such as CRISPR-Cas9, have revolutionized fields from immunotherapy to cancer research, significant challenges persist, including off-target effects and the need for precise functional annotations. Understanding these complex biological systems requires integrated approaches that elucidate gene functions and their regulatory mechanisms within broader molecular networks.

This section examines the evolution of omics technologies, their applications, and the computational frameworks addressing their inherent challenges. Particular emphasis is placed on gene expression analysis, systems biology approaches, and network-based methodologies to provide a comprehensive understanding of cellular functions and their implications for research and medicine.

1.1.1 Genomics provides foundation for advances research

The genome, comprising complete Deoxyribonucleic acid (DNA) sequence, contains all the information necessary for an organism to develop and function. Located in the cell nucleus and organized into chromosomes. Genes, discrete sections of DNA, codes for Ribonucleic acid (RNA) molecules that facilitate protein synthesis outside the nucleus. Understanding the genome has become essential for explaining how sequence defect can lead to disruptions in biological pathways and disease manifestations.

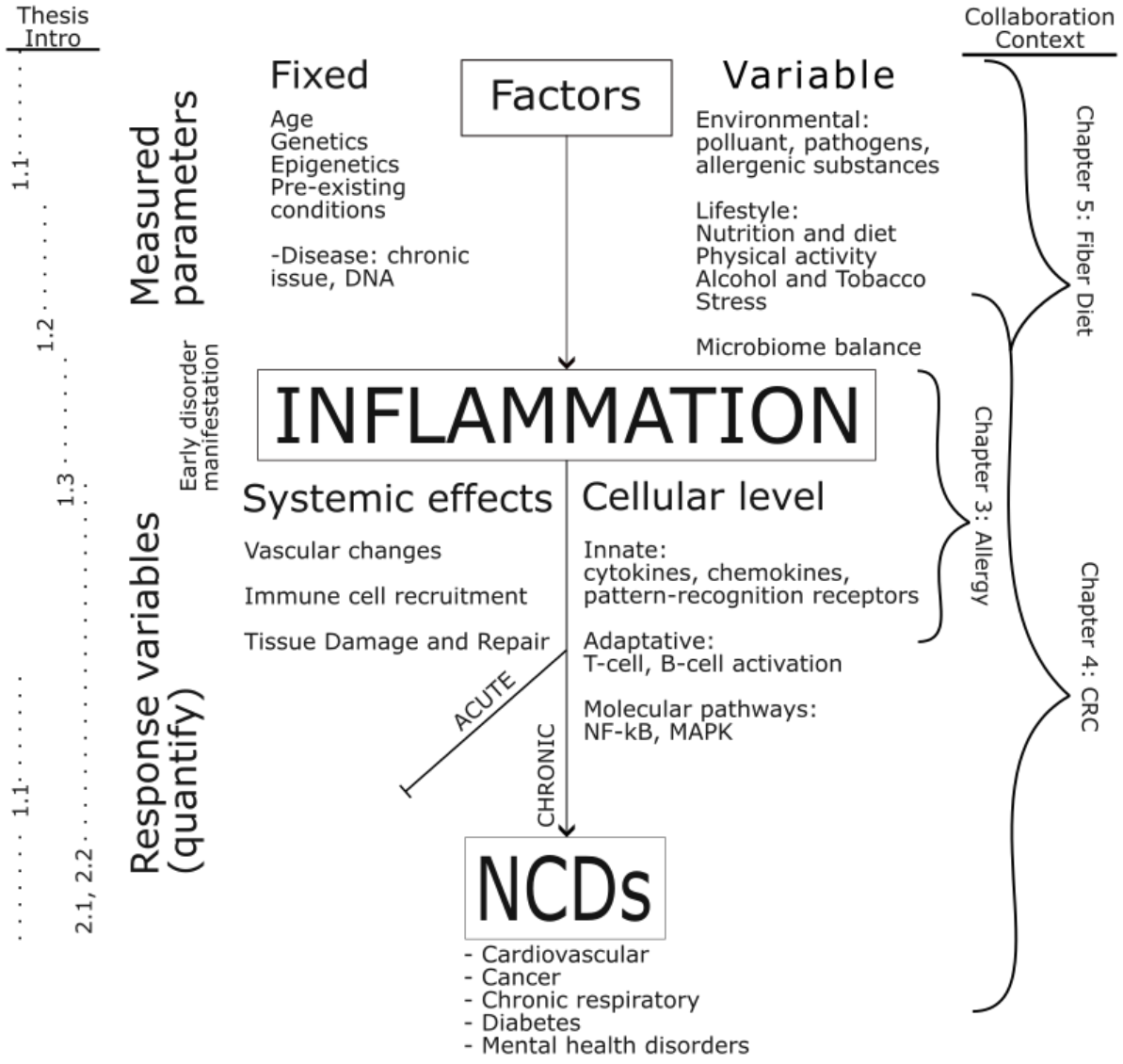


Figure 1.1 Inflammation as a central hub linking biological, environmental, and lifestyle factors to Noncommunicable diseases. The diagram illustrates how fixed (*e.g.* genetics, epigenetics) and variable (*e.g.* diet, microbiome, pollutants) factors influence inflammatory pathways at the cellular and systemic levels. Acute and chronic inflammation drive distinct pathological outcomes, with chronic inflammation contributing to NCD development (*e.g.* cardiovascular disease, cancer, diabetes). Targeted interventions, such as dietary modifications (Dietary fibers) and allergy management, can modulate these inflammatory processes.

Genome sequencing encompasses methods to capture comprehensive DNA information, of an organism, at a specific time point. This capability has shifted research toward data-driven approaches within the fields of "bioinformatics". Whole Genome Sequencing (WGS), a widely adopted method, aims to determine "precise" sequence of nucleic bases (A, T, G, C) by detecting single nucleotide variants, insertions or deletions, copy number changes, and large structural variants. This technology has proven particularly valuable in identifying genetic factors driving diseases, as evidenced by a 62.5% diagnostic rate in previously undiagnosed patients and the discovery of potentially pathogenic variants were discovered for mitochondrial disease (IGF2/INS-IGF2) and Klippel-Trenaunay-Weber syndrome (FBN3) [1]. To understand current challenges and gap, it is essential to start from the historical first revolution in the field of genome sequencing.

Sanger sequencing sets the Gold standard for genomics studies

First generation sequencing is exemplified by the Sanger method, developed in the 1970s by Federick Sanger. This pioneering technique, also known as the chain termination method, involves the selective incorporation of chain-terminating dideoxynucleotides during DNA replication, yielding DNA fragments of various lengths. These fragments undergo separation by gel electrophoresis and subsequent analysis deduce the DNA sequence [2]. The Human Genome Project (HGP) [3], operational from 1990 and 2003, represented a groundbreaking global scientific initiative that addressed significant challenges, such as characterizing the high number of genes in the human genome ($\approx$ 20000 coding genes) and more than 3 billion nucleotides. The project was achieved completion ahead of the initial schedule, successfully delivering the first first human genome sequence covering 99% of gene-containing regions with 99.99% accuracy. While the HGP relied on the Sanger sequencing method to determine relatively small fragment, considerable efforts focused on overlapping these fragments to create longer sequences and eventually reconstruct each chromosomes.

The scientific community recognized that to fully leverage the acquired knowledge in the future, more efficient, cheaper and more convenient methods needed to be developed. This realization catalyzed the development of next-generation sequencing technologies, which revolutionized the field of genomics.

Emerging technologies in sequencing offers new opportunities

Second and third generation sequencing technologies have transformed genomics research

by drastically improving speed and reducing costs [4]. Second generation sequencing, also known as Next generation sequencing (NGS) enables parallel sequencing of millions of DNA fragments (short reads from 100 to 300 bp) simultaneously. The field is dominated by platforms such as Illumina, Ion Torrent and Roche 454 pyrosequencing, facilitating high-throughput sequencing of genomes, transcriptomes and metagenomes [5]. However, NGS faces certain limitations, such as Polymerase Chain Reaction (PCR) amplification bias and difficulty sequencing repetitive regions due to short read lengths [6]. To address these challenges, Third generation sequencing (TGS) technologies, such as Pacbio SMRT sequencing and Oxford Nanopore sequencing, have emerged, capable of generating much longer reads, often over 10kb in length, without the need of PCR amplification [6]. TGS offers several advantages over NGS, including enhanced capability in resolving complex genomic regions, identifying structural variations and generating high-quality de novo genome assemblies.

The selection of appropriate technologies depends on project-specific requirements, such as the scale, type of data needed, error tolerance, and available budget [7]. In a recent study focused on Thalassemia applications [8], researchers determined that TGS currently has higher error rates and cost compared to NGS, which is useful for variant calling. Consequently, TGS is more appropriate for studying variants on homologous genes and CNV calling, while NGS remains the more cost effective solutions for large sample volumes.

Genomics applications across diverse research domains

Modern sequencing technologies have become indispensable tools in various medical applications [9]. In the context of clinical diagnostics, these technologies facilitate rare disorder evaluation by enabling accurate diagnosis and informing therapeutic strategies. Additionally, they prove crucial in understanding neoplasms, which result from abnormal tissue growth. These conditions frequently present genetic alterations, and their identification can guide Personalized medicine (PM) and treatment plans. Through comprehensive genetic analysis of a neoplasm, clinicians can implement effective targeted therapies, thereby enhancing patient prognosis. Furthermore, these technologies support prenatal screening by evaluating chromosomal aneuploidies and other genetic disorders, providing vital information for informed decision-making and early intervention when necessary.

Two predominant approaches have emerged in clinical genomics: exome sequencing and multi-gene panel. The selection between these methods depends on various factor, including clinical indication, suspected genetic aetiology and the available resources:

Exome sequencing encompasses the analysis of protein-coding regions, named exons, providing a cost-effective and a targeted approach for identifying clinically relevant genetic alterations.

Multi gene panels involve the parallel sequencing of a predefined set of genes, offering greater efficiency and focus compared to whole exome approaches.

For example, if a specific genetic disorder is suspected, a multi-gene panel targeting the genes associated with that disorder may be the more appropriate approach. On the other hand, if the genetic cause is unknown or multiple genetic disorders are suspected, exome sequencing may be a more comprehensive approach.

The application of sequencing technologies has significantly advanced our understanding of evolutionary biology, particularly in elucidating genetic relationships between organisms at both inter- and intra-species levels. At the inter-species level, substantial progress has been made in phylogenetics, notably in the identification of speciation genes and comparative analysis of orthologous genes across species. This has enhanced our ability to reconstruct evolutionary histories and understand the mechanisms driving species divergence [10, 11].

At the intra-species level, these technologies enable temporal analyses of genetic variation. At a given time point, population genetics approaches can reveal patterns of genetic diversity and structure within species. Over time, these analyses extend to studying genome evolution, developmental biology, and population-level epidemiology [12]. In microbial systems, this temporal dimension is particularly informative due to their rapid generation times and adaptability.

Microbial applications expand to clinical, veterinary, food industry, and environmental applications. In the context of the antibiotic resistance crisis, microbial genomics has become an essential tool for investigating pathogenic mechanisms, tracking genetic drift, and identifying novel drug targets [13]. Recent microbiome studies have revealed crucial host-microbe interactions, particularly in the gut-brain axis, with significant implications for understanding neurodegenerative disorders [14].

Overcoming technical challenges in sequencing

Omics sciences represent a translational field that bridges molecular biology, computational methods, and statistical analysis. This integration faces multiple technical challenges across different domains that need to be addressed systematically.

The biological hardware challenges primarily comes from the physical processes of genetic material manipulation. Sequencing techniques present significant challenges in terms of accuracy and data management. Sequencing errors can occur due to incorrect base calls in specific sequence contexts, such as homopolymer runs or GC-rich regions, potentially leading to false positive variant calls [15]. Short sequencing reads often encounter mapping ambiguities in repetitive regions of the reference genome, complicating the accurate calling of variants in these challenging regions. Additionally, biases in sequencing coverage can arise from factors such as GC content and PCR amplification during library preparation, significantly impacting read depth and affecting variant calling [16].

The biological software challenges involve converting raw molecular signals into organized digital information. The distinction between true genetic variants from sequencing errors remains a challenge, particularly for rare variants [17]. In clinical settings, calling somatic mutations in oncology, remains challenging due to tumor heterogeneity, normal cell contamination and a diverse range of variant allele frequencies. Implementation of best practices, including multiple biopsies and paired tumor-normal designs, can substantially improve accuracy [18].

The computational software challenges involve developing and optimizing statistical methods and algorithms for data analysis. These include addressing the computational complexity of variant calling in repetitive regions, managing mapping ambiguities with short-read data, and optimizing analysis pipelines for different variant types, from Single-nucleotide polymorphism (SNP) to complex structural variants [19].

The computational hardware challenges center around the infrastructure required for processing and storing datasets. Several "disciplines" generate vast amounts of data, including astronomy research, social media platforms (including Youtube and Twitter) and Biology research [20]. The exponential growth in data volume poses massive storage and computational challenges, particularly in the context of metagenomics and Genome-Wide Association Study (GWAS). To illustrate, a single human genome comprises more than 3 billion of nucleotides, requiring approximately 3 gigabytes of storage [21]. Current projections by the International Data Corporation indicate that the worldwide need for data storage will escalate to 175 zettabytes as we approach 2025 [22].

The ability to decode the genetic code of an organism opens new avenue for understanding its biology, from 3D structure and protein binding to precision medicine, through comprehensive examination of the relationships between mutations, phenotypes, and genotypes.

The field of bioinformatics has evolved to extract meaningful insights from this data, driven by advancements in both physical technologies (high-performance computing, nanopore sequencing, cryo-eletron microscopy) and computational approaches (from classical and Bayesian statistics to a large range of machine learning techniques).

1.1.2 Functional analysis drives insights from Big Data

The vast quantity of genomic data generated by modern sequencing studies has created an imperative need to understand gene roles of theses genes and their expression profiles under varying conditions. This section explores how functional analysis has evolved to meet these challenges and its crucial role in current research environment.

Linking functional analysis to biomedical applications

The relationship between genome and phenotype exhibits remarkable complexity and multifaceted interactions. While the genome plays a crucial role in determining the phenome (the complete set of observable characteristics expressed by an organism), the relationship is not straightforward, as multiple processes and environmental factors influence the phenotypic expression. For instance, studying differences and predisposition in sports requires measuring different omics data to better understand the effects from the micro to the macro scale [23]. A gene's function refers to the biological activities across multiple scales, from molecular to cellular to organism level. Understanding this multi-level complexities is crucial for elucidating disease mechanisms.

Historical context highlights the evolution of functional analysis

The foundations of functional analysis trace back to classical genetics, with (Mendel's Laws) formulating the principle of inheritance through his pioneering work with pea plants [24]. However, the field has since recognized that not all the genes comply with the strict conditions of two allelic variants with discrete effects on trait and complete dominance. The discipline evolved significantly in the 1900s with the development of the chromosome theory of inheritance [25]. Modern understanding reveals that the relationship between genes and functions is is vastly more intricate, and the field has evolved through advancements in genetics, molecular biology, and gene editing technologies. Contemporary methods include DNA

transgenes approach, transgenes detection, gene expression assessment through transcription levels measurement, and proteins expression detection. More on the following resource: History of Transgenesis [26].

Bioinformatics approaches for functional analysis

Modern functional analysis use technologies for gene manipulation and study. Gene knockout technologies aim to disable a specific gene, providing crucial insights into their roles in organism development, physiology or pathology [27]. Several technologies have emerged:

RNAi Employs small RNA molecules to inhibit gene expression

Zinc Finger Nuclease (ZFNs) Engineered proteins designed to bind and cut specific DNA locations, then disrupting gene's function

Transcription Activator-Like Effector Nucleases (TALENs) similar than ZFN

CRISPR/Cas9 Currently the predominant method utilizing guide RNA molecules to direct Cas9 protein for precise genetic modifications

These technologies have revolutionized various fields, such as developing immunotherapy such as CAR T cell-based treatment [28]. However, significant challenges persist including off-target effect [29], inefficiency of precise knock-in, or PAM restriction for CRISPR-CAS9 [28]. While functional analysis provides important biological insights, it is recognized that biological functions rely on complex molecular systems, necessitating integration with gene expression analysis for comprehensive understanding.

1.1.3 Gene expression analysis explains cellular functions

Gene expression analysis facilitates understanding of the transcriptome, encompassing the set of RNA transcripts produced under specific conditions. This approach enables investigation of gene expression profiles that differ between individuals. Researchers can identify patterns and signatures while studying the transcriptional activity of genes in diseases, as demonstrated in studies of pesticide exposure [30]. In response to varying conditions, genes may be up-regulated or down-regulated. The identification of these Differentially expressed genes (DEG) provides valuable insights into dysfunctional pathways, or disease mechanisms.

Evolution of gene expression analysis techniques

The progression of expression analysis techniques represents a significant advancement in molecular biology. Northern blots representing one of the earliest methods for RNA analysis, involves separating RNA fragments based on their size using gel electrophoresis, transferring them to a membrane, and detecting specific sequences using labeled DNA probes. While this technique established fundamental principles, its low-throughput nature has led to its replacement by more advanced methods.

The development of Reverse Transcription PCR (RT-PCR) marked a significant advancement in the field [31]. This technique's key lies in the amplification and quantification of RNA through initial conversion to complementary DNA (cDNA). Further refinement led to Real-time RT-PCR, which enables quantitative measurement of RNA through continuous monitoring of the amplification process using fluorescent probes [32].

A paradigm shift occurred with the introduction of microarrays, which incorporate thousands of DNA probes immobilized on a solid surface, enabling high-throughput parallel analysis of gene expression [33]. The technique employs fluorescence-labeled cDNA hybridization to the array, and the signal intensity correlates with the expression level. This development represented a major advancements over the PCR-based methods by facilitating genome-wide expression profiling.

In current research, RNA has largely replaced microarrays as the preferred method for transcriptomics, involving direct sequencing of cDNA libraries for expression quantification. A significant recent innovation is Single-cell RNA sequencing (scRNA-seq), which enables expression levels measurement in individual cells rather than bulk samples, revealing cell to cell heterogeneity and facilitating identification of rare cell types. This technological advancement has proven instrumental in building cell atlases, studying development, and characterizing complex tissues [34].

Resources and Databases for gene expression research

The filed of gene expression analysis is supported by several crucial developments in databases and tools, particularly those essential for analyzing large-scale human genomics data.

- Gene Expression Omnibus (GEO): a public repository that archives microarray, next-generation sequencing, and other high throughput functional genomics data [35].

- Genotype-Tissue Expression (GETx): a large scale RNA sequencing study that accumulated massive quantities of data across different human tissues and pathologies [36].

- The Cancer Genome Atlas (TCGA): a cancer genomic program that "characterized over 20000 primary cancer and matched normal samples spanning 33 cancer types" [37].

- Expression Atlas (EMBL-EBI): provides information on "gene and protein expression across species and biological conditions such as different tissues, cell types, developmental stages and diseases" [38].

Transformative applications in research and medicine

Gene expression profiling in multicellular organisms facilitates the identification of co-regulated genes sets and jointly up- or down- regulated, while also enabling the discovery of essential housekeeping genes crucial for the organism function [39]. Comparative analysis of developmental gene expression patterns between normal and abnormal states, such as in Down syndrome [40], can elucidate genes and pathways potentially responsible for disorders when miss-regulated.

Recent research in depressive disorder [41] has led to the identification of DEG related to antidepressant treatment and response. Gene expression data has also demonstrated significant utility in drug re-purposing and drug combination prediction, enabling the identification of potential new application of existing drugs or synergistic drug combinations through analysis of complementary or opposing gene expression patterns [42]. Furthermore, microarray data has proven valuable in drug toxicity prediction [43].

Challenges in interpreting gene expression data

The inherent complexity of gene expression data presents multiple analytical challenges. A fundamental challenge lies in distinguishing between causative and reactive differences in gene expression, *i.e.* specifically determining whether differential expression drives a phenotype or merely responds to other factors. While methodological approaches such as Likelihood Causality Model Selection (LCMS) have been proposed, they face limitations when dealing with highly correlated genes or complex networks [44].

Gene expression datasets frequently exhibit high levels of technical and biological noise/variation, potentially obscuring true signals. While some strategies such as Random Matrix Theory have been proposed, these approaches risk eliminating meaningful gene relationships. Alternative methodologies like Knowledge Independent Network Construction attempt to address noise by assessing each gene-gene pair independently, rather than applying global threshold to entire gene sets [45].

The simultaneous measurement of thousands of genes requires rigorous multiple hypothesis testing correction to effectively control false positive results and ensure statistical relevance. The scientific community has adopted several methods to control significance, including false discovery rate estimation, permutation testing and null hypothesis comparison [46]. Enhanced confidence in significant gene targets can be achieved through integration of results across multiple studies and datasets. An additional strategic approach involves cross-referencing expression data with epigenetic information to develop a more holistic view of gene regulation.

New approaches shaping gene expression studies

The emergence of advanced computational methods, particularly single-cell technologies with RNA and genomics focus, has enabled profiling at the individual cell resolution, providing unprecedented insights into cellular heterogeneity and states [47, 48]. The field of MO offers the potential to leverage simultaneous measurement of multiple omics layers to investigate these intricate interactions [49].

Real-time analysis platforms have evolved to meet these complex demands. Systems such as Argonaut [50] and STAGEs [51] provide comprehensive integrated environments for multiple aspects of analysis, including data management, Quality Control (QC), normalization, statistical analysis, and visualization of gene expression and MO experiments. Notable example include platforms like GENEASE [52], which facilitate integrated analysis across multiple data types.

Machine Learning (ML) has emerged as a powerful approach in this domain, with models such as Enformer [53] demonstrating superior accuracy in gene expression prediction compared to previous approaches. Feature selection methods, exemplified by MRMR (Maximum Relevance Minimum Redundancy), provide enhanced capabilities for differential expression analysis, identifying the most informative genes while minimizing redundancy [54].

Additional algorithmic advances include HAPLEXD, which has improved gene expression prediction from haplotype sequence, [55], while other ML approaches continue to enhance prediction accuracy across various contexts [56].

1.1.4 Biological network databases expand to meet research needs

Biological network serve as sophisticated framework for modeling and understanding complex cellular interactions and processes. Rather than examining pathways in isolation, network approaches conceptualize biological systems as highly interconnected networks of interacting components. From a mathematical perspective, biological networks are represented as graphs, where nodes represent biological entities such as genes, proteins or metabolites, and edges capture the interactions or relationships between them. This representation enables the application of graph theory and network analysis [57] methods from mathematics and computer science to elucidate the structure and dynamics of these networks.

Small-World properties define of biological networks

A fundamental characteristic of biological networks is their exhibition of the "small-world" property [58], first described by Watts and Strogatz in 1998. This property manifests through two key characteristics: a high clustering coefficient, indicating that nodes tend to form tightly connected neighborhoods, and a low characteristic path length, demonstrating that most nodes can be reached from every node through a small number of connections. Theses features facilitate efficient communication between nodes while preserving modular structure.

Scale-Free topologies for understanding biological systems

Another fundamental characteristic, described by Barabasi and Albert in 1999, is the free scale topology of biological networks [59]. This property is characterized by node degrees following a power law distribution, where a few highly connected hub nodes maintain network cohesion while most nodes maintain a limited number of connections.

Methods unraveling complexity in biological networks

The aforementioned properties enable computational analysis of large-scale biological networks. Key analytical approaches include the identification of important nodes based on centrality measures such as degree, betweenness, closeness and eigenvector centrality [60].

Additional methodologies focus on detecting functional modules and communities using clustering methods [61], comparing networks through similarity and distance measures, modeling network perturbations and robustness, and predicting links or interactions [62]. The scientific community has developed various tools for network visualization, topological analysis and integration of omics datasets, with Cytoscape emerging as one of the most widely adopted and versatile platforms. Several major categories of biological networks capture different aspects of cellular interaction:

Metabolic networks represent the complete set of metabolic and physical processes that determine the physiological and biochemical properties of a cell [63]. The nodes represents metabolites such as carbohydrates, lipids, amino acids and nucleotides. The edges are the chemical reactions representing the process between two metabolites such as catalisation by an enzyme. Analysis of these networks can provide insights into how selection acts on the metabolic pathways. Perturbations in conserved subsystems of metabolic networks influence the functions of the entire network across different species more strongly than other subsystems [64].

Protein-Protein Interaction networks represents the interactions between proteins, as the nodes in a cell [65]. Their interactions are exclusively described as undirected edges, and they are essential to most cellular processes. They tend to exhibit small-world network properties, helping to identify functional modules. Analyzing them also helps revealing protein functions, discover disease mechanisms and find potential new drug targets [66].

Gene regulatory networks represent how genes would activate or repress each other's expression [67]. Genes are represented by nodes and the directed edges indicate the regulatory interactions, with transcription factors as key players. They exhibit a hierarchical scale-free topology, with few highly connected hub genes and many genes with few connections.

Signaling networks represent how cells sens and reacts to their environment via cascades of protein-protein interactions and post-translational modifications. Typically, they integrate the above mentioned networks, nodes are proteins and metabolites and edges are physical and regulatory interactions. They are characterized by their modularity where they have functional modules such as Kinase cascades [68], feedback loops [69].

Advancements in biological network databases

The field has witnessed significant advancement in computational tools and methodologies designed to extract biological insights from network data [70]. Graph Neural Networks (GNN) have emerged as powerful deep learning methods for graph-structured data. In the context of Protein-protein interaction networks (PPI), GNNs facilitate tasks such as node classification for protein function prediction and link prediction for novel PPIs identification [71], employing methods including GraphSAGE and node2vec [62].

Notable developments include tool such as SWIM (switch miner), which enables identification of "switch genes" that may regulate major changes during biological transitions or disease progression [72]. The extensive availability of biological network data hgas facilitated investigation and prediction of protein function and drug interaction using Drug-Drug interaction network (DDI) [62]. While current applications remain limited to specific datasets such as TCGA, extensive development of network-based integrative MO approaches presents promising opportunities for resolving complex disease [73].

To support network-based analyses, the scientific community has developed numerous specialized databases, allowing for curation and storage of networks obtained from various experimental sources [62]:

PPI databases STRING: Functional protein association networks, Human Protein Reference Database (HPRD), IntAct, BioGRID.

Pathways databases KEGG: Molecular pathways and modules, Reactome, WikiPathways: Community-curated.

Regulatory Network databases TRRUST: Transcriptional regulatory networks, miRTarBase: microRNA-target interactions.

Network analysis drives advances in systems biology applications

Network-based methodologies have demonstrated significant utility in prioritizing important omics features and subnetworks associated with specific phenotypes or disease states, contributing to the discovery of novel biomarkers and disease subtypes [74]. The integration of networks representing different omics layers enables investigation of the interplay and crosstalk between these layers. Advanced approaches, such as iOMICSPASS [75], incorporate

prior knowledge and inference of causal regulatory relationships, facilitating identification of potential molecular drivers of biological processes such as Transcription factors and target genes.

MO data integration through network approaches enables capture of temporal and spatial dynamics of biological systems, advancing understanding of the evolution of cellular states and molecular mechanisms underlying physiological and disease processes. Networks approaches provide and intuitive framework for visualizing and interpreting complex omics data by aligning the experimental results with known and curated molecular interaction networks, thereby generating mechanistic insights and hypotheses.

The research community has developed numerous specialized tools for various applications. GWAB tool enhances the detection of weak associations in GWAS data, while Netter facilitates gene network inference. MetaNetVar provides sophisticated capabilities for genomic variant analysis, and MUFFINN specializes in identifying cancer genes from somatic mutation data [70]. These methodological advances have yielded significant mechanistic insights into diseases, as exemplified by MRNETSnapshot's focus on WNT1-regulated network in neurodegeneration, GENIE3's identification of regulators in melanoma invasive cell states, and GRNBoost2's application in kidney fibrosis research [76].

Addressing challenges in shaping future systems biology

The field faces significant challenges in managing the heterogeneity and complexity of generated data. Each data type presents unique characteristics [77] in term of scale, noise levels, and missing values, complicating direct comparison and integration. Furthermore, relationships between data layers exhibit intricate complexity due to sophisticated regulatory mechanisms.

The dynamic nature of the biological networks present another fundamental challenge. Current collected data represents an incomplete interactome, and most network capture only static snapshots of inherently dynamic systems. Addressing these limitations requires the development of more advanced computational models that integrate temporal and spatial information, such as dynamic Bayesian networks and agent-based models.

Result interpretation and validation constitute significant challenges, particularly due to the size and complexity of the networks, especially in MO data context where meaningful pattern prove difficult to identify [78]. Moreover, experimental validation of predicted results often requires substantial resources [79]. These challenges necessitate the development of

more experimental validation strategies and enhanced visualization and exploration tools.

To address these challenges, researchers have proposed several strategic directions. The development of more advanced computational methods for data integration and network inference, utilizing deep learning and graph neural networks, shows promise in addressing data heterogeneity and non-linear relationships between omics layers. Another promising approach involves contextual modeling through the incorporation of prior knowledge into the network models, potentially reducing false positive results and improving the interpretability through tissue-specific or disease-specific network analysis.

Finally, the field would benefit from enhanced Collaborative Research initiatives that bring together interdisciplinary expertise. By fostering collaboration between experimental and computational biologists, along with for other field, we can accelerate the research and unlock the full potential of the dataset to understand and treat complex diseases.

1.2 Omics approaches explore one dimension at a time

1.2.1 Omics as a gateway to molecular biology

The field of omics encompasses diverse biological disciplines focused on comprehensively studying distinct molecular layers within an organism. These disciplines have evolved through the implementation of high-throughput technologies, as previously discussed in section 1.1. Each omics field systematically investigates a specific molecular layer, providing unique insights into biological systems. The primary omics fields include:

Genomics As previously described, genomics investigates genes and the genome to elucidate the relationships between genetic elements and their influence on phenotypic expression.

Transcriptomics This fields examines the entire set of RNA molecules, including messenger-RNA, ribosomal-RNA, transfer-RNA and other non-coding-RNAs.

Proteomics This domain investigates proteins to gain insights into their structure, function and cellular interactions. The proteome represents the complete set of expressed proteins at a specific time point.

Metabolomics This filed conducts large-scale studies of small molecules, commonly referred to as metabolites, within cells, or entire organisms. The metabolome encompasses the set of metabolites present in a biological system.

Phenomics This discipline examines physical and biochemical traits of the organism, particularly focusing on their susceptibility to change due to genetic mutation and environmental factors.

With continuous advancements in research methodologies and measurement technologies, the omics landscape has expanded significantly, giving rise to numerous specialized fields. The following sections will elaborate on specific omics approaches that are related to the collaborative efforts this PhD project.

1.2.2 Addressing gene variant challenges at cohort scale

Genome-Wide Association Study (GWAS) represent a key methodology related of this PhD project, with applications discussed in section 5.2. This powerful analytical approach systematically identifies genetic variants associated with specific traits or diseases across the genome [80]. GWAS methodology typically involves genotyping numerous SNPs in a population cohort and conducting statistical analysis to identify significant associations between phenotypes of interest and genetic variants.

The typical GWAS workflow includes several critical stages: First, investigators carefully select individuals with the disease and corresponding controls in "binary condition" [81]. This is followed by DNA isolation, genotyping and rigorous QC procedures. Subsequently, researchers conduct statistical testing for associations between validated SNPs that meet QC thresholds and are relevant to the outcome. The final phase requires replication of identified associations in independent populations and/or experimental validation of functional implications.

GWAS has demonstrated remarkable success in identifying numerous genetic variants associated with diverse human diseases and traits, providing crucial insights into the underlying biological pathways [80]. The outcomes of theses studies serve multiple applications, including: estimating heritability, calculating genetic correlations, developing clinical risk predictions, informing drug development strategies, and inferring potential causal relationships between risk factors and heath outcomes. GWAS offers several key advantages:

Unbiased approach GWAS is a hypothesis-free approach, screening millions of genetic variants without prior assumptions about traits associations. This methodology facilitates the discovery of novel associations that might have remained unexplored through traditional hypothesis-driven research [82].

Gene and Pathway identification The approach helps in identifying genes and biological pathways involved in complex diseases, thereby providing insights into the disease mechanisms and potential drug targets, which ultimately enhances understanding of underlying biological processes and therapeutic interventions possibilities [80].

Statistical Power through scale The substantial patient cohorts in modern GWAS studies enable robust statistical power to detect associations between genetic variant and phenotypes, facilitating the identification of smaller effect sizes and rare variants that might be overlooked in limited studies [83].

Collaborative potential The filed benefits from large-scale collaboration and resource sharing among researcher worldwide, facilitating the pooling of resources and expertise to address complex research questions. Notable examples include the International Genomics of Alzheimer's Project (IGAP) and the Psychiatric Genomics Consortium (PGC).

Clinical Application GWAS findings contribute to precision medicine through the development of Polygenic risk score (PRS), which aggregate the effects of multiple genetic variants to predict individual disease risk. These scores have demonstrated significant potential in identifying high-risks individuals for various conditions, including coronary artery disease and autoimmune diseases [84, 85].

GWAS and its limitations in genetic research

Despite its advantages, GWAS faces several significant challenges. The "missing heritability" problem persists, as most SNPs demonstrate small effect on disease susceptibility, thereby limiting improvements in risk prediction through genetic testing [86]. The challenge of non-causal associations remains significant, as GWAS identifies associations that require validation through fine mapping and functional studies to establish causations [87].

A significant constraint lies in the fact that approximately 30 percents of GWAS associations occur in non-coding regions, complicating the interpretation of their relevance to the trait of interest. Furthermore, functional consequences require validation through animal models to gain insight into disease mechanisms [83]. However, the integration of GWAS with other omics data, particularly epigenetic information, may provide additional insights.

The clinical application of PRS scores faces limitations due to population specificity, with results often showing limited transferability across different populations [80]. This limitation is particularly significant given that most GWAS have been conducted primarily in European populations, restricting the generalization of findings to other ethnic groups.

1.2.3 scRNA-seq for transcriptome decoding

scRNA-seq represent a method for detecting and quantifying mRNA molecules at unprecedented resolution, enabling molecular analysis at the individual cell level. This advanced technique provides granular insights into cellular responses, revealing heterogeneity that would be masked in bulk RNA sequencing approaches [88]. The analytical framework for scRNA-seq consists in three primary stages:

1. Raw data processing and QC.

2. Basic data analysis, including clustering and cell type labeling.

3. Advanced data analysis tailored to address specific research objectives.

The exponential growth in scRNA-seq applications has driven the development of specialized computational approaches designed to handle vast data volumes generated by these experiments. In the context of cancer research, scRNA-seq has proven particularly valuable for characterizing cellular heterogeneity, enabling precise molecular discrimination between "normal" and "malignant" cell populations [89]. This capability provides critical comparative insights between cell subsets, enhancing our understanding of disease mechanisms.

In the pharmaceutical domain, scRNA-seq has emerged as a powerful tool for informed decision-making processes. The technology facilitates enhanced biomarker identification for patient stratification and enables more precise therapeutic targeting strategies [90]. For comprehensive guidance on methodological best practices in scRNA-seq analysis, readers are directed to the forthcoming book chapter: Galati et al. 2025, titled "Best practices in single-cell RNA-seq data analysis".

1.2.4 Metabolomics: Mapping metabolic pathways

The metabolome refers to the complete repertoire of small molecule chemicals, known as metabolites, present within biological sample. These metabolites can be categorized into two distinct classes: endogenous metabolites, which are produced by the organism itself (including amino acids, organic acids, nucleic acids, fatty acids, amines, sugars, vitamins, co-factors, pigments, antibiotics ...), and exogenous metabolites, which originate from external sources (such as drugs, environmental contaminants, food additives, toxins and other xenobiotics).

The metabolome serves as a crucial interface between an organism's genome and its environment, making it an ideal proxy for the phenotype expression. This dynamic field has found diverse applications across multiple sectors. In pharmaceutical research and healthcare, metabolomics facilitates biomarker discovery, enables drug safety screening, and advances Personalized medicine approaches. In agricultural applications, the field contributes to stress testing, crop protection and genetic engineering initiatives [91]. Within biomedical research, metabolomics plays a pivotal role in biomarkers identification and elucidation of molecular mechanisms [92].

Three principal technological platforms dominate metabolite measurement initiatives: nuclear magnetic resonance (NMR) spectroscopy and two variant of mass spectrometry (LC-MS and GC-MS) [93]. In the specific context of food allergy and intolerance research, targeted diagnostic approaches have emerged. For instance, Micro Array Diagnostics (MADx) have developed diagnostic tools such as ALEX^2, an ELISA-based (enzyme-linked immunosorbent assay) in vitro multiplex allergy test, which enables simultaneous measurements of total Immunoglobulin E (IgE) and specific IgE against an extensive panel of allergen extracts and molecular allergens. This is particularly relevant for studying IgE-mediated allergies, where the immune system produce responses to specific allergens by producing IgE antibodies. These antibodies subsequently bind to allergens, triggering the release of histamine and other chemicals mediators responsible for allergic reactions symptoms [94].

Metabolomic data analysis typically employs a combination of statistical approaches, incorporating both univariate and multivariate methodologies such as ANOVA (analysis of variance), PCA (Principal Component Analysis) and PLS-DA (Partial Least-Squares Discriminant Analysis), to identify significant metabolic signature. Subsequent pathway and network analyses utilize established tools and databases, including KEGG and MetaboAnalyst, with the latter serving as a comprehensive analytical toolkit [95].

1.2.5 Microbiome studies: Investigating host-microorganisms interplay

The microbiome constitutes a complex ecological community of microorganism inhabiting a specific environment, encompassing bacteria, archaea, fungi, algae, small protists and their associated elements including phages, viruses, plasmids. This dynamic ecosystem demonstrates remarkable plasticity, responding to various environmental stimuli such as exercise, diet, medication. The human body hosts distinct and specialized microbial communities across various anatomical sites, including the skin, gastrointestinal tract, respiratory tract, and oral cavity [96].

These microbial communities perform essential functions in human physiology, particularly in fundamental processes such as digestion, metabolism, immune system development, and protection against pathogenic organisms [97]. A healthy microbiome is characterized by high diversity and optimal balance in symbiotic relationship with the host organism [98]. Perturbations to this delicate balance have been implicated in numerous pathological conditions, including Inflammatory Bowel Disease, Irritable Bowel Syndrome, Colorectral cancer (CRC), obesity and metabolic disorders such as type 2 diabetes [99]. The microbiome influences through multiple mechanisms, including inflammation modulation, alterion of the gut barrier function, and complex interactions with the host immune system.

Microbiome research methodologies have been revolutionized by advances in high-throughput sequencing technologies. Two predominant approaches for investigating microbiome composition have emerged: 16S rRNA gene sequencing and Shotgun metagenomic [100]. Additional insights into microbiome function are gained through complementary omics approaches, including metatranscriptomics, metaproteomics and metabolomics.

The field has witnessed a substantial development in computational tools and statistical methods specifically designed to analyze the vast quantities of generated data, addressing the unique challenges posed by high-dimensionality, sparsity, and compositional nature of microbiome data. Machine learning approaches are rapidly evolving to facilitate microbiome-based disease prediction and biomarker discovery [101]. Various experimental models have been established to investigate host-microbiome interactions, including germ-free mice for in-vivo testing, and organ-on-chip systems, and anaerobic culturing techniques for in-vitro testing.

The future of microbiome research holds promising therapeutic strategies, including the development of targeted probiotics and prebiotics for microbial composition modulation [102],

implementation of fecal microbiota transplantation protocols for microbiome restoration, advancement of personalized nutrition strategies based on an individual microbiome profile [103], creation of engineered microbial systems for specific functions [104], and utilization of microbiome-derived metabolites and postbiotics as therapeutic agents.

However, the translation of microbiome research findings into clinical applications faces significant challenges, primarily due to substantial inter-individual variability, the intricate nature of host-microbe interactions, and the necessity for extensive clinical trials [105].

1.2.6 Barriers to integration in single omics

While individual omics layer provide valuable insights, their true potential lies in their contribution to systems biology through integration of fundamental data layers for comprehensive understanding. However, the integration of various omics layers presents significant challenges [106]. Omics datasets are characterized by high-dimensionality, such as described previously for microbiome data where there are more variables than samples, which complicates analysis and interpretation. Data quality exhibits considerable variability and heterogeneity, with inherent noise potentially obscuring "true" biological signals, necessitating careful preprocessing and normalization techniques. Furthermore, the diversity of platforms and technologies generates datasets requiring harmonization for cross-study comparability.

Challenges highlight the need for Multi-Omics approaches

Understanding the limitations and capabilities of single omics approaches provides crucial foundation for exploring MO integration analyses, which will be discussed in the subsequent section. These integrative approaches combine diverse omics layers to achieve a comprehensive and holistic view of the biological systems, potentially overcoming some single-omics limitation and providing more accurate insights into biological complexity.

1.3 Multi-Omics provides holistic biological framework

1.3.1 Multi-Omics integration drives biological understanding

Multi-Omics (MO), also known as integrative omics or pan-omics, represents a comprehensive analytical paradigm that integrates multiple omics datasets. This complex approach enables researchers to mine complex biological big data, thereby uncovering novel associa-

tions between biological entities, identifying novel markers, and elucidating the underlying mechanisms of physiology and diseases.

A principal application of MO lies in the study of human disease, particularly in critical areas such as cancer, neurodegenerative disorders, aging, and drug target discovery [107]. Through comprehensive characterization of molecular alterations at different levels and their associations with these conditions, researchers can identify potential therapeutic targets and develop personalized treatment strategies.

The field of system biology has been instrumental in developing frameworks and computational tools to combine and analyze large-scale biological data. However, the exponential growth in data quantity and rapid evolution of technologies present significant challenges in term of data management, integration, and interpretation. Successfully addressing these challenges would enhance the power of MO models and improve the accuracy of biological insights [108]. As the field continues to advance, MO integration holds great promise for PM, drug development, and understanding disease mechanisms.

1.3.2 Foundational principles highlighting data integration

MO integration facilitates the analysis of heterogeneous datasets, enabling statistical analysis and the application of machine learning techniques. Notably, unsupervised techniques predominates in discovering molecular/disease subtypes and different patterns [109].

Current MO methods can be classified into two types: horizontal and vertical integration [110]. While vertical integration encompasses different and various omics datasets as previously described, horizontal integration focuses on the integration of diverse datasets within a single omics type.

1.3.3 Integration techniques in Multi-Omics research

The complexity and dynamic nature of this research field necessitates understanding not only the statistical foundations of the methods but also their purposes, limitations, and most crucially, the appropriate use cases for the datasets under investigation. This section aims to provide comprehensive explanations to enhance understanding of different use case and methods that are published in the literature that could be considered into a tailored analysis. It is important to note that tool accessibility varies, with some available as web based service or

through programming language (primarily R, but also Python and Matlab). While numerous reviews have thoroughly explained each facet of these tools, this chapter aims to present the information in an accessible and structured manner.

When implementing an integration analysis, researchers can employ several distinct strategies to manipulate datasets. The main options include and are described in [111]:

Early integration which involves concatenating datasets into a single matrix.

Mixed integration which requires independent transformation/mapping of each dataset before combination.

Intermediate integration which transforms datasets into common and omics-specific representations.

Late integration where analyses are conducted separately and final predictions are combined.

Hierarchical integration which utilizes prior knowledge (such as regulatory relationships) to guide the integration process.

Recent literature confirms these MO approaches in the metabolomics context, specifically addressing the "How and When" aspects within the workflow of dataset integration [112]. Of particular significance, the research emphasizes the importance of hypothesis-driven research and delineates different approaches such as multi-staged and meta-dimensional integration.

From a ML perspective, tools can be categorized according to standard statistical framework [113]:

Supervised approaches focus on predicting one or more target associated with a given sample. This category subdivides into classifier (predicting sample classes such as pathogenic versus and non-pathogenic samples) and regressors (estimating quantities, such as pathogenicity risk level). Both methodologies commonly employ support vector machines (SVM) and artificial neural networks (ANN).

Unsupervised approaches concentrate on exploring the data structure deconstructing its variation or correlation. These methods primarily fall into association algorithms (uncovering latent rules or trend in the data) or clustering algorithms (partitioning samples based on hidden characteristics).

The reduction and simplification of large MO datasets facilitates interpretation. Principal Component Analysis stands as one of the widely adopted approaches, reducing datasets to lower dimensions while preserving maximum of the original variance among variables. Factor Analysis serves to decompose data using latent relationships that describes correlations between variables. Matrix factorization functions to 'denoise' the original dataset, with Nonnegative Matrix Factorization specifically employed in uncovering ecological interaction networks from metagenomics samples.

In the context of clustering, k-means and hierarchical clustering remain the most prevalent methods in Life Sciences [114]. While relatively simplistic in their approaches, they offer straightforward interpretation and comprehension for collaborators who generated the data but may not specialized in statistics compared to other more advanced available approaches.

Among academic literature classification proposition, three different classes are distinct by their approaches to handle datasets:

Concatenation based combining different omics dataset into one matrix as described as early integration earlier.

Transformation based which is related to the intermediate integration from earlier review article.

Model based handling multiple omics simultaneously with statistical or ML models, such as MoGCN [115].

Further classification focused on algorithmic aspects presents four primary categories [116]:

Network-Free Non-Bayesian encompassing sequential analysis and including sparse Multi-Block Partial Least Square (sMB-PLS) regression or iPAC which is an unsupervised method of CNV and gene expression data.

Network-Free Bayesian often utilizing prior distributions to model data. For instance, iCluster uses a Gaussian latent variable model, and Dirichlet Multinomial Allocation (DMA) uses mixture models.

Network-based Non-Bayesian using molecular interaction networks or correlation-based networks, through different strategies *i.e.* similarity network fusion (SNF).

Network-based Bayesian leveraging probabilistic models such as Paradigm tool algorithm.

1.3.4 Beyond omics: Integrating non-omics data with Multi-Omics

The integration of MO data presents several significant challenges that must be addressed for successful implementation [117]:

Data heterogeneity needs distinct approaches

A fundamental challenge in MO integration relies on the diversity of data types, standards, and formats across omics research methods [118]. Each omics technology generates distinct outputs with specific standards, which significantly complicates the prior integration process. This heterogeneity introduces additional complexity when scaling, normalizing, and transforming the data to meet the requirements of statistically robust analysis. In this context, batch effect management becomes increasingly critical.

Leveraging missing data

The challenge of missing values impacts MO analysis for multiple reasons [119]. Current technological limitations present the measurement of all the bio-molecules, resulting in inherently missing of filtered information. furthermore, the distribution of missing observations and their proportions often vary substantially among different datasets. While this issue affects all omics disciplines, it manifests most prominently in metabolomics due to the extent and complexity of the chemical space and the relative newness of the field [120].

Scalability and computational bottlenecks

The process of merging datasets and applying integration strategies demands substantial computational resources for processing and analysis [121]. This challenge becomes particularly acute when dealing with big data and numerous features. The computational complexity increases exponentially as more feature are considered, leading to vast combinations that must be evaluated [122].

Multi-omics results remains complex

The intricate nature of biological systems manifests in complex MO results that present significant interpretational challenges [120]. The quality of input data proves crucial, as substandard data inevitably produces unreliable or disputable results.

While the subsequent collaboration sections will exemplify strategies and methods in greater detail, it is important to note that each approach presents its own specific challenges. For instance, intermediate integration typically necessitates robust pre-processing due to data heterogeneity, while late integration may fail to capture inter-omics interaction since each omics is analyzed separately before integration.

1.3.5 Multi-Omics contributes to NCDs studies

Recent advances in MO studies have successfully identified specific biomarkers across various conditions:

Pancreatic cancer A study of 175 patients from the TCGA database integrated gene mutation expression, methylation level distribution and m RNA expression data [123]. Through weighted correlation network analysis, researchers identified nine hub genes as prognostic biomarkers: MST1R, TMPRSS4, PTK6, KLF5, CGN, ABHD17C, MUC1, CAPN8, and B3GNT3.

Colorectal cancer A study revealed elevated metabolic levels of oleic acid and FA (18:2) in patients compared to healthy controls, suggesting their potential as as plasma biomarker for early diagnosis [124]. Additional review summarizes recent findings and biomarker proposed by MO integrations [125].

Cervical cancer A microarray dataset analysis study from GEO identified several Differentially expressed genes as potential biomarkers utilizing MO database validation approaches: MCM4, NUSAP1, CDCA5, CDC45, DTL and CDT1 [126].

Toxicology Understanding gene responses to toxicant, and other stressors [127]. This understanding of individual omics layers proves essential for the construction and interpretation of biological networks.

1.3.6 Multi-Omics analysis enables future clinical applications

The emergence of single-cell MO technologies has fundamentally transformed molecular cell biology research [128]. This revolutionary advancement facilitates comprehensive cellular characterization by integrating data from various cellular layers (*i.e.* omics) and characterizing cell states and activities.

Through the integration of diverse cellular data, researchers can now achieve unprecedented insights into cellular function and behavior. This integrated approach provides a more nuanced understanding of cellular heterogeneity and dynamics, offering valuable insights for both fundamental research and clinical applications.

Chapter 2

Translational applications in clinical research

2.1 Bridging scientific outcomes to clinical value

Translational research represents a critical bridge between fundamental scientific discoveries and their practical clinical applications, with the ultimate goal of improving human health and patient well-being. The commonly used "bench-to-bedside" paradigm effectively illustrates this process, wherein laboratory discoveries are translated into innovative therapeutic interventions and clinical recommendations. This research approach operates bidirectionally, fostering close collaboration among researchers from diverse disciplines across multiple organizational levels.

2.1.1 Bioinformatics link science and clinical applications

Building upon the foundations established in previous sections, bioinformatics has emerged as a crucial interdisciplinary field that develops methodologies and tools for analyzing and interpreting the exponentially growing volume of biological data. The significance of bioinformatics can be attributed to several key capabilities:

Data Interpretation Advanced processing of omics data, big data analysis, identification of potential biomarkers.

Data Integration Integration of heterogeneous datasets (*e.g.* omics, molecular, clinical, metadata).

Predictive analysis Development of disease progression model, prediction of treatment outcomes, and patient responses assessment in the context of precision medicine, and "digital twins" development.

Drug Research Simulation of drug-target interactions, thereby significantly reducing the temporal and financial investments compared to traditional drug discovery approaches.

2.1.2 Translating bioinformatics insights to clinical practice

The translation of bioinformatics research into clinical practice represent a crucial step in leveraging computational insights for enhanced patient care. This translation manifests across several critical domains:

Diagnostic tools Development of high-precision diagnostic tests based on validated molecular signature, enabling more accurate and specific disease identification [129].

Therapeutics Enhanced understanding of disease mechanisms facilitates the development of novel therapeutic agents and enables strategic re-purposing of existing drugs [129].

Personalized medicine personalized treatment optimized for specific subtype/subgroups of disease/patients [130].

Clinical trials Enhancing statistically robust clinical trial design, facilitating efficient and meaningful data collection [131].

Public health Population scale analysis enabling the identification of disease risk factors to inform evidence-based public health policies [132].

2.1.3 Enrichment analysis unlocks new biological insights

Enrichment analysis represent a crucial methodological approach for translating extensive gene lists, identified through various experimental setups, into meaningful insights about biological processes, pathways or functions [133]. This analytical approach facilitates the interpretation and derivation of meaningful biological conclusions.

The reliability and interpretation of enrichment analyses are influenced by multiple factors, including the choice of the statistical methodology, algorithmic approaches, and the quality of reference data sources. To ensure robust results, it is imperative to either collaborate with

domain experts investigator with extensive biological knowledge or utilize carefully cured database, while adhering to established statistical practices for post-hoc corrections, including the application of adjusted p-values, Confidence Intervals.

Research has demonstrated that Gene Set Enrichment Analysis (GSEA) effectively addresses challenges in subjective gene filtering and ineffective gene-to-gene comparisons, while facilitating the selection of appropriate animal models [134]. Furthermore, enrichment approaches have demonstrated significant utility in multivariate and MO analyses [135].

Extensive literature reviews reveals numerous successful applications of enrichment analysis in identifying therapeutic targets and elucidating molecular pathways, leading to novel strategic approaches in disease intervention.

2.1.4 In silico modeling simulates disease and drug responses

In-silico modeling has emerged as a powerful computational approach for predicting drug responses and simulating disease progression patterns. Through the integration of in vitro and in vivo experimental data, researchers can test extensive arrays of parameters and explore diverse experimental configurations, particularly when considering the organism as a whole.

A comprehensive review of modeling approaches demonstrates their broad application across various medical conditions, including cancer, immunological disorders, neurological diseases, and infectious diseases [136]. Notable success has been achieved in drug response prediction for Glioblastoma, where drug-mutation associations demonstrated 85% accuracy [137].

The disease Maps Project represents a significant collaborative initiative directed at enhancing our understanding of disease specific mechanisms [138]. This comprehensive project involves multidisciplinary experts, including clinicians and biologists who validate core mechanisms for accurate representation. The project framework extends standards formats such as SBGN, SBML and BioPAX.

Through open participation and systematic community engagement, the project maintains rigorous standards for data curation, representations and accessibility. The initiative emphasizes a modular approach for representing molecular cascades through interoperable and reusable sub-maps. Notable implementations include COVID-19, Parkinson disease maps, and the Atlas of Cancer Signaling Network. Additional modeling frameworks have emerged, including integration of compound databases for examination target-mediated modulation of disease pathways [139].

2.1.5 Navigating challenges and opportunities in translational research

Ensuring reproducibility strengthens bioinformatics workflows

Reproducibility represent a pillar principle in bioinformatics, essential for ensuring the reliability and verifiability of research findings. While implementing reproducibility in daily research practices presents significant challenges, the "Five pillars of reproducible computational research" emphasize the importance of literate programming, code version control, compute environment control, persistent data sharing and documentation [140].

Despite established guidelines, several factors impede reproducibility, including the complexity of bioinformatics workflows, dependence on specialized software, and absence of standardized formats for specific analyses.

The analysis and integration of highly heterogeneous datasets present both significant methodological challenges and unprecedented opportunities for discovery. While numerous integration strategies exist to address various analytical limitations the inherent complexity of these approaches often creates barriers to reproducibility and method translation across different studies and datasets [141].

Increasing complexity in modern drug development

The drug development process represents a substantial investment of time and resources, characterized by significant uncertainty regarding the probability of successful outcomes [142]. Research in specific domains, particularly nervous system disorders, manifests more complexity due to unknown physio pathology in various conditions. The limitations of animal models in fully replicating disorder characteristics, combined with significant patient heterogeneity, present substantial challenges in translating preclinical findings to clinical trials. Furthermore, the absence of standardized, widely-adopted biomarkers and diagnostic tools for objective disease detection and biological state measurement compounds these challenges across most disease categories.

Regulatory and administrative barriers challenge progress

Clinical trials frequently encounter significant regulatory and administrative challenges, particularly manifested in prolonged intervals between protocol approval and trial implementation in action [143]. Such delays stem not only from regulations requirements but also from

institutional internal processes and bottlenecks. The Occluded Artery Trial (OAT) exemplifies these challenges, experiencing a three-year delay between the initial NIH committee meeting and trial commencement. These prolonged timelines not only escalate research costs but may also compromise the clinical relevance of trial, as medical practices can evolve significantly during the delay period.

Target validation presents significant challenges

Early-stage inadequacies in drug target validation have been directly correlated with substantial clinical failures and reduced drug approval rates [144, 145]. The validation of molecular targets as causal factors in disorders, coupled with confirmation of drug target validity, continues to present significant challenges. These challenges are exacerbated by inherent biological disparities observed across in silico, in vitro and in vivo models. The comprehensive validation of molecular targets often remains incomplete until successful drug development, creating a circular challenge that particularly impacts first-in-class drug discovery and development processes.

Ethical considerations in translational bioinformatics

The ethical implementation of translational bioinformatics necessitates careful consideration of patient data privacy, and security. The inherently sensitive nature of patient data, particularly genetic information, presents significant privacy implications due to its capacity to reveal comprehensive insights into individual health status, traits, and potential health risks. The exponential growth of large-scale genomics databases introduces heightened privacy vulnerability concerns, particularly regarding personal health information. This challenge is further complicated by the open-source nature of numerous bioinformatic tools and databases. Therefore, robust data system security becomes crucial for preserving patient trust and maintaining the integrity of the biomedical research.

Consent and genomics data utilization [146, 147]. Informed consent defines ethical research, ensuring comprehensive participant awareness regarding data utilization and facilitating informed participation decision. Current bioinformatic research present unique challenges to traditional informed consent frameworks due to its dynamic nature, where data may be repurposed for multiple future studies and integration analyses not initially anticipated during data collection. The concept of dynamic consent, enabling

ongoing patient control over data utilization, has emerged as a potential solution to these challenges. Furthermore, the sharing of patient genomic datasets introduces additional privacy concerns, particularly regarding potential re-identification, necessitating meticulous data processing protocols and comprehensive consents considerations that align with patients preferences.

Personalized medicine The advent of PM represents a paradigm shift in healthcare delivery, offering potential for optimizing therapeutic interventions through targeted treatment approaches and reducing adverse effects [148]. This advancement, while promising, raises ethical considerations regarding healthcare equity and accessibility. The implementation of PM approaches introduces complex challenges in healthcare delivery systems, particularly concerning equitable access to advanced therapeutic strategies. A critical ethical consideration emerges from the potential exacerbation of existing healthcare disparities. This disparity manifests in two primary dimensions: socioeconomic access and genetic representation. Populations with greater financial resources may disproportionately benefit from tailored therapeutic approaches, while genetic-based interventions may preferentially serve demographic groups whose genetic profiles have been extensively studied in research.

2.2 The i2TRON project links inflammation to Noncomunicable diseases

2.2.1 Mechanisms of inflammation in acute and chronic diseases

The mechanisms of acute inflammation are complex biological responses to various stimuli, including pathogens, damaged cells or irritants [149]. This intricate cascade of reactions is initiated when the immune system, particularly macrophages and dendritic cells, recognizes harmful stimuli through pattern recognition receptors. Subsequently, these sentinel cells trigger the inflammatory response by releasing inflammatory mediators, including cytokines, chemokines, and eicosanoids, which triggers a series of vascular changes. These changes encompass blood vessel dilatation, increases vascular permeability, and enhanced immune cell recruitment to the inflamed site.

The acute inflammatory response is characterized by the strategic recruitment of neutrophils and monocytes, manifesting in the acute inflammation symptoms: redness, swelling, heat, pain, and potential loss of function. These cellular and molecular events converge to eliminate the harmful agent through mechanisms, including phagocytosis, release of reactive oxygen species, and antibody production. Following the elimination of the treat, anti-inflammatory mediators initiate the resolution phase, dampening the inflammatory response and promoting tissue repair to restore homeostasis.

While acute inflammation is self-limiting and protective, its persistence can lead to chronic inflammation, potentially resulting in tissue damage and contributing to various non, including diabetes and cancer. Chronic inflammation can arise from multiple sources, including:

- Persistent autoimmune response to self antigens.
- Prolonged exposure to irritants, such as industrial chemicals.
- Inability to eliminate the initial inflammatory agent.

Research has identified several modifiable risk factors associated with chronic inflammation, including obesity, smoking, stress and sleep deprivation. [150] The chronic inflammatory response is distinguished from acute inflammation by its distinct cellular composition, predominantly involving macrophages, ultimately leading to progressive tissue damage and fibrosis [151].

2.2.2 Aetiology of Colorectal cancer

Chronic inflammation represent on fundamental driver of Colorectral cancer (CRC) development, operating through multiple interconnected mechanisms including immune dysfunction, oxidative stress, DNA damage, and epigenetic modifications [152]. The inflammatory pathways implicated in CRC pathogenesis share significant overlap with inflammatory bowel disease, suggesting common molecular mechanisms underlying these conditions.

The relationship between inflammation and CRC is particularly evident in patients with IBD, specifically ulcerative colitis and Crohn's disease, who demonstrate an elevated risk of CRC development. This association is underpinned by shared inflammatory pathways including MMP10, LCN2, REG1A, REG3A, DUOX2 [153]. Furthermore, the inflammatory microenvironnement in CRC is characterized by distinct protein expression pattern in inflammatory cells, particularly macrophages and neutrophils [154].

Recent investigations have highlighted the crucial role of gut microbiome alterations in CRC pathogenesis. Several bacterial species have been implicated as significant risk factors, notably *Fusobacterium nucleatum*, *Enterotoxigenic Bacteroides fragilis*, and pks+ *E. coli* [155]. These microorganisms influence CRC development through multiple mechanism, including bacterial genotoxicity, biofilm formation, oxidative stress and dysbiosis [156].

As key feature in CRC progression is the epithelial-mesenchymal transition (EMT), where epithelial cells acquire mesenchymal phenotype, quantifiable through EMT scoring systems. This transition represents a crucial step in tumor progression and metastasis [157].

Nutritional and life style factors significantly influence CRC development through various inflammatory pathways. For instance, Fish-derived n-3 polyunsaturated fatty acids (PUFAs) demonstrate anti-neoplastic effects in CRC by modulating multiple cellular processes, including cell proliferation, apoptosis, angiogenesis and metastasis [158]. Additionally, dietary components such as heme iron have shown to modulate multiple pathways implicated in CRC aetiology [159]. The gut microbiota further mediates these nutritional influences through the metabolism of Short-Chain Fatty Acids (SCFA), which can modulate pro-inflammatory responses and potentially reduce CRC risks [160].

2.2.3 Aetiology of Peanut allergy

Peanut Allergy (PA) represents a significant public health concern, manifesting as immune-mediated responses ranging from mild symptoms to severe anaphylactic reactions. The increasing prevalence of PA in developed countries presents a substantial challenge to health-care systems [161]. The complex aetiology of PA involves intricate interactions between genetic predisposition, environmental influences including dietary patterns and life style factors, and skin barrier function [162].

The immunological basis of PA centers on type 1 hypersensitivity reactions. Upon initial exposure, the immune system generates peanut-specific IgE antibodies, which bind to effector cells, primarly mast cells and basophils, These cells contain granules rich in histamine and other inflammatory mediators. Subsequent allergen exposure triggers degranulation, releasing these inflammatory substances and initiating the allergic cascade.

The molecular complexity of peanut allergens is reflected in their classification into four major protein families, each with distinct structural and immunological properties [163]:

Cupin superfamilly Ara h 1 and Ara h 3.

Prolamin superfamily Ara 2, Ara h 6, Ara h 7 and Ara h 9.

Profilin familly Ara h 5.

Bet v-1-related poteins Ara h 8.

These protein families evolved to serve various plant functions, particularly protein storage and defense mechanisms. Understanding their molecular implications is crucial for elucidating cross-reactivity patterns among different allergen sources.

Post-translational modifications of peanut proteins, both enzymatic and non-enzymatic, significantly inflence their allergenicity. Advanced Glycation End-products (AGEs) modifications on peanut allergens can activate the receptor for AGE (RAGE), initiating a cascade that involves dendritic cells activation, T helper 2 (Th2) cell responses, and subsequent IgE production, ultimately leading to mast cells and eosinophils activation.

The thermal processing of peanut significantly impacts their immunogenicity and allergenicity profiles [164]. Additionally, research has identified associations between IgG-mediated food intolerance and increased gut permeability, potentially allowing food antigens to enter circulation. This mechanism may trigger the production of food-specific IgG antibodies, resulting in diverse manifestations including gastrointestinal symptoms, neurological effects, chronic fatigue and dermatological changes such as hair loss [165]. However, the presence of food-specific IgG does not necessarily indicate clinical allergy or intolerance, as it may represent a normal physiological response.

2.2.4 SCFA metabolism and early manifestation of inflammation

SCFA, primarly acetate, propionate and butyrate, are crucial metabolites produced through the anaerobic fermentation of Dietary fibers (DF) (non-digestible carbohydrates) by the gut microbiome. The absorption kinetics of these compounds are remarkable, with approximately 95% being rapidly absorbed by the colonocytes, while the remaining 5% are excreted in feces [166]. Of particular significance is butyrate, which serves as the primary energy sources for colonocytes, providing 60-70% of their energy requirements. These metabolites participate in diverse physiological processes, including energy metabolism, intestinal homeostasis maintenance, and immune system modulation [167].

The molecular mechanisms underlying SCFA function primarily operate through two major pathways: the activation of G Protein-Coupled Receptors (GPCRs) and inhibition of Histone Deacetylases (HDAC) [167]:

GPCR The pathway involves specific receptors, notably FFAR2 (GPR41), FFAR3 (GPR43) and GPR109A, which are expressed across various cells types including intestinal epithelial cells and immune cells such as monocytes, eosinophils and neutrophils [168]. This receptor activation modulates critical leukocyte functions, including cytokine production, chemotaxis and pathogen clearance capabilities.

HDAC Its inhibition, primarily mediated by butyrate, represents a significant epigenetic regulatory mechanism. This inhibition results in altered gene expression patterns, characterized by the suppression of pro-inflammatory cytokines and enhancement of anti-inflammatory responses. Furthermore, this pathway influences cellular processes including proliferation, differentiation and apoptosis.

The immunomodulatory effects of SCFAs extend beyond these primary pathways to encompass several additional mechanisms:

Cytokines Regulation SCFAs demonstrate the capacity to suppress pro-inflammatory cytokines including TNF-α, IL-2, and IL-6 while simultaneously enhancing the production of anti-inflammatory IL-10 [169].

Regulatory T-Cell differentiation Through FFAR2 activation, SCFAs promote the development of an anti-inflammatory micro-environment [170].

Leucocytes function modulation SCFAs influence monocytes and neutrophil functions through FFAR2-mediated mechanisms, affecting both chemoaxis and cellular activation [171, 168].

Pattern recognition receptor modulation SCFAs demonstrate the ability to modulate Toll-Like Receptor (TLR) signaling [172].

Inflammasome regulation SCFAs influence NLRP3 inflammasome activity through interactions with FFAR3 and GPR109A receptors [172].

Recent investigations have revealed additional interactions between SCFAs and various immune cell populations, including B cells and innate lymphoid cells, which play crucial roles in antibody production and immune regulation [173]. These findings further emphasize the extensive influence of SCFAs on immune system function and inflammatory processes.

2.2.5 Advancing inflammation research

The overarching objective of this research project, as outlined in table 2.1, centers on exploring diverse collaborative datasets to advance our understanding of inflammatory mechanisms in chronic diseases. Through the integration of cohort data and the application of tailored bioinformatic approaches, this project seeks to elucidate specific inflammatory mechanisms manifesting in both clinical and experimental contexts.

Our precision medicine approach fundamentally aims to advance the understanding of patient-specific inflammatory responses through the comprehensive analysis of symptomatic manifestations, molecular signatures, and genetic factors. This multi-layered investigation of inflammatory processes has potential to enhance clinical decision-making by providing more nuanced diagnostic insights.

The research strategy relies on several key objectives:

Integration of MO data The project leverages bioinformatics integration approaches to analyze complex datasets and experimental design from our collaborators. This comprehensive analysis enables the investigation of inflammatory mechanisms at multiple biological levels, providing a better holistic understanding of disease progression.

Mechanism characterization Through detailed examination of specific inflammatory mechanisms in various clinical and experimental settings, we aim to identify precursor events that may contribute to the development of complex diseases. this mechanistic understanding us crucial for early intervention strategies.

Stratification Analysis In collaborative investigation, we explore analytical methods to stratify patient populations and identify distinct feature subgroups. This approach enables the detection of complex relational pathways and potential disease subtypes, contributing to more personalized therapeutic strategies.

Translation to Clinical practice The ultimate aim is to translate the research findings into actionable clinical insights. By evaluating inflammation across diverse molecular and cellular layers, we seek to provide clinicians with more comprehensive diagnostic tools and improved decision-making frameworks.

This integrated approach to studying inflammatory processes has the potential to reveal novel insights into disease mechanisms and contribute to the development of more targeted therapeutic interventions. Through careful analysis of patient-specific responses and molecular signatures, we aim to advance the field of precision medicine in inflammatory diseases.

Topic	**Allergy**	**CRC**	**Fiber Diet**
Collaboration	LIH MTA	FSTM MDM	LIH NutriHealth
Public datasets	–	CRC, PDAC	–
Local datasets	APSIS	–	LuxFico
Methods	section 3.2	section 4.2	section 5.2
Data analysis	correlation, bootstrapping	scRNA analysis	GWAS
Evaluation	networks, comparative analysis	Pathway enrichment, CCC	genetic, immune
Results	section 3.3	section 4.3.1, section 4.3.4	Under embargo
Study output	Immune phenotypes	Cell type stratification	Inflammatory pathways
Modeling output	–	Ligand/Receptor targets	PRS score model
Validation targets	Inflammation markers	WNT5A interaction	Immune composition
Translational output	–	CMS4 group characterization	Fiber diet insights
Status	Targets investigations	Experimental Validation	Embargo until end of 2025
Impact	Personalized allergy insights	CRC tumor subtyping	Refined patient stratification
Publication	mid-2025	Complete draft	Draft (waiting for embargo)

Table 2.1 Thesis structure: Multi-modal data integration for disease mechanisms. The table outlines key topics, collaborations, datasets, analytical approaches, and expected outcomes across three research areas: allergy, Colorectral cancer (CRC), and Dietary fibers (DF). It highlights the use of multi-omics methods (*e.g.* GWAS, scRNA-seq, PRS), network analyses, and pathway enrichment to uncover immune and inflammatory mechanisms. The study aims to refine patient stratification, identify novel biomarkers, and generate translational insights, with expected publication timelines indicated.

Part II

Applications, methods and results

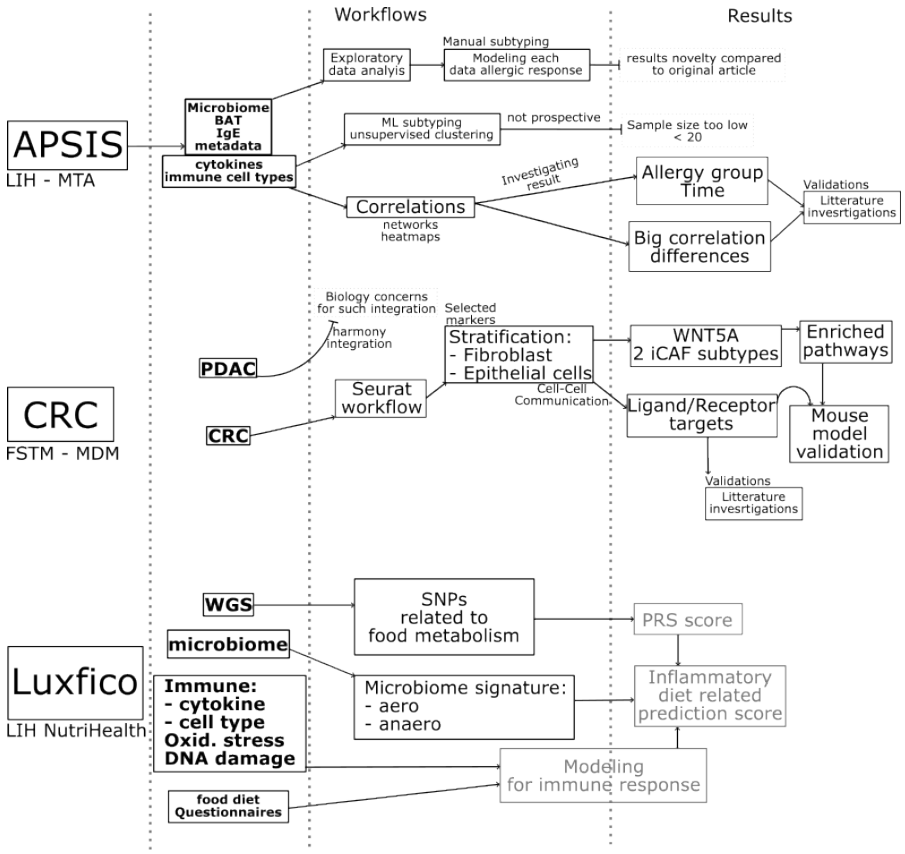


Figure 2.1 Multi-institutional investigation of inflammation and Noncommunicable diseases (NCD). The table outlines collaborative research efforts with three Luxembourg-based collaborators, integrating multi-omics data (*e.g.* microbiome, cytokines, immune cell types, genetics) to study inflammation-related mechanisms. Key workflows include exploratory data analysis, network-based correlation studies, stratification of immune and epithelial cells, and pathway enrichment analysis. Results contribute to refining patient subtyping, identifying immune-inflammatory signatures, and developing predictive models, including inflammatory diet-related risk scores.

This part provides an overview of the workflows, methodologies, and results derived from the collaborative projects forming the backbone of this thesis work. Each chapter elaborates on specific aspects of these investigations, highlighting how various collaborations have contributed to advancing the understanding of complex biological conditions.

The figure 2.1 illustrates the interconnected workflows and outputs across three major research projects: APSIS, CRC, and Luxfico. Each project explores distinct yet complementary scientific questions, leveraging diverse datasets and analytical frameworks.

APSIS exploration in section 3.2.4 focused on allergic response modeling, while having microbiome data, the analysis focus was on cytokine profiles and immune cell types for subtyping and network-based analyses. The results emphasize key temporal and group-specific differences in allergic responses.

CRC research in section 4.2.2 utilizing Seurat workflows, this project investigates colorectal cancer, focusing on cellular stratification (fibroblasts and epithelial cells) and intercellular communication. It identifies specific inflammatory pathways (*e.g.* WNT5A and iCAF subtypes), which is then validated by mouse models and literature-based findings.

Luxfico investigation in chapter 5 this study delves into MO data integration, encompassing microbiome, WGS, and immune profiling. Outputs include SNP identification for food metabolism, microbiome signatures (aerobic vs anaerobic states), and predictive modeling for immune responses and inflammatory diet-related risks.

By integrating methodologies such as exploratory data analysis, unsupervised clustering, and pathway enrichment, these projects collectively push the boundaries of translational research, particularly in inflammation and immune system regulation. Each contribution enriches the MO and data science-driven insights central to this thesis.

Chapter 3

Mechanisms of Peanut allergy development in early childhood

3.1 Early-life events influence food allergy

Peanut Allergy (PA) represents a significant public health concern, with particularly high prevalence among pediatric population in Western countries. The epidemiological landscape of PAdemonstrates notable geographical and demographic variations, with prevalence estimates ranging from 2% to 5% among school-age children in the US [174]. Recent studies have documented prevalence rates higher in Western nations compared to Asian countries [175].

The diagnostic framework for PAhas multiple approaches. The Skin Prick Test (SPT) serves as an initial diagnostic tool, characterized by high sensitivity but limited specificity. A positive SPT result is defined by a wheal diameter exceeding the negative control by ≥3mm [176]. Specific IgE (sIgE) testing provides quantitative measurement of allergen-specific antibodies in serum, with values ranging from <0.35 to >100 kUA/L. A threshold of 15kUA/L for peanut-specific IgE demonstrates 95% positive predictive value for clinical reactivity [177].

Oral Food challenge (OFC) represents the gold standard diagnostic methodology for evaluating both IgE and non-IgE mediated food allergies, addressing the limitations inherent in conventional diagnostic approaches. The procedure uses a systematic protocol of graduated allergen exposure under controlled clinical conditions, typically requiring 2-3 hours for dose administration followed by an observation period [178].

The diagnostic utility of OFC extends beyond basic allergy confirmation, encompassing multiple clinical applications including the monitoring of allergy resolution, determination of

reaction thresholds, and assessment of cross-reactive potential in sensitized individuals. Recent research has established diagnostic parameters, with optimal testing thresholds identified at specific IgE levels ≤2 kU/L and SPT results ≤5 mm [179].

Clinical outcome data demonstrates favorable success rates across major allergens, with passage rates ranging from 62% to 88%. However, failed challenges present significant clinical concerns, with 84% of failures manifesting objective symptoms, 23% producing multisystemic reactions, and 15% necessitating epinephrine intervention. These safety considerations underscore the importance of proper clinical infrastructure, including experienced medical personnel, emergency equipment availability, and standardized protocols [179]. Recent advances in molecular diagnostics have introduced Component Resolved Diagnostics, enabling the quantification of IgE antibodies specific to individual peanut protein components. Among these, Ara h 2 demonstrates the highest predictive value for clinical reactivity, while Ara h 1, 3, 6, and 9 serve as secondary diagnostic indicators [176].

The clinical significance of PAis underscored by its association with severe allergic reactions. Studies indicate that 59.2% of peanut-allergic pediatric patients experience severe reactions [175], with approximately 10% of pediatric anaphylaxis cases occurring at school [180].

3.2 Methodological challenges in multi-omics integration

3.2.1 Study design considerations

This study was conducted as part of the APSIS project, which aimed to identify molecular signatures specific to peanut allergy in children. The study design dollowed a systematic investigation framework integrating multiple high-dimensional immunological datasets collected during OFC.

The investigation was executed through a collaboration between the Luxembourg Institute of Health (LIH) and the Luxembourg Center for Systems Biomedicine (LCSB). As referenced in Klueber et al. [181], this research builds upon previously published analyses of this cohort data.

The study has a longitudinal observational design with data collection at three distinct timepoints:

1. T0: Baseline measurements before OFC.

2. T1: Measurements at the onset of an allergic reaction.

3. T1 + 1 hour: Follow-up measurements one hour after the allergic reaction.

The investigation framework incorporated multiple analytical approaches:

Clinical assessment Comprehensive evaluation of patient history, allergic symptoms, and reactions patterns.

Biological profiling Analysis of: cytokines levels, microbiome composition, specific IgE measurements, Basophil Activation Test (BAT) responses, CyTOF cell type analysis.

The design specifically addressed the challenges of data integration across heterogeneous datasets while accounting for the relatively small sample size inherent in specialized allergy studies. This approach allowed for a comprehensive investigation of the immunological signatures associated with PA while maintaining rigorous scientific standards.

3.2.2 Participants selection and cohort characteristics

The study cohort comprised 35 children aged between 3 and 16 years. Participants were recruited through clinical referral based on suspected peanut allergy, as indicated by their medical history and positive specific IgE (sIgE) to peanut, confirmed either through SPT or serum analysis. As described [181], the participant selection process ensured a representative sample of pediatric peanut allergy cases.

For analytical purposes, participants were stratified into multiple comparison groups based on their clinical responses and characteristics. The primary stratification distinguished between patients who demonstrated allergic reactions during the oral food challenge (reactive group) and those who did not show clinical reactivity (non-reactive group). This classification was fundamental for subsequent immunological investigations and formed the basis for comparative analyses.

3.2.3 Data collection strategies

The data collection strategy was designed to capture both broad and specific aspects of the immune response to peanut allergen exposure. More technical details in [181].

Cytokines Assessed the inflammatory response profile, quantifying IFN-γ, IL-10, IL-13, IL-17A/F, IL-22, IL-33, IL-4, IL-5, IL-9, TNF-α.

Microbiome obtained from 16 allergic participants and 4 adult controls. Shotgun sequencing provided total count numbers and relative abundances of microbial communities, resolved to the species level where possible.

IgE profiling Specific IgE levels against 300 allergens were measured using the MADx platform.

BAT performed to assess cellular reactivity to peanut protein. The analysis measured CD63+ basophil release in response to stimulation with 100ng peanut protein, providing a functional assessment of cellular response capabilities. This test offered insights into the immediate hypersensitivity mechanisms involved in peanut allergy.

Cell types conducted using Cytometry by Time of Flight (CyTOF), enabling detailed characterization of cellular populations. The analysis recorded percentages of different cell types per participant across multiple timepoints, providing dynamic information about cellular immune responses throughout the challenge protocol.

The temporal organization of data collection followed a structured timeline with three distinct sampling points. At T0 (before OFC), samples were collected for cell type analysis, cytokines profiling, microbiome analysis, BAT, and clinical data. At T1 (onset of allergic reaction), samples were obtained for cell type analysis, cytokines measurement, and metabolite profiling. The final timepoint, T1 + 1 hour (one hour post-reaction), focused on cytokines measurements to track the resolution phase of the immune response.

Given the heterogeneous nature and low sample size of the datasets, the study emphasized the following points. 9 IDs matched across the datasets.

- Microbiome data which is available for 16 allergic participants and 4 controls.
- Cell types and metabolites data which are collected from 19 participants.
- Clinical data, MADx (allergens), and Cytokines which are available from 35 participants.

3.2.4 Immunological signature investigation

To investigate the complex interplay between immune cell populations and cytokines expression in peanut-allergic patients, we developed a systematic analytical framework incorporating multiple statistical approaches. The analysis focused on two distinct timepoints (T0 and T1) across 19 patients, stratified by their allergic response status (positive or negative reaction).

Data Preprocessing and Quality Control

Cytokine data underwent center-scaling normalization to achieve a mean of zero and unit variance, facilitating comparative analyses across different measurement scales. The immune cell population data, originally expressed as percentages of total cells, were retained in their native format as they already represented normalized proportions.

Statistical Framework

Our analytical approach had three complementary statistical methods:

Correlation Analysis We employed Spearman's rank correlation coefficient to assess relationships between immune parameters, chosen for its robustness to non-normal distributions and outliers. Statistical significance was evaluated at $\alpha = 0.05$, with correlation coefficients filtered for magnitude ($|p| > 0.6$) to focus on strong biological associations.

Bootstrap-Based Inference To address the challenges of small sample sizes, we implemented a bootstrap resampling strategy with 2,500 iterations, determined through stability analysis. This approach provided robust estimates of correlation coefficients and their confidence intervals. Correlations were considered reliable only when their 95% confidence intervals did not cross zero, implementing a conservative filter against potentially weak associations.

Comparative Analysis Framework We developed a structured comparative analysis to examine: Temporal changes within patient groups (T0 vs T1), Group differences at each timepoint (positive vs negative reactors), Differential correlation patterns across conditions using Δ ($Feature2 - Feature1$) and fold calculations ($Feature2/Feature1$).

Network Visualization of immune interactions

Significant correlations were visualized as network graphs, with nodes representing immune parameters and edges indicating correlation strength. Edge colors denoted positive (blue) or negative (red) correlations, with edge thickness proportional to correlation magnitude. This visualization approach facilitated the identification of immune parameter clusters and key regulatory relationships.

Software Implementation

All analyses were performed in R (version 4.3.1), utilizing specialized packages for correlation analysis (corrplot), network visualization (Cytoscape). Custom functions were developed for correlation filtering, Bootstrap calculations and comparative analyses, ensuring reproducibility and systematic evaluation across conditions.

3.3 Immune parameter integration reveals distinct correlation patterns

Our integrative analysis revealed distinct patterns of immune parameter correlations that varied both in time and between patient response groups. The correlation structures demonstrated notable differences between positive and negative groups, particularly showing significant relationships between immune cell types and cytokines parameters.

T0 - Positive patients shows Th2-oriented responses to allergy

Based on the correlation analysis at T0 for patients exhibiting positive reactions, several significant immunological relationships emerged, as visualized in both the correlation matrix heatmap and the network visualization figure 3.1 and figure 3.2. These complementary visualizations reveal complex immunoregulatory networks, with the heatmap quantifying correlation strengths ($|0.6|$, $p < 0.05$) and the network diagram illustrating the intricate cellular and molecular cross-talk (ranging from −1 in red to +1 in blue).

Several key immunological relationships emerged from this analysis. Strong coordination between innate immune components was evident, particularly between Classical Monocytes and mDC ($r = 0.902$), and between pDC and CM CD4+ T cells ($r = 0.909$). These relationships suggest robust innate-adaptive immune coordination.

The cytokines clustered revealed significant inter-correlations among IFN-γ, IL-5, and IL-22 (r ranging from 0.7 to 0.8), indicating concurrent activation of distinct T helper cell subsets. A notable antagonistic relationship emerged between Plasma Cells and IL-33 ($r = -0.860$), suggesting a potential regulatory mechanism in humoral immunity.

In the context of cellular immunity, we observed several significant relationships. Basophils demonstrated strong positive correlation with CD4+ Th2 cells ($r = 0.832$), supporting their established role in Type 2 immune responses. Conversely, IL-17A/F showed inverse

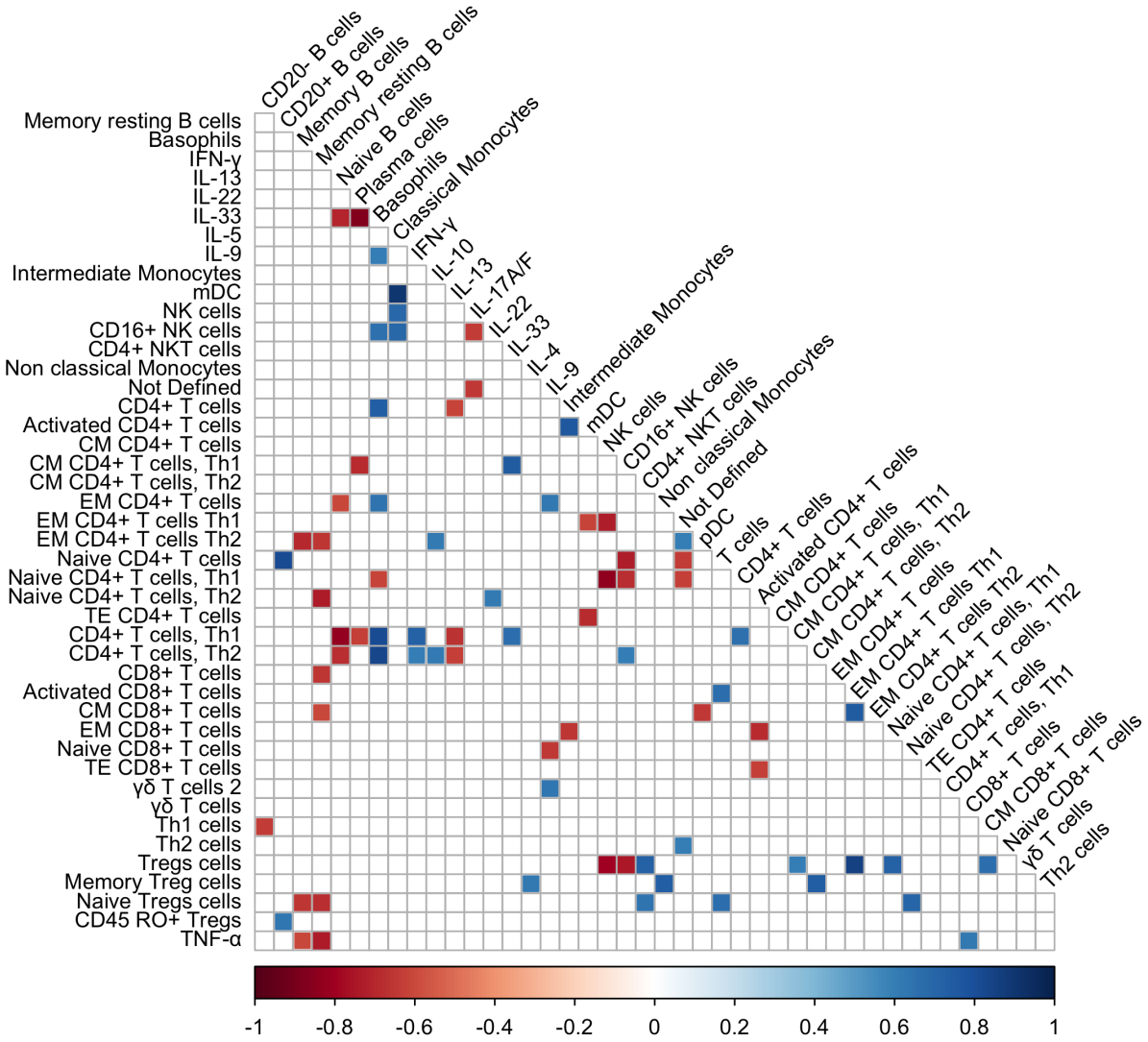


Figure 3.1 Correlation analysis of immune cell populations and cytokines in the "T0 Pos" patient group. The heatmap displays significant correlations between immune cell subsets and cytokines levels, highlighting key interactions within the immune system architecture. Positive (blue) and negative (red) correlations reveal potential immunoregulatory patterns, aiding in the identification of immune signatures associated with inflammatory responses.

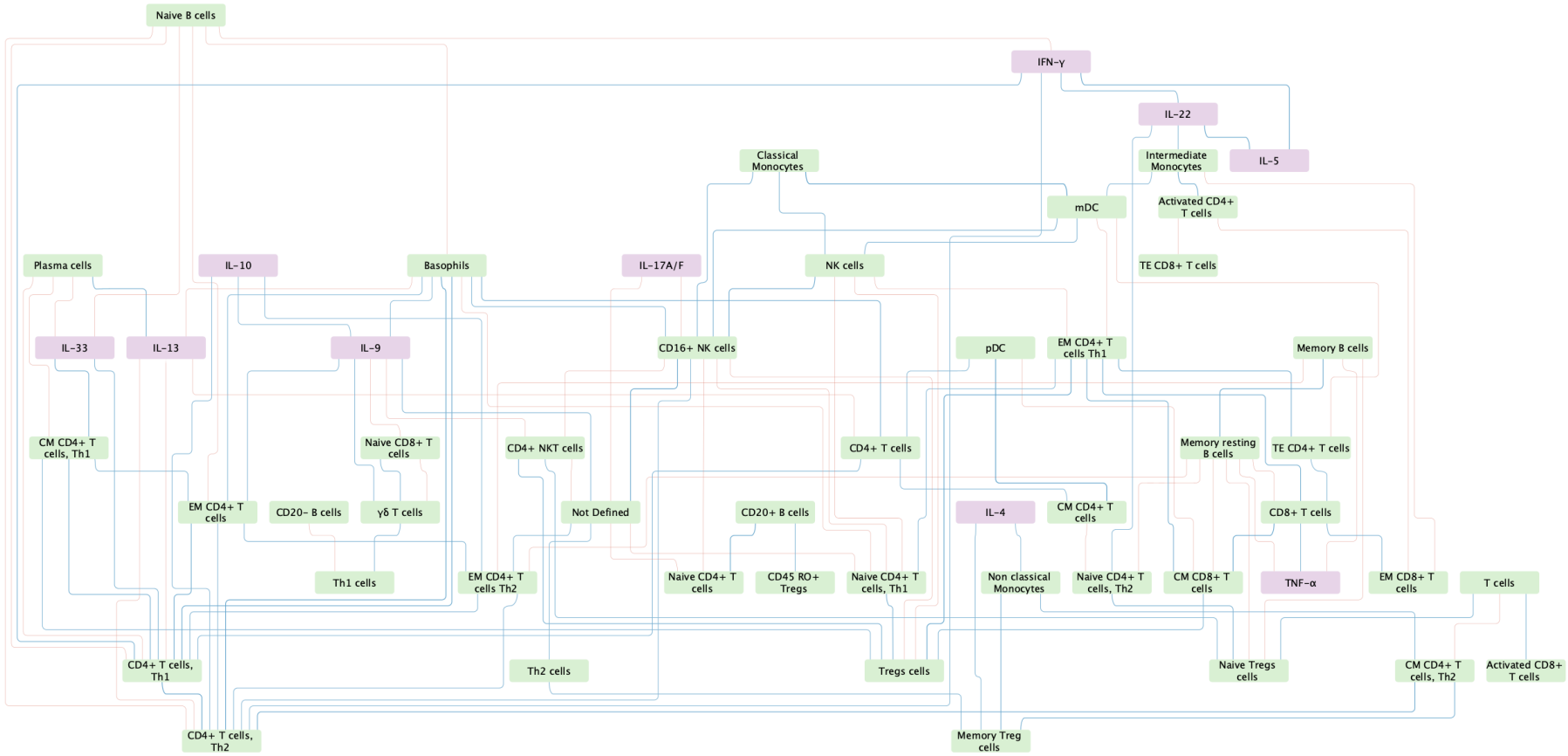


Figure 3.2 Network representation of significant correlations in the "T0 Pos" patient group. The graph visualizes key associations between immune cell subsets and cytokines, highlighting the complexity of the immunological landscape. Nodes represent immune components (immune cell types in green and cytokines in pink), while edges indicate significant correlations, with color-coded connections distinguishing positive (blue) and negative (red) interactions. This network-based approach provides insights into immune system dynamics and potential regulatory mechanisms in inflammation.

correlation with CD16+ NK cells (r = -0.637), suggesting potential cross-regulation between Th17-associated responses and cytotoxic NK cell activity. NK cells also displayed strong negative correlation with naive Th1 cells (r = -0.839), indicating possible reciprocal regulation.

Notably, we observed unexpected relationships that challenge classical paradigms. The negative correlation between EM CD4+ Th1 cells and regulatory T cells suggests complex co-regulatory mechanisms in immune homeostasis. Additionally, the correlation between Th1 and Th2 responses challenges the traditional Th1/Th2 dichotomy, potentially indicating a more nuanced immune response than previously appreciated.

These findings provide insights into both established immune mechanisms and novel relationships that warrant further investigation. The strong correlations observed between monocyte/mDC and pDC/CM CD4+ T cell populations may indicate promising targets to validate in the further analysis steps.

T1 - Positive patients exhibit enhanced Th2-mediated immune network activation

Examination of the correlation patterns at the T1 timepoint revealed changes in the relationships between immune cell populations compared to the baseline T0 state. As shown in figure 3.3 and figure 3.4, the correlation matrix demonstrates a significant reconfiguration of cellular relationships, characterized by an elaborate and highly interconnected immune network typical for an active allergic response.

The T cell compartment exhibited strong interactions, with EM CD4+ T cells showing high positive correlations with both Th1 cells (r = 0.958) and Th2 cells (r = 0.923). Th1 and Th2 cells displayed strong positive correlation (r = 0.888), challenging the classical paradigm of Th1/Th2 similarly as the T0 observation.

In the innate immune compartment, classical Monocytes maintained strong correlation with mDC (r = 0.860), indicating preserved innate immune coordination. Basophils demonstrated significant positive correlations with both Th2 cells (r = 0.867) and pDC (r = 0.853), further supporting their role in Th2-polarized immune responses and suggesting potential cross-talk with dendritic cells.

The cytokine network analysis revealed several unexpected relationships. IL-13, typically associated with Th2 responses, showed strong negative correlation with Th2 cells (r = -0.870), potentially indicating a novel feedback inhibition mechanism or context-specific regulation. Similarly, IL-33 exhibited strong negative correlation with $\gamma\delta$T cells (r = -0.867). The anti-inflammatory cytokines IL-10 and IL-22 demonstrated positive correlation (r = 0.748),

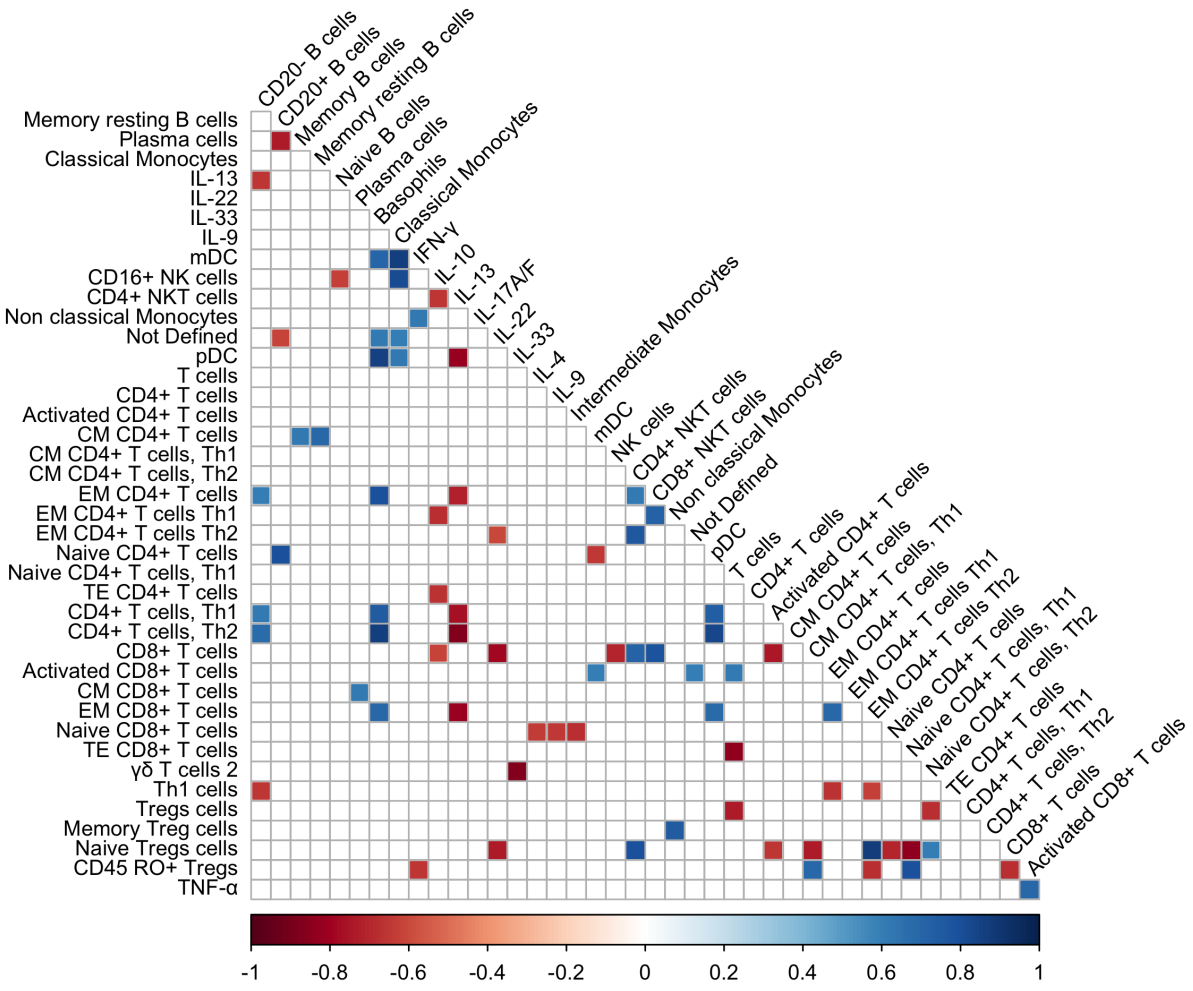


Figure 3.3 Correlation analysis of immune cell populations and cytokines in the "T1 Pos" patient group. The heatmap displays significant correlations between immune cell subsets and cytokines levels, highlighting key interactions within the immune system architecture. Positive (blue) and negative (red) correlations reveal potential immunoregulatory patterns, aiding in the identification of immune signatures associated with inflammatory responses.

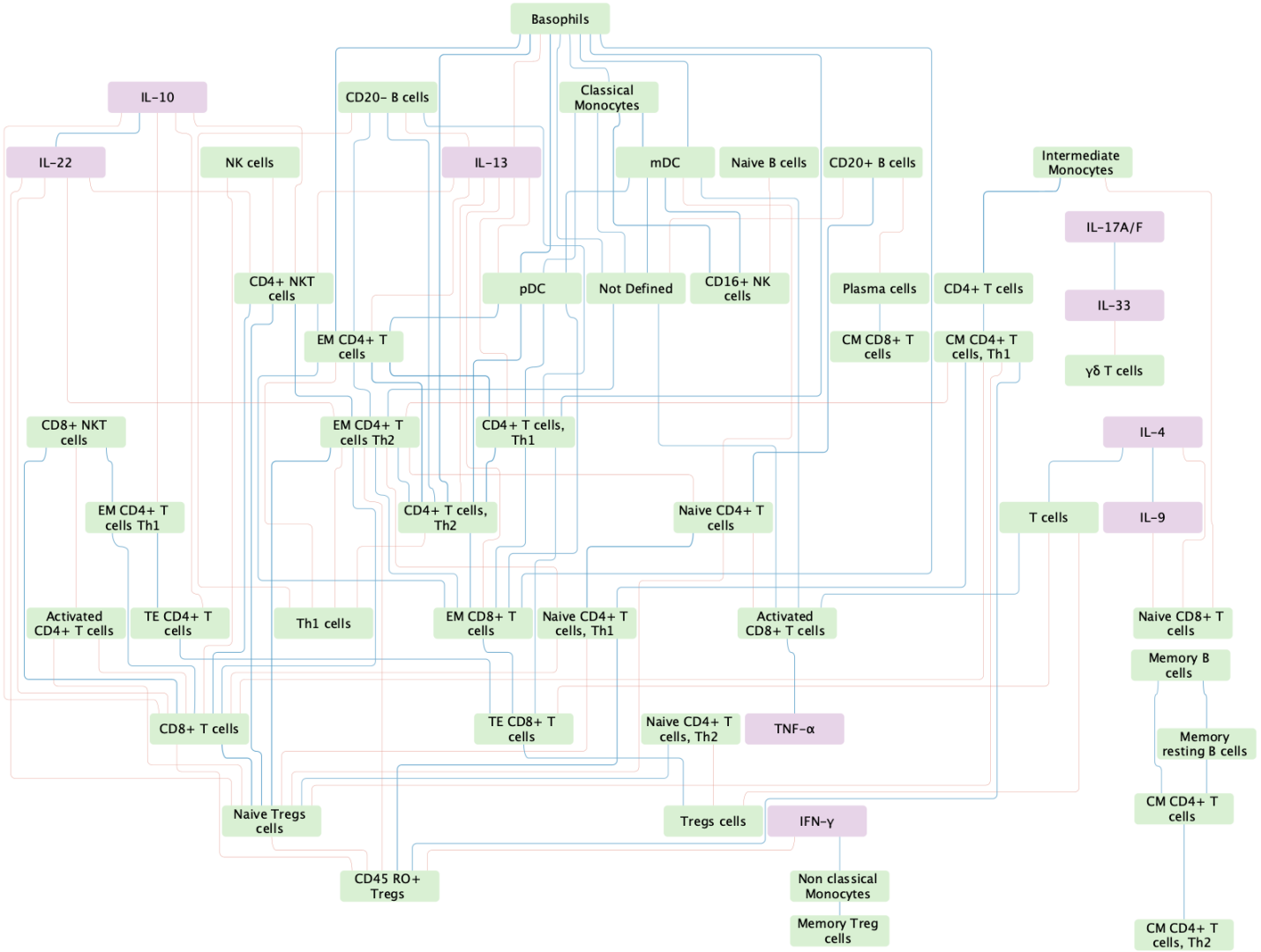


Figure 3.4 Network representation of significant correlations in the "T1 Pos" patient group. The graph visualizes key associations between immune cell subsets and cytokines, highlighting the complexity of the immunological landscape. Nodes represent immune components (immune cell types in green and cytokines in pink), while edges indicate significant correlations, with color-coded connections distinguishing positive (blue) and negative (red) interactions. This network-based approach provides insights into immune system dynamics and potential regulatory mechanisms in inflammation.

suggesting coordinated immunoregulatory activity.

Analysis of regulatory T cell dynamics revealed complex patterns of immune homeostasis. Naive Tregs showed positive correlation with EM CD4+ Th2 cells (r = 0.874) but negative correlation with Naive CD4+ Th1 cells (r = -0.832), suggesting regulation of Th1/Th2 responses. An unexpected negative correlation emerged between T cells and TE CD8+ T cells (r = -0.833), potentially indicating regulatory exhaustion or compartmentalization of the cellular immune response.

These findings reveal novel interactions within the immune network at T1, particularly highlighting the complexity of T cell subset relationships, cytokines-mediated regulation, and innate immune cell cooperation. The paradoxical correlations observed between Th1-Th2 cells and IL-13-Th2 cells warrant further mechanistic investigation, as they challenge current understanding of allergic response regulation. The maintained strong correlations between innate immune components (monocytes/mDC) alongside enhanced T cell subset coordination suggest a highly orchestrated immune response with potential therapeutic implications.

T0 - Negative patients demonstrate balanced immune regulation

Analysis of the correlation patterns in T0 negative patients (non-reactors) revealed distinctive network characteristics that contrasted with those observed in T0 positive patients. The network structure demonstrated sparser network with reduced Th2-associated connectivity, suggesting a more balanced immune environment prior to OFC.

In the cellular immune compartment, basophils demonstrated complex interactions with T cell subsets. A remarkably strong positive correlation emerged between basophils and EM CD4+ T cells (r = 0.976), suggesting potential basophil involvement in EM T cell maintenance or function. Conversely, basophils showed equally strong negative correlation with Naive CD4+ Th1 cells (r = -0.976), indicating potential regulatory cross-talk that may influence naive Th1 cell differentiation pathways.

The Th1 compartment exhibited distinct coordination patterns. CM CD4+ Th1 cells and Naive CD4+ Th1 cells demonstrated strong positive correlation (r = 0.952), suggesting synchronized activity or shared differentiation pathways within Th1 subsets. However, IFN-γ showed negative correlation with CM CD4+ Th1 cells (r = -0.857), potentially reflecting differential cytokines production profiles between central memory and effector Th1 populations.

The focus on cytokines revealed several significant relationships characteristic of balanced immune regulation. IL-22 and IL-33 exhibited strong positive correlation (r = 0.874), consistent

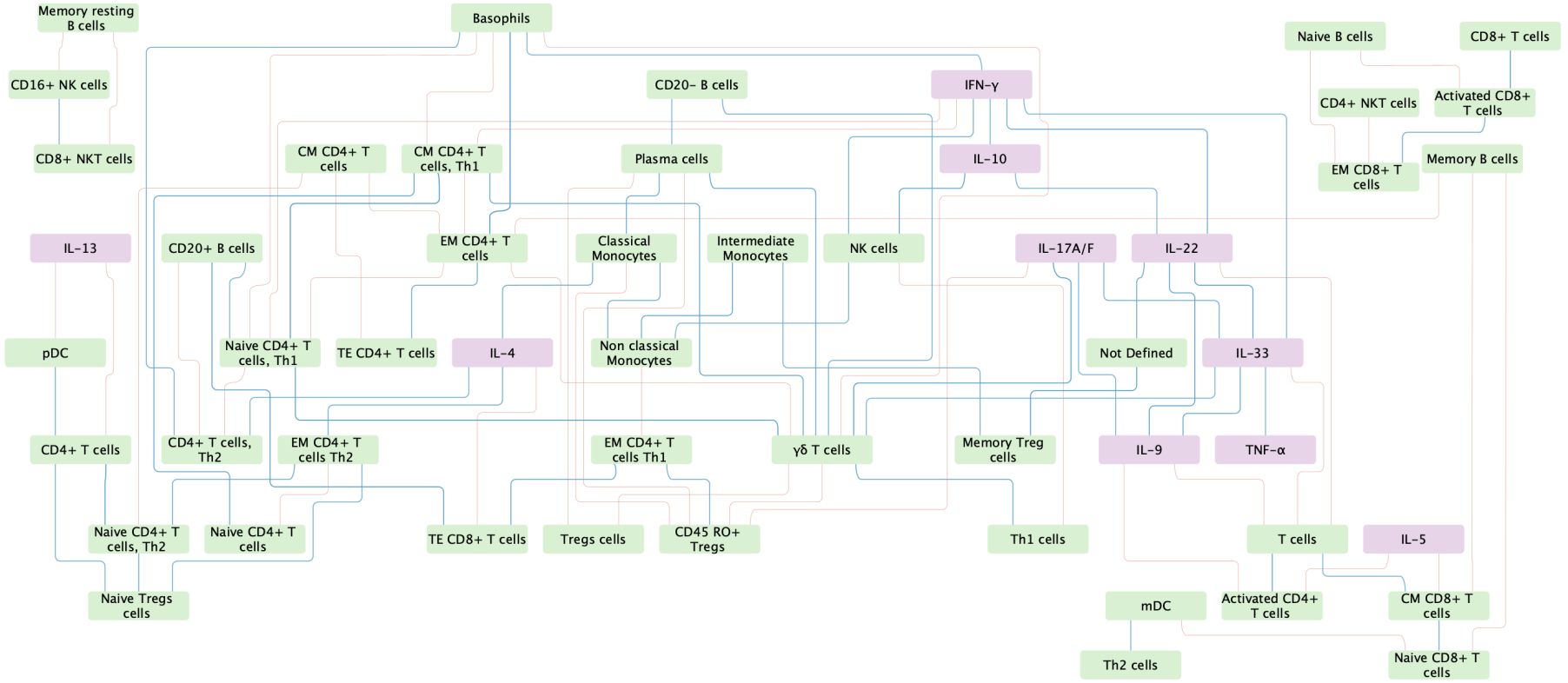


Figure 3.5 Network representation of significant correlations in the "T0 Neg" patient group. The graph visualizes key associations between immune cell subsets and cytokines, highlighting the complexity of the immunological landscape. Nodes represent immune components (immune cell types in green and cytokines in pink), while edges indicate significant correlations, with color-coded connections distinguishing positive (blue) and negative (red) interactions. This network-based approach provides insights into immune system dynamics and potential regulatory mechanisms in inflammation.

with their complementary roles in maintaining mucosal immunity and tissue homeostasis. IL-17A/F showed strong positive correlation with $\gamma\delta$ T cells (r = 0.905), aligning with the established role of $\gamma\delta$ T cells as primary IL-17 producers in inflammatory responses. Additionally, IL-33 and IL-9 demonstrated positive correlation (r = 0.826), suggesting potential synergistic activity in type 2 immune responses.

Analysis of immunoregulatory relationships revealed notable patterns. Plasma cells showed strong negative correlation with Tregs (r = -0.905), suggesting potential counterbalance between humoral immunity and immunosuppression. IL-9 exhibited negative correlation with T cells (r = -0.905), indicating possible inhibitory mechanisms in T cell responses.

Cell-cell interactions revealed additional regulatory complexity. Memory B cells demonstrated strong negative correlation with CM CD8+ T cells (r = -0.929), suggesting distinct regulation of memory responses in humoral and cellular immunity. $\gamma\delta$ T cells emerged as central coordinators, showing multiple strong associations, including negative correlation with EM CD4+ T cells and positive correlation with IL-17A/F, highlighting their diverse regulatory functions.

These findings illuminate a complex network of cellular and molecular interactions characterizing the immune status of negative responders. The observed patterns suggest tightly regulated immune homeostasis, with balanced activity between innate cells (basophils, $\gamma\delta$T cells), adaptive immune subsets (Th1 cells, Tregs), and their associated cytokines networks. This balanced state may contribute to the absence of allergic responses in these individuals.

T1 - Negative patients maintain regulatory network dominance post-challenge

Analysis of the T1 negative cohort revealed a distinctive immunological profile characterized by maintained network stability and enhanced regulatory features following OFC. Unlike the other T1 positive network, these patients demonstrated a preserved network architecture without developing extensive Th2-associated correlations, suggesting effective immune regulation during allergen exposure.

The T helper cell compartment exhibited coordination patterns, particularly within Th1 subsets. CM CD4+ Th1 cells and Naive CD4+ Th1 cells showed strong positive correlation (r = 0.976), suggesting synchronized differentiation pathways or functional synergy in Th1-mediated responses. This Th1 axis was further reinforced by significant positive correlation between CD20+ B cells and CM CD4+ Th1 cells (r = 0.929), indicating potential implication in antibody class switching or cell-mediated immune coordination.

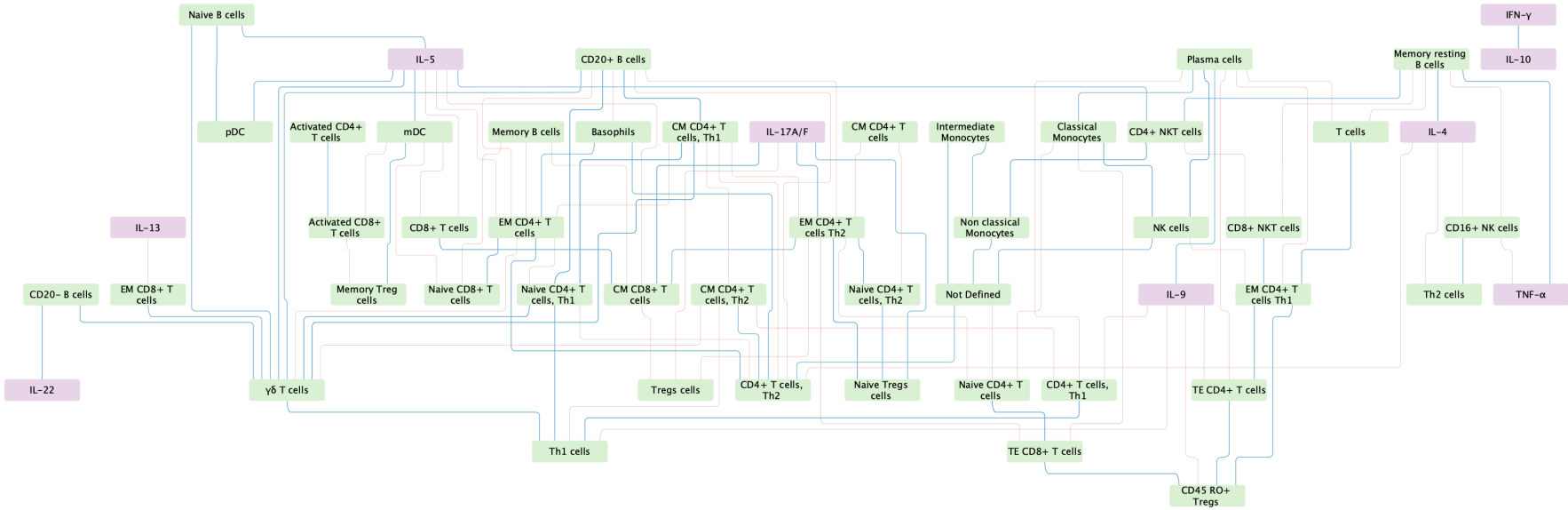


Figure 3.6 Network representation of significant correlations in the "T1 Neg" patient group. The graph visualizes key associations between immune cell subsets and cytokines, highlighting the complexity of the immunological landscape. Nodes represent immune components (immune cell types in green and cytokines in pink), while edges indicate significant correlations, with color-coded connections distinguishing positive (blue) and negative (red) interactions. This network-based approach provides insights into immune system dynamics and potential regulatory mechanisms in inflammation.

Several significant negative correlations emerged, highlighting important regulatory mechanisms. mDC demonstrated strong negative correlation with Activated CD8+ T cells (r = -0.952), suggesting potential mDC-mediated regulation of cytotoxic T cell responses. The relationship between IL-17A/F and Tregs showed strong negative correlation (r = -0.929), reflecting critical balance between pro-inflammatory and immunosuppressive mechanisms. Similarly, IL-5 exhibited negative correlation with CD8+ T cells (r = -0.905), suggesting potential suppression of cytotoxic T cell activity in Th2-skewed environments.

The cytokines focus revealed several unexpected relationships that suggest complex regulatory mechanisms. IFN-ɣ and IL-10 showed strong positive correlation (r = 0.905), an unusual association between pro- and anti-inflammatory mediators that may indicate sophisticated feedback regulation of inflammatory responses. IL-4, despite its canonical role in Th2 responses, demonstrated negative correlation with Th2 cells (r = -0.805), potentially reflecting negative feedback mechanisms or context-specific regulation. IL-9 showed negative correlation with Th1 cells (r = -0.786), supporting its role in antagonizing Th1-mediated responses.

Notable cell-cell interactions included strong positive correlation between Plasma cells and NK cells (r = 0.880), suggesting potential synergy between humoral and innate cytotoxic responses. The negative correlation between IL-5 and CD8+ T cells (r = -0.904) further supported the hypothesis that Th2-associated environments may suppress cytotoxic T cell activity.

These findings reveal complex regulatory networks characterized by intricate balance between innate and adaptive immune components. The maintenance of strong Th1-associated correlations, coupled with evidence of active immunoregulatory mechanisms, suggests that negative responders maintain effective immune homeostasis during allergen challenge. The observed patterns of cytokines-mediated regulation and cellular cross-talk provide insights into mechanisms that may prevent development of allergic responses in these individuals.

Correlation changes highlight immune responses shifts

To better understand how immune interactions evolve between different clinical groups and at distinct timepoints, we generated and compared multiple tables of correlation coefficients.

The table 3.1 highlights the most pronounced positive or negative shifts in correlation values, while the table 3.2 focuses on fold changes, thereby helping distinguish immunological relationships that are either strengthened or weakened in particular groups over time. While

		Pos		Neg		Δ Score				
Cluster 1	Cluster 2	T0	T1	T0	T1	Pos	Neg	T0	T1	Max
Basophils	EM CD4+ T cells	0.64	0.80	0.98	0.74	0.15	-0.24	-0.33	0.06	0.33
Basophils	Naive CD4+ T cells, Th1	-0.61		-0.98				0.36		0.36
IL-13	Th2 cells	-0.62	-0.87			-0.24				0.24
IL-17A/F	γδ T cells	-0.58		0.91				-1.48		1.48
IL-4	Th2 cells			0.73	-0.73		-1.46			1.46
mDC	Activated CD8+ T cells		0.61		-0.95				1.56	1.56
CM CD4+ T cells, Th1	EM CD4+ T cells	0.65		-0.88	-0.86		0.02	1.53		1.53
CM CD4+ T cells, Th1	Naive CD4+ T cells, Th1		0.70	0.95	0.98		0.02		-0.28	0.28
EM CD4+ T cells	Th2 cells	0.62	0.92		0.81	0.30			0.11	0.30
Naive CD4+ T cells, Th2	Naive Tregs cells	0.71	0.60	0.81	0.81	-0.10		-0.10	-0.21	0.21

Table 3.1 Distinct immune cell and cytokines correlation patterns across conditions. The table presents filtered Δ correlation scores, highlighting shifts in immune cell interactions between conditions (T0, T1). Positive and negative correlation clusters reveal distinct immunological signatures, with notable changes in T-cell subsets, cytokines, and antigen-presenting cells. These findings contribute to identifying immune modulation dynamics and potential biomarkers in inflammation-related processes.

Comparison		Pos		Neg		Fold Score			
Cluster 1	Cluster 2	T0	T1	T0	T1	Pos	Neg	T0	T1
CD20+ B cells	Naive CD4+ T cells, Th1			0.74	0.91		1.23		
Memory B cells	CM CD8+ T cells			-0.93	-0.74		0.80		
Basophils	EM CD4+ T cells	0.64	0.80	0.98	0.74	1.24	0.76	0.66	1.08
Basophils	Naive CD4+ T cells, Th1	-0.61		-0.98				0.63	
IFN-γ	IL-10	0.59		0.76	0.91		1.19	0.78	
IL-13	Th2 cells	-0.62	-0.87			1.39			
IL-17A/F	$\gamma\delta$ T cells	-0.58		0.91				-0.64	
IL-4	Th2 cells			0.73	-0.73		-1.00		
mDC	Activated CD8+ T cells		0.61		-0.95				-0.64
NK cells	EM CD4+ T cells Th1	-0.72	-0.59		-0.76	0.82			0.78
CD4+ NKT cells	Naive Tregs cells	0.64	0.79			1.23			
Not Defined	Naive CD4+ T cells	-0.64	-0.59		-0.76	0.92			0.77
pDC	CD4+ T cells	0.61		0.79				0.77	
CM CD4+ T cells, Th1	EM CD4+ T cells	0.65		-0.88	-0.86		0.97	-0.74	
CM CD4+ T cells, Th1	Naive CD4+ T cells, Th1		0.70	0.95	0.98		1.02		0.72
EM CD4+ T cells	Th1 cells	0.78	0.96			1.23			
EM CD4+ T cells	Th2 cells	0.62	0.92		0.81	1.48			1.14
Naive CD4+ T cells, Th2	Naive Tregs cells	0.71	0.60	0.81	0.81	0.85	1.00	0.87	0.74
$\gamma\delta$ T cells	Th1 cells	0.64		0.81	0.81		1.00	0.80	

Table 3.2 Fold change analysis of correlation scores highlighting group-specific immune dynamics. The table displays fold changes in correlation scores between conditions (T0, T1) and groups for immune cell subsets and cytokines. These results provide insights into differential immune regulation and potential biomarkers associated with inflammatory responses to the allergy.

the comprehensive interpretation of these findings necessitates specialized immunological expertise, my role as a bioinformatician involves also to conducting systematic analyses deciphering the complexity of these relationships. I have investigated patterns from the target list, proposing my understanding of the intricate immunological networks complexity and their temporal dynamics revealed by these results.

The analysis revealed dynamic shifts in immune cell-cytokines interactions, marked by significant Δ between time points T0 and T1. Notably, IL-17A/F and $\gamma\delta$ T cells exhibited the largest disruption (Δ = 1.48), reflecting a sharp decline in negative correlation strength. A complete polarity reversal was observed for IL-4 and Th2 cells (Δ = 1.46), suggesting altered regulatory dynamics. Emergent interactions, such as those between mDC and Activated CD8+ T cells (Δ = 1.56), highlighted new associations at T1, while pairs like Basophils and EM CD4+ T cells demonstrated moderate but consistent strengthening of positive correlations (Δ = 0.33). These findings underscore the plasticity of immune cell networks under the studied conditions, with Δ scores serving as critical indicators of biologically relevant shifts in interaction patterns.

Beyond looking at raw changes in correlation coefficients (such as the simple difference from T0 to T1), we evaluated "fold changes" to highlight relationships whose magnitude of shift may be noteworthy, see table 3.2 By combining these two metrics (Δr from table 3.1 and the fold change from table 3.2), we can better distinguish between modest alterations in correlation and those that may reflect more biologically significant re-wiring of immune cell networks.

The focus on fold change reveals dynamic temporal shifts in immune cell-cytokine interactions, characterized by distinct correlation fold scores between time points T0 and T1. Notably, Basophils exhibited strengthened positive correlations with EM CD4+ T cells (T0: 0.64; T1: 0.80) alongside a significant fold score increase (Pos: 1.24), suggesting early and sustained pro-inflammatory signaling. Th2-associated cytokines displayed divergent regulation: IL-13 showed amplified positive correlation with Th2 cells (fold score: 1.39), while IL-4 demonstrated a marked reversal (fold score: −1.00), highlighting nuanced regulatory roles. Persistent negative correlations between Memory B cells and CM $CD8^+$ T cells (T0: −0.93; T1: −0.74) underscored antagonistic cross-talk. These findings emphasize Th1/Th2 axis plasticity and time-resolved immune coordination, providing mechanistic insights into group-specific inflammatory and regulatory pathways.

3.4 Decoding immune system interactions in allergy reaction

Methodological considerations and analysis limitations

The correlation analysis presented in this study represents a focused examination of immune system interactions in the context of peanut allergy, specifically highlighting statistically robust relationships identified through stringent filtering criteria ($r > 0.6$, $p < 0.05$). While this approach enables the identification of strong pairwise relationships, it is essential to acknowledge several methodological considerations and limitations inherent to this analytical framework. First, the complexity of the immune system extends far beyond pairwise interactions. The current analysis captures only direct correlational relationships between parameters, potentially overlooking higher-order interactions and complex regulatory networks. This limitation is particularly relevant in immunology, where cellular and molecular interactions often involve multiple mediators and feedback loops operating simultaneously. The statistical power of our analysis is constrained by the cohort size, a common limitation in clinical immunology studies. While bootstrap analysis helps mitigate the impact of outliers and provides more robust correlation estimates, as evidenced in table 3.3, the relatively small sample size may still influence the detection of weaker but biologically relevant correlations. This limitation particularly affects the interpretation of temporal changes between T0 and T1 timepoints, where individual variations might have substantial impact on correlation coefficients. The volume of results generated, even after applying stringent filtering criteria, necessitated selective presentation focused on the most statistically significant relationships. The visualization through heatmaps and network diagrams, while informative, represents a curated view of the immune system's complexity. This selective approach, while necessary for interpretability, may inadvertently obscure broader patterns or subtle interactions that could be biologically relevant.

Integration with primary analysis and future directions

The correlation patterns identified in this study should be interpreted in conjunction with the primary analysis [181] of pairwise comparisons between patient groups. This integrative perspective is crucial for several reasons. The coregulation patterns observed through correlation analysis may provide mechanistic insights into the differential expression patterns identified in the primary analysis. However, the relationship between these two analytical approaches

Comparison		r score	
Cluster 1	Cluster 2	"Classic"	Bootstrap
CD20- B cells	Th1 cells	-0.64	-0.60
CD20+ B cells	Naive CD4+ T cells	0.81	0.78
CD20+ B cells	CD45 RO+ Tregs	0.64	–
Memory B cells	Memory resting B cells	0.79	0.75
Memory B cells	EM CD4+ T cells, Th2	-0.69	-0.64
Memory B cells	Naive Treg cells	-0.66	–
Memory B cells	TNF-α	-0.60	–
Memory resting B cells	EM CD4+ T cells, Th2	-0.66	-0.63
Memory resting B cells	Naive CD4+ T cells, Th2	-0.73	-0.70
Memory resting B cells	CM CD8+ T cells	-0.65	-0.61
Memory resting B cells	CM CD8+ T cells	-0.61	–
Memory resting B cells	Naive Treg cells	-0.68	-0.66
Memory resting B cells	TNF-α	-0.73	-0.70
Naive B cells	Basophils	-0.61	–
Naive B cells	IFN-γ	-0.61	–
Naive B cells	IL-33	-0.71	-0.68
Naive B cells	EM CD4+ T cells	-0.61	–
Naive B cells	Naive CD4+ T cells, Th1	-0.82	-0.80
Naive B cells	Naive CD4+ T cells, Th2	-0.68	-0.64
Plasma cells	IL-13	0.73	–

Table 3.3 Comparative correlation analysis using classical and bootstrap approaches. The table presents the top 20 correlation pairs between immune cell subsets and cytokines, comparing results from the classical correlation method and bootstrap validation. Strong positive and negative correlations highlight key immunological interactions, with bootstrap analysis providing robustness checks and complementary validation to classical filtering for the interpretation of the targets.

is not always straightforward, necessitating careful interpretation and validation. Several complementary in silico approaches could enhance our understanding of the observed correlations: Advanced computational approaches could theoretically complement our correlation analysis to better understand the complex immune system interactions. However, the limited sample size of our cohort ($n < 30$) poses significant constraints on the application of such methods. While traditional pathway analysis tools and databases are limited in the context of allergy research due to the scarcity of curated allergy-specific pathways, the sample size further restricts our analytical options. Advanced dimensionality reduction techniques (*e.g.* t-SNE, UMAP) or clustering methods, while powerful for larger datasets, may not provide reliable insights with our current sample size as they typically require larger populations to establish stable, reproducible patterns. This limitation particularly affects our ability to identify robust coordinated patterns of immune responses beyond pairwise correlations.

Given these constraints, our focus on carefully filtered correlation analysis, supported by bootstrap validation for interpretation, represents a simple but robust analytical strategy for the available data. Future studies with larger cohorts would be needed to effectively employ more sophisticated computational approaches for understanding how different immune cell populations and their associated cytokines form functional modules during allergic responses. The validation of these findings would ideally require an independent cohort with similar immunological parameters. In the absence of such validation data, careful literature investigation becomes crucial for contextualizing our findings within the broader framework of allergic responses. The bootstrap analysis provides some validation by identifying correlations that are robust to sampling variation, but external validation remains important for confirming the biological significance of these findings. Network-based analytical approaches, while traditionally developed for gene expression analysis, could be adapted for immunological parameters with careful consideration of the biological context. However, such adaptations would require methodological innovations to account for the distinct nature of immunological data, where the relationships between parameters may not follow the same principles as gene co-expression networks.

Biological complexity and expert knowledge integration

The interpretation of immunological correlations requires careful consideration of biological context and expert knowledge. While bioinformatic analysis can identify statistically significant relationships, the biological relevance of these correlations must be evaluated within the framework of known immunological mechanisms and potential novel interactions. The networks and heatmaps presented in this study suggest complex patterns of immune regulation that extend beyond simple linear relationships. The simultaneous activation of multiple T helper subsets, as observed in our analysis, challenges traditional paradigms of T helper cell differentiation and highlights the need for more sophisticated models of immune regulation. The involvement of innate immune cells, particularly dendritic cells and basophils, in coordinating adaptive immune responses aligns with current understanding of allergic mechanisms. However, the specific patterns of correlation observed in our study suggest potentially novel regulatory relationships that warrant further investigation through targeted experimental approaches. The bootstrap analysis provides additional confidence in certain correlations, particularly those maintaining significance across resampled datasets. However, the biological interpretation of these relationships requires integration of immunological

expertise to distinguish between correlations that represent known mechanisms and those that might indicate novel regulatory pathways.

Literature investigation for presented results

The following discussion attempts to interpret the findings presented in the results section from a bioinformatics and data science perspective, with particular emphasis on the methodological approaches employed to analyze this complex dataset. As a bioinformatician, my expertise focus in developing and implementing computational strategies to explore large-scale biological data, rather than in specialized immunological interpretation. While I do not possess the clinical expertise to definitively assess the biological significance of these findings, I have focused on establishing an comprehensive analytical workflows and applying appropriate statistical methods to extract meaningful patterns from the data. This methodological foundation allows for systematic exploration of the results, while acknowledging that the biological interpretations offered here should be considered as data-driven hypotheses requiring validation by domain experts.

For T0 Pos, the immunological relationships observed both corroborate and extend current understanding of immune network dynamics. The strong coordination between classical monocytes and mDC aligns with their established role in antigen presentation and T cell priming [182, 183], while the pDC/CM CD4+ T cell axis may reflect their combined role in maintaining antiviral immunity [182]. The cytokines triad of IFN-γ, IL-5 and IL-22 suggests concurrent Th1/Th2/Th22 activation, consistent with recent evidence of helper T cell plasticity in inflammatory contexts [184, 185]. The inverse Plasma Cell/IL-33 relationship echoes findings that IL-33 can suppress humoral responses through BAFF regulation [186], while the Basophil-Th2 correlation supports their recognized role in type 2 immunity [187]. Notably, the observed Th1/Th2 co-activation challenges classical models that Th1 cells would counterbalance and protect against Th2-mediated allergic responses, potentially reflecting microenvironment-specific immune adaptations. These network interactions highlight previously underappreciated crosstalk between innate sensors and adaptive effectors, particularly the regulatory interplay between NK cell subsets and T helper populations [188], suggesting new layers of immune homeostasis control that need further mechanistic investigation.

Further analysis on T1 Pos and the immune network dynamics reveals additional complex interactions that both challenge and refine classical immunological paradigms. The observed immune network dynamics reveal complex interactions that challenge classical immunological

paradigms while suggesting novel regulatory mechanisms. The strong positive correlations between EM CD4+ T cells and both Th1/Th2 subsets, coupled with direct Th1-Th2 coordination, suggest these lineages may cooperate rather than antagonize in certain contexts, potentially through transcriptional regulators like STAT5/GATA3 [189, 190]. This aligns with emerging evidence of hybrid Th1/Th2 states in chronic inflammation [191]. The paradoxical negative correlation between IL-13 and Th2 cells mirrors findings in cutaneous leishmaniasis where IL-13 dominates over IL-4 in suppressing Th1 responses, potentially indicating a compensatory Th2 cell feedback mechanism. Similarly, IL-33's negative association with $\gamma\delta$ T cells supports recent evidence of IL-33-mediated Treg control over innate lymphocyte activation [192]. The coordinated anti-inflammatory activity of IL-10/IL-22 reflects their shared STAT3 signaling but highlights underappreciated pro-inflammatory potential in specific microenvironments [193, 194]. Regulatory T cell dynamics, particularly naive Treg-Th2 coordination versus Th1 opposition, align with models of Treg specialization for subset-specific suppression [195, 196], while the T cell-TE CD8+ negative correlation may reflect exhaustion mechanisms observed in chronic antigen exposure [197]. These findings collectively suggest an integrated immune architecture where innate-adaptive crosstalk (monocyte-mDC) and cytokines feedback loops maintain homeostasis through non-canonical pathways, warranting mechanistic studies to elucidate the cellular circuits underlying these network-level phenomena.

The focus on Δ analysis revealed dynamic temporal shifts in immune cell-cytokines network interactions, with $\gamma\delta$ T cells and IL-17A/F exhibiting high Δ score, consistent with their established role in bridging innate/adaptive immunity and driving neutrophilic inflammation through IL-17-mediated epithelial activation [198, 199]. The complete polarity reversal between IL-4 and Th2 cells aligns with GATA3-mediated transcriptional regulation of Th2 plasticity and IL-4's capacity to modulate IgE class-switching in B cells [200, 201]. Emergent interactions like mDC-Activated CD8+ T cell associations may reflect IL-3/IL-4 crosstalk enhancing CD8+ effector functions, as demonstrated in basophil-T cell cooperation models [202]. While moderate strengthening of basophil-EM CD4+ T cell correlations could indicate $\gamma\delta$ T cell-mediated IL-4 production influencing Th2 polarization [199, 203]. These Δ-quantified interactions underscore the context-dependent plasticity of immune coordination, where cytokines-mediated cellular crosstalk dynamically reshapes effector responses.

Studies provide insights for the observed temporal dynamics highlighted in by the fold change comparison. The strengthened basophil–EM CD4+ T cell correlation aligns with evidence that basophils coordinate early Th2 polarization and Th17/Th1 cytokines expression,

facilitating sustained pro-inflammatory signaling [204]. The divergent regulation of IL-13 (amplified) versus IL-4 (reversed) reflects functional specialization within Th2 pathways: IL-13 primarily drives alternative macrophage activation and tissue repair, while IL-4 shows tighter regulatory control through negative feedback loops that prevent excessive IgE responses [205]. This dichotomy is further reinforced by compensation mechanisms in IL-4/IL-13 double-knockout models, where residual Th1 activity persists through IL-5/eosinophil axes [206]. The persistent negative correlation between memory B cells and CM CD8+ T cells reflect on findings that B cell depletion impairs CD8+ T cells memory formation, suggesting competition for survival signals like IL-15 or TNF-α in lymphoid niches [207]. Notably, the Th1/Th2 plasticity observed in fold-change dynamics resonates with recent advances that have challenged the traditional Th1/Th2 paradigm, revealing a more nuanced understanding of CD4+ T cell biology. Research has demonstrated that helper T cells exhibit remarkable plasticity, with the ability to dynamically transition between Th1 and Th2 states through complex transcriptional network modifications [208]. This flexibility is further evidenced by the identification of hybrid phenotypes, where allergen-specific T cells simultaneously express both Th1-associated IFN-γ and Th2-associated IL-4/IL-13 cytokines [208].

In conclusion, this study conducted a systematic literature analysis of targets identified from the Positive cohort at both baseline (T0) and allergy manifestation (T1) timepoints. The integration of delta (Δ) analysis and fold change metrics enabled the prioritization of candidates from an extensive dataset into a manageable subset of targets. This computational approach provides a foundation for future experimental validation. The ongoing collaboration will facilitate the validation of these predicted targets through appropriate in vitro or in vivo models. These findings warrant further investigation into the molecular mechanisms underlying the observed changes in the Positive cohort.

Contributions

This work was conducted under the primary supervision of Prof. Dr. Reinhard Schneider and Dr. Venkata Satagopam at the Luxembourg Center for Systems Biomedicine (LCSB). I, Mathias Galati, developed the methodological framework, performed the computational analyses, and implemented the integration approaches for understanding immune system interactions in food allergy.

The research was carried out in collaboration with the Luxembourg Institute of Health (LIH), coordinated through Dr. Annette Kuehn's research group, which provided the APSIS

datasets. While Dr. Kuehn provided expertise in allergy research and immunology, and Dr. Rebecca Czolk and Theresa-Maria Böhm contributed insights into the experimental design and biological context. Dr. Feng Hefeng provided methodological guidance at the intersection of bioinformatics and immunology, ensuring the computational approaches aligned with biological relevance. The primary analysis and interpretation of results presented in this thesis were independently conducted by Mathias Galati. This included the comprehensive initial screening and prioritization of the extensive set of molecular targets. The findings and literature review presented here form the foundation for future refinement through additional immunological expertise as the work progresses toward publication.

The successful completion of this work was made possible through the continuous supervision of Prof. Dr. Schneider and Dr. Satagopam, who provided strategic direction and ensured the overall scientific quality of the research, particularly in the development and validation of the computational methods.

Chapter 4

WNT5A and fibroblast roles in colorectal cancer

Abstract

The multifactorial and heterogeneous nature of CRC is characterized by various cellular populations and molecular interactions contributing to its progression. Despite the growing research interest over WNT5A, details and significance about its various mechanisms in CRC remain unclear. This study implemented single-cell RNA sequencing analysis by exploring clinical CRC samples and characterizing two distinct subpopulations of Cancer-associated fibroblasts (CAF): inflammatoryCAF (iCAF) and myofibroblastic CAF (myCAF). To gain deeper insight into cellular interplays and WNT5A expression, we subtyped fibroblasts cells and explored their enriched pathway. We applied Ligand Interaction Analysis and Network and NicheNet to investigate the ligand-receptor interactions specific to fibroblasts expressing WNT5A in the CRC TME, particularly focusing on the relationship among iCAF, myCAF) and other cell types. Our outcomes propose a better understanding of the expression of WNT5A in CAF, highlighting its contribution to the complex cellular mechanism that promotes cancer growth. The recognition of WNT5A signaling in CAF and its involvement in cellular interactions within the TME could leverage the possibility of targeting associated molecular pathways for therapeutic strategies in precision medicine.

4.1 WNT5A modulates fibroblast activity in colorectal cancer

4.1.1 Background on Colorectral cancer

CRC ranks as the third most prevalent cancer globally and is the fourth primary cause of all cancer-related deaths, contributing to 9% of all cases [209]. Despite advances in early screening programs, surgical procedures, and chemotherapeutic drug development, CRC is still a significant public health challenge. Smoking, obesity, and red meat consumption notably increase the risk for CRC incidence [210]. Recent researches target the heterogeneity of CRC, focusing on the variety of biological behaviors, the genetic modifications, and the perturbations in signaling pathways.

Germline genetic mutations cause various hereditary CRC syndromes, while somatic mutations accumulated over time lead to the development of sporadic colon cancer. A particular mutation frequently found among Ashkenazi Jewish individuals could potentially precipitate familial colon cancer [211]. Lynch syndrome, a recognized hereditary CRC syndrome, results from DNA mismatch repair gene mutations primarily in MLH1, MSH2, MSH6, PMS2 and EPCAM, which constitutes approximately 1-4% of all CRC cases [212]. Familial adenomatous polyposis (FAP), as another example of a hereditary CRC syndrome, arises from APC gene mutations and is characterized by the production of numerous adenomatous polyps in the colon and rectum [213]. MUTYH-associated polyposis, juvenile polyposis syndrome, Peutz-Jeghers syndrome, and serrated polyposis syndrome are additional hereditary CRC syndromes. Identifying specific genetic mutations linked to these hereditary CRC syndromes carries crucial clinical implications as it enables the implementation of strategies for early cancer detection and prompts consideration of prophylactic surgery in certain instances [214].

Significant progress in early detection, risk assessment, prevention, and treatment of CRC has resulted directly from understanding the genetics surrounding it [215]. However, there remains an urgent need to explore more thoroughly the mechanisms of CRC development and progression, especially in non-inherited cases. The discovery of new genetic mutations and signaling pathways has sped up the creation of innovative therapeutic approaches targeted at these specific molecular changes; this could potentially enhance treatment results for patients suffering from CRC. Understanding CRC genetics has provided insights for recognizing the Tumor microenvironment (TME) as a crucial factor in initiating, progressing and responding

to treatment of CRC. The TME has intricate interactions between various cell types such as cancer-associated fibroblasts (CAF), immune cells, and epithelial cells, that collectively contribute towards promoting cancer progression [216].

4.1.2 Role of fibroblasts in the Tumor microenvironment

Primarily, we recognize fibroblasts for their essential roles as components of connective tissue, with primary responsibility for producing the extracellular matrix and mediating tissue repair. However, recent advancements in the field redefine our understanding of these cells; illuminating their intricate involvement in host immunity and defense mechanisms [217].

Fibroblasts functions are context-dependent, often varying based on the tissue localisation. These cells in different tissues manifest unique roles, especially in their interactions with lymphocytes, thereby tailoring immunity to the specific tissue context [218]. Specialized fibroblasts, such as the fibroblastic reticular cells in secondary lymphoid organs, demonstrate a pronounced impact on immune responses. These cells are important in regulating immune responses, orchestrating mechanisms that can initiate, modulate, or even suppress immunity [219].

Fibroblasts play a role in immunity through their ability to identify and react to pathogens through inflammatory processes [217]. They take roles in the synthesis of antimicrobial peptides, growth factors, and other immune molecules. Such adaptive strategies modulate the body's defense mechanisms, particularly against bacterial and viral pathogens [220]. Moreover, fibroblasts extend their influence by interacting with various immune cells. In certain pathological contexts, such as Inflammatory Bowel Disease, their modulatory role on inflammation has serious implications, influencing both defensive mechanisms and therapeutic interventions [217]. The expression and activation of Toll-like receptors (TLR) within fibroblasts contribute significantly to this process. While expressing several TLR (from 1 to 10), they act as the frontline defense for identifying pathogens and initiating the production of cytokines and chemokines [220].

However, the role of fibroblasts is not limited to healthy states. Pathological fibroblasts emerge as key players in various diseases as they maintain tissue homeostasis and repair balance, but can also propel disease progression. For instance, they drive fibrosis and inflammation in diseases such as Rheumatoid Arthritis and even in COVID-19. Moreover, oncologists have highlighted the importance of CAF, which critically influence immune responses and can dictate the outcome of certain immunotherapies [218]. Recent studies

indicate that therapeutic targeting of fibroblasts, along with the investigation of their diverse functional states, not only provides valuable insights but also has the potential to facilitate precision treatments in diseases such as RA, IBD, and various cancers [219].

CAF's significant role in the TME emerges from their roles in signaling pathways and diverse cellular mechanisms [221], which can be attributed to their origins in a variety of cells such as mesenchymal stem cells (MSCs), mesothelial cells, adipocytes, hematopoietic stem cells, and fibrocytes [222]. Thus, pointing towards CAF's crucial participation in the development of fibrosis and evolution of the TME . CAF are characterized by the presence of specific markers, namely α-smooth muscle actin (α-SMA) and fibroblast activation protein (FAP), which serve as identification tools and perform various functional roles [222]. The distinct spatial distributions, characteristics, and functional attributes of the subtypes iCAF and myCAF exhibit conspicuous heterogeneity within theCAFpopulation. iCAF, classified by their absence of FAP- and αSMA-, significantly contrast with myCAF that possess FAP+ and αSMA+ [222, 223]. The varying functional roles in the TME are underscored by their distinct molecular profiles, highlighting one of the significant functions exhibited by CAF as they contribute in extracellular matrix (ECM) remodeling. Tumor cell invasion and metastasis are facilitated by the secretion of ECM components and matrix-degrading enzymes such as matrix metalloproteinases (MMPs), thus emphasizing the essential role played by the ECM in tumor progression [224]. CAF are crucial for facilitating angiogenesis, a critical process necessary for the survival and dissemination of tumors. They fulfill this role by releasing pro-angiogenic factors, notably vascular endothelial growth factor (VEGF), thereby inducing the formation of novel blood vessels [225]. CAF significantly influence the immune response within the TME by creating an immune-suppressive environment through their production of specific cytokines and chemokines, such as TGF-β. CAF play a crucial role in manipulating the immune environment within the tumor by recruiting immunosuppressive cells such as regulatory T cells and myeloid-derived suppressor cells [226]. Moreover, recent revelations have revealed the presence of antigen-presenting CAF known as apCAF, which are distinguished by their expression of markers like MHC II and CD74. Elucidating the specific roles and functions of CAF in CRC could pave the way for developing innovative therapeutic strategies that target these cells, thereby enhancing cancer treatment outcomes, while emphasizing their potential involvement in immunomodulation within the TME necessitates distinguishing between various functionalities of CAF subtypes [222].

4.1.3 Significance of WNT5A in cancer progression

The Wingless and Int-1 (Wnt) signaling pathway acts as a key controller of diverse cellular processes, particularly in tissue homeostasis, repair, and tumor development. It has a noticeable impact in CRC as an overactive pathway [227]. The Wnt pathway's crucial role in the development of CRC is firmly established, as around 70% of cases display an APC gene mutation [228]. This abnormal activation of Wnt signaling has been consistently noted in various human malignancies, particularly prominent within CRC [229]. The R-spondin (RSPO) family is involved in gene fusions that accelerate Wnt-dependent tumor initiation in CRC, an observation of great significance [230].

The Wnt pathway, specifically the Wnt/β-catenin signaling, plays a fundamental role in both embryonic development and adult homeostasis. The pathway's involvement spans a range of cellular activities, from the expansion of cells to the intricate tasks of cell diversity and epithelial-mesenchymal transition (EMT), ultimately directing the movement, infiltration, and spread of cells [231]. Considering its connection with various growth-related processes, this pathway becomes an interesting focus for therapeutic investigations, particularly in addressing CRC research [232].

WNT5A expression in CAF has been correlated with more advanced stages of CRC, marking it as a potential target explaining CRC pathophysiology [233]. WNT5A's distinguished characteristic among other members of the Wnt family lies in its unique capability to activate and inhibit the canonical Wnt pathway, while being one of the major drivers of both the non-canonical pathways [234, 235, 236]. Its significance is manifested in embryonic development where it assumes an indispensable role in coordinating cell polarity and establishing the axis of the body [237]. WNT5A binds to a variety of receptors such as FZD, ROR, PTK7, and RYK, resulting in multiple ligand/receptor combinations capable of activating either the canonical or non-canonical Wnt pathways [238]. WNT5A emerges as an interesting oncology target, exhibiting a dual-faced role with both anti-tumor and pro-tumor properties. Initially, CRC experiences WNT5A as a hindering influence that represses the crucial canonical Wnt pathway necessary for early tumorigenesis. Nevertheless, as CRC advances, augmented expression of WNT5A becomes a precursor of unfavorable prognostics. WNT5A acts as a tumor suppressor in liquid tumors and CRC, whereas it exhibits tumorigenic properties in the majority of solid tumors [239, 238]. In malignancies such as melanoma and ovarian cancer, heightened levels of WNT5A indicate advanced stages primarily due to its vital role in the Planar Cell Polarity pathway crucial for metastasis [240, 241]. The intricate role of WNT5A

in oncology fluctuates depending on the tumor type. WNT5A's intricate biology could be linked with its therapeutic potential, highlighting its significance in developmental biology and as a pivotal aspect of oncology research.

4.1.4 scRNA-seq solutions to investigate immunogenetics

scRNA-seq has emerged as a powerful tool for investigating ligand-receptor interactions in CRC, providing invaluable insights into TME and the roles of immune cells in cancer development. By analyzing CCC through ligand-receptor interactions it becomes apparent that there is a stronger correlation between CRC subtype CMS1 and an abundance of key players like monocytes and macrophages within TME [242]. Immunotherapies targeting ligand-receptor interactions (LRIs) have experienced rapid advancements in the treatment of colorectal cancer. Nevertheless, the patterns of LRIs in CRC, their impact on the TME, and their clinical significance remain to be fully elucidated [243]. Investigating ligand-receptor interactions between cells in diseased tissues using scRNA-seq offers valuable insights into the pathogenesis and progression of the disease [244]. In recent years, there have been several computational techniques developed for predicting CCC events through ligand-receptor interactions using scRNA-seq data. Two examples of these methods are NicheNet and LIANA. NicheNet is a computational approach that utilizes pre-existing knowledge, and it has been iused in groundbreaking research to evaluate the efficacy and mechanisms of novel cancer immunotherapies [245]. LIANA (Ligand-receptor Inference by Associating scRNA-seq and Networks) is a versatile tool offering a unified interface for various CCC analysis resources and methods. LIANA enables users to access and apply resources and methods in any combination, presenting a consensus ranking and simplifying the process of analyzing CCC from scRNA-seq data [246]. Both tools have been developed in response to the ongoing advancements in scRNA-seq technology, which have shown a growing interest in understanding intercellular crosstalk and led to the creation of many other tools and resources for investigating CCC.

In this study, our aim is to examine the expression of the WNT5A gene in CAF and predict its receptor bindings in the context of CRC. We aim to enhance knowledge of fibroblast involvement in the TME by gaining a deeper understanding of the role of WNT5A in fibroblasts, despite its poorly understood mechanisms and downstream effects on these cells. The development of new therapeutic strategies targeting the WNT5A signaling pathway improving treatment outcomes for CRC patients, may be promoted by this enhanced comprehension.

4.2 Characterization using scRNA and CCC

4.2.1 Data collection and processing

The dataset used in this analysis is derived from a prior study conducted by [247]. The demographic characteristic reveals that the age distribution among these patients ranges from the 50 to 90. Two patients are between the ages of 50-55 years, one patient is in the age group of 76-80 years, three patients fall within the range of 80-85 years, and one patient is aged between 86-90 years. The cohort has a gender distribution of four females and three males. In terms of tumor staging, the cohort showcases a spectrum ranging from stage I to stage IVA. Two patients exhibit a diagnosis at stage IIA, while two others present with stage IIIB. Additionally, one patient demonstrates each of stages I, IIB, and IVA. Predominantly, three patients exhibit tumors situated in the sigmoid region, while two patients have them positioned specifically in the caecum and one patient has it located within the ascending colon. Additionally, another patient exhibits a tumor at the left rectosigmoid. As the phenotype subtype, all tumors are of moderately differentiated adenocarcinomas. However, one tumor is located in the right caecum and is described as moderately differentiated adenocarcinoma with an amalgamation of glandular and mucinous growth patterns that are further accompanied by moderate budding. Six tumors are characterized as Microsatellite Stable from a molecular standpoint, while one of the two tumors situated in the right caecum displays a profile of Microsatellite Instability-High.

4.2.2 Computational Analysis

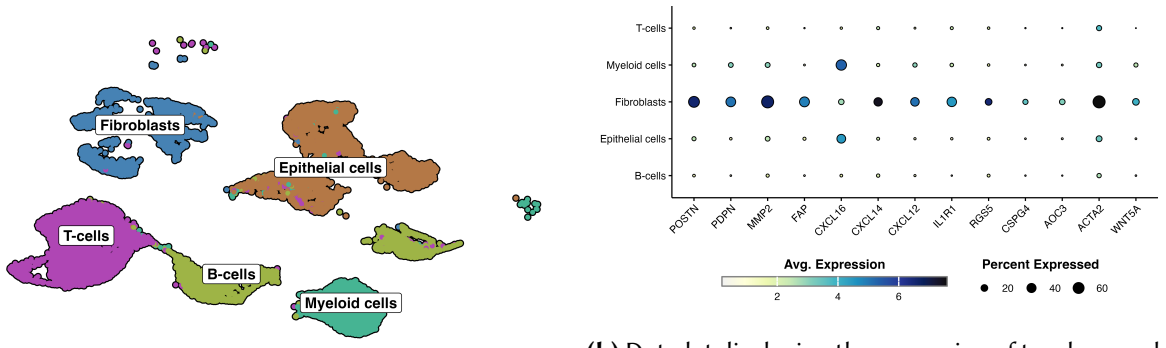
In order to identify fibroblast cells within our dataset, we used the Seurat package in the R programming language. Initial quality control and filtering measures were implemented to ensure the integrity and reliability of the obtained data. Specifically, our study only considered cells with a minimum of 401 unique molecular identifiers (UMIs) and a maximum of 15% of reads mapping to mitochondrial RNA. Data normalization was carried out using the logNormalize method with a scale factor of 1^6. Subsequently, we selected 2000 highly variable genes utilizing a feature selection function and scaled the data by regressing on confounding factors, such as the number of UMIs and the percentage of mitochondrial RNA.

Principal component analysis (PCA) was applied to the dataset across 20 dimensions, followed by the employment of a clustering algorithm with a resolution of 0.4 and a shared-nearest neighbor graph between cells to group the cells into distinct clusters. The smart local

moving (SLM) algorithm and Uniform Manifold Approximation and Projection (UMAP) were utilized for cluster visualization, while cell types were annotated using the SingleR package and the BlueprintEncode reference.

To identify subtypes of fibroblast cells, we conducted a clustering analysis on a subset of the fibroblast cells previously identified. The same preprocessing and analysis steps were employed with identical parameters. We carried out a Differentially expressed genes (DEG) analysis using the MAST algorithm and a specific selection of marker genes, allowing us to identify two subtypes of fibroblast cells: iCAFand myCAF. This procedure was also repeated for epithelial cells, employing an appropriate list of marker genes to discern the following cell types: Tumor Cycling Cells (TCC), Revival Stem Cells (RSC), Fetal Stem Cells (FSC), Paneth, Goblets, Enterocytes, Transit amplifying cells (TA), cycling TA, Tumor Goblet Cells (TGC), Tuft cells, and Enteroendocrine cells (EE). We explored the expression of various markers using the feature plot function within the Seurat package, identifying two iCAF clusters expressing WNT5A (WNTiCAF) and identified them as WNTiCAF1 and WNTiCAF2. DEG between the WNTiCAF against the other cluster of WNTiCAF cells and the myCAF cells, using the MAST algorithm. The selection criteria employed involved stringent filters, including an average log2 fold change greater than 1.5 or less than -1.5, and an adjusted p-value threshold of 10^{-32}, Additionally, pathway enrichment analysis using org.Hs.eg.db reference was conducted to pinpoint pathways that were enriched (with an overlap of 7 genes at least) within the WNTiCAF cells.

We used the LIANA (Linking Interactions and Network Analysis) package and the NicheNet package to research potential signaling pathways and interactions among different cell types, in order to understand the complexities of CCC within our dataset. Our analysis with the LIANA package primarily focused on Macrophages, CD4+ T-cells, CD8+ T-cells, Neutrophils, Monocytes, WNTiCAF, iCAF, RSC, HSC, FSC, TCC and myCAF cell types. We aimed to identify significant interactions based on the highest ranking scores that can be likened to p-values in terms of their significance. We prioritized WNTiCAF cells as the primary sender cells in order to gain a comprehensive understanding of their involvement in intercellular communication. Through the NicheNet analysis, we focused our investigation on examining the interactions between these cells and the aforementioned cell types. We explored deeper into the ligand-receptor (LR) pairs, with a particular emphasis on WNT5A as the ligand within our scope of interest. This enabled us to precisely identify distinct potential regulatory mechanisms activated by WNT5A in the sphere of WNTiCAF interactions with



(a) UMAP visualization of major cell populations, highlighting fibroblasts, epithelial cells, T cells, B cells, and myeloid cells.

(b) Dot plot displaying the expression of two key markers per cluster, with color intensity indicating average expression levels and dot size representing the percentage of cells expressing each marker.

Figure 4.1 Characterization of global cell types including fibroblasts.

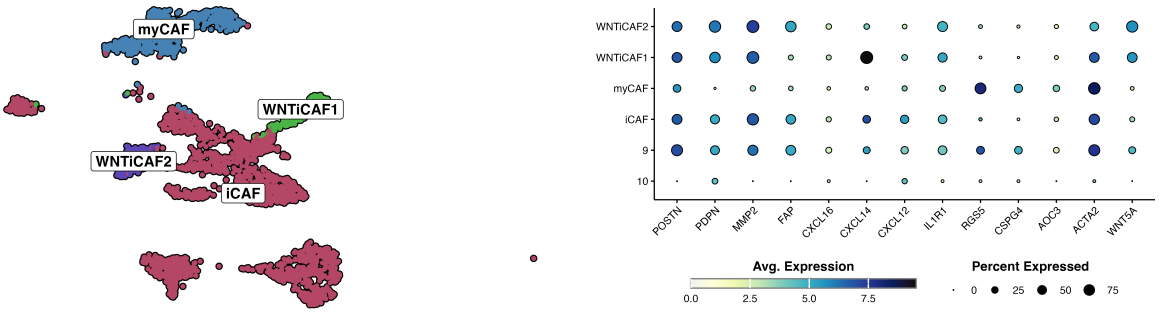
interacting cells.

4.3 Fibroblasts drive inflammation and cancer progression

4.3.1 Characterization of CAF clusters

In the present study, we investigated the identification and characterization of fibroblast and epithelial cell subtypes, as well as the analysis of CCC, in the context of cancer development and progression. Our results provide insights into the cellular heterogeneity as we identified and characterized two distinct subtypes of fibroblast cells, iCAF and myCAF. These subtypes displayed unique gene expression profiles and biological properties, which were further investigated by the Differentially expressed genes (DEG) analysis and extensive enrichment analysis.

The figure 4.1 illustrates the single-cell transcriptomic landscape of identified cellular populations in our study, as visualized by a UMAP. On figure 4.1a, each point corresponds to an individual cell, and they are color-coded based on their assigned cell types, highlighting the notable cluster of fibroblasts. On figure 4.1b, a dot plot showcases the expression patterns of two significant marker genes for each identified cell cluster. The size of each dot represents the proportion of cells within the cluster expressing the respective marker gene, while the color intensity denotes the averaged expression level. This dot plot facilitates the simultaneous representation of the marker genes prevalence and abundance across different cell populations,



(a) UMAP visualization of fibroblast subclusters, highlighting distinct populations, including two clusters with WNT5A overexpression (WNTiCAF1 and WNTiCAF2), alongside myCAF and iCAF subtypes.

(b) Dot plot displaying the expression of two key markers per cluster, with color intensity representing average expression levels and dot size indicating the percentage of cells expressing each marker.

Figure 4.2 Characterization of stratified fibroblasts

thereby helping in the robust characterization of cellular heterogeneity in the sampled tissue.

The figure 4.2 goes deeper into the fibroblast heterogeneity identified in the preceding analysis. figure 4.2a presents a UMAP focusing on the stratified fibroblasts, with the two clusters demonstrating an overexpression of WNT5A explicitly highlighted. The figure 4.2b, a dot plot, displays two key marker genes associated with each fibroblast sub-cluster. Similar to figure 4.1, this representation enables the visualization of differential marker expression across the fibroblast subpopulations, which can hint at their varying functional roles.

The combined visualization provided in figure 4.2 emphasizes the intricate layers of fibroblasts heterogeneity, with specific attention to the WNT5A-overexpressing clusters. This detailed stratification underscores the role of distinct fibroblast subsets in the biological system under our scope, highlighting the importance of single-cell resolution in investigating the complexity of cellular environments.

The figure 4.3 presents nebula plots, a form of spatial gene expression visualizations, highlighting the WNT5A expression across the cells as mapped onto the two UMAPs previously described. Each nebula plot uses color intensities to denote the expression level of WNT5A, with lighter shades representing higher expression.

In the left nebula plot, the WNT5A expression pattern is based on the UMAP of the global cell types from figure 4.1. This plot allows us to gauge the relative WNT5A expression across all cell types, with a particular focus on the fibroblasts.

The right nebula plot overlays the WNT5A expression onto the stratified fibroblasts UMAP from figure 4.2. By doing so, it visually emphasizes the two fibroblast clusters that overexpress WNT5A, as previously identified, providing a clear illustration of their distinct

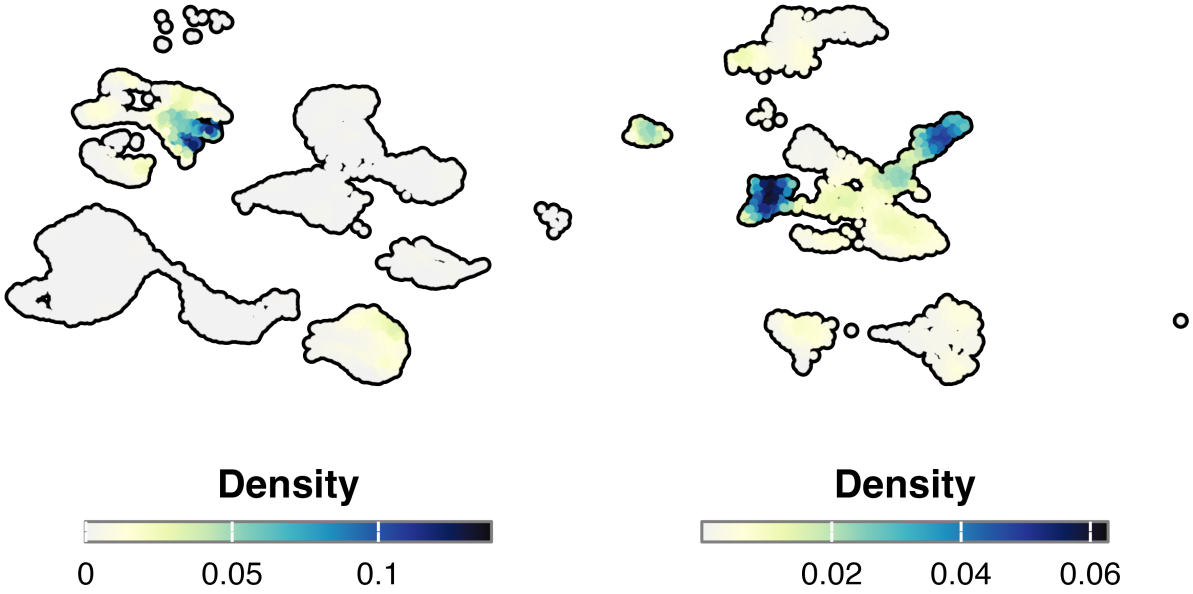


Figure 4.3 Nebula plots illustrating WNT5A expression across fibroblast clusters. Density plots displaying the spatial distribution of WNT5A expression within the dataset, mapped onto the previously identified UMAP. Higher expression levels (dark blue) indicate regions with enriched WNT5A activity, aligning with the stratified fibroblast populations

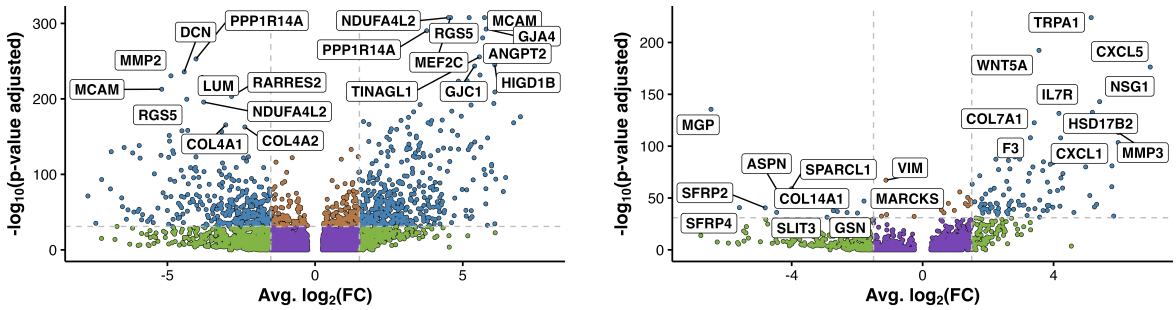
molecular signature within the broader fibroblast population.

Taken together, the nebula plots in figure 4.3 give context to the WNT5A expression within our dataset, further illuminating its potential role in shaping the observed cellular heterogeneity and the functionality of specific fibroblast subtypes.

The figure 4.4 presents volcano plots depicting the results of Differentially expressed genes (DEG) analysis that involves first all the identified fibroblasts cells (figure 4.4a) and both WNT5A-overexpressing fibroblast clusters (figure 4.4b).

The figure 4.4a plot, including both WNTiCAF clusters, provides an overview of DEG from the combined WNTiCAF clusters as compared with other fibroblasts. Although WNT5A does not appear among the top 10 labeled genes in this aggregate view, its overexpression is captured in the underlying data and is a key signature of the WNTiCAF population.

The figure 4.4b plot breaks down WNTiCAF clusters, revealing cluster-specific expression profiles. In this more focused view, WNT5A is among the upregulated genes that define the unique transcriptional within each WNTiCAF subgroup. This highlights the important role of WNT5A in driving cluster-specific phenotypes and potentially influencing underlying mulecular pathways.



(a) Volcano plot displaying DEG across all fibroblasts, highlighting the top 10 genes, including both WNTi-CAF clusters.

(b) Volcano plot focusing on DEG specific to WNTi-CAF clusters, emphasizing the top 10 most significant markers.

Figure 4.4 Comparative volcano plots of Differentially expressed genes in fibroblast populations. Log2 fold change (x-axis) represents gene expression differences, while the adjusted -log10(p-value) (y-axis) indicates statistical significance.

4.3.2 Fibroblasts clusters enrichment comprehension

WNTiCAF clusters against the other fibroblast cells

The figure 4.2a presents a UMAP that specifically highlights the two WNT5A overexpressing fibroblast clusters (WNTiCAF). This UMAP visualization effectively sets the stage for the enrichment analysis by emphasizing the spatial distribution and relative density of the WNTiCAF clusters within the broader context of the fibroblasts cells. The distinct clustering patterns and their separation from other fibroblasts cell populations underscore the unique transcriptional signatures of these fibroblast subsets, further justifying their selection for deeper functional exploration through enrichment analysis.

A comprehensive Gene Ontology (GO) enrichment analysis (figure 4.5) was performed on the 103 DEG identified and revealed several pertinent biological processes, molecular functions, and cellular components that could potentially be dysregulated within our study. Among immune-related pathways, we observed enrichment for “Leukocyte migration”, “Leukocyte chemotaxis”, “Cell chemotaxis”, “Cellular response to molecule of bacterial origin”, “Cellular response to biotic stimulus”, and “Response to lipopolysaccharide”. These findings suggest an activated immune microenvironment and possible chronic inflammatory processes in CRC.

Developmental and differentiation pathways were also significantly enriched, including “Muscle cell differentiation”, “Mesenchyme development”, “Mesenchymal cell differentiation”, “Regulation of vasculature development”, “Regulation of angiogenesis”, “Cell fate commitment”, “Ossification”, and “Pattern specification process”. This highlights potential disruptions in cell lineage determination and tissue homeostasis, which could contribute to tumor pro-

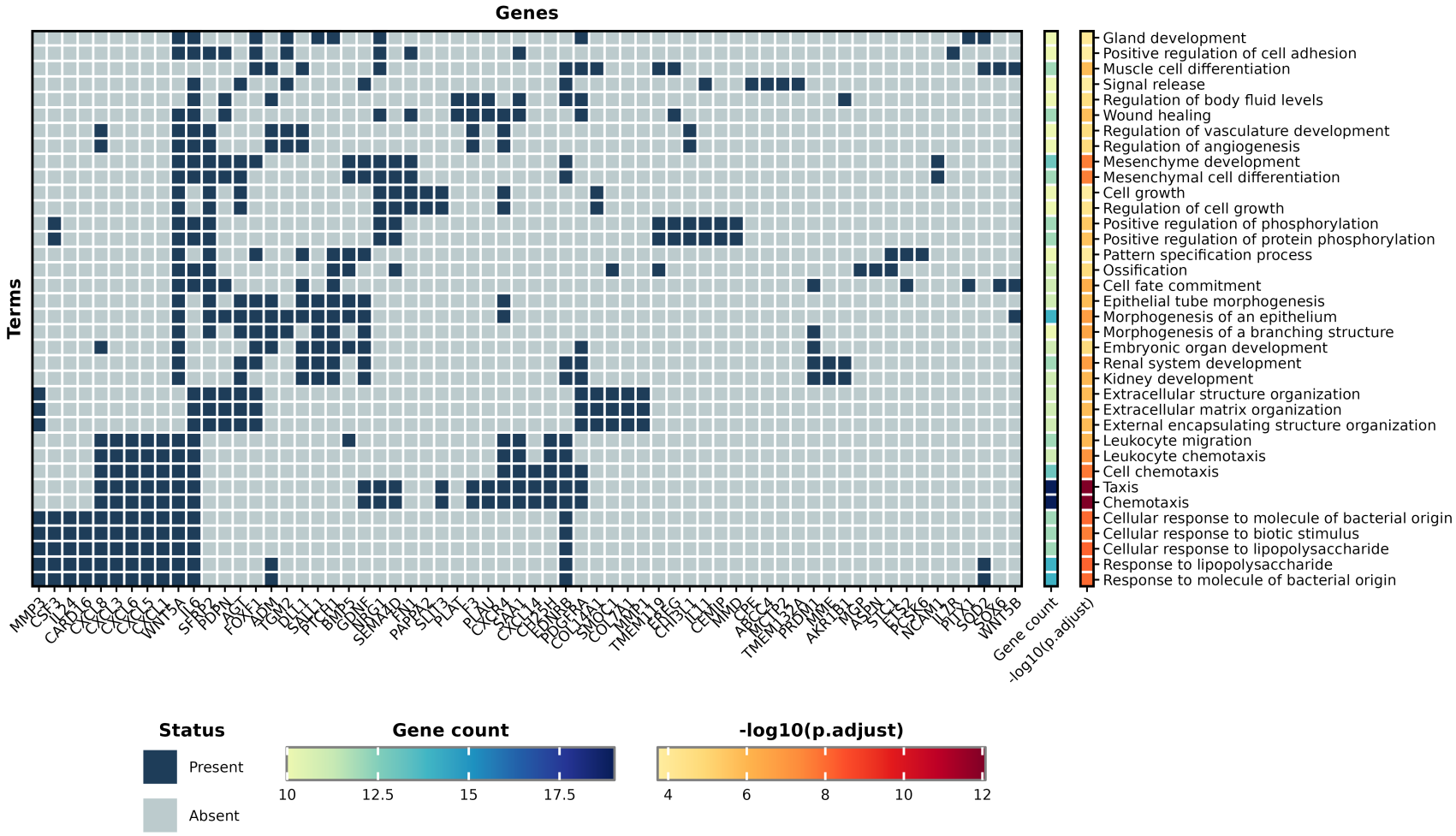


Figure 4.5 Gene Ontology (GO) enrichment analysis of WNTiCAF clusters. The heatmap displays significantly enriched GO terms associated with both WNTiCAF clusters. Rows represent genes, and columns correspond to enriched biological processes. The color scale indicates the number of genes associated with each term, while the top bar represents the statistical significance of enrichment (-log10 adjusted p-value). This analysis highlights key pathways involved in fibroblast function and WNT5A-associated processes.

gression.

Additionally, several processes related to tissue architecture and extracellular matrix remodeling were enriched, such as “Extracellular matrix organization”, “Extracellular structure organization”, and “Morphogenesis of a branching structure”. These results underscore a likely remodeling of the TME, which may impact cancer cell invasion and progression.

Overall, our GO enrichment analysis provides a comprehensive insight into the interplay between immune activation, developmental dysregulation, and extracellular matrix remodeling in CRC, reinforcing the complex pathophysiology underlying the disease.

Molecular function (GO | MF) terms enriched in these experimental conditions (figure 4.6) suggest roles in signal transduction, protein-protein interactions, enzymatic activity, and receptor binding. Specifically, these include “Signaling receptor binding”, “Enzyme binding”, “Transmembrane signaling receptor activity”, “Peptidase activity”, “Identical protein binding”, “Ion binding”, “Heparin binding”, and “Glycosaminoglycan binding”. The enrichment of these functions highlights potential alterations in extracellular matrix interactions, enzymatic processing, and receptor-mediated signaling, which may be relevant to the dysregulated cellular communication observed in CRC.

Moreover, the enriched terms in the Biological Processes (GO | BP) (figure 4.7) further illustrate the intricate interplay of transcriptional regulation in CRC, particularly through pathways such as “Negative regulation of transcription by RNA polymerase II” and “Positive regulation of transcription by RNA polymerase II”. These findings emphasize potential disruptions in gene expression control that may contribute to tumorigenesis.

Additionally, pathways associated with neuronal development — including “Neuron projection development”, “Neuron projection morphogenesis”, “Axonogenesis”, and “Axon development” — suggest a possible role for neurogenic mechanisms in CRC progression, potentially linked to neural remodeling within the TME.

Finally, the presence of angiogenesis-related terms such as “Regulation of angiogenesis” and “Positive regulation of angiogenesis” underscores the vascular adaptations crucial for tumor growth and metastasis. Together, these enriched pathways highlight the multifaceted nature of CRC pathophysiology, integrating transcriptional control, neural plasticity, and vascular remodeling as key biological components of tumor progression.

WNTiCAF1 cluster DEG enrichment focus

Another comprehensive GO enrichment analysis was performed on the 74 DEG expressed

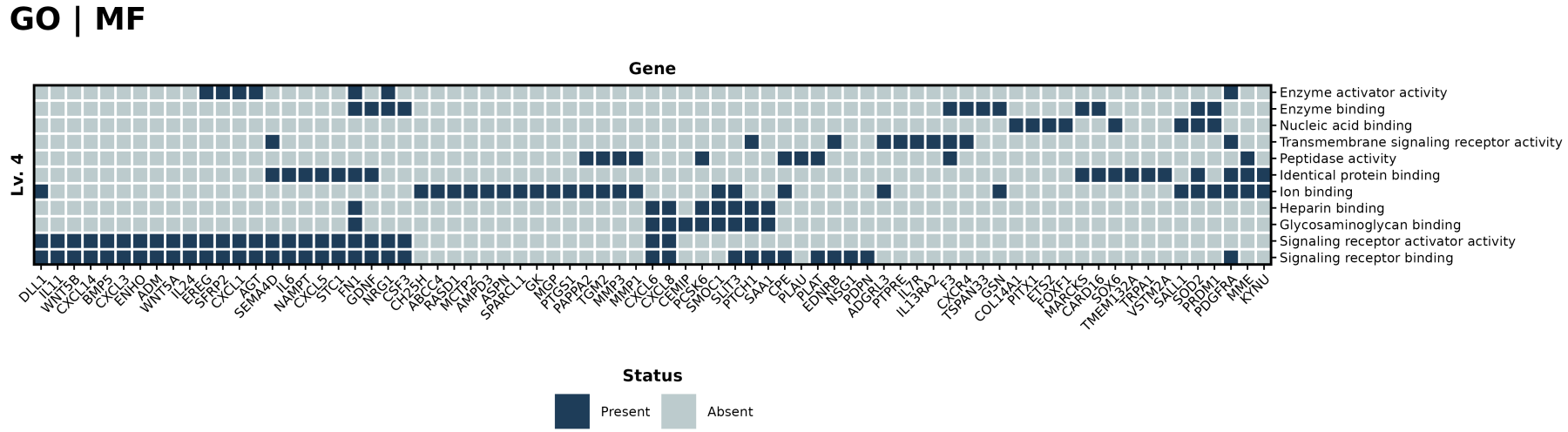


Figure 4.6 Enriched molecular functions (MF) in WNTiCAF clusters. The heatmap presents significantly enriched GO molecular function terms associated with both WNTiCAF clusters. Rows represent genes, and columns correspond to specific molecular functions. Dark blue cells indicate gene-term associations, while light gray cells represent absence. This analysis highlights key molecular activities, such as signaling receptor binding and enzymatic activity, that may underlie WNTiCAF functional properties.

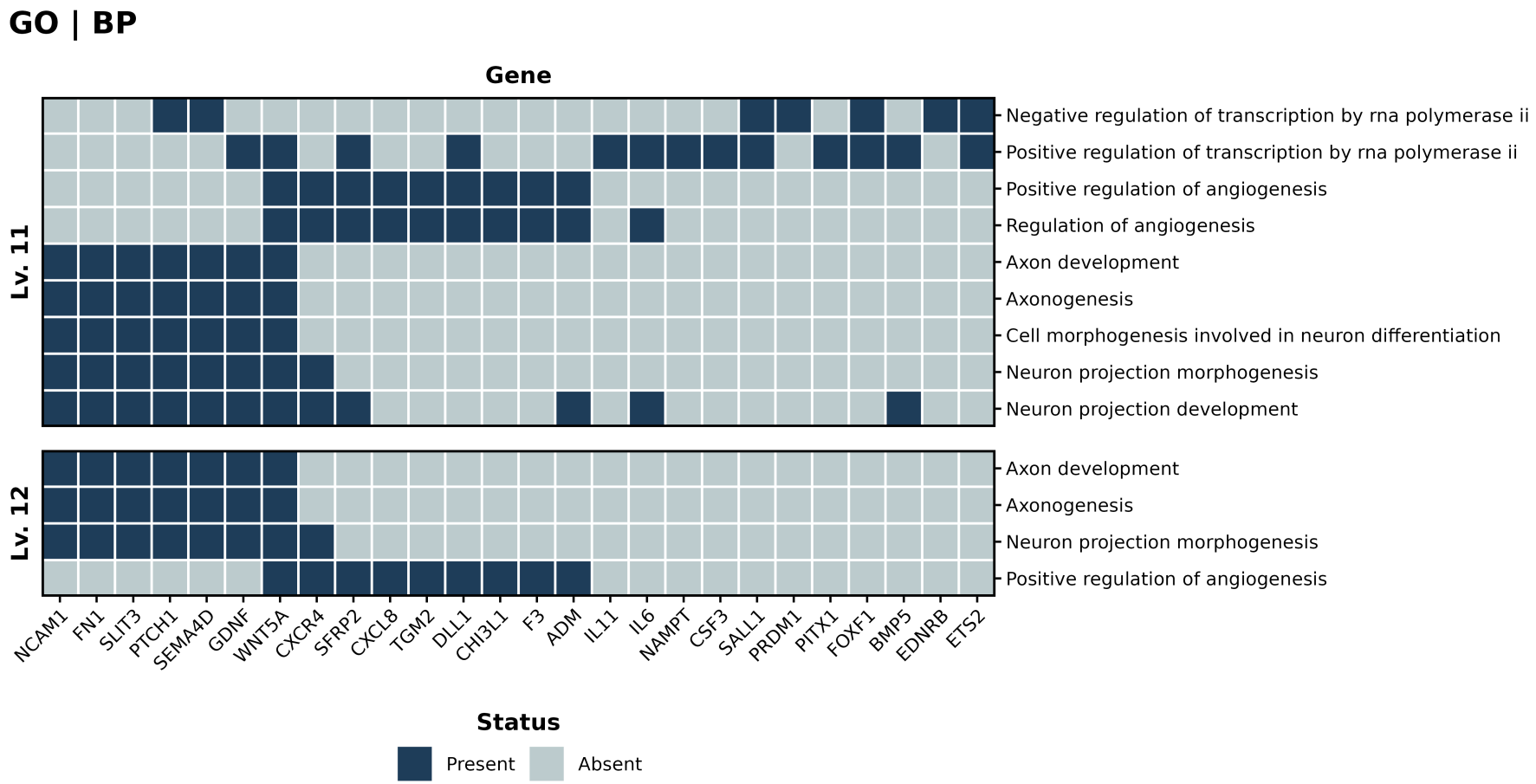


Figure 4.7 Enriched biological processes (BP) in WNTiCAF clusters. he heatmap presents significantly enriched GO biological processes associated with both WNTiCAF clusters. Rows represent genes, while columns correspond to biological processes. Dark blue cells indicate gene-term associations, while light gray cells represent absence. Processes are grouped by hierarchical levels (Lv. 11 and Lv. 12), highlighting pathways involved in neuron projection, axonogenesis, angiogenesis regulation, and transcriptional control.

genes identified focusing in this part on the WNTiCAF1 cluster (figure 4.2a). The results of the analysis were insightful, underscoring the broad impact of alterations in the expression patterns of these genes on biological processes and cellular functionalities

The analysis revealed pertinent GO processes involving morphogenesis and development. The identified processes (figure 4.8) included “mesenchyme development”, “mesenchymal cell differentiation”, “connective tissue development”, and “cell fate commitment”, all of which are relevant to tumorigenesis, particularly in the remodeling of the TME and metastatic dissemination. Additionally, “wound healing” and “epithelial tube morphogenesis” suggest potential implications in tissue regeneration and repair mechanisms often co-opted by cancer cells. “Regulation of cell growth”, “chemotaxis”, and “regulation of apoptotic signaling pathways” further highlight key aspects of tumor progression, including aberrant proliferation, migration, and evasion of apoptosis. These findings underscore the dysregulation of developmental and signaling pathways that may contribute to the aggressive phenotypes observed in CRC.

In terms of molecular function (GO | MF) (figure 4.9), several activities stood out that could play a role in CRC pathogenesis. The highlighted functions were primarily centered around receptor signaling, including “signaling receptor binding and activator activity” and “transmembrane signaling receptor activity”, which are crucial in mediating intercellular communication and driving cellular responses. Additionally, functions related to protein binding, such as identical protein binding and enzyme binding, were also found, suggesting the potential for altered protein-protein interactions within the CRC cellular milieu. Lastly, the “DNA-binding transcription factor activity, RNA polymerase II-specific”, and “nucleic acid binding” indicate changes in the transcriptional regulation machinery which could further contribute to oncogenic transformation.

As for biological processes (GO | BP) (figure 4.10), we observed over representation of genes implicated in the “regulation of transcription by RNA polymerase II”, including both its positive and negative regulation. This suggests that transcriptional dysregulation may be an essential feature of the WNTiCAF1 cluster. Additionally, genes associated with “neuron projection morphogenesis”, “axon development”, and “axonogenesis” were prominently represented. This enrichment in neuronal processes may indicate a previously unappreciated link between transcriptional regulation and axonal growth in the context of WNTiCAF1, warranting further investigation.

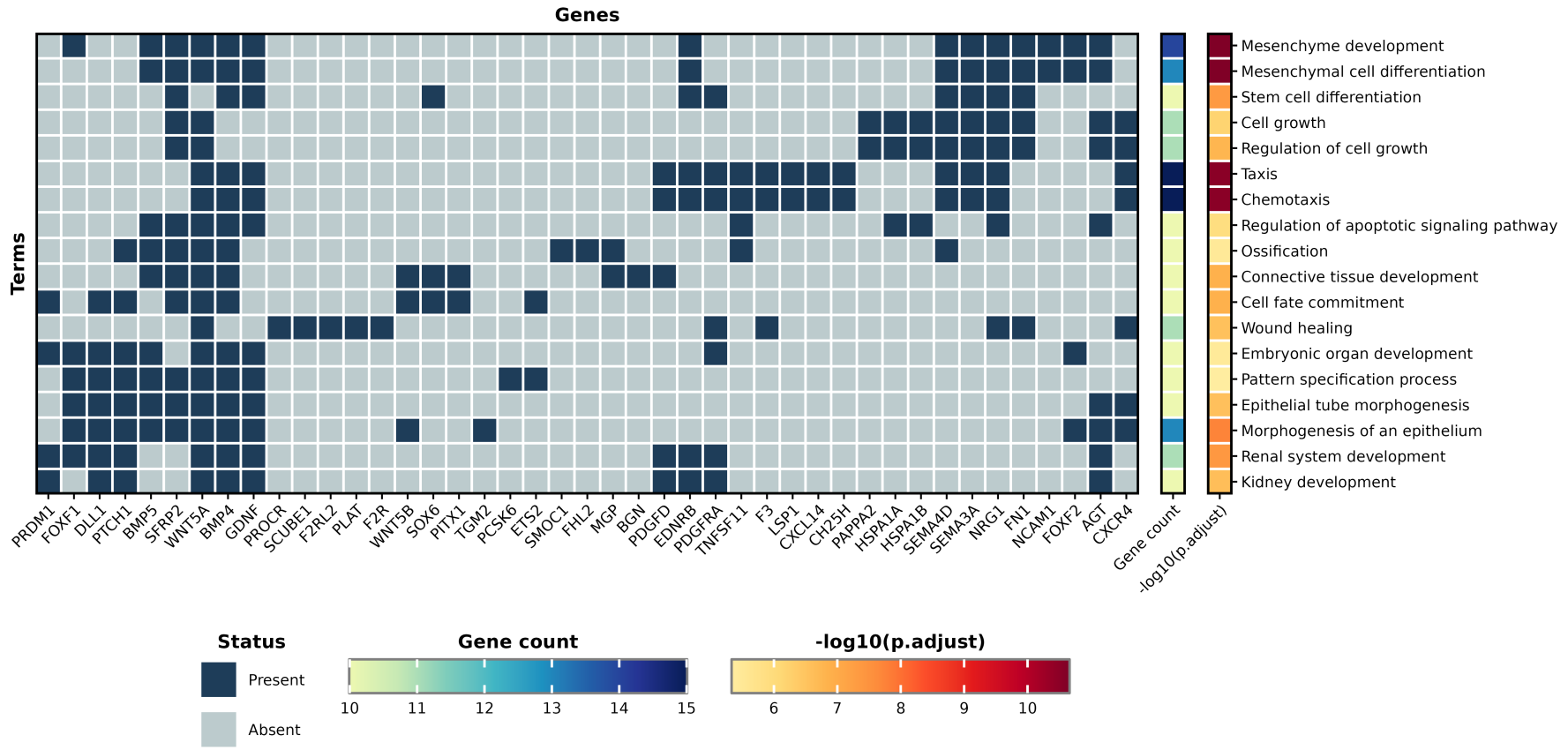


Figure 4.8 GO term enrichment analysis in WNTiCAF1 cluster. The heatmap displays Gene Ontology (GO)GO biological processes significantly enriched in the WNTiCAF1 cluster. Rows represent genes, and columns correspond to enriched GO terms. Dark blue cells indicate gene-term associations, while light gray cells represent absence. The top bar represent gene count per term, while the color gradient indicates the statistical significance (-log10 adjusted p-value) of enrichment. Key enriched pathways include mesenchymal differentiation, tissue development, wound healing, and epithelial morphogenesis.

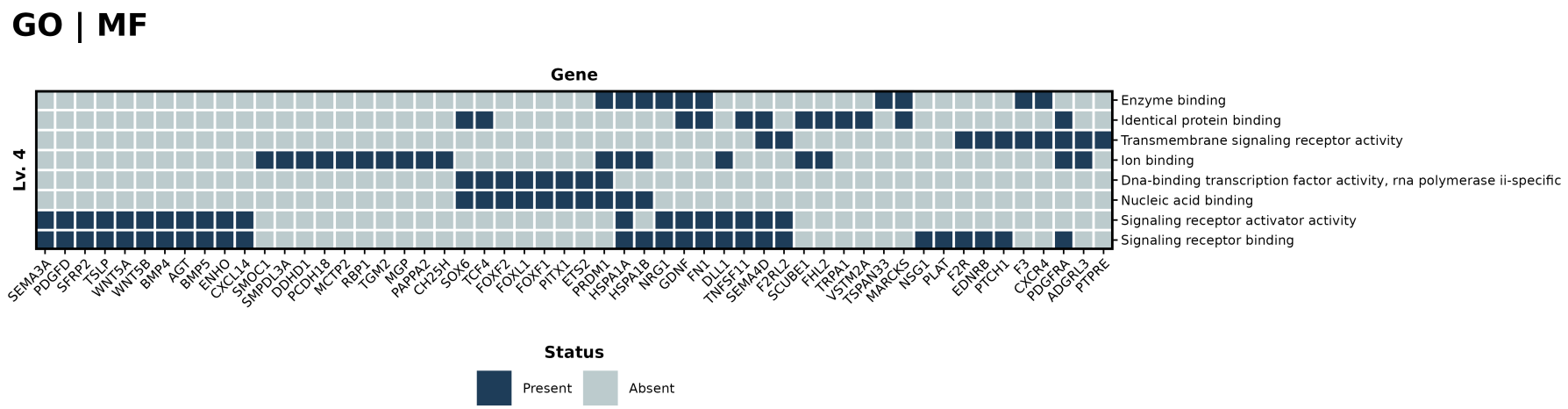


Figure 4.9 Enriched molecular functions (MF) in the WNTiCAF1 cluster. The heatmap presents significantly enriched GO molecular function terms associated with the WNTiCAF1 cluster. Rows represent genes, and columns correspond to molecular functions. Dark blue cells indicate gene-term associations, while light gray cells represent absence. Key enriched functions include signaling receptor activity, enzyme binding, transcription factor activity, and ion binding, highlighting the molecular mechanisms underlying WNTiCAF1 functionality.

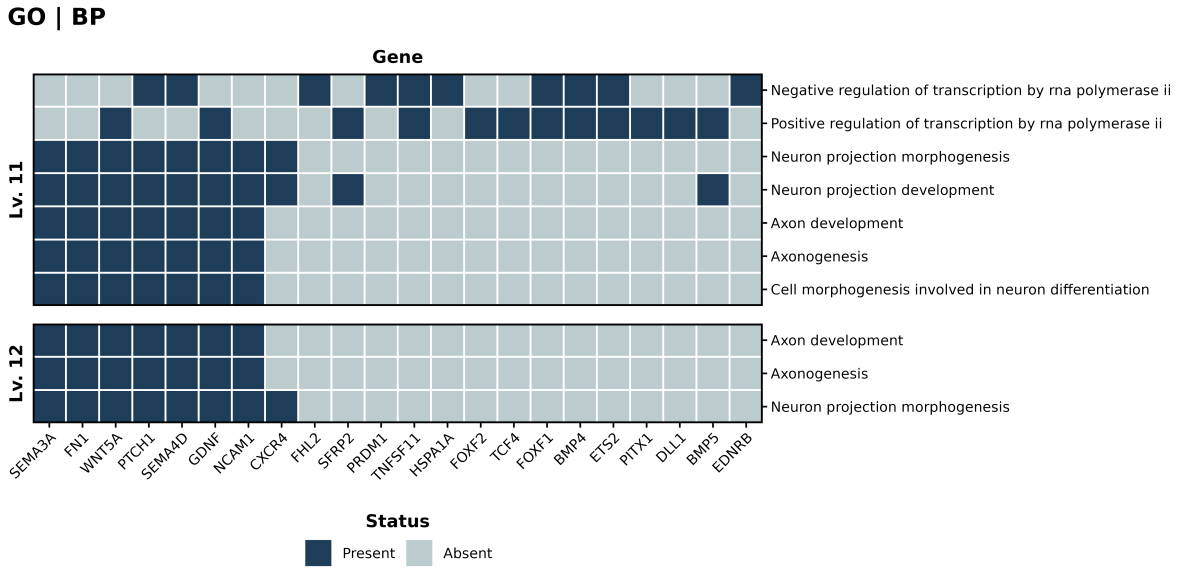


Figure 4.10 Enriched biological processes (BP) in the WNTiCAF1 cluster. The heatmap presents significantly enriched GO biological processes associated with the WNTiCAF1 cluster. Rows represent genes, and columns correspond to enriched biological processes. Dark blue cells indicate gene-term associations, while light gray cells represent absence. Key enriched pathways include transcriptional regulation, neuron projection morphogenesis, axonogenesis, and cell differentiation, highlighting the functional landscape of WNTiCAF1 fibroblasts.

WNTiCAF2 cluster DEG enrichment focus

Our enrichment analysis on WNTiCAF2 (figure 4.2a identified 87 genes and revealed significant insights into the biological processes, molecular functions, and GO categories linked with CRC pathology.

Enrichment analysis of GO terms (figure 4.11) revealed a notable over-representation of categories related to microbial response, including, "response to molecule of bacterial origin", "cellular response to lipopolysaccharide", "cellular response to molecule of bacterial origin", and "cellular response to biotic stimulus". These categories highlight a critical immune response to bacterial components, a key factor in the colorectal environment. Additionally, GO terms associated with leukocyte activity, such as "leukocyte migration", "leukocyte chemotaxis", "myeloid leukocyte migration", and "cell chemotaxis", indicate a dynamic immune cell movement, which may be instrumental in modulating the inflammatory landscape.

Beyond immune responses, the analysis also revealed an enrichment of GO terms related to tissue remodeling and structural organization, including "extracellular matrix organization", "extracellular structure organization", and "external encapsulating structure organization". These pathways suggest active stromal remodeling processes that may contribute to TME. Furthermore, the presence of "regulation of inflammatory response" and "developmental maturation" underscores the interplay between immune activation and tissue differentiation, potentially influencing disease progression and therapeutic responses.

The molecular function (GO | MF) (figure 4.12) results highlighted functions predominantly related to binding and enzymatic activity. Identified categories such as "Signaling receptor binding", "Nucleoside phosphate binding", "Heterocyclic compound binding", "Ribonucleotide binding", "Ion binding", and "Nucleic acid binding" suggest a strong involvement of molecular interactions and regulatory networks in CRC. Additionally, "Peptidase activity", "Enzyme binding", and "Signaling receptor activator activity" denote potential alterations in enzymatic functions and proteolytic processes, which may contribute to disease progression and cellular microenvironment dynamics.

The biological processes (GO | BP) (figure 4.13) unveiled are predominantly centered around transcriptional regulation and angiogenesis, both of which play crucial roles in tumorigenesis and cancer progression. The identified terms, such as "positive regulation of transcription by RNA polymerase II", highlight potential transcriptional dysregulation. Additionally, the processes "regulation of angiogenesis" and "positive regulation of angiogenesis" emphasize alterations in blood vessel formation, a key factor in tumor growth and metasta-

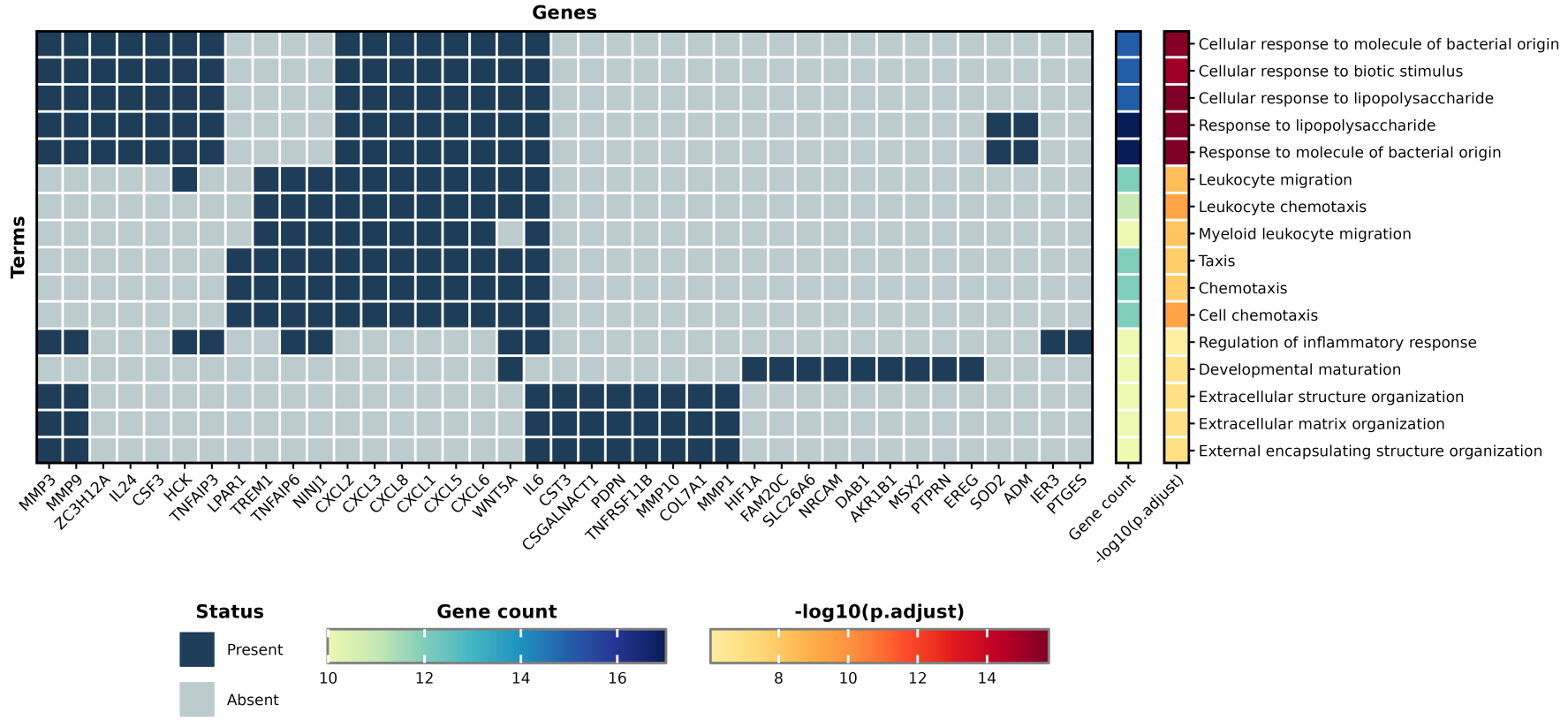


Figure 4.11 GO term enrichment analysis in the WNTiCAF2 cluster. The heatmap presents significantly enriched GO biological processes associated with the WNTiCAF2 cluster. Rows represent genes, and columns correspond to enriched GO terms. Dark blue cells indicate gene-term associations, while light gray cells represent absence. The top bar denotes gene count per term, and the color gradient indicates the statistical significance (-log10 adjusted p-value) of enrichment. Key enriched pathways include immune response regulation, extracellular matrix organization, chemotaxis, and leukocyte migration, highlighting the inflammatory role of WNTiCAF2 fibroblasts.

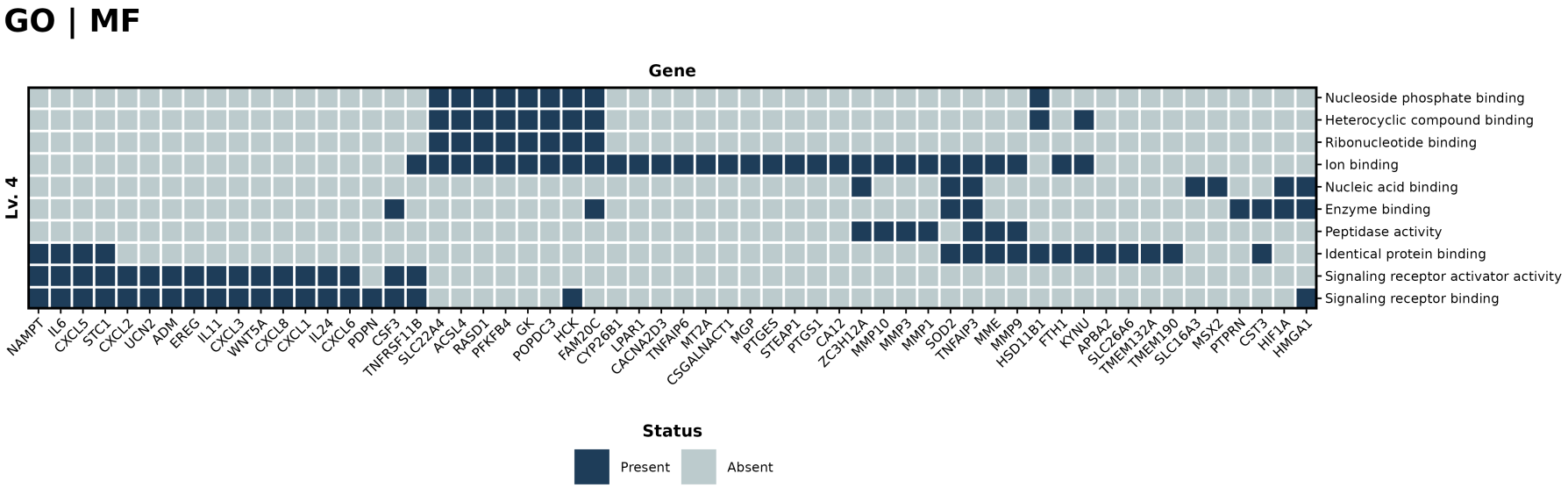


Figure 4.12 Enriched molecular functions (MF) in the WNTiCAF2 cluster. The heatmap presents significantly enriched GO molecular function terms associated with the WNTiCAF2 cluster. Rows represent genes, while columns correspond to molecular functions. Dark blue cells indicate gene-term associations, while light gray cells represent absence. Key enriched functions include nucleotide binding, signaling receptor activity, enzyme activation, and protein binding, highlighting the functional properties of WNTiCAF2 fibroblasts.

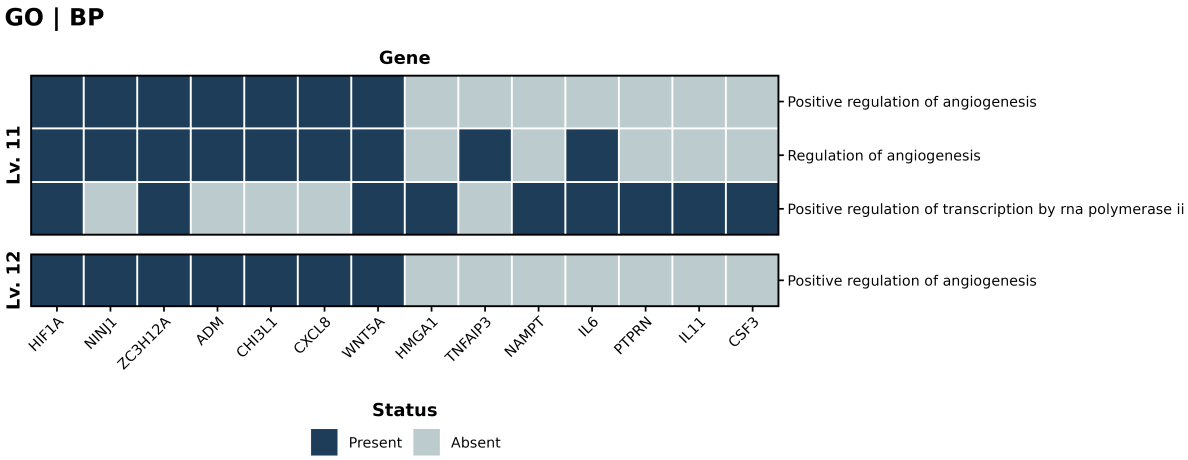


Figure 4.13 Enriched biological processes in the WNTiCAF2 cluster. The heatmap presents significantly enriched GO biological processes associated with the WNTiCAF2 cluster. Rows represent genes, while columns correspond to enriched biological processes. Dark blue cells indicate gene-term associations, while light gray cells represent absence. Key enriched pathways include the regulation of angiogenesis and transcriptional control, suggesting a role for WNTiCAF2 fibroblasts in vascular remodeling and gene expression regulation.

sis. Collectively, these results provide insights into the molecular mechanisms potentially disrupted in CRC and offer a view for further investigations.

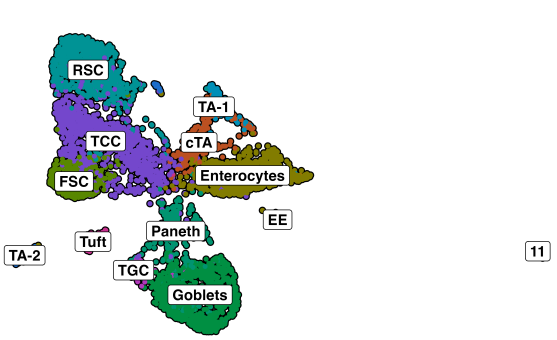
4.3.3 Characterization of epithelial cells in the TME

In addition to fibroblast cells, we identified and characterized 13 distinct epithelial cell subtypes: Tumor Cycling Cells (TCC), Revival Stem Cells (RSC), Fetal Stem Cells (FSC), Paneth, Goblets, Enterocytes, Transit amplifying cells (TA), Tumor Goblet Cells (TGC), Tuft cells, and Enteroendocrine cells (EE), shown as a UMAP in figure 4.14. A dot plot displays significant marker genes that helped in the characterization of these epithelial cells. This provides a comprehensive view of the gene expression profiles that define each epithelial subpopulation.

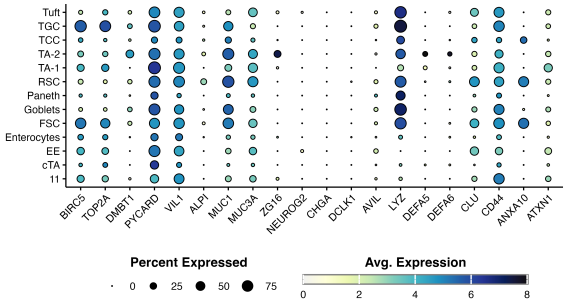
4.3.4 TME cellular interactions involving WNT5A-expressing CAFs

Utilizing the LIANA and NicheNet packages, we investigated CCC within the TME. Our analysis revealed intricate intercellular signaling networks involving multiple ligand-receptor pairs between fibroblast, epithelial, and myeloid cell populations.

The figure 4.15 showcases a network visualization of predicted cellular interactions among the global cell types identified in our study, as inferred by the Liana computational framework. Each node in the network represents a different cell type, and edges between nodes symbolize



(a) UMAP visualization of epithelial cell clusters, highlighting distinct populations such as RSC, TCC, FSC, enterocytes, Paneth cells, Goblets, and tuft cells.



(b) Dot plot showing significant marker expression for each epithelial cluster. Dot size represents the percentage of cells expressing the marker, while color intensity indicates average expression levels.

Figure 4.14 Characterization of stratified epithelial cell clusters.

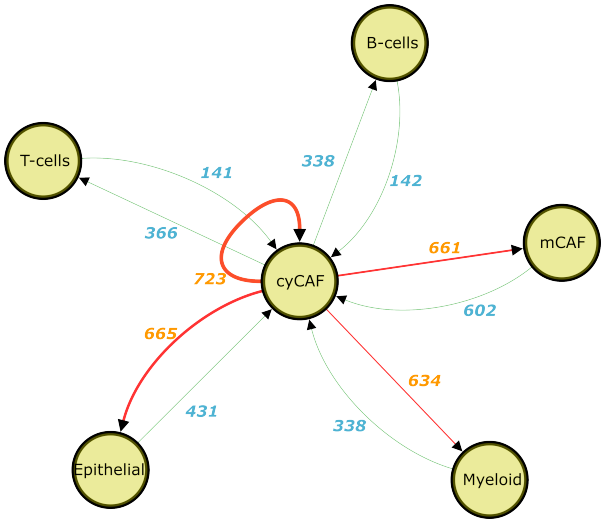


Figure 4.15 Predicted cell-cell interaction network across global cell types. Network representation of ligand-receptor interactions predicted by Liana, illustrating communication between cyCAF and other major cell types, including epithelial cells, T cells, B cells, myeloid cells, and mCAF. Edge thickness and color indicate interaction strength, with red representing strong interactions and blue indicating weaker ones. This analysis provides insights into the signaling dynamics shaping the cellular microenvironment.

predicted cellular interactions, with their thickness and colors corresponding to the total number of interactions predicted. This network diagram provides a global overview of potential CCC patterns within the tissue, highlighting the intricate interplay among various cell types that collectively contribute to tissue function and homeostasis. In the network diagram of figure 4.15, significant crosstalks between the iCAF cells, epithelial cells, Myeloid cells, myCAF cells, and WNTiCAF cells themselves are particularly notable. They represent potential lines of communication or interactions, suggesting these cell populations may be involved in coordinated cellular functions or regulatory mechanisms within the tissue microenvironment.

The figure 4.16 showcases a Circos plot that represents the top 10 ranked interactions (also filtered by the LIANA consensus rank) where WNTiCAF cells serve as ligand-producing cells. The plot displays these interactions with the following cell types of interest: Macrophages, CD4+ T-cells, CD8+ T-cells, Neutrophils, Monocytes, WNTiCAF, iCAF, myCAF, RSC, HSC, FSC, and TCC. Each segment on the outer circle of the Circos plot corresponds to one of these cell types. The inner part of the circle contains colored ribbons (according to the receptor-producing cell type color), each connecting the two cell types. The width of these ribbons at each end indicates the strength of the ligand-receptor interaction between the corresponding cell types, with thicker ribbons denoting stronger interactions. Next to the tip of each ribbons, a colored box indicates the corresponding LIANA consensus rank.

The resulted circos figure 4.16 Circos plot allows for an efficient and visually intuitive representation of complex interaction networks. It also provides an insightful overview of the top interactions involving WNTiCAF cells, highlighting their potential role in coordinating CCC within the studied biological system. This global view of the WNTiCAF cell interactions can inform future investigations into the functional implications of these cellular crosstalks.

The analysis conducted using NicheNet revealed the following receptors to bind with WNT5A as a ligand from the identified WNTiCAF cluster: LRP6, TFRC, SCARB2, RYK, M6PR, LRP5, LDLR, IGF2R, FZD7, FZD5, FZD1, FRZB, ANTXR1, DAB2, PTPRK, and EGFR. In contrast, Liana detected these receptors: FZD1, FZD4, FZD5, FZD7, FZD8, RYK, LRP6, LRP5, PTPRK, ANTXR1, LDLR, ROR1, ROR2, PTK7, MCAM. A comparative analysis of the results obtained from both tools was performed to establish a consensus of the detected receptors.

The overlapping receptors identified by both NicheNet and Liana include: FZD1, FZD5, FZD7, RYK, LRP6, LRP5, PTPRK, ANTXR1, and LDLR. These consensus receptors signify potential targets for further exploration in the context of proteins of interest.

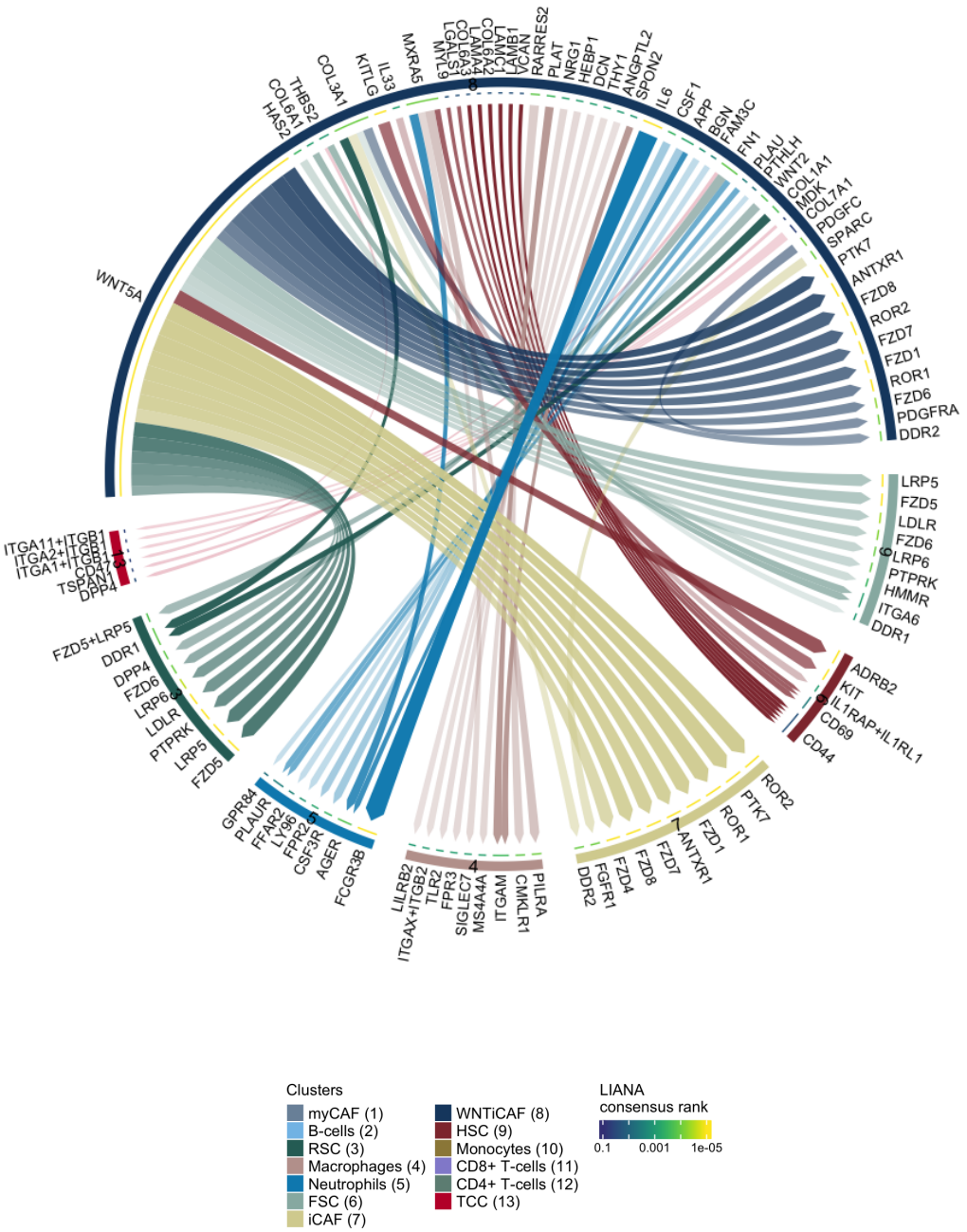


Figure 4.16 Circos plot of top-ranked ligand-receptor interactions involving WNTiCAF cells. The plot visualizes the top 10 ranked ligand-receptor interactions where WNTiCAF cells act as either ligands or receptors, interacting with various immune and stromal cell types, including macrophages, CD4+ T cells, CD8+ T cells, neutrophils, monocytes, myCAF, iCAF, RSC, HSC, FSC, and TCC. Edge thickness represents interaction strength, while colors indicate different cell clusters. This analysis provides insights into the signaling network underlying cellular crosstalk in the Tumor microenvironment.

FZD1, FZD5, and FZD7 belong to the Frizzled family of proteins, which serve as 7-transmembrane receptors in the Wnt signaling pathway [248]. These receptors are crucial for various cellular activities such as cell differentiation, growth, and migration. Although the cysteine-rich domains (CRDs) of FZD5, FZD7, and FZD8 have a similar α-helical dimer structure, their peptide-binding regions display differences in amino acid composition [249]. FZD1, FZD5, and FZD7 are involved in synapse formation and plasticity, and studies have shown that their expression levels decrease in the hippocampus during the early stages of Alzheimer's disease [250]. Researchers have observed FZD5 methylation in the early stages of acute myeloid leukemia (AML) development. FZD5 antibodies have been employed to inhibit receptor activity in pancreatic and CRC cell lines, suggesting that targeting FZD5 could be a potential preventive measure for AML and other cancers [251]. FZD7 has been linked to potential therapeutic targets for CRC [252].

The RYK gene encodes a receptor-like tyrosine kinase, which functions as a receptor for Wnt proteins required for cell-fate determination, axon guidance, and neurite outgrowth in different organisms [253]. While RYK, as a Wnt receptor, has been shown to interact with noncanonical Wnt ligand Wnt5a in gastric cancer [254], more studies are needed to elucidate the specific role of the RYK gene and its interactions with the Wnt signaling pathway in the context of colorectal cancer.

LRP5 and LRP6 are closely related low-density lipoprotein receptor-related proteins, which are crucial for mediating Wnt signaling pathways [255]. As co-receptors for canonical Wnt ligands, they both activate the Wnt/β-catenin signaling pathway but exhibit unique characteristics by interacting with distinct protein partners [255]. In the context of CRC, LRP5 regulation occurs through promoter methylation, while LRP6 is regulated via post-translational modifications involving protein ubiquitination and degradation. CRC displaying high TRAP1 expression are marked by the concurrent upregulation of active β-Catenin, LRP5, and LRP6 [256]. Additionally, LRP5 has been implicated in the promotion of cancer stem cell properties and resistance to chemotherapy in CRC [257]. While LRP6 knockdown in CRC cell lines with mutations in downstream Wnt signaling genes did not significantly impact Wnt signaling activity, it notably reduced Wnt signaling activity in esophageal squamous cell carcinoma cell lines lacking Wnt signaling-related mutations [258]. Research has shown that using single-domain antibody fragments to target LRP5/6 can effectively mitigate the development of intestinal cancers [259].

Protein tyrosine phosphatase receptor-type kappa (PTPRK) is a tumor suppressor that

also functions as an inhibitor of Wnt signaling in human cancer cells, as it has been shown in the Spemann organizer of Xenopus embryos. It operates through the transmembrane E3 ubiquitin ligase ZNRF3, which is a negative regulator of Wnt signaling that facilitates the degradation of Wnt receptors [260]. PTPRK dephosphorylates and exposes the 4Y motif, which promotes the internalization and lysosomal targeting of ZNRF3 and Wnt receptors, leading to a reduction in Wnt signaling. Conversely, an unidentified tyrosine kinase counteracts PTPRK by phosphorylating the 4Y motif, impairing ZNRF3/Wnt receptor internalization, and ultimately increasing Wnt signaling [261]. In CRC, the PTPRK-RSPO3 (P:R) fusion is a well-known structural variation that activates Wnt signaling and promotes tumorigenesis [262]. Active studies investigates therapies targeting the Wnt pathway for tumors with RSPO2/3 fusions. CRC with PTPRK-RSPO3 fusions have demonstrated a response to RSPO3-taxane treatment [263].

ANTXR1, also known as Anthrax Toxin Receptor 1, is a type I membrane protein that has a role in maintaining extracellular matrix balance, promoting angiogenesis, and stimulating cell growth [264]. It has been shown to regulate the reactivation of fetal hemoglobin (HbF) in various cell types by interacting with LRP6, facilitating the nuclear entry of β-catenin, and activating the Wnt/β-catenin signaling pathway [265]. Regarding CRC, ANTXR1 has been investigated as a potential biomarker for the detection and prognosis of this disease [266].

Members of the low-density lipoprotein receptor (LDLR) family, such as LRP5 and LRP6, are involved in a multitude of physiological processes, including lipid metabolism and oxidation. LRP6, a crucial effector of the canonical Wnt signaling pathway, has been associated with an array of diseases, including cancer [267]. The extracellular domain of LRP6 has been demonstrated to interact with Wnt-1 and associate with Frizzled (Fz) in a Wnt-dependent fashion [268]. Another LDLR family member, LRP1, has been reported to interact with human frizzled-1 (HFz1) and down-regulate the canonical Wnt signaling pathway. Furthermore, the LDLR family, including LRP5/6, has been shown to function as Wnt co-receptors by interacting with Fz receptors [269].

In conclusion, our study presents a thorough characterization of fibroblast and epithelial cell subtypes and their intercellular communication within the TME. The identified receptors offer promising therapeutic possibilities for CRC, owing to their involvement in diverse signaling pathways and cellular processes related to cancer progression. Further research is needed to validate their potential as effective targets and develop specific inhibitors or therapeutic agents that can modulate their activity in CRC patients. Our findings contribute to

a more profound understanding of the complex cellular heterogeneity and signaling networks in cancer, potentially paving the way for the development of novel therapeutic strategies targeting specific cell subtypes and intercellular communication pathways.

4.4 Hypothesis generation from Single-cell analysis

As scRNA-seq is a relatively new technology, various decisions must be made by the researcher during data processing. One crucial aspect is quality control (QC), where the determination of specific cutoff values for filtering cells with low gene expression can significantly affect the final results. Additionally, the selection of an appropriate normalization method is critical in scRNA-seq analysis, as it can greatly impact the findings. In our study, we utilized log normalization, a classical and reliable approach that stabilizes the variance of expression values and shapes the data into a Gaussian-like distribution. This method effectively mitigates batch effects and reduces the impact of technical variability. However, alternative normalization methods, such as the sctransform method introduced by Hafemeister and Satija [270], could also be explored. This innovative approach eliminates technical variability and confounded biological heterogeneity without the need for heuristic steps, stabilizes variance, and enhances downstream analytical tasks. The choice of normalization method should be tailored to the specific goals and requirements of the study and the dataset characteristics. In our case, we opted for log normalization, as it is a reliable, well-established method that met our study's needs.

The choice of clustering resolution and algorithm greatly influence the analysis of scRNA-seq data, playing a crucial role in exploring the diversity of cell types and states within the dataset. Researchers have proposed various clustering algorithms, including traditional hierarchical methods and innovative deep learning approaches such as scDeepCluster [271], to classify cells based on their gene expression patterns. Resolutions in clustering approaches determine the level of detail for cell clusters, potentially allowing to highlight new cell types while also presenting a risk of excessive clustering with imbalanced clusters. Finding a balance here is crucial when some clusters are close together at lower resolutions, thereby hiding finer differences. The choice of algorithm and resolution directly impacts the resulting quantity and composition of cell clusters, making it imperative to capture rare cell types in conditions like cancer where a small population of cells could be responsible for a disease outcome, while also avoiding artificial division of cell populations into separate biased clusters.

These decisions heavily influence the interpretation of results, such as understanding tissue heterogeneity, identifying novel cell types, due to inherent biases and assumptions in each algorithm. Benchmarking multiple methods is crucial in determining their strengths and limitations across various contexts, as exemplified by the scCCESS study which examined 14 different algorithms or resolutions [272].

The rapid evolution of scRNA-seq has highlighted the importance of accurate cell type labeling. Automation is increasingly used due to the complicated manual annotation, with tools like ScType and scAnnotate leveraging extensive databases and machine learning techniques [273, 274, 275].

However, while automation offers speed and general classification, it is worth recognizing its limitations. For detailed and specific cell type identification, particularly at a fine level, manual labeling remains indispensable, because subtle distinctions and nuances in cell types might be overlooked by automated systems. Additionally, every labeling method, automated or manual, can introduce biases. It is essential to critically assess the accuracy and reliability of chosen methods. While automation can handle wide annotations and provide an initial framework, manual validation and refinement are often required to ensure the highest level of accuracy, especially when exploring novel cell types in context such as cancer research.

LIANA provides a versatile platform for L/R analysis which uses separate methods and resources. Its efficient implementation facilitates intercellular communication studies by integrating with many other tools, such as CellPhoneDB, CellChat, and NATMI. In addition, LIANA's is a flexible toolbox with customizable resources that enhance its capabilities. LIANA's consensus ranking harmonizes prediction scores from these various tools. Furthermore, the integration of OmniPath deepens insights into intercellular communication complexities and strengthens the outcomes [246].

NicheNet, which specifically analyzes the intracellular implications of CCC, is valuable in predicting ligand-target associations and evaluating their associated functional activity. Its well-structured integration with ligand-receptor inference further enhances the applicability of NicheNet, enriching its CCC predictions. The methodology and documentation is extensive, providing useful guidance on various aspects ranging from Seurat analysis to model optimization [276].

However, the capabilities of such tools present crucial challenges due to the variation in methods, scoring, and resources. Such variability frequently complicates direct comparative analyses. These tools, because of their diverse analytical approaches, establish intricate

scoring systems that can be complicated to compare. Researchers find frameworks like LIANA useful for comparing, processing, and achieving consensus results from various tools. These predictions are of great importance as they are closely connected to the accuracy of the underlying models and databases. Relying on such predictions requires awareness of their limitations and possible sources of error. The coverage and accuracy of the databases could be an origin of errors, therefore caution is necessary in interpreting these results. Remembering that bioinformatic predictions are only stepping stones in the research framework, further experimental validations are necessary to confirm both the proposed L/R bindings and validate the study's findings.

Our findings offer novel perspectives regarding the expression and functionality of WNT5A in fibroblast cells, as well as its involvement in CRC. The characterisation of two distinct subtypes of fibroblasts, iCAF and myCAF, based on gene expression analysis emphasizes the importance of investigating fibroblast cells in the context of cancer. Findings on the enriched pathways in the WNTiCAF cells potentially have therapeutic implications and offer targets for cancer treatment. We highlight the necessity for a more profound exploration into the functional functions of these cell subtypes, specifically with regards to their interactions among other cells in the TME. Our identification of iCAF and myCAF, alongside 13 epithelial cell subtypes, explore the functional diversity inherent in these cells. This finding looks for mechanistic impact on tumor progression, metastasis, and therapeutic response. Moreover, our study showcases the potential for innovative therapeutic strategies by targeting distinct cell subtypes or intercellular communication pathways. Focusing future research on the development of targeted therapies that use these pathways or the properties of these cell subtypes might lead to more personalized and effective treatment options for cancer patients, such as the observed crosstalk between iCAF cells and cancer cells facilitated through WNT5A targeting [239].

Biomarkers are a key to analyze disease pathophysiology and molecular signatures, thus directing targeted treatments. KRAS, BRAF, and TP53 mutations are key biomarkers for prognosis and monitoring cancer states [277]. Our study identified marker genes and molecular signatures that could be further investigated as novel predictive markers. Analyzing separate cellular subtypes within the TME opens opportunities for early cancer diagnosis, tracking disease progression, and categorizing patients for personalized precision medicine approaches. Understanding more of these complex cellular interplays in the TME can be facilitated by investigating using a different approach such as proteomics [278]. Meanwhile,

immunotherapies hold promise in the use of biomarkers allowing for tailored medicinal solutions. Integrating different approaches, interdisciplinary cooperation, and clinical application, would improve the understanding of cancer biology while proposing innovative therapeutic approaches that hold promise for patient outcomes.

Contributions

This work was conducted under the primary supervision of Prof. Dr. Reinhard Schneider and Dr. Venkata Satagopam at the Luxembourg Center for Systems Biomedicine (LCSB), in collaboration with the Molecular Disease Mechanisms (MDM) group at the Faculty of Science, Technology and Medicine (FSTM). I, Mathias Galati developed the computational framework, implemented the analytical methods, and performed the bioinformatics analyses of the single-cell RNA sequencing data.

The MDM group collaboration is led by Dr. Elisabeth Letellier, who provided essential immunological expertise and guidance. Dr. Eric Koncina and Dr. Anthoula Gaigneau contributed their specialized bioinformatics expertise to the development and optimization of the scRNA-seq analysis pipeline, ensuring methodological rigor and scientific relevance. Their input was particularly valuable in establishing robust analytical strategies and validating the computational approaches. This work was conducted in parallel with Dr. Rubens Begaj's doctoral research, and his expertise served in particular in cell type fine stratification and the identification of relevant cellular and molecular targets. The integration of his insights enhanced the biological interpretation of the computational findings.

The successful completion of this work was made possible through the continuous guidance of Prof. Dr. Schneider and Dr. Satagopam, who provided strategic direction and ensured the overall scientific quality of the research.

Chapter 5

Molecular mechanisms linking Dietary fibers to gut inflammation

5.1 Dietary fibers, Gut microbiota, and Molecular inflammation

5.1.1 Introduction to Dietary Fiber and Health

Dietary fibers (DF) are plant-derived carbohydrates that are resistant to digestion by human enzymes. They are two main types [279]: soluble and insoluble fibers. Soluble dissolves in water and help lower blood cholesterol and glucose levels, which common sources including oats, beans, and apples. Insoluble fiber which do not dissolve in water, promote the movement of material through the digestive system and increase stool bulk. These are typically found in whole-wheat flour, nuts, beans, vegetables such as potatoes.

Additionally, resistant starch [280], a specific type of starch that resists digestion and reaches the colon, is fermented by the gut microbiota, and functions similarly to soluble fiber.

DF modulates the gut microbiota, impacting overall health. Different types of DF have distinct effects:

- Soluble fiber:
  - Inulin is known to improve metabolic function and regulate intestinal immunity by being fermentated in the colon, producing SCFA such as acetate, butyrate and propionate. These SCFAs contribute to regulate immune cell differentiation and host immunity. Inulin has shown potential benefits in preventing diseases such

as inflammatory bowel disease, type 2 diabetes and some cancers. [281]

  - Pectin helps reducing glucose absorption and cholesterol levels, offering benefits for disease like diabetes. Its anti-inflammatory properties are beneficial for conditions such as infant regurgitation and diarrhea [282].

- Insoluble fiber:

  - Cellulose supports gut health by promoting intestinal barrier function and providing anti-inflammatory benefits [283].
  - Resistant Starch which is not digested in the small intestine but is fermented in the colon, promoting the growth of beneficial bacteria and enhancing SCFA production.

The effects of DF on health vary depending on the type of fiber and individual microbiota composition. This suggests that personalized dietary interventions may be more effective than a one-size-fits-all approach. For instance, the mediterranean diet, rich in grains, vegetables and fruits, provides high levels of DFs and significantly lowers plasma LDL cholesterol levels, reducing the risk of cardiovascular diseases [284].

5.1.2 Interindividual Variability in Response to Dietary Fibers

The response to DF is limited by interindividual variability, which includes genetic factors, gut microbial composition and lifestyle factors. Genetic variations can affect the enzymes involved in the food metabolism, while the microbiome plays a critical role in fermenting the fiber into SCFA impacting glucose metabolism and inflammation. Lifestyle factors such as diet and physical activity also contribute to this variability.

Nutrigenetics explores how genetic variants influence individual responses to DF. Specific Single-nucleotide polymorphism (SNP) can affect the efficiency of fiber metabolism and the SCFA production by gut bacteria. The Wnt pathways is on example of a genetic component reviewed in relation to DF metabolism [285].

5.1.3 Interaction between Genetics, Microbiome and Dietary Fibers

The holobiont concept, introduced by Lynn Margulis, redefines organisms as dynamic networks comprising a host and its associated microbiota (bacteria, fungi, viruses, and microbial eukaryotes) [286, 287]. This framework challenges traditional evolutionary models by empha-

sizing symbiotic relationships as drivers of adaptation and speciation [288]. The hologenome theory further posits that holobionts act as units of selection, with host-microbiota genetic covariance influencing evolutionary trajectories [287, 289]. Methodologically, advances in multi-omics and network analysis have revealed intricate microbial interactions, such as microalgae shaping phyllosphere communities in plants through biofilm-mediated hubs [290, 289]. However, debates persist about whether holobionts represent true evolutionary individuals or transient ecological communities, particularly given horizontal gene transfer and environmental microbial acquisition [288]. Clinically, dysbiosis in human gut microbiomes has been linked to inflammatory diseases, underscoring the translational potential of holobiont research [291]. These insights highlight the holobiont as a critical lens for studying host-microbe coevolution, ecosystem dynamics, and precision medicine, while emphasizing the need for integrated computational approaches to analyze heterogeneous biological datasets [290, 288, 289].

5.1.4 Impact of Dietary Fibers on inflammation and oxidative stress

DF contributes to reducing inflammation and oxidative stress through various mechanisms: Pectin and Toll-like receptors (TLRs) [292]: Pectin interacts with immune receptors such as TLR2, inhibiting the pro-inflammatory TLR2-TLR1 pathway while maintaining the tolerogenic TLR2-TLR6 pathway. Fermentation and SCFA production [293]: Each SCFAs from the gut microbiota is produced through specific bacterial metabolic pathways, and they play key roles in reducing the inflammation.

- Acetate is the most abundant SCFA in the colon and is produced through a process called acetogenesis, which involves a large variety of bacteria. Acetate is formed from acetyl-CoA, a product of glycolysis, via pathways such as phosphotransacetylase-acetate kinase.

- Propionate is primarly produced through the succinate pathway, which is prevalent in Bacteroidetes and some Firmicutes species, and through the propanediol pathway, which is characteristic by certain Lachnospiraceae.

- Butyrate is produced by Firmicutes, including specific Clostridium species, and synthesized through the conversion of acetyl-CoA to butyryl-CoA. Butyrate has been shown to reduce inflammation through several mecanisms such as, promoting the differentiation

of naive T cells into regulatory T cells (Tregs) by enhancing histone acetylation at specific gene loci such as Foxp3 [294]. Additionally, Butyrate contribute to the health of the intestinal barrier as an energy source for colonocytes [295].

SCFAs can also inhibit inflammatory pathways by activating G protein-coupled receptors (GPCRs) such as GPR41 and GPR43 on immune cells. This activation leads to a reduction in the production of pro-inflammatory cytokines and modulation of immune cell activity.

Reduction of systemic inflammatory markers: Cytokines, essential small proteins for cell signaling, are important in the immune response and are classified into several types based on their specific roles and sources [296]. Interleukins (IL) for example regulate immune and inflammatory responses: IL-1 promotes inflammation, fever, and the acute phase response. While IL-2 stimulates T cell proliferation. IL-6 is implicated in fever and acute phase protein production. Tumor necrosis factor (TNF) such as TNF-α, produced by macrophages, induce fever, apoptotic cell death, and inflammation. With TNF-α being particularly involved in systemic inflammation and the death of certain tumor cells. In term of functions and mechanisms, cytokines bind to specific receptors on target cells, initiating intracellular signaling cascades that modify gene expression and cellular activity. They are a key player in immune regulation by balancing humoral and cell-mediated immune responses and influencing immune cell differentiation and growth. Additionally, they can act both as pro- and anti- inflammatory agents that can initiate and resolve inflammatory processes.

Interferons (INF) are primarly involved in antiviral responses and immune system modulation, with INF-α and IFN-β inhibiting viral replication and IFN-γ activating macrophages and enhancing antigen presentation.

Chemokines, another type of cytokine type, direct immune cells to sites of infection or injury. For instance, CCL2 (MCP-1) recruits monocytes to inflammation sites, and CXCL8 (IL-8) attracts neutrophils [297]. Colony-stimulating factors (CSF), like G-CSF and GM-CSF, stimulate blood cell production, particularly granulocytes and macrophages.

SCFAs exhibit strong anti-inflammatory properties, inhibiting the production of pro-inflammatory cytokines such as IL-6 and TNF-α. Some studies have shown an inverse relationship between DF intake and inflammatory markers, notably IL-6, TNF-α receptor 2 (TNF-α-R2), and C-reactive protein CRP) [298], eventhough CRP association is less consistent than the other two markers [299].

Moreover, the fermentation process reduces gut permeability, which decrease the translocation of lipopolysaccharides from the gut into the bloodstream, a process that would otherwise trigger systemic inflammation through cytokines like IL-6.

In adolescents, higher cereal and grain fiber intake is associated with lower leptin levels, potentially contributing to reduced chronic inflammation [298].

Consumption of DF is associated with improved oxidative stress markers, thus linked to increased activity of antioxydant enzymes such as glutathione peroxidase (GPx), which helps maintain cellular redox homeostasis [300]. Genetic variations as SNPs can have both prooxidant and antioxidant roles [301]. XDH, CYBA, CYP1A1, PTGS2, NOS and MAO have prooxidant involvment. In contrast to SOD, CAT, GPX, GSS, GLUL, GSR, GSTM1, GSTM5, GSTP1, TXN and HMOX1 that help to maintain the oxidative balance. Mutations affecting enzimes or proteins involved in the dietary antioxidant uptake can impact the metabolism function, such as the MnSOD gene polymorphism (Val16Ala) that affects mitochondrial targeting the oxidative stress [302].

Various pathways are involved in oxidative stress. NF-κB signaling pathway is activated in response to oxidative stress and regulates cellular proliferation and apoptosis. it can induce the expresion of antioxidative enzymes such as SOD2 and GPX4. Pi2K/AKT pathway modulates vascular tone by regulating nitric oxide production through endothelial nitric oxide synthase (eNOS) phosphorylation. Other pathways involved iclude ferroptotic, apoptotic, FoxO, and ErbB, which regulate cellular responses to oxidative stress by promoting the expression of antioxidative enzymes such as SOD2 and GPX1 [301].

5.1.5 Polygenic Risk Score and Personalized Nutrition

PRS for disease prediction and risk assessment

Polygenic risk score (PRS) aggregate genome-wide genetic variants to quantify an individual's genetic predisposition to complex diseases, enabling enhanced risk stratification and personalized preventive care [303, 304]. By integrating PRS with conventional risk factors, clinicians can identify high-risk subgroups more accurately as exemplified for prediction of first-onset cardiovascular disease [305]. This genomic stratification supports tailored interventions, including dietary modifications, to mitigate disease risk. Evidence highlights the role of optimal DF intake (25–29g/day) in reducing cardiometabolic and colorectal cancer risks, with fiber modulating gut microbiota and improving glycemic control [306]. Gene-diet interaction

studies demonstrate that genetic susceptibility to conditions like obesity can be attenuated by high-fiber diets, emphasizing the potential for PRS-guided nutritional strategies [307]. For example, adherence to Mediterranean diets neutralizes diabetes risk in genetically predisposed individuals [307], while personalized nutrition counseling improves outcomes in chronic disease cohorts [308]. However, PRS clinical translation faces challenges, including reduced predictive accuracy in non-European populations and the need for standardized protocols to integrate genetic, clinical, and environmental data [304]. Despite these limitations, combining PRS with DF optimization represents a promising avenue for precision prevention, particularly in multi-omics frameworks that address genetic heterogeneity and lifestyle interactions [309, 310].

Personalized nutrition as a preventive strategy

Personalized medicine and research is centered around stratifying individuals into responders and non responders to DF involving the different

DF intake plays a significant role in reducing inflammation and oxidative stress, contributing the health benefits. Reduction of inflammation has been linked with modulation of gut microbiota with the influence of rich fiber diet, where SCFA like butyrate contribute with their anti-inflammatory properties. For instance, a study have show these effect using high-fat diet-fed rats and a diet focused on oats and tartary buckwheat [311]. Increased dietary fiber is significantly associated with reduced systemic immune and inflammatory biomarkers, such as SII, SIRI and hs-CRP [312]. The risk of chronic disease can be reduced with high fiber content by promoting a healthy gut microbiome and providing antioxidants.

5.2 Strategies and multi-omics integration in the Holobiont

5.2.1 Methods for preventive inflammation monitoring

The Luxfico study is a randomized, cross-over study including 30 adult individuals. The cohort's mean age is 33.83 (range 23 to 46) and BMI is around 23.04 kg/m^2. All patients were free from clinically diagnosed chronic conditions and inflammatory diseases at enrollment.

The study included two controlled one-week dietary interventions separated by a washout period to minimize carryover effects. Dietary Interventions consisted of two distinct DF regimes:

High-DF diet ranging from 33,4 to 51,5 g/day (average intake around 41.17 g/day).

Low-DF diet ranging from 17,3 to 25 g/day (average intake of 20.31 g/day).

Blood, urine, and stool samples were collected at baseline, during, and after each dietary intervention period. WGS was performed on polymorphonuclear cells (PBMCs), and previously identified, from the literature, SNP related to dietary fiberDF metabolism and SCFA pathways pathways were analyzed. These SNP were examined both individually and through the construction of Polygenic risk score (PRS) to explore their potential contributions to inflammatory (*e.g.* IL-1β, TNF-α, IFN-γ, IL-4, IL-6, IL-10, IL-12p70, IL-22) and oxidative stress markers (*e.g.* DNA/RNA damage, malondialdehyde (MDA), F2-isoprostanes, and total antioxidant capacity via FRAP and ABTS assays).

The Luxfico study employs a multi-dimensional approach to investigate the complex interactions between genetics, nutrition, and gut inflammation.

Genomic analysis The focus is on identifying SNP in molecular transporters such as monocarboxylate transporters (MCTs) and sodium-coupled monocarboxylate transporters (SMCTs), as well as in G-protein-coupled receptors (GPCRs) that influence SCFA metabolism. PRS scores are constructed to evaluate the cumulative effects of theses targeted SNP on inflammatory markers and oxidative stress biomarkers, including MDA and F2-isoprostanes.

Immuno-profiling The aim is to model cytokines signature and immune cell profiles to explore the links between diet, genetic predispositions and inflammation.

Multi-omics integration Building upon the previous analyses, the integration of the microbiome dataset would allow to investigates how bacterial profiles would influence the phenotype. It has been reported in the literature that specific microbes are associated with the metabolism of some SCFA [99], such as *Prevotella* and propionate production, and *Roseburia* with butyrate production. The study considers to investigate gene-microbiome interactions such as the impact of vitamin D receptor (VDR) variants on *Parabacteroides* abundance and lactase (LCT) genes variants on *Bifidobacteria* levels.

The different level analyses integrations enable a progressive comprehension of the genetic interaction with the direct key aspects of the SCFA metabolism. This approach allows for a stratified investigation which identify and predict genetic-environment interactions and the manifestation of the phenotype.

The deliverables of the Luxfico study include a predictive inflammatory score that could be applied in both clinical and research settings. The insights gained could reflect on informed personalized dietary recommendations tailored to patients genetic, microbiome and dietary profiles. Which would contribute to more tailored and effective strategies for managing patient inflammation and oxidative stress before any more complex complications.

5.2.2 Genomic analysis of functional pathways in SCFAs metabolism

The genomic analysis will be conducted on WGS data obtained from peripheral blood mononuclear cells (PBMCs), using an Illumina NovaSeq 6000 platform. For this analysis, we selected PCR-free library preparation to minimize potential amplification biases and ensure better depth of coverage and genotype quality of distribution. The sequencing will target a mean coverage around 30x, which is expected to provide more than 99.77% sensitivity for homozygous and more than 99.82% for heterozygous SNP detection, with positive predictive values exceeding 99.98% and 99.07%, respectively.

The analysis pipeline will consist of several key steps. While analysis is planned, tool selections are flexible regarding reproducibility and facility to use. Initially, raw sequencing data (in fastq) will undergo QC assessment and preprocessing, using FastQC [313], and Cutadapt [314]. Following alignment to the GRCh38 reference genome with Bowtie2 [315], variant calling will be performed with a focus on previously identified SNP related to DF metabolism and SCFA pathways, using tools such as GATK's HaplotypeCaller, FreeBayes [316]. Of particular interest are variants in genes encoding MCTs, SMCTs, and GPCRs that influence SCFA metabolism. Detected variants will be filtered based on their quality score, and interpreted using variant functional predictor tools such as SnpEff [317].

The statistical analysis framework will employ a composed approach of both traditional genetic association methods and ML such as Random Forest for quantifying feature contribution to the phenotype RF example that finds SNP [318]. This dual analytical strategy enable both hypothesis testing and discovery of complex genetic patterns that may influence individual responses to DF interventions.

Variant Selection and Curation For the construction of PRS, we will employ a systematic approach to variant selection and weighting. Variants will be curated from established databases including OMIM, ClinVar, and GWAS catalogs, with a specific focus on SNP associated with SCFA metabolim, transport genes, SCFA-signaling GPCR and metabolic disorders, inflammatory pathway.

Scoring system and Weight Assignments A scoring system has been designed to incorporate both the presence of variants and their reported effect sizes from previous studies. We have developed a novel classification scale ranging from -3 to +7, where negative scores indicate "benign, experimentally verified" variants and positive scores indicate "potentially verified" variants, with +7 representing the strongest evidence for impact on metabolic disease.

PRS calculation Several approaches exist for calculation PRS, such as the Classical Clumping and Thresholding (C+T), the Shrinkage methods and Bayesian Approaches. The equation below represent C+T method which perform relatively well while keeping algorithmic simplicity simplicity [319].

$$PRS_j = \sum_{i=0}^{M} x_{i,j}\beta_i \tag{5.1}$$

- β_i: effect size for variant i
- $x_{i,j}$: the effect allele count for sample j at variant i
- M: the number of variants

ML Refinement The previously described targeted approach will be complemented by PCA to capture major axes of genetic variation while maintaining computational integrity. Additional feature selection will be able to explore additional variants, such as elastic net regression, which combines L1 and L2 regularization, to identify sparse sets of predictive variants [320]. Complex genetic interactions will be modeled using gradient boosted decision trees (XGBoost) and random forests, which can capture non-linear relationships interactions between variants [321]. These methods will be applied within a cross-validation framework to ensure robust identification of interaction effects.

Machine Learning for refining PRS score

To further refine the C+T PRS score and validate these initial SNP effects, we use ML feature selection and importance metrics.

Running the ML models Train the prediction model (such as gradient boosting machine such as XGBoost, or a random forest) using the selected variants as input features and the target phenotype, which include inflammatory marker levels as the output variable.

Extract Features importance From the trained ML model, extract feature importance measures. Variants with higher importance contribute more to the predictive power of the model and are likely to be biologically relevant.

Weight adjustment based on ML results If a variant scored +3 (moderate evidence) but the ML model indicates high importance, its score might be incremented to +4 or +5. Conversely, a variant initially scored +5 but deemed unimportant by ML feature importance might be downgraded to +3 or +2.

A possible formula for recalibrating scores could be:

$$S_i' = S_i \times f(I_i) \tag{5.2}$$

where S_i is the original score, and $f(I_i)$ is a function of ML -derived importance I_i. For example, if the model ranks variant i in the top 10% of importance, $f(I_i)$ could be > 1 to increase its weight.

Interaction effects and Composite Feature ML methods can highlight interaction effects between variants. If certain variants only show importance when combined, we can create composite features. For example, if variants A and B together are more predictive than either alone, we might create a combined score component $S_{A,B}$ and add it to the final PRS. This step captures gene-gene interaction effects.

The PRS is not static. We continuously refine it using new data and insights. After adjusting weights based on ML outputs, we:

1. Recalculate PRS for all individuals.
2. Reassess associations between the updated PRS and the phenotypic measures.

3. Validate on an independent dataset or through bootstrapping and cross-validation to ensure that refinements improve predictive performance and are not overfitted to a single cohort.

Model development will follow a rigorous cross-validation protocol such as k-fold cross-validation to prevent overfitting and ensure generalization, or leave-one-out cross-validation. Performance evaluation will utilize multiple metrics such as R^2, Mean Squared Error, and concordance index for continuous traits with particular attention to the stability of predictions across different cross-validation folds.

This integrated approach ensures that the PRS is biologically relevant, enriched by various cured databases, and dynamically modeled through data-driven ML insights (immunological datasets). The result will reflect to a more nuanced and potentially more accurate risk stratification tool for metabolic and inflammatory phenotypes.

Immuno-profiling for inflammation assessment

The objective is to determine how genetic variation (SNP or PRS), from previous step of the analysis, modulates the immune and oxidative stress responses measured by the immuno-profiling aspect of the study. The statistical approaches could start with a first modeling using mixed-effects models with genetic covariates. However, the large number of variants to include would represent a significant analysis burden and a large sample size is required to detect gene-by-environment interactions.

$$\begin{aligned} \text{Biomarker}_{ij} = \gamma_0 &+ \gamma_1 \text{PRS}_i + \gamma_2 \text{Time}_{ij} + \gamma_3 \text{Treatment}_{ij} \\ &+ \gamma_4 (\text{PRS}_i \times \text{Treatment}_{ij}) + \gamma_5 \text{Sequence}_i \\ &+ \gamma_6 \text{Period}_j + (1 + \text{Time} \mid \text{Patient}_i) + \epsilon_{ij} \end{aligned} \tag{5.3}$$

Variables/Terms Explained:

Biomarker_{ij} The level or concentration of a particular inflammatory or oxidative stress biomarker for individual *i* at time point *j*. For example, IL-6 concentration at the second follow-up visit for participant 12. Data Type would be continuous (often log-transformed for normality).

γ_0 The fixed-effect intercept (population-level intercept). It represents the expected biomarker level when all other predictors (Time, Treatment) are at their reference or baseline categories (or zero, in the case of continuous predictors).

$\gamma_1 \mathbf{PRS}_i$ A fixed-effect term for the PRS of individual i, calculated previously. γ_1 indicates how much the biomarker level changes for each unit increase in the PRS, controlling for time, treatment, and random effects.

$\gamma_2 \mathbf{Time}_{ij}$ represent either a continuous time variable or discrete visits (V1, V2, V3 with V1 as baseline reference).

$\gamma_3 \mathbf{Treatment}_{ij}$ A fixed effect for the dietary intervention or treatment condition at time j for patient i. If *Treatment* is coded as a categorical variable (such as 2 = High Fiber 1 = Low Fiber 0 = Control), γ_2 represents the difference in expected biomarker level under the active treatment compared to the reference condition.

$\gamma_4 (\mathbf{PRS}_i \times \mathbf{Treatment}_{ij})$ gene-by-diet interaction term, determining whether PRS modifies dietary response.

$\gamma_5 \mathbf{Sequence}_i$ which sequence the participant belongs to (High $\longrightarrow$ Low vs Low $\longrightarrow$ High); can be included as a fixed effect if you want to account for possible systematic differences in the order participants received diets.

$\gamma_6 \mathbf{Period}_j$ included as a fixed effect to adjust for any systematic difference between the first and second time a person undergoes an intervention. Distinguishes between the first time a person receives *any* intervention (Period 1) vs the second time they receive an intervention (Period 2).

$(\mathbf{1 + Time \mid Subject}_i)$ Random effects structure, indicating that each subject i can have: A random intercept (such as a subject-specific baseline shift in biomarker level) or a random slope for *Time* (*i.e.* a subject-specific rate of change over time). This accounts for within-subject correlation of repeated measurements and allows each participant to have a unique trajectory over time.

ϵ_{ij} Residual error term for individual i at time j. Typically assumed to be independent and normally distributed with mean 0 and variance σ^2.

Modeling baseline (V1) as a separate reference time or use baseline-corrected outcomes (*e.g.* change from baseline) can be both valid, but typically we include the raw repeated measures and let the random intercept handle baseline differences.

Depending on the research focus, Sequence or Period may not be included, as some experimental designs can handle these via randomization and washout to minimize carryover. However, for a sensitivity check, it might be good practice to include them as additional fixed effects.

The model does not explicitly require a "washout" variable, timepoint-based modeling inherently captures changes after each washout. However, time-since-diet-coded or washout-coded variables could be encoded if we suspect remaining partial carryover.

With n=40 total and 20 per sequence group, this might constrain power for detecting gene-by-environment (PRS × Treatment) interactions unless effect sizes are moderate or large.

A Bayesian mixed-effects approach is equally valid and may allow for partial pooling and better handling of small sample sizes, especially with prior information about PRS effects.

As a conclusion, the PRS model in a Mixed-Effects Framework fits well with the two-arm, crossover, repeated-measures nature of LUXFICO. It allows to estimate how a genetic predisposition (in the form of a PRS modifies the response to different fiber diets within individuals over time, capturing both fixed (treatment, PRS, interactions) and random (subject-level baseline, subject-level slopes) effects. Minor design-specific adaptations (such as adding sequence or period effects) may be considered to fully respect the crossover structure, but the core model structure is accurate for this experimental setup.

Multi-omics integration for inflammation predictive scoring

The integration of genomic, immunological and microbiome datasets will be done through a systematic multi-layered analytical framework designed to elucidate the complex interactions between host genetics, dietary intervention and inflammatory responses. This integration strategy will implement MO largely covered tools such as DIABLO (Data Integration Analysis for Biomarker discovery using Latent variable approaches for Omics studies)[322] , which enables the identification of MO signatures while accounting for the inherent relationships between different data modalities.

The analytical pipeline will begin with appropriate and validated data processing approach for each layer. Building upon the last 2 previous analyses chapter, microbiome data will be integrated after appropriate normalization/transformation. The framework will adopt a three step approach:

Pairwise relationships Canonical correlation analysis will identify strong associations between features across the different biological layers

Feature selection sparse partial least squares discriminant analysis (sPLA-DA) will identify key features from each dataset that contribute to the discrimination between different dietary responses.

Multi-omics integration Methods such as DIABLO [322] or SNF [323] will help to build comprehensive models to explain the interactions and associations between the different biological layers.

To evaluate the relative contributions of different biological components to the observed phenotypes, we will develop a novel scoring system that incorporates the previous analyses. The idea is to classify patients into 9 distinct profiles based on the following characteristics:

Consumer status based on DF intake (high DF vs Low DF).

Fermenter status derived from microbiota composition and SCFA production capability from the DF.

Metabolizer status determined by genetic variants or metabolic markers in relevant pathways.

$$\text{Gut Inflammation} = \text{Consumer} + \text{Fermenter} + \text{Metabolizer} \tag{5.4}$$

The immune landscape of each profile will be characterized using cytokines expression levels and immune cell compositions, enabling validation of the scoring system's accuracy in predicting inflammatory states.

Patient Stratification Model for precision medicine

By considering the three factors in a binary classification (high = +, low = -), we derive nine possible patient profiles, from health-associated (anti-inflammatory) to pro-inflammatory/metabolic disease-prone states as examplified in table 5.1.

This classification framework provides a structured approach to understanding how diet, microbiota function, and host metabolism collectively shape inflammatory responses. The subsequent analyses will assess how immune parameters (*e.g.* cytokines profiles and immune cell distributions) align with the predicted inflammatory risk across the nine profiles, refining the model's predictive value in stratifying patients based on gut inflammation potential.

Consumer	Fermenter	Metabolizer	
+	+	+	
+	+	-	
+	-	+	healthy →
+	-	-	
-	+	+	
-	+	-	
-	-	+	← unhealthy
-	-	-	

Table 5.1 Stratification of patient profiles based on dietary intake, microbiota fermentation, and metabolic capacity. Patients are classified into nine distinct profiles based on three key factors: Consumer status (DF intake), Fermenter status (microbiota composition and SCFA production), and Metabolizer status (genetic or metabolic markers). Each category is assigned a binary value (+ high, - low), with the first profile representing the most health-associated state and the ninth corresponding to a pro-inflammatory/metabolic disease profile.

Contributions

This work was conducted under the primary supervision of Prof. Dr. Reinhard Schneider and Dr. Venkata Satagopam at the Luxembourg Center for Systems Biomedicine (LCSB), in collaboration with the NutriHealth group at the Luxembourg Institute of Health (LIH) under the framework of the i2TRON Doctoral Training Unit funded by the Luxembourg National Research Fund (FNR).

The research was carried out in close collaboration with Dr. Torsten Bohn's Nutrihealth group at Luxembourg Institute of Health (LIH). Dr. Guilherme Meyers provided essential expertise in human biology, particularly in understanding the intricate relationships between nutrition, biological pathways, and the microbiome. His contributions were instrumental in refining the modeling design through close collaboration with me, Mathias Galati. The in-silico experimental framework developed in this thesis was established within the legal framework agreed upon between LCSB and LIH's research groups, particularly regarding data integration and analysis methodologies.

I, Mathias Galati, am developing the computational methods, implementing the analytical framework, and performing the bioinformatics analyses. The successful integration of nutritional and molecular data is possible through the continuous guidance of Prof. Dr. Schneider and Dr. Satagopam, who provided strategic direction and ensured the overall scientific quality of the research.

Part III

Discussion and perspectives

Chapter 6

Key findings and implications

6.1 Expanding the scope of translational research

Translational research serves as a bridge between laboratory discoveries and clinical applications, facilitating the transformation of scientific insights into tangible healthcare solutions. In the context of MO and big data, adherence to best practices standards ensures research reproducibility, ethical conduct, and meaningful impact. This section examines the fundamental concepts of reproducibility, data sharing and integration, and their practical implementation across research ecosystem.

Ensuring scientific rigor and reproductibility

Scientific rigor is crucial to producing valid and reliable results, particularly in the complex landscape of MO research. The integration of multiple data types and analytical approaches demands meticulous attention to methodological detail and validation procedures. Core aspects of maintaining scientific rigor include:

Robust experimental design In the CRC collaboration utilizing scRNA-seq, meticulous consideration of cellular heterogeneity and systematic sampling methodologies ensured accurate representation of the tumor micro-environment's complexity. However, challenges such as batch effect and technical variability remain prevalent. Implementation of strategic approaches, including randomized sample processing protocols and incorporation of comprehensive control samples, can effectively mitigate these technical constraints.

Data Analysis Advanced computational methods have evolved to address the complexities of high-dimensional data analysis. While tools like Seurat facilitates scRNA-seq analysis, they require careful parameter optimization. There exists a significant risk of model overfitting and bias introduction when using ML approaches. For instance, in immunological dataset integration on peanut allergyPeanut Allergy, the selection of appropriate data transformation and normalization methodologies proved crucial for preventing skewed results and ensuring accurate biological interpretation.

Interpretation of Results Drawing meaningful conclusions necessitates a comprehensive understanding of both biological context and inherent data limitations. Fostering robust collaborative relationships and maintaining continuous feedback loops with research partners facilitates the development of optimized analysis workflows and appropriate dataset use.

Data Validity Practices First, the implementation of replication studies represents a fundamental aspect of scientific validation. While replication across diverse cohorts strengthens result validity, resource and founding constrains frequently impose practical limitation. The Luxfico study, characterized by a limited patient population, necessitates validation of its proof-of-concept findings in larger, more diverse cohorts. The peer review process serves as another essential component of data validation. Although peer review aims to ensure quality, current literature reporting practices may sometimes fail to identify methodological shortcoming. Maintaining transparent reporting and documentation workflows significantly benefits the broader scientific community by enabling thorough methodology assessment and result verification [324]. Finally, strict adherence to established protocols allows for comprehensive data validity. Implementation of standardized protocols enhances inter-study comparability and facilitates integration with published biomarker findings. However, the rapid pace of technological advancement often outpaces protocols development, necessitating continuous updates and refinements. For instance the scRNA-seq collaboration in section 4.2 and initiated in 2022 exemplifies this challenge, as Seurat has undergone continuous development and currently operates at version 5 (as of 29 Oct 2024) [325].

Navigating regulatory "pathways" in biomedical research

The translation of research findings into clinical applications represents a complex journey

through regulatory frameworks that are essential for ensuring patient safety and treatment efficacy. Understanding and successfully navigating regulatory requirements proves crucial for revolutionizing medicine through research applications. The process of bringing new diagnostics and therapeutics to the bedside involves complex approval processes, overseen by regulatory bodies such as the Food and Drug Administration (FDA) or European Medicines Agency (EMA). Notably, the FDA has implemented progressive reforms to accelerate therapeutic development while maintaining rigorous safety standards. Their 2019 guidance document on adaptive trial designs marks a significant advancement in regulatory flexibility, offering frameworks for pre-planned modifications to ongoing clinical trials based on interim data analysis. these adaptive approaches enhance statistical efficiency and ethical considerations across various trial phases, from initial dose-finding studies to confirmatory trials, ultimately facilitating a more nuanced understanding of treatment effects [326].

Early engagement with regulators authorities constitutes a key strategy for successful translation. Consulting with these bodies during initial stages of research and development enables the identification of potential issues before they become significant obstacles. This proactive approach allows researchers to address regulatory concerns systematically while maintaining research momentum.

The implementation of Good Clinical Practice (GCP) guidelines represent an important aspect of regulatory compliance. While adherence to these guidelines facilitates smoother approval processes, it necessitates comprehensive oversight and documentation systems. This regulatory framework ensures the protection of human subjects while maintaining data integrity throughout the research process [327, 328].

Comprehensive reporting of methods and results stands as a cornerstone of regulatory compliance. While this requirement can present significant administrative challenges, adherence to established standards such as FAIR principles [329] and STAR Methods for publishing provides a robust framework for documentation. These standardized approaches ensure transparency and reproducibility while meeting regulatory expectations.

The International Council for Harmonization of Technical Requirements for Pharmaceuticals for Human Use (ICH) plays a pivotal role in standardizing regulatory requirements across different regions. This harmonization facilitates international collaboration and streamlines the process of bringing potential new treatments to global markets. The ICH guidelines provide a comprehensive framework that addresses various aspects of drug development, from quality control to clinical safety, ensuring consistent standards across different regulatory

jurisdictions [330].

Strengthening collaborative frameworks in translational science

The inherently interdisciplinary nature of translational research necessitates robust collaborative frameworks that facilitates effective integration of diverse expertise and resources. These frameworks serve as essential foundations for accelerating scientific discoveries and their practical applications in clinical settings.

Cross-disciplinary collaboration brings successful translational research, integrating expertise from bioinformatics, clinical medicine, and biological sciences. In the projects described in this manuscript, partnerships with domain experts significantly enhanced the development of appropriate workflow designs and the identification of meaningful results for reporting. However, these collaborative efforts often encounter challenges from communication barriers between disciplines and divergent priorities among research teams. Establishing common ground and shared objectives becomes crucial for overcoming these obstacles.

The sharing of bioinformatic resources extends beyond mere data exchange, encompassing tools, methodologies, and analytical frameworks. While open sharing of these resources promotes innovation and accelerates scientific progress, it raises important considerations regarding data privacy and intellectual property protection. The delicate balance between maintaining openness and safeguarding sensitive information becomes particularly challenging when coordinating between different laboratories and their respective legal departments.

Cloud computing infrastructure has emerged as a critical component in supporting collaborative data analysis efforts. These platforms effectively address the computational demands associated with MO data processing and analysis. However, the implementation of cloud-based solutions introduces additional complexities related to data security, cost management, and compliance with regulatory frameworks such as GDPR. High-Performance Computing (HPC) infrastructure must carefully balance these considerations while maintaining efficient research operations [331].

Project management and organizational hierarchy shapes collaborative success. While hierarchical structures provide clear leadership channels and accountability, the need to foster creativity and innovation in translational research demands flexibility in organizational approaches. Agile methodologies offer adaptability in project management but may sometimes lack clear directional focus. Principal Investigators (PIs) serves as key architects in structuring collaborations that promote meaningful investigations while maintaining project coherence.

The complexity of translational settings, characterized by cross-disciplinary interactions and exploratory research paths, requires careful balance between structured oversight and creative freedom.

Advancing training and education in Translation medicine

The research landscape presents unprecedented challenges for the next generation of scientist, who must navigate fast paced evolution of technologies, methodologies, and increasingly complex cross-disciplinary opportunities. This evolution demands innovative approaches to research education and professional development.

The i2TRON doctoral training program, at the origin of this PhD thesis, exemplifies structured approaches to research education in the modern era. While such programs provide essential foundational knowledge and structured learning pathways, they may not fully encompass all emerging areas of scientific landscape. The program's framework, although comprehensive, highlight a crucial need for supplementary educational strategies. Self-directed learning initiatives and continuous professional development resources have become essential components in researchers for the dynamic nature of future science. However, these structured programs, despite their merits in educational frameworks, often face challenges in effectively fostering cross disciplinary collaborations.

The imperative to maintain current knowledge presents a significant challenge in the research environments. Researcher must continuously update their understanding of new tools, methodologies, and theoretical frameworks to remain effective in their fields. This requirement for constant learning creates a substantial cognitive burden, as the volume of new volume of information can become overwhelming. Moreover, researchers face the additional challenge of distinguishing between reliable and questionable resources in an era of rapid information dissemination. Success in current research environments requires not only technical proficiency but also the ability to synthesize knowledge across disciplines. Educational programs must therefore evolve to foster both specialized expertise and interdisciplinary competence, preparing researchers to address complex challenges than span traditional disciplinary boundaries — as exemplified by the i2TRON DTU initiative.

Enhancing patient-centered research strategies

The success of translational research fundamentally depends on effective partnerships among divers stakeholders, each contributing unique perspectives and capabilities to the research

ecosystem. Understanding the roles, responsibilities, and interactions among these stakeholders proves crucial for advancing medical innovations form laboratory discoveries to clinical applications:

Academic institutions serves as primary drivers of innovation in translational research, generating novel insights and methodologies. However, these institutions often face resource limitations that can constrain large-scale implementation of promising discoveries. This resource constraint highlights the necessity of strategic partnerships with other stakeholders in the translational research landscape.

Industry partners plays a vital role by providing essential funding and infrastructure support. While their involvement accelerates research progress and implementation, their profit-oriented objectives may sometimes conflict with principles of open sciences. This tension necessitate careful balance between commercial interest and scientific transparency.

Healthcare providers contribute invaluable clinical insights derived from direct patient care experience. However, established clinical practices and institutional inertia can create resistance to implementing novel approaches. This resistance highlights the importance of demonstrating clear clinical value and developing effective implementation strategies.

Patients occupy a central position in translational research, yet their involvement in decision-making processes has historically been limited. Recent initiatives, such as Patient and Public Involvement (PPI), represent significant progress toward more inclusive research practices. These programs actively incorporate patient perspectives into study design, implementation, and evaluation, ensuring research outcomes align patient needs and preferences.

The alignment of stakeholder goals remain challenging in translational research. Conflicting objectives can impede effective collaboration and slow research progress. To address this challenge, transparent and regular communication among all parties has become essential. Communication media, including patient associations and social networks, facilitate more inclusive and effective dialogue among all stakeholders.

Intellectual property (IP) considerations are center in stakeholder relationships. While necessary to incentivize innovation and investment, negotiations regarding IP rights often involve complex and time-consuming processes. These negotiations must balance protecting commercial interests with ensuring research findings benefit the broader medical community.

The evolution of partnerships in translational research reflects a growing recognition of the need for more integrated approaches to medical innovation. Success in this domain increasingly depends on developing frameworks that align diverse stakeholder interests while maintaining focus on improving patient outcomes. these frameworks must accommodate both the practical needs of implementation and the ethical imperatives of medical research.

Ethical standards in translational applications

The implementation of robust ethical considerations standards throughout the research life cycle represent a significant consideration in translational research, particularly when working with human patients and their sensitive data. These considerations must be integrated from the initial project design phase through to completion, ensuring responsible research practices at every stage.

Informed consent present unique challenges in the context of biomedical research. data The complexity of data reuse and biobanking creates significant hurdles in obtaining truly informed consent, as patients may struggle to fully comprehend the potential future applications of their data. Recent studies indicate varying levels of patient willingness to participate in research. Additionally, patients may give partial consent, creating added complexities for the downstream analyses. While approximately 60% of patients express willingness to share their genomic data data internationally, 44% express concerns about potential re-identification risks. This disparity highlights the delicate balance between advancing research objectives and protecting patient privacy [332].

Data security measures are critical in maintaining ethical research practices. While robust security protocols are essential for protecting patient information from unauthorized access or breaches, their implementation often introduces operational complexities that can impede research progress. Nevertheless, adherence to stringent security standards remains necessary for maintaining research credibility and protecting participant privacy.

Management of critical findings presents complex ethical challenge in translational research. The discovery of unexpected results necessitates careful consideration of when and how to communicate such findings to patients. This consideration becomes particularly relevant in different research contexts: for patients with existing diagnoses,

participation often stems hope for innovative treatments, while in exploratory studies like the fiber diet collaboration, genetic analyses may reveal previously unknown risk factors for complex diseases. Clear policies and careful evaluation of potential impacts on patients must guide these decisions.

Ethical committees Institutional review boards and ethics committees provide essential oversight. However, these bodies may face limitations when evaluating novel in-silico approaches, particularly in the context of rapid development of ML tools. This knowledge gap can result in delayed research approvals and potentially restrict the applications to cutting-edge modeling strategies.

The evolution of ethical considerations in translational research reflects the growing complexity of biomedical investigation. Success in this domain requires careful balance between advancing scientific knowledge and protecting participant interests. This balance demands continuous refinement of ethical frameworks to addresses emerging challenges while maintaining rigorous standards for patient protection.

The application of ethical standards must adapt to technological advancement while maintaining core principles of patient protection. This adaptation requires ongoing dialogue among researchers, ethics committees, and patients to develop frameworks that adequately address both traditional and emerging ethical challenges.

Data-Driven Approaches and Challenges

The reproducibility crisis in scientific research has highlighted critical needs for more stringent practices in data management and analysis methodologies [333]. This challenge becomes particularly acute in the context of MO research, where the complexity of integrated data layers amplifies potential reproducibility issues [121].

Multiple factors contribute to the current reproducibility crises in scientific research. Selective reporting of results, inadequate data sharing practices, and methodological shortcomings represent primary challenges that must be addressed [334, 335]. In MO research settings, these challenges are inevitable by the inherent complexity of integrating multiple data types. Each additional data layer introduces new opportunities for methodological errors and increases the difficulty of ensuring reproducible results.

The FAIR principles (Findable, Accessible, Interoperable, and Reusable) enhances research reproducibility [336]. However, the practical application of these principles presents challenges in MO research. The vast diversity of data types encountered in integrated studies necessitates tailored approaches to data standardization. This standardization must address not only the primary data but also the associated metadata, such as health records, ensuring continuity in the comprehensive documentation of experimental conditions and analytical methods [337].

The development of workflows and pipeline systems offers promising solutions to reproducibility challenges. these tools can help ensure consistent application of analytical methods across different studies and research groups. However, their implementation requires careful consideration of documentation requirements and version control to maintain reproducibility over time.

Success in addressing reproducibility challenges requires a coordinated effort across research community. This effort must encompass improvements in methodological transparency, data sharing practices, and standardization of analytical approaches, while maintaining the flexibility necessary for scientific innovation. The goal is to establish a research ecosystem that promotes both scientific rigor and creative exploration.

Critical Reflections and Future Directions

While established best practices provide essential frameworks for conducting high-quality research, significant challenges persist in translational research that warrant careful consideration and innovative solutions. These challenges span multiple domains and require thoughtful approaches to balance competing demands while maintaining research quality and impact.

Standardization vs Innovation While adherence to standardized methods ensures consistency and reproducibility, overly rigid protocols can constrain scientific creativity and limit the exploration of novel approaches. Researchers leaders must carefully balance these competing needs, establishing framework that promote methodological rigor while maintaining sufficient flexibility to explore emerging technologies.

Data integration Challenges The inherent MO complexity of combining diverse data types, each with distinct scales and noise characteristics, demands considerations. Computational methods for integration are under constant evolution, reflecting the dynamic

nature of the field. However, the absence of "well-rounded" solution, or simplified analysis tools underscores the need for careful consideration of methodological choices in each research context.

Ethical dilemmas in data sharing While open data policies promote transparency and scientific progress, they must be balanced against privacy concerns, especially regarding genomic data's inherently identifiable nature. Recent research has highlighted the unique challenges of genomic data privacy, demonstrating that these data can have implications not only for individual participants but also for their biological relatives. This situation necessitates careful consideration of privacy protections that extend beyond individual participants [338].

Educational gaps The filed increasingly demands professionals who possess expertise in both computational and biological sciences [339]. Current evidence suggest that interdisciplinary training programs require enhancement, with particular emphasis on practical applications and research experience. Programs must evolve to prepare future researchers for effectively bridge the educational gaps [340].

Regulatory lag Development costs for new therapeutics continue to rise, with recent estimates suggesting expenses exceeding 2 billon dollars per drug [341]. The extended timeline from discovery to patient access, often spanning up to 15 years, highlights the need for more efficient regulatory processes that maintain safety standards while accelerating therapeutic development.

6.2 Multi-omics integration for identifying new therapeutic targets

In this thesis, three primary collaborative studies have been conducted to elucidate the complex interplay between immunology responses on cellular, molecular, genetic and nutrition levels. Investigations leveraged various aspects of MO integration, advanced computational analyses and cross-disease approach in oncology, allergology and nutritional medicine contexts. The different conditions and methodology have provided multi-layered insights on complex inflammatory processes in chronic diseases, which underscore the significance of data integration in life sciences for a better development of translational research.

Cell-Cell Communication in CRC To explore the complex interplay between immunology responses at the cellular level, we collaborated with Dr. Rubens Begaj and Dr. Elisabeth Letellier, to develop a scRNA-seq approach for focusing on the role of WNT5A signaling in colorectal cancer. Extending the research through CCC analysis using LIANA and NicheNet, we identified two fibroblast subtypes (iCAF and myCAF) characterized by high WNT5A expression. This work suggests potential for targeted strategies aiming to disrupt WNT5A-mediated signaling and boost anti-tumor immune responses. The scRNA-seq and CCC analyses were central for uncovering these novel therapeutic targets.

Immunological data integration during Allergy Building upon the immunology insights from CCC analysis in CRC, we shifted our focus to explore the complex immunological mechanisms of Peanut Allergy. In collaboration with Dr. Rebecca Czolk (NextImmune DTU), Theresa-Maria Böhm (PhD student NextImmune2 DTU), we used different approaches, including cytokine and cell type analysis, to identify distinct immune signatures differentiating PA patients between their reactions manifestation, particularly during Oral Food challenge.

Genetic and Dietary contributions to inflammation risk To further expand our understanding of the factors contributing to inflammation, we collaborate with Dr. Guilherme Meyers and Dr. Torsten Bohn to investigate the interplay between genetic variants, DF intake, and inflammatory markers. This study aims to bridge the gap between genetic predispositions and environmental factors, such as diet, in modulating inflammatory responses. By integrating genetic, microbiome, and dietary data, we developed a strategy to establish an inflammatory risk score that stratify patients based on their potential for personalized dietary recommendations to mitigate inflammation, particularly in individuals with genetic predispositions affecting fiber metabolism.

Cross-Disease insights and translational relevance A recurring pathway target across these initiatives is the central role of inflammation in driving disease progression. Despite the distinct contexts of colorectal cancer, food allergy, and dietary interventions, we explored noteworthy inflammatory pathways, underscoring the interconnected nature of these processes. A MO integration, as P-integration, comparing different cohorts would facilitate the understanding of cellular and molecular interactions essen-

tial and shared for cross-disease therapeutic strategies. These insights contribute to developing the broader field of translational medicine, supporting the development of more precise and personalized healthcare interventions.

6.2.1 Multi-omics in perspective of the research collaborations

This section critically examines the techniques used, highlights their strengths and limitations, and discusses the broader implications of these approaches for translational research.

Strengths

Data integration allowed us to capture the complex interactions of biological processes. Correlation networks helped to visualize long lists of interactions. Dimension reduction methods such as PCA, or UMAP were able to capture meaningful biological signals despite the heterogeneity and high dimensionality of the data.

Machine learning models such as Random Forest enabled the identification of key features in the group identified by unsupervised approaches over diverse datasets. Feature selection methods hold great promise for building predictive models, such as the inflammatory risk score described in this manuscript, and incorporating genetic, dietary and microbiome data.

Challenges

The data heterogeneity among the different datasets posed significant challenge. As secondary usage, some dataset are not raw and already processed by the primary study, thus limiting the possibilities for modeling.

For the correlation based network, even though bootstrapping , some limitation remain such for difficulties to detect non-linear interactions, or small sample size limit the detection of complex dependencies.

Generalization

One of the advantages of the integrative approaches in this thesis are their transferability. As long as data meets assumption of a model, they can be adapted. In P-integration, same data layers would unravel shared mechanisms which could be relevant. For example, APSIS and Luxfico cohorts immunological and microbiome data integration would be interesting to explore the possible shared inflammatory and metabolic pathways.

Predictive models such as the inflammation risk score from Luxfico cohort can be

flexible, allowing for the incorporation of additional data types or features as they become available. The modularity of data-driven models supports potential extension to other cohorts, including those with different genetic backgrounds or environmental exposures, either for refinement and enrichment.

Reproducibility

The integrative analyses were designed with a focus on reproducibility, employing standardized methods for data processing, statistical testing and visualization. For some case, bootstrapping and cross-validation techniques further enhanced the robustness of the findings. However, reproducibility remains a challenge, particularly in the complex of MO datasets.

Computational consideration

The different ML method proved effective in revealing interesting biological insights but they required careful tuning of parameters to avoid over-fitting or loss of biological information. For instance in scRNA-seq, the resolution parameter in clustering step significantly influenced the identification of subpopulations, thus impacting the CCC analysis.

6.3 Cross-disease research advances personalized medicine

Building on the integrative bioinformatics and MO analyses conducted, several procedures and future directions can be proposed to enhance disease stratification, patient monitoring, and early intervention strategies. This section outlines the further steps and explores the implications for PM and precision healthcare.

The findings from this thesis highlight the potential for stratifying patients based on their inflammatory profiles, starting from cytokines signatures, cellular interactions to molecular and genetic markers:

Early disease monitoring scRNA-seq and CCC analysis identified several patterns and signature that could be associated in tumor development and inflammatory responses. Validating these communication pathways, a diagnostic panel could be build with clinical datasets for monitoring patients in early stages of CRC.

Risk Stratification The development of a predictive risk score, incorporating genetic data (SNP), microbiome composition and dietary factors, offers a novel approach for stratifying patients based on their predisposition to inflammation. This model could be applied in clinical settings to identify individuals at high risk of developing severe inflammatory responses, enabling early intervention, such as dietary modifications or anti-inflammatory therapies, designed for the patients specific profile.

Longitudinal monitoring By leveraging MO integration, a dynamic monitoring system could be established to track changes in a patient's inflammatory profile over time. For example, serial measurement of cytokines levels combined with single-cell transcriptomics data could provide multi-modal insights into the development of inflammation, guiding treatment and life style adjustments and improving patient outcomes.

6.3.1 Expanding the scope of personalized medicine

Given the results on the three diseases investigated in this manuscript, it would be beneficial to apply these approaches to additional disease contexts to investigate the inflammatory responses. Potential other disease models could include chronic inflammatory conditions such as rheumatoid arthritis or inflammatory bowel disease, where cytokines signaling could express similar patterns or microbiome interactions could influence the disease progression. Expanding the investigation scope to include these diseases could help identify relevant pathogenic and common inflammatory signatures and further elucidate the early development of chronic disease.

While several biomarkers have been identified, further validation is required to translate these hypotheses. The next logical step would involve wet-lab validation using appropriate techniques such as ELISA for cytokines quantification, flow cytometry for cell profiling, and knockouts to assess the functional impact of specific genes. Additionally, the ultimate validation would involve prospective cohort studies to assess the predictive value of these biomarkers in diverse patient population.

From the diet collaboration, to establish the clinical utility of the identified biomarkers and predictive models, clinical trials would be preferred. For instance, a randomized controlled trial could be designed to evaluate the effectiveness of personalized dietary intervention based on the inflammatory risk score developed in the thesis. Patients could be stratified based on their genetic and microbiome profiles, and the impact of tailored fiber intake on inflammatory

biomarker quantification and body responses could be assessed.

Suggestions and improvements for such investigations would include different aspects improving the quality and rapidity of the hypotheses driven analyses. One of the main challenges in MO integration is the variability in data quality across different datasets. Developing an automated QC platform that standardizes quality control metrics and integrates them into an unified report would streamline the analysis process and reduce the risk of batch effects and data inconsistencies. In a similar way, a complete pipeline that facilitates the assessment of statistical tests (such as differential expression, correlation analysis) and integrates robust methodologies such as bootstrapping and cross-validation would enhance the reliability of the findings and the comparability to other studies. This pipeline could be designed to automatically suggest the most appropriate tests based on the data characteristics, simplifying the workflow for researchers. Another idea would be to design enhanced interpretation tools to facilitate decision-making, developing user-friendly tools that provide integrated explanations and reports of the results. For example, interactive visualizations highlighting key features and offer context specific interpretations (based on pathway enrichment analysis for example) with interactive selections, could make the results more accessible and dynamic in collaborations, improving the decision flow and follow up interventions.

Challenges and limitations

The integration of MO data in this thesis revealed several significant challenges that are particularly pertinent to immunology and allergy research. A fundamental limitation emerged from the initial experimental design of collaborative projects, which were primarily conceived for classical A/B testing rather than comprehensive MO modeling. This constraint exemplifies a common challenge in the field: the re-purposing of existing datasets for advanced integrative analyses often faces inherent limitations based on their original design parameters.

In the context of allergy studies, we encountered specific challenges related to the relative scarcity of reference datasets compared to more extensively studied diseases such as cancer. This limitation significantly impacted our ability to perform comprehensive enrichment analyses and validate findings through external data sources. While technical aspects of Single-cell RNA sequencing (scRNA-seq) analysis were well-supported through established workflows like Seurat, the primary challenge lay in developing novel, biologically relevant signatures for characterizing cell types and identifying significant targets.

Legal and administrative constraints posed another significant challenge, particularly evident in the fiber diet collaboration. These barriers resulted in limited access to detailed dataset information and original publication results, significantly impacting the pace of development and validation of our findings. This experience highlights a broader challenge in MO research: the balance between data protection requirements and the need for transparent, reproducible science.

The economic aspects of MO studies emerged as a fundamental limiting factor. While reusing existing datasets represents a cost-effective approach, the limitations inherent in their original design often restrict the application of advanced MO methodologies. Furthermore, the scientific community's skepticism towards complex MO analyses necessitates extensive validation of findings. Traditional enrichment analyses, while useful as filtering tools, are often not considered sufficient validation, creating a requirement for additional experimental verification that may be beyond the scope or resources of many studies.

Chapter 7

Future directions in inflammation and multi-omics research

7.1 Integrated research shapes the future of personalize medicine

The integration of MO approaches in clinical practice presents transformative potential for healthcare delivery, particularly in disease prevention and PM. Our research suggests that the primary impact of these integrative approaches will manifest in enhanced disease screening capabilities and the development of comprehensive molecular footprinting. This advancement enables earlier disease detection and more precise mechanistic understanding of pathological processes, fundamentally shifting healthcare paradigms from reactive to preventive approaches.

The translation of MO research into clinical applications necessitates significant technological and methodological adaptations. While current research primarily focuses on target identification and pharmaceutical collaborations, the future implementation demands the development of practitioner-friendly analytical tools. These tools must bridge the complexity of MO analysis with the practical constraints of clinical settings, enabling healthcare providers to effectively utilize sophisticated molecular data in patient care.

A significant paradigm shift is anticipated in therapeutic approaches, moving beyond traditional pharmaceutical interventions toward more personalized and preventive strategies. This evolution encompasses tailored dietary modifications, targeted metabolite supplementation, and lifestyle interventions based on individual molecular profiles. However, the widespread

implementation of comprehensive molecular profiling, such as full microbiome sequencing, remains constrained by cost considerations. Nevertheless, signature-based analyses offer promising intermediate solutions for patient monitoring and lifestyle optimization.

The advancement of digital technologies, particularly the widespread use of smartphones, presents opportunities for patient data collection and monitoring. However, this potential is tempered by significant challenges in data security and accessibility for healthcare providers. The development of robust clinical decision support systems capable of processing and interpreting large-scale patient data remains a critical need.

Recent advances in Large Language Models show promise in facilitating communication between various stakeholders in translational science, particularly in summarizing complex reports and explaining intricate molecular concepts. However, the current limitations of these models, including their susceptibility to hallucinations and potential for false statements, preclude their direct application in clinical decision-making processes.

The evolution of MO research necessitates the development of comprehensive infrastructure and analytical tools to support clinical implementation. A crucial advancement would be the establishment of integrated disease databases incorporating multiple molecular data layers, enabling researchers and clinicians to identify and analyze specific patient subgroups within large cohorts. These databases would facilitate the development of robust MO signatures for distinct disease phenotypes, significantly enhancing our understanding of disease heterogeneity and treatment response patterns.

The field requires standardized analytical frameworks comparable to established tools like Seurat for Single-cell RNA sequencing (scRNA-seq). Such frameworks would need to accommodate various MO integration methodologies while maintaining user accessibility and computational efficiency. The development of a unified analytical platform would enable systematic comparison of different integration approaches, fostering reproducibility and methodological standardization across the field.

A critical step toward clinical implementation is the development of automated, user-friendly pipelines suitable for deployment in healthcare settings. These systems must balance sophisticated analytical capabilities with operational simplicity, enabling non-bioinformaticians to effectively utilize MO approaches in clinical decision-making. Initial implementation of such systems would likely focus on well-funded disease areas, such as oncology, where existing infrastructure and resources can support the substantial computational and technical requirements. This staged approach would provide valuable insights

for broader implementation across other disease domains, while allowing for refinement of analytical protocols and user interfaces based on real-world clinical feedback.

7.2 Advancing inflammation treatment through continued investigation

The continuation of comprehensive insights into inflammatory mechanisms necessitated the exploration of multiple analytical strategies, many of which faced limitations inherent to data complexity, experimental design, and translational feasibility. Below, the challenges encountered in alternative approaches are discussed, also the barriers to multi-omics integration, and the indispensable role of biological expertise in navigating these complexities.

Leveraging Big Data for allergy research

While A/B testing (*e.g.* case/control comparisons) remains a gold standard for identifying statistically significant differences in clinical studies (*e.g.* cytokines levels between cohorts), its utility is confined to surface-level insights (mean/variance shifts). For instance, group differences highlight disease-associated trends but failed to explain whether these shifts were drivers of pathology or secondary effects of confounding factors (*e.g.* medication use). This limitation underscores the need for more hypothesis-driven, focused mechanistic studies to complement exploratory analyses.

Correlation isn't Causality. Correlation analyses between cytokines and immune cell populations revealed co-regulation patterns in section 3.3. While causal inference approaches (*e.g.* Mendelian randomization, for investigating gene causality) can be considered, they required genetic or longitudinal data unavailable in our cohort. This exemplifies a broader challenge: clinical datasets can lack the temporal resolution or multi-modal layers needed for causal modeling.

Manual subgrouping based on delayed treatment response was explored to stratify patients (*e.g.* "early" vs "late" responders). However, the binary design of our primary dataset lacked granular temporal metadata, rendering such stratification statistically underpowered. Future studies incorporating continuous monitoring (*e.g.* consistent cytokines profiling over the experience) could enable these hypotheses, bridging gaps between clinical phenotypes and mechanistic drivers.

Unsupervised clustering (*e.g.* k-means on cytokines profiles) aimed to identify patient subtypes with distinct immune signatures. However, the limited cohort size (n = 40) and high dimensionality of features led to unstable clusters. Even when clusters emerged, biological interpretation relied heavily on manual metadata scrutiny (*e.g.* correlating clusters with comorbidities), emphasizing that algorithmic outputs alone cannot replace domain expertise.

The integration of multi-omics data (*e.g.* immune, microbiome, metabolomic layers) holds theoretical promise for unraveling systemic drivers of inflammation. However, practical implementation faced difficulties in the framework of this thesis. Only a small number of patients had complete multi-omics profiles due to sample collection times and inconsistent biomarker panels. Subgroup analyses in small cohorts then produced unreliable effect sizes, rendering the possible approaches — such as Bayesian ones — speculative at best.

These challenges highlight a translational paradox: while multi-omics frameworks are biologically compelling, their success hinges on harmonized and consistent large-scale data infrastructure.

As exemplified in section 3.4, statistical outputs is requiring rigorous biological validation to distinguish signal from noise. An iterative dialogue between data and domain knowledge is not a limitation but a necessity, ensuring findings are mechanistically plausible and clinically actionable.

Future research directions should address several key barriers identified in this work. First, establishing prospective cohorts with embedded multi-omics analyses will be crucial. These cohorts should implement standardized collection protocols, including fixed timepoints and centralized laboratory processing, to minimize technical and biological heterogeneity across samples.

To move beyond correlative findings, future studies should incorporate causal inference frameworks that integrate genetic and interventional data. This approach will strengthen our ability to identify true mechanistic relationships within the complex inflammatory networks we have observed.

Additionally, the development of collaborative data repositories represents a critical step forward. Such shared infrastructure would enable the harmonization of currently fragmented datasets across institutions, facilitating meta-analyses with sufficient statistical power to detect subtle biological signals.

While this thesis has highlighted the inherent challenges in studying complex inflammatory processes, it ultimately reinforces a fundamental principle in inflammation research:

breakthrough discoveries emerge from the synergistic application of analytical rigor and deep biological expertise. This combination remains essential for advancing our understanding of inflammatory mechanisms and their therapeutic implications.

High-Resolution genetic studies in CRC

Single-cell RNA sequencing (scRNA-seq) has fundamentally transformed our ability to characterize tumor heterogeneity, enabling unprecedented resolution in identifying rare cancer cell subtypes and complex cell-cell communication networks. The field continues to evolve with the integration of complementary multi-omics approaches, including ATAC-seq for chromatin accessibility profiling, RNA velocity analysis, and spatial transcriptomics, which collectively provide deeper insights into tumor evolution and microenvironmental interactions. These advanced techniques offer powerful analytical capabilities — pseudotime analysis can now elucidate developmental trajectories connecting progenitor — such as states to metastatic phenotypes, while gene regulatory network analysis helps identify master transcription factors driving colorectal cancer progression.

In our investigation of cross-disease mechanisms, we employed Harmony-based integration to compare CRC and pancreatic ductal adenocarcinoma (PDAC) profiles. However, this approach revealed a fundamental challenge: while cross-disease comparisons can illuminate universal oncogenic pathways, they risk obscuring context-specific biological mechanisms. Moving forward, we propose that studies should prioritize intra-cancer integrative analyses, such as harmonizing colorectal cancer datasets across different stages or molecular subtypes (CMS1-4), with particular attention to rigorous batch correction and careful metadata alignment, especially regarding treatment history.

A crucial consideration in Single-cell RNA sequencing is the impact of analytical parameter selection on results interpretation. For instance, modifications to clustering resolution or choice of dimensionality reduction methods (UMAP versus t-SNE) can significantly alter the identified cell populations. While our initial analysis pipeline adhered to established best practices through the Seurat toolset, we recognize that addressing new hypotheses may necessitate parameter adjustments. This presents a methodological dilemma: strict adherence to original workflows ensures reproducibility but potentially overlooks biologically significant subpopulations. We advocate for a balanced approach of principled flexibility, emphasizing transparent documentation of parameter choices and validation against orthogonal datasets such as spatial transcriptomics.

Fiber Diet and immunology modulation: A roadmap for future work

Current Progress and Anticipated Single-nucleotide polymorphism (SNP) analyses presented in section 5.2 remains in development, preliminary work establishes a pipeline for linking dietary patterns to immunologic outcomes via genetic profiling. By mid-2025, we aim to : Identify SNP associated with immune responses to high-fiber diets. And benchmark these genetic signatures against public genetic reports in databases such as OMIM, ClinVar.

A core objective of this work is to evaluate how predictive models of fiber-mediated immunologic responses align with existing public health nutrition guidelines. For instance, the European Food Safety Authority recommends a daily fiber intake 25 g for adults [342], while regional guidelines could emphasize fiber sources (*e.g.* whole grains vs legumes). Our models will assess whether genetic subpopulations require tailored thresholds, challenging one-size-fits-all recommendations. To draft our guidelines, it is expected to be partnering with public local organizations, and integrate the outcomes — such as SNP profiles — into electronic health records to flag at-risk patients during dietary counseling.

Looking ahead, several critical research directions emerge for advancing our understanding of Dietary fibers's impact on health outcomes. Longitudinal studies spanning 5-10 years will be essential to track the relationships between fiber intake, microbiome composition shifts, and immune markers, providing robust validation of causal relationships. Additionally, intervention trials present an opportunity to test personalized fiber prescriptions in controlled clinical settings. Perhaps most importantly, expanding our dietary fiber study to include SNP-diet database aggregations from underrepresented populations, particularly African and Indigenous cohorts, represents a crucial step toward addressing nutritional health disparities on a global scale.

References

[1] H.-Y. Liu et al. “Diagnostic and clinical utility of whole genome sequencing in a cohort of undiagnosed Chinese families with rare diseases”. In: *Scientific Reports* 9.1 (Dec. 18, 2019). Publisher: Nature Publishing Group, p. 19365. DOI: 10.1038/s41598-019-55832-1.

[2] J. M. Heather and B. Chain. “The sequence of sequencers: The history of sequencing DNA”. In: *Genomics* 107.1 (Jan. 1, 2016), pp. 1–8. DOI: 10.1016/j.ygeno.2015.11.003.

[3] L. Hood and L. Rowen. “The Human Genome Project: big science transforms biology and medicine”. In: *Genome Medicine* 5.9 (Sept. 13, 2013), p. 79. DOI: 10.1186/gm483.

[4] H. Satam et al. “Next-Generation Sequencing Technology: Current Trends and Advancements”. In: *Biology* 12.7 (July 2023). Number: 7 Publisher: Multidisciplinary Digital Publishing Institute, p. 997. DOI: 10.3390/biology12070997.

[5] N. Gupta and V. K. Verma. “Next-Generation Sequencing and Its Application: Empowering in Public Health Beyond Reality”. In: *Microbial Technology for the Welfare of Society* 17 (Sept. 13, 2019), pp. 313–341. DOI: 10.1007/978-981-13-8844-6_15.

[6] X. Liu, J. Zheng, J. Ding, J. Wu, F. Zuo, and G. Zhang. “When Livestock Genomes Meet Third-Generation Sequencing Technology: From Opportunities to Applications”. In: *Genes* 15.2 (Feb. 2024). Number: 2 Publisher: Multidisciplinary Digital Publishing Institute, p. 245. DOI: 10.3390/genes15020245.

[7] V. Meslier, B. Quinquis, K. Da Silva, F. Plaza Oñate, N. Pons, H. Roume, M. Podar, and M. Almeida. “Benchmarking second and third-generation sequencing platforms for microbial metagenomics”. In: *Scientific Data* 9.1 (Nov. 11, 2022). Publisher: Nature Publishing Group, p. 694. DOI: 10.1038/s41597-022-01762-z.

[8] S. Hassan, R. Bahar, M. F. Johan, E. K. Mohamed Hashim, W. Z. Abdullah, E. Esa, F. S. Abdul Hamid, and Z. Zulkafli. "Next-Generation Sequencing (NGS) and Third-Generation Sequencing (TGS) for the Diagnosis of Thalassemia". In: *Diagnostics* 13.3 (Jan. 2023). Number: 3 Publisher: Multidisciplinary Digital Publishing Institute, p. 373. DOI: 10.3390/diagnostics13030373.

[9] J. B. Krier, S. S. Kalia, and R. C. Green. "Genomic sequencing in clinical practice: applications, challenges, and opportunities". In: *Dialogues in Clinical Neuroscience* 18.3 (Sept. 30, 2016), pp. 299–312. DOI: 10.31887/DCNS.2016.18.3/jkrier.

[10] F. Wu, L. A. Mueller, D. Crouzillat, V. Pétiard, and S. D. Tanksley. "Combining Bioinformatics and Phylogenetics to Identify Large Sets of Single-Copy Orthologous Genes (COSII) for Comparative, Evolutionary and Systematic Studies: A Test Case in the Euasterid Plant Clade". In: *Genetics* 174.3 (Nov. 2006), pp. 1407–1420. DOI: 10.1534/genetics.106.062455.

[11] J. Sun, F. Lu, Y. Luo, L. Bie, L. Xu, and Y. Wang. "OrthoVenn3: an integrated platform for exploring and visualizing orthologous data across genomes". In: *Nucleic Acids Research* 51 (W1 July 5, 2023), W397–W403. DOI: 10.1093/nar/gkad313.

[12] R. D. Clark, K. A. Catalano, K. S. Fitz, E. Garcia, K. E. Jaynes, B. N. Reid, A. Sawkins, A. A. Snead, J. C. Whalen, and M. L. Pinsky. "The practice and promise of temporal genomics for measuring evolutionary responses to global change". In: *Molecular Ecology Resources* n/a (n/a 2023). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.13789. DOI: 10.1111/1755-0998.13789.

[13] G. M. Rossolini and M. C. Thaller. "Coping with antibiotic resistance: contributions from genomics". In: *Genome Medicine* 2.2 (Feb. 25, 2010), p. 15. DOI: 10.1186/gm136.

[14] J. S. Loh, W. Q. Mak, L. K. S. Tan, C. X. Ng, H. H. Chan, S. H. Yeow, J. B. Foo, Y. S. Ong, C. W. How, and K. Y. Khaw. "Microbiota–gut–brain axis and its therapeutic applications in neurodegenerative diseases". In: *Signal Transduction and Targeted Therapy* 9.1 (Feb. 16, 2024). Publisher: Nature Publishing Group, pp. 1–53. DOI: 10.1038/s41392-024-01743-1.

[15] N. Stoler and A. Nekrutenko. "Sequencing error profiles of Illumina sequencing instruments". In: *NAR Genomics and Bioinformatics* 3.1 (Mar. 1, 2021), lqab019. DOI: 10.1093/nargab/lqab019.

[16] M. A. Taub, H. Corrada Bravo, and R. A. Irizarry. “Overcoming bias and systematic errors in next generation sequencing data”. In: *Genome Medicine* 2.12 (Dec. 10, 2010), p. 87. DOI: 10.1186/gm208.

[17] D. C. Koboldt. “Best practices for variant calling in clinical sequencing”. In: *Genome Medicine* 12.1 (Oct. 26, 2020), p. 91. DOI: 10.1186/s13073-020-00791-w.

[18] S. Zverinova and V. Guryev. “Variant calling: Considerations, practices, and developments”. In: *Human Mutation* 43.8 (2022). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/humu.24311, pp. 976–985. DOI: 10.1002/humu.24311.

[19] R. Poplin, J. M. Zook, and M. DePristo. “Challenges of Accuracy in Germline Clinical Sequencing Data”. In: *JAMA* 326.3 (July 20, 2021), p. 268. DOI: 10.1001/jama.2021.0407.

[20] Z. D. Stephens, S. Y. Lee, F. Faghri, R. H. Campbell, C. Zhai, M. J. Efron, R. Iyer, M. C. Schatz, S. Sinha, and G. E. Robinson. “Big Data: Astronomical or Genomical?” In: *PLoS Biology* 13.7 (July 7, 2015), e1002195. DOI: 10.1371/journal.pbio.1002195.

[21] T. A. Brown. “The Human Genome”. In: *Genomes. 2nd edition.* Wiley-Liss, 2002.

[22] Y. Dong, F. Sun, Z. Ping, Q. Ouyang, and L. Qian. “DNA storage: research landscape and future prospects”. In: *National Science Review* 7.6 (June 1, 2020), pp. 1092–1107. DOI: 10.1093/nsr/nwaa007.

[23] V. Ginevičienė, A. Utkus, E. Pranckevičienė, E. A. Semenova, E. C. R. Hall, and I. I. Ahmetov. “Perspectives in Sports Genomics”. In: *Biomedicines* 10.2 (Feb. 2022). Number: 2 Publisher: Multidisciplinary Digital Publishing Institute, p. 298. DOI: 10.3390/biomedicines10020298.

[24] J. B. Wolf, A. C. Ferguson-Smith, and A. Lorenz. “Mendel’s laws of heredity on his 200th birthday: What have we learned by considering exceptions?” In: *Heredity* 129.1 (July 2022). Publisher: Nature Publishing Group, pp. 1–3. DOI: 10.1038/s41437-022-00552-y.

[25] O. Harman and E. Lamm. “History of Classical Genetics”. In: *eLS*. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470015902.a0003094.pub2. John Wiley & Sons, Ltd, 2015, pp. 1–6. DOI: 10.1002/9780470015902.a0003094.pub2.

[26] T. L. Saunders. "The History of Transgenesis". In: *Transgenic Mouse: Methods and Protocols*. Ed. by M. A. Larson. New York, NY: Springer US, 2020, pp. 1–26. DOI: 10.1007/978-1-4939-9837-1_1.

[27] C. Tong, Y. Liang, Z. Zhang, S. Wang, X. Zheng, Q. Liu, and B. Song. "Review of knockout technology approaches in bacterial drug resistance research". In: *PeerJ* 11 (Aug. 17, 2023). Publisher: PeerJ Inc., e15790. DOI: 10.7717/peerj.15790.

[28] Q. Zhang, J. Yang, E. N. E. A. Manoharan, A. B. Yu, and M. C. Milone. "CRISPR/Cas9-mediated Gene Knockout Followed by Negative Selection Leads to a Complete TCR Depletion in ortho CAR19 T Cells". In: *Bio-Protocol* 12.15 (Aug. 5, 2022), e4485. DOI: 10.21769/BioProtoc.4485.

[29] S. A. Smirnikhina, A. A. Anuchina, and A. V. Lavrov. "Ways of improving precise knock-in by genome-editing technologies". In: *Human Genetics* 138.1 (Jan. 1, 2019), pp. 1–19. DOI: 10.1007/s00439-018-1953-5.

[30] F. Giambò et al. "Genetic and Epigenetic Alterations Induced by Pesticide Exposure: Integrated Analysis of Gene Expression, microRNA Expression, and DNA Methylation Datasets". In: *International Journal of Environmental Research and Public Health* 18.16 (Jan. 2021). Number: 16 Publisher: Multidisciplinary Digital Publishing Institute, p. 8697. DOI: 10.3390/ijerph18168697.

[31] K. J. Livak and T. D. Schmittgen. "Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2−ΔΔCT Method". In: *Methods* 25.4 (Dec. 1, 2001), pp. 402–408. DOI: 10.1006/meth.2001.1262.

[32] Y. Mo, R. Wan, and Q. Zhang. "Application of Reverse Transcription-PCR and Real-Time PCR in Nanotoxicity Research". In: *Nanotoxicity: Methods and Protocols*. Ed. by J. Reineke. Totowa, NJ: Humana Press, 2012, pp. 99–112. DOI: 10.1007/978-1-62703-002-1_7.

[33] S.-H. Shiu and J. O. Borevitz. "The next generation of microarray research: applications in evolutionary and ecological genomics". In: *Heredity* 100.2 (Feb. 2008). Publisher: Nature Publishing Group, pp. 141–149. DOI: 10.1038/sj.hdy.6800916.

[34] D. Jovic, X. Liang, H. Zeng, L. Lin, F. Xu, and Y. Luo. "Single-cell RNA sequencing technologies and applications: A brief overview". In: *Clinical and Translational Medicine* 12.3 (2022). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ctm2.694, e694. DOI: 10.1002/ctm2.694.

[35] E. Clough and T. Barrett. “The Gene Expression Omnibus Database”. In: *Statistical Genomics: Methods and Protocols*. Ed. by E. Mathé and S. Davis. New York, NY: Springer, 2016, pp. 93–110. DOI: 10.1007/978-1-4939-3578-9_5.

[36] K.-F. Tung, C.-Y. Pan, C.-H. Chen, and W.-c. Lin. “Top-ranked expressed gene transcripts of human protein-coding genes investigated with GTEx dataset”. In: *Scientific Reports* 10.1 (Oct. 1, 2020). Publisher: Nature Publishing Group, p. 16245. DOI: 10.1038/s41598-020-73081-5.

[37] *The Cancer Genome Atlas Program (TCGA) - NCI*. Archive Location: nciglobal,ncienterprise. May 13, 2022. URL: https://www.cancer.gov/ccg/research/genome-sequencing/tcga (visited on 11/20/2024).

[38] I. Papatheodorou et al. “Expression Atlas: gene and protein expression across multiple studies and organisms”. In: *Nucleic Acids Research* 46 (D1 Jan. 4, 2018), pp. D246–D251. DOI: 10.1093/nar/gkx1158.

[39] D. Greco, P. Somervuo, A. D. Lieto, T. Raitila, L. Nitsch, E. Castrén, and P. Auvinen. “Physiology, Pathology and Relatedness of Human Tissues from Gene Expression Meta-Analysis”. In: *PLOS ONE* 3.4 (Apr. 2, 2008). Publisher: Public Library of Science, e1880. DOI: 10.1371/journal.pone.0001880.

[40] L. R. Chapman, I. V. P. Ramnarine, D. Zemke, A. Majid, and S. M. Bell. “Gene Expression Studies in Down Syndrome: What Do They Tell Us about Disease Phenotypes?” In: *International Journal of Molecular Sciences* 25.5 (Jan. 2024). Number: 5 Publisher: Multidisciplinary Digital Publishing Institute, p. 2968. DOI: 10.3390/ijms25052968.

[41] A. K. Nøhr, M. Lindow, A. Forsingdal, S. Demharter, T. Nielsen, R. Buller, I. Moltke, M. Vitezic, and A. Albrechtsen. “A large-scale genome-wide gene expression analysis in peripheral blood identifies very few differentially expressed genes related to antidepressant treatment and response in patients with major depressive disorder”. In: *Neuropsychopharmacology* 46.7 (June 2021). Publisher: Nature Publishing Group, pp. 1324–1332. DOI: 10.1038/s41386-021-01002-9.

[42] K. E. Regan-Fendt, J. Xu, M. DiVincenzo, M. C. Duggan, R. Shakya, R. Na, W. E. Carson, P. R. O. Payne, and F. Li. “Synergy from gene expression and network mining (SynGeNet) method predicts synergistic drug combinations for diverse melanoma genomic subtypes”. In: *npj Systems Biology and Applications* 5.1 (Feb. 26, 2019). Publisher: Nature Publishing Group, pp. 1–15. DOI: 10.1038/s41540-019-0085-4.

[43] S. Wang and Q. Cheng. “Microarray Analysis in Drug Discovery and Clinical Applications”. In: *Bioinformatics and Drug Discovery*. Ed. by R. S. Larson. Totowa, NJ: Humana Press, 2006, pp. 49–65. DOI: 10.1385/1-59259-964-8:49.

[44] E. E. Schadt et al. “An integrative genomics approach to infer causal associations between gene expression and disease”. In: *Nature Genetics* 37.7 (July 2005). Publisher: Nature Publishing Group, pp. 710–717. DOI: 10.1038/ng1589.

[45] J. J. R. Burns, B. T. Shealy, M. S. Greer, J. A. Hadish, M. T. McGowan, T. Biggs, M. C. Smith, F. A. Feltus, and S. P. Ficklin. “Addressing noise in co-expression network construction”. In: *Briefings in Bioinformatics* 23.1 (Jan. 1, 2022), bbab495. DOI: 10.1093/bib/bbab495.

[46] S. Das and S. N. Rai. “Statistical Approach for Biologically Relevant Gene Selection from High-Throughput Gene Expression Data”. In: *Entropy* 22.11 (Nov. 2020). Number: 11 Publisher: Multidisciplinary Digital Publishing Institute, p. 1205. DOI: 10.3390/e22111205.

[47] B. Hwang, J. H. Lee, and D. Bang. “Single-cell RNA sequencing technologies and bioinformatics pipelines”. In: *Experimental & Molecular Medicine* 50.8 (Aug. 2018). Publisher: Nature Publishing Group, pp. 1–14. DOI: 10.1038/s12276-018-0071-8.

[48] A. S. E. Cuomo, A. Nathan, S. Raychaudhuri, D. G. MacArthur, and J. E. Powell. “Single-cell genomics meets human genetics”. In: *Nature Reviews Genetics* 24.8 (Aug. 2023). Publisher: Nature Publishing Group, pp. 535–549. DOI: 10.1038/s41576-023-00599-5.

[49] J. Lee, D. Y. Hyeon, and D. Hwang. “Single-cell multiomics: technologies and data analysis methods”. In: *Experimental & Molecular Medicine* 52.9 (Sept. 2020). Publisher: Nature Publishing Group, pp. 1428–1442. DOI: 10.1038/s12276-020-0420-2.

[50] D. R. Brademan, I. J. Miller, N. W. Kwiecien, D. J. Pagliarini, M. S. Westphall, J. J. Coon, and E. Shishkova. “Argonaut: A Web Platform for Collaborative Multi-omic Data Visualization and Exploration”. In: *Patterns* 1.7 (Oct. 9, 2020), p. 100122. DOI: 10.1016/j.patter.2020.100122.

[51] C. W. T. Koh, J. S. G. Ooi, E. Z. Ong, and K. R. Chan. “STAGEs: A web-based tool that integrates data visualization and pathway enrichment analysis for gene expression studies”. In: *Scientific Reports* 13.1 (May 2, 2023). Publisher: Nature Publishing Group, p. 7135. DOI: 10.1038/s41598-023-34163-2.

[52] S. Ghandikota, G. K. K. Hershey, and T. B. Mersha. "GENEASE: real time bioinformatics tool for multi-omics and disease ontology exploration, analysis and visualization". In: *Bioinformatics* 34.18 (Sept. 15, 2018), pp. 3160–3168. DOI: 10.1093/bioinformatics/bty182.

[53] Ž. Avsec, V. Agarwal, D. Visentin, J. R. Ledsam, A. Grabska-Barwinska, K. R. Taylor, Y. Assael, J. Jumper, P. Kohli, and D. R. Kelley. "Effective gene expression prediction from sequence by integrating long-range interactions". In: *Nature Methods* 18.10 (Oct. 2021). Publisher: Nature Publishing Group, pp. 1196–1203. DOI: 10.1038/s41592-021-01252-x.

[54] M. Abbas and Y. EL-Manzalawy. "Machine learning based refined differential gene expression analysis of pediatric sepsis". In: *BMC Medical Genomics* 13.1 (Aug. 28, 2020), p. 122. DOI: 10.1186/s12920-020-00771-4.

[55] B. A. Alpay, P. Demetci, S. Istrail, and D. Aguiar. "Combinatorial and statistical prediction of gene expression from haplotype sequence". In: *Bioinformatics* 36 (Supplement_1 July 1, 2020), pp. i194–i202. DOI: 10.1093/bioinformatics/btaa318.

[56] N. Al taweraqi and R. D. King. "Improved prediction of gene expression through integrating cell signalling models with machine learning". In: *BMC Bioinformatics* 23.1 (Aug. 6, 2022), p. 323. DOI: 10.1186/s12859-022-04787-8.

[57] A. Ma'ayan. "Introduction to Network Analysis in Systems Biology". In: *Science Signaling* 4.190 (Sept. 13, 2011). Publisher: American Association for the Advancement of Science, tr5–tr5. DOI: 10.1126/scisignal.2001965.

[58] D. J. Watts and S. H. Strogatz. "Collective dynamics of 'small-world' networks". In: *Nature* 393.6684 (June 1998). Publisher: Nature Publishing Group, pp. 440–442. DOI: 10.1038/30918.

[59] A.-L. Barabási and R. Albert. "Emergence of Scaling in Random Networks". In: *Science* 286.5439 (Oct. 15, 1999). Publisher: American Association for the Advancement of Science, pp. 509–512. DOI: 10.1126/science.286.5439.509.

[60] M. Koutrouli, E. Karatzas, D. Paez-Espino, and G. A. Pavlopoulos. "A Guide to Conquer the Biological Network Era Using Graph Theory". In: *Frontiers in Bioengineering and Biotechnology* 8 (Jan. 31, 2020). Publisher: Frontiers. DOI: 10.3389/fbioe.2020.00034.

[61] C. Gaiteri, D. R. Connell, F. A. Sultan, A. Iatrou, B. Ng, B. K. Szymanski, A. Zhang, and S. Tasaki. "Robust, scalable, and informative clustering for diverse biological networks". In: *Genome Biology* 24.1 (Oct. 12, 2023), p. 228. DOI: 10.1186/s13059-023-03062-0.

[62] G. Muzio, L. O'Bray, and K. Borgwardt. "Biological network analysis with deep learning". In: *Briefings in Bioinformatics* 22.2 (Mar. 1, 2021), pp. 1515–1530. DOI: 10.1093/bib/bbaa257.

[63] M. Beguerisse-Díaz, G. Bosque, D. Oyarzún, J. Picó, and M. Barahona. "Flux-dependent graphs for metabolic networks". In: *npj Systems Biology and Applications* 4.1 (Aug. 14, 2018). Publisher: Nature Publishing Group, pp. 1–14. DOI: 10.1038/s41540-018-0067-y.

[64] C. Ramon and J. Stelling. "Functional comparison of metabolic networks across species". In: *Nature Communications* 14.1 (Mar. 27, 2023). Publisher: Nature Publishing Group, p. 1699. DOI: 10.1038/s41467-023-37429-5.

[65] N. Safari-Alighiarloo, M. Taghizadeh, M. Rezaei-Tavirani, B. Goliaei, and A. A. Peyvandi. "Protein-protein interaction networks (PPI) and complex diseases". In: *Gastroenterology and Hepatology From Bed to Bench* 7.1 (2014), p. 17.

[66] G. C. K. W. Koh, P. Porras, B. Aranda, H. Hermjakob, and S. E. Orchard. "Analyzing Protein–Protein Interaction Networks". In: *Journal of Proteome Research* 11.4 (Apr. 6, 2012). Publisher: American Chemical Society, pp. 2014–2031. DOI: 10.1021/pr201211w.

[67] F. Emmert-Streib, M. Dehmer, and B. Haibe-Kains. "Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks". In: *Frontiers in Cell and Developmental Biology* 2 (Aug. 19, 2014). Publisher: Frontiers. DOI: 10.3389/fcell.2014.00038.

[68] R. M. Gordley, L. J. Bugaj, and W. A. Lim. "Modular engineering of cellular signaling proteins and networks". In: *Current Opinion in Structural Biology*. Engineering and design • Membranes 39 (Aug. 1, 2016), pp. 106–114. DOI: 10.1016/j.sbi.2016.06.012.

[69] J. J. Tyson and B. Novák. "Functional Motifs in Biochemical Reaction Networks". In: *Annual Review of Physical Chemistry* 61 (Volume 61, 2010 May 5, 2010). Publisher: Annual Reviews, pp. 219–240. DOI: 10.1146/annurev.physchem.012809.103457.

[70] A. Oulas, G. Minadakis, M. Zachariou, K. Sokratous, M. M. Bourdakou, and G. M. Spyrou. “Systems Bioinformatics: increasing precision of computational diagnostics and therapeutics through network-based approaches”. In: *Briefings in Bioinformatics* 20.3 (May 21, 2019), pp. 806–824. DOI: 10.1093/bib/bbx151.

[71] H. Zhou, W. Wang, J. Jin, Z. Zheng, and B. Zhou. “Graph Neural Network for Protein–Protein Interaction Prediction: A Comparative Study”. In: *Molecules* 27.18 (Jan. 2022). Number: 18 Publisher: Multidisciplinary Digital Publishing Institute, p. 6135. DOI: 10.3390/molecules27186135.

[72] G. Fiscon, F. Conte, L. Farina, and P. Paci. “Network-Based Approaches to Explore Complex Biological Systems towards Network Medicine”. In: *Genes* 9.9 (Sept. 2018). Number: 9 Publisher: Multidisciplinary Digital Publishing Institute, p. 437. DOI: 10.3390/genes9090437.

[73] J. Yan, S. L. Risacher, L. Shen, and A. J. Saykin. “Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data”. In: *Briefings in Bioinformatics* 19.6 (Nov. 27, 2018), pp. 1370–1381. DOI: 10.1093/bib/bbx066.

[74] F. E. Agamah, J. R. Bayjanov, A. Niehues, K. F. Njoku, M. Skelton, G. K. Mazandu, T. H. A. Ederveen, N. Mulder, E. R. Chimusa, and P. A. C. ’t Hoen. “Computational approaches for network-based integrative multi-omics analysis”. In: *Frontiers in Molecular Biosciences* 9 (Nov. 14, 2022). Publisher: Frontiers. DOI: 10.3389/fmolb.2022.967205.

[75] H. W. L. Koh, D. Fermin, C. Vogel, K. P. Choi, R. M. Ewing, and H. Choi. “iOmicsPASS: network-based integration of multiomics data for predictive subnetwork discovery”. In: *npj Systems Biology and Applications* 5.1 (July 9, 2019). Publisher: Nature Publishing Group, pp. 1–10. DOI: 10.1038/s41540-019-0099-y.

[76] G. Galindez, S. Sadegh, J. Baumbach, T. Kacprowski, and M. List. “Network-based approaches for modeling disease regulation and progression”. In: *Computational and Structural Biotechnology Journal* 21 (Jan. 1, 2023). Publisher: Elsevier, pp. 780–795. DOI: 10.1016/j.csbj.2022.12.022.

[77] M. Zitnik et al. *Current and future directions in network biology*. June 11, 2024. DOI: 10.48550/arXiv.2309.08478.

[78] F. R. Pinu, D. J. Beale, A. M. Paten, K. Kouremenos, S. Swarup, H. J. Schirra, and D. Wishart. “Systems Biology and Multi-Omics Integration: Viewpoints from the Metabolomics Research Community”. In: *Metabolites* 9.4 (Apr. 2019). Number: 4 Publisher: Multidisciplinary Digital Publishing Institute, p. 76. DOI: 10.3390/metabo9040076.

[79] K. Azer and I. Leaf. “Systems biology platform for efficient development and translation of multitargeted therapeutics”. In: *Frontiers in Systems Biology* 3 (Sept. 18, 2023). Publisher: Frontiers. DOI: 10.3389/fsysb.2023.1229532.

[80] E. Uffelmann, Q. Q. Huang, N. S. Munung, J. de Vries, Y. Okada, A. R. Martin, H. C. Martin, T. Lappalainen, and D. Posthuma. “Genome-wide association studies”. In: *Nature Reviews Methods Primers* 1.1 (Aug. 26, 2021). Publisher: Nature Publishing Group, pp. 1–21. DOI: 10.1038/s43586-021-00056-9.

[81] A. Chandra, D. Mitry, A. Wright, H. Campbell, and D. G. Charteris. “Genome-wide association studies: applications and insights gained in Ophthalmology”. In: *Eye* 28.9 (Sept. 2014). Publisher: Nature Publishing Group, pp. 1066–1079. DOI: 10.1038/eye.2014.145.

[82] R. J. F. Loos. “15 years of genome-wide association studies and no signs of slowing down”. In: *Nature Communications* 11.1 (Nov. 19, 2020). Publisher: Nature Publishing Group, p. 5900. DOI: 10.1038/s41467-020-19653-5.

[83] V. Tam, N. Patel, M. Turcotte, Y. Bossé, G. Paré, and D. Meyre. “Benefits and limitations of genome-wide association studies”. In: *Nature Reviews Genetics* 20.8 (Aug. 2019). Publisher: Nature Publishing Group, pp. 467–484. DOI: 10.1038/s41576-019-0127-1.

[84] A. P. Patel et al. “A multi-ancestry polygenic risk score improves risk prediction for coronary artery disease”. In: *Nature Medicine* 29.7 (July 2023). Publisher: Nature Publishing Group, pp. 1793–1803. DOI: 10.1038/s41591-023-02429-x.

[85] C. Khunsriraksakul, H. Markus, N. J. Olsen, L. Carrel, B. Jiang, and D. J. Liu. “Construction and Application of Polygenic Risk Scores in Autoimmune Diseases”. In: *Frontiers in Immunology* 13 (June 27, 2022). Publisher: Frontiers. DOI: 10.3389/fimmu.2022.889296.

[86] J. H. Moore, F. W. Asselbergs, and S. M. Williams. "Bioinformatics challenges for genome-wide association studies". In: *Bioinformatics* 26.4 (Feb. 15, 2010), pp. 445–455. DOI: 10.1093/bioinformatics/btp713.

[87] J. A. Riancho. "Genome-Wide Association Studies (GWAS) in Complex Diseases: Advantages and Limitations". In: *Reumatología Clínica (English Edition)* 8.2 (Mar. 1, 2012). Publisher: Elsevier, pp. 56–57. DOI: 10.1016/j.reumae.2011.07.007.

[88] T. S. Andrews, V. Y. Kiselev, D. McCarthy, and M. Hemberg. "Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data". In: *Nature Protocols* 16.1 (Jan. 2021). Publisher: Nature Publishing Group, pp. 1–9. DOI: 10.1038/s41596-020-00409-w.

[89] L. Li et al. "What are the applications of single-cell RNA sequencing in cancer research: a systematic review". In: *Journal of Experimental & Clinical Cancer Research* 40.1 (May 11, 2021), p. 163. DOI: 10.1186/s13046-021-01955-1.

[90] B. Van de Sande et al. "Applications of single-cell RNA sequencing in drug discovery and development". In: *Nature Reviews Drug Discovery* 22.6 (June 2023). Publisher: Nature Publishing Group, pp. 496–520. DOI: 10.1038/s41573-023-00688-4.

[91] D. S. Wishart. "Emerging applications of metabolomics in drug discovery and precision medicine". In: *Nature Reviews Drug Discovery* 15.7 (July 2016). Publisher: Nature Publishing Group, pp. 473–484. DOI: 10.1038/nrd.2016.32.

[92] V. Gonzalez-Covarrubias, E. Martínez-Martínez, and L. del Bosque-Plata. "The Potential of Metabolomics in Biomedical Applications". In: *Metabolites* 12.2 (Feb. 2022). Number: 2 Publisher: Multidisciplinary Digital Publishing Institute, p. 194. DOI: 10.3390/metabo12020194.

[93] A.-H. Emwas et al. "NMR Spectroscopy for Metabolomics Research". In: *Metabolites* 9.7 (July 2019). Number: 7 Publisher: Multidisciplinary Digital Publishing Institute, p. 123. DOI: 10.3390/metabo9070123.

[94] S. Anvari, J. Miller, C.-Y. Yeh, and C. M. Davis. "IgE-Mediated Food Allergy". In: *Clinical Reviews in Allergy & Immunology* 57.2 (Oct. 1, 2019), pp. 244–260. DOI: 10.1007/s12016-018-8710-3.

[95] Y. Chen, E.-M. Li, and L.-Y. Xu. "Guide to Metabolomics Analysis: A Bioinformatics Workflow". In: *Metabolites* 12.4 (Apr. 2022). Number: 4 Publisher: Multidisciplinary Digital Publishing Institute, p. 357. DOI: 10.3390/metabo12040357.

[96] E. Dekaboruah, M. V. Suryavanshi, D. Chettri, and A. K. Verma. "Human microbiome: an academic update on human body site specific surveillance and its possible role". In: *Archives of Microbiology* 202.8 (Oct. 1, 2020), pp. 2147–2167. DOI: 10.1007/s00203-020-01931-x.

[97] G. A. Ogunrinola, J. O. Oyewale, O. O. Oshamika, and G. I. Olasehinde. "The Human Microbiome and Its Impacts on Health". In: *International Journal of Microbiology* 2020.1 (2020). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1155/2020/8045646, p. 8045646. DOI: 10.1155/2020/8045646.

[98] M. J. Bull and N. T. Plummer. "Part 1: The Human Gut Microbiome in Health and Disease". In: *Integrative Medicine (Encinitas, Calif.)* 13.6 (Dec. 2014), pp. 17–22.

[99] G. Ramos Meyers, H. Samouda, and T. Bohn. "Short Chain Fatty Acid Metabolism in Relation to Gut Microbiota and Genetic Variability". In: *Nutrients* 14.24 (Jan. 2022). Number: 24 Publisher: Multidisciplinary Digital Publishing Institute, p. 5361. DOI: 10.3390/nu14245361.

[100] E. C. Gotschlich, R. A. Colbert, and T. Gill. "Methods in microbiome research: Past, present, and future". In: *Best Practice & Research Clinical Rheumatology*. Microbiome in Musculoskeletal Science 33.6 (Dec. 1, 2019), p. 101498. DOI: 10.1016/j.berh.2020.101498.

[101] L. J. Marcos-Zambrano et al. "Applications of Machine Learning in Human Microbiome Studies: A Review on Feature Selection, Biomarker Identification, Disease Prediction and Treatment". In: *Frontiers in Microbiology* 12 (Feb. 19, 2021). Publisher: Frontiers. DOI: 10.3389/fmicb.2021.634511.

[102] E. L. Gulliver et al. "Review article: the future of microbiome-based therapeutics". In: *Alimentary Pharmacology & Therapeutics* 56.2 (2022). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/apt.17049, pp. 192–208. DOI: 10.1111/apt.17049.

[103] D. Vandeputte. "Personalized Nutrition Through The Gut Microbiota: Current Insights And Future Perspectives". In: *Nutrition Reviews* 78 (Supplement_3 Nov. 30, 2020), pp. 66–74. DOI: 10.1093/nutrit/nuaa098.

[104] R. Omer, M. Z. Mohsin, A. Mohsin, B. S. Mushtaq, X. Huang, M. Guo, Y. Zhuang, and J. Huang. "Engineered Bacteria-Based Living Materials for Biotherapeutic Applications". In: *Frontiers in Bioengineering and Biotechnology* 10 (Apr. 28, 2022). Publisher: Frontiers. DOI: 10.3389/fbioe.2022.870675.

[105] C. M. Cullen et al. "Emerging Priorities for Microbiome Research". In: *Frontiers in Microbiology* 11 (Feb. 19, 2020). Publisher: Frontiers. DOI: 10.3389/fmicb.2020.00136.

[106] E. López de Maturana, L. Alonso, P. Alarcón, I. A. Martín-Antoniano, S. Pineda, L. Piorno, M. L. Calle, and N. Malats. "Challenges in the Integration of Omics and Non-Omics Data". In: *Genes* 10.3 (Mar. 2019). Number: 3 Publisher: Multidisciplinary Digital Publishing Institute, p. 238. DOI: 10.3390/genes10030238.

[107] C. Chen, J. Wang, D. Pan, X. Wang, Y. Xu, J. Yan, L. Wang, X. Yang, M. Yang, and G.-P. Liu. "Applications of multi-omics analysis in human diseases". In: *MedComm* 4.4 (2023). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mco2.315, e315. DOI: 10.1002/mco2.315.

[108] T. Ivanisevic and R. N. Sewduth. "Multi-Omics Integration for the Design of Novel Therapies and the Identification of Novel Biomarkers". In: *Proteomes* 11.4 (Dec. 2023). Number: 4 Publisher: Multidisciplinary Digital Publishing Institute, p. 34. DOI: 10.3390/proteomes11040034.

[109] N. Vahabi and G. Michailidis. "Unsupervised Multi-Omics Data Integration Methods: A Comprehensive Review". In: *Frontiers in Genetics* 13 (Mar. 22, 2022). Publisher: Frontiers. DOI: 10.3389/fgene.2022.854752.

[110] Y. Zheng et al. "Multi-omics data integration using ratio-based quantitative profiling with Quartet reference materials". In: *Nature Biotechnology* 42.7 (July 2024). Publisher: Nature Publishing Group, pp. 1133–1149. DOI: 10.1038/s41587-023-01934-1.

[111] M. Picard, M.-P. Scott-Boyer, A. Bodein, O. Périn, and A. Droit. "Integration strategies of multi-omics data for machine learning analysis". In: *Computational and Structural Biotechnology Journal* 19 (Jan. 1, 2021), pp. 3735–3746. DOI: 10.1016/j.csbj.2021.06.030.

[112] T. Jendoubi. "Approaches to Integrating Metabolomics and Multi-Omics Data: A Primer". In: *Metabolites* 11.3 (Mar. 2021). Number: 3 Publisher: Multidisciplinary Digital Publishing Institute, p. 184. DOI: 10.3390/metabo11030184.

[113] G. Zampieri, S. Vijayakumar, E. Yaneske, and C. Angione. “Machine and deep learning meet genome-scale metabolic modeling”. In: *PLOS Computational Biology* 15.7 (July 11, 2019). Publisher: Public Library of Science, e1007084. DOI: 10.1371/journal.pcbi.1007084.

[114] K. Zheng, Y. Hai, Y. Xi, Y. Zhang, Z. Liu, W. Chen, X. Hu, X. Zou, and J. Hao. “Integrative multi-omics analysis unveils stemness-associated molecular subtypes in prostate cancer and pan-cancer: prognostic and therapeutic significance”. In: *Journal of Translational Medicine* 21.1 (Nov. 7, 2023), p. 789. DOI: 10.1186/s12967-023-04683-6.

[115] X. Li, J. Ma, L. Leng, M. Han, M. Li, F. He, and Y. Zhu. “MoGCN: A Multi-Omics Integration Method Based on Graph Convolutional Network for Cancer Subtype Analysis”. In: *Frontiers in Genetics* 13 (Feb. 2, 2022). Publisher: Frontiers. DOI: 10.3389/fgene.2022.806842.

[116] M. Bersanelli, E. Mosca, D. Remondini, E. Giampieri, C. Sala, G. Castellani, and L. Milanesi. “Methods for the integration of multi-omics data: mathematical aspects”. In: *BMC Bioinformatics* 17.2 (Jan. 20, 2016), S15. DOI: 10.1186/s12859-015-0857-9.

[117] S. Tarazona, A. Arzalluz-Luque, and A. Conesa. “Undisclosed, unmet and neglected challenges in multi-omics studies”. In: *Nature Computational Science* 1.6 (June 2021). Publisher: Nature Publishing Group, pp. 395–402. DOI: 10.1038/s43588-021-00086-z.

[118] D. Gomez-Cabrero, I. Abugessaisa, D. Maier, A. Teschendorff, M. Merkenschlager, A. Gisel, E. Ballestar, E. Bongcam-Rudloff, A. Conesa, and J. Tegnér. “Data integration in the era of omics: current and future challenges”. In: *BMC Systems Biology* 8.2 (Mar. 13, 2014), p. I1. DOI: 10.1186/1752-0509-8-S2-I1.

[119] J. E. Flores, D. M. Claborne, Z. D. Weller, B.-J. M. Webb-Robertson, K. M. Waters, and L. M. Bramer. “Missing data in multi-omics integration: Recent advances through artificial intelligence”. In: *Frontiers in Artificial Intelligence* 6 (Feb. 9, 2023). Publisher: Frontiers. DOI: 10.3389/frai.2023.1098308.

[120] M. T. Odenkirk, D. M. Reif, and E. S. Baker. “Multiomic Big Data Analysis Challenges: Increasing Confidence in the Interpretation of Artificial Intelligence Assessments”. In: *Analytical Chemistry* 93.22 (June 8, 2021). Publisher: American Chemical Society, pp. 7763–7773. DOI: 10.1021/acs.analchem.0c04850.

[121] M. Krassowski, V. Das, S. K. Sahu, and B. B. Misra. "State of the Field in Multi-Omics Research: From Computational Needs to Data Mining and Sharing". In: *Frontiers in Genetics* 11 (Dec. 10, 2020). Publisher: Frontiers. DOI: 10.3389/fgene.2020.610798.

[122] M. Bersanelli, E. Mosca, D. Remondini, E. Giampieri, C. Sala, G. Castellani, and L. Milanesi. "Methods for the integration of multi-omics data: mathematical aspects". In: *BMC Bioinformatics* 17.2 (Jan. 20, 2016), S15. DOI: 10.1186/s12859-015-0857-9.

[123] F. Jiang, X. Huang, F. Zhang, J. Pan, J. Wang, L. Hu, J. Chen, and Y. Wang. "Integrated Analysis of Multi-Omics Data to Identify Prognostic Genes for Pancreatic Cancer". In: *DNA and Cell Biology* 41.3 (Mar. 2022). Publisher: Mary Ann Liebert, Inc., publishers, pp. 305–318. DOI: 10.1089/dna.2021.0878.

[124] F. Xing, R. Zheng, B. Liu, K. Huang, D. Wang, R. Su, and S. Feng. "A new strategy for searching determinants in colorectal cancer progression through whole-part relationship combined with multi-omics". In: *Talanta* 259 (July 1, 2023), p. 124543. DOI: 10.1016/j.talanta.2023.124543.

[125] I. Ullah, L. Yang, F.-T. Yin, Y. Sun, X.-H. Li, J. Li, and X.-J. Wang. "Multi-Omics Approaches in Colorectal Cancer Screening and Diagnosis, Recent Updates and Future Perspectives". In: *Cancers* 14.22 (Jan. 2022). Number: 22 Publisher: Multidisciplinary Digital Publishing Institute, p. 5545. DOI: 10.3390/cancers14225545.

[126] X. Li, M. A. Abdel-Maksoud, I. Iqbal, A. Mubarak, M. A. Farrag, M. Haris, S. Alghamdi, Q. U. Ain, and S. Almekhlafi. "Deciphering cervical cancer-associated biomarkers by integrated multi-omics approach". In: *American Journal of Translational Research* 14.12 (Dec. 15, 2022), p. 8843.

[127] R. Recio-Vega, R. A. Facio-Campos, S. I. Hernández-González, and E. Olivas-Calderón. "State of the Art of Genomic Technology in Toxicology: A Review". In: *International Journal of Molecular Sciences* 24.11 (June 1, 2023), p. 9618. DOI: 10.3390/ijms24119618.

[128] A. Baysoy, Z. Bai, R. Satija, and R. Fan. "The technological landscape and applications of single-cell multi-omics". In: *Nature Reviews Molecular Cell Biology* 24.10 (Oct. 2023). Publisher: Nature Publishing Group, pp. 695–713. DOI: 10.1038/s41580-023-00615-w.

[129] K. Regan and P. R. Payne. “From Molecules to Patients: The Clinical Applications of Translational Bioinformatics”. In: *Yearbook of Medical Informatics* 10.1 (Aug. 13, 2015), pp. 164–169. DOI: 10.15265/IY-2015-005.

[130] R. C. Wang and Z. Wang. “Precision Medicine: Disease Subtyping and Tailored Treatment”. In: *Cancers* 15.15 (July 28, 2023), p. 3837. DOI: 10.3390/cancers15153837.

[131] S. K. Gill, A. F. Christopher, V. Gupta, and P. Bansal. “Emerging role of bioinformatics tools and software in evolution of clinical research”. In: *Perspectives in Clinical Research* 7.3 (2016), pp. 115–122. DOI: 10.4103/2229-3485.184782.

[132] A. Sivadas and V. Scaria. “Population-scale genomics-Enabling precision public health”. In: *Advances in Genetics* 103 (2019), pp. 119–161. DOI: 10.1016/bs.adgen.2018.09.001.

[133] D. W. Huang, B. T. Sherman, and R. A. Lempicki. “Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists”. In: *Nucleic Acids Research* 37.1 (Nov. 25, 2008), p. 1. DOI: 10.1093/nar/gkn923.

[134] C. Weidner, M. Steinfath, E. Wistorf, M. Oelgeschläger, M. R. Schneider, and G. Schönfelder. “A Protocol for Using Gene Set Enrichment Analysis to Identify the Appropriate Animal Model for Translational Research”. In: *Journal of Visualized Experiments : JoVE* 126 (Aug. 16, 2017), p. 55768. DOI: 10.3791/55768.

[135] M. Paczkowska, J. Barenboim, N. Sintupisut, N. S. Fox, H. Zhu, D. Abd-Rabbo, M. W. Mee, P. C. Boutros, and J. Reimand. “Integrative pathway enrichment analysis of multivariate omics data”. In: *Nature Communications* 11.1 (Feb. 5, 2020). Publisher: Nature Publishing Group, p. 735. DOI: 10.1038/s41467-019-13983-9.

[136] D. Barh, V. Chaitankar, E. C. Yiannakopoulou, E. O. Salawu, S. Chowbina, P. Ghosh, and V. Azevedo. “*In Silico* Models”. In: *Animal Biotechnology*. Ed. by A. S. Verma and A. Singh. San Diego: Academic Press, Jan. 1, 2014, pp. 385–404. DOI: 10.1016/B978-0-12-416002-6.00021-3.

[137] S. C. Pingle et al. “In silico modeling predicts drug sensitivity of patient-derived cancer cells”. In: *Journal of Translational Medicine* 12.1 (May 21, 2014), p. 128. DOI: 10.1186/1479-5876-12-128.

[138] A. Mazein et al. “Systems medicine disease maps: community-driven comprehensive representation of disease mechanisms”. In: *npj Systems Biology and Applications* 4.1 (June 2, 2018). Publisher: Nature Publishing Group, pp. 1–10. DOI: 10.1038/s41540-018-0059-y.

[139] L. Vinhoven, M. Voskamp, and M. M. Nietert. “Mapping Compound Databases to Disease Maps—A MINERVA Plugin for CandActBase”. In: *Journal of Personalized Medicine* 11.11 (Nov. 2021). Number: 11 Publisher: Multidisciplinary Digital Publishing Institute, p. 1072. DOI: 10.3390/jpm11111072.

[140] M. Ziemann, P. Poulain, and A. Bora. “The five pillars of computational reproducibility: bioinformatics and beyond”. In: *Briefings in Bioinformatics* 24.6 (Nov. 1, 2023), bbad375. DOI: 10.1093/bib/bbad375.

[141] V. Gligorijević and N. Pržulj. “Methods for biological data integration: perspectives and challenges”. In: *Journal of The Royal Society Interface* 12.112 (Nov. 6, 2015). Publisher: Royal Society, p. 20150571. DOI: 10.1098/rsif.2015.0571.

[142] F. o. N.
bibinitperiod N. S. Disorders, B. o. H. S. Policy, and I. o. Medicine. “Drug Development Challenges”. In: *Improving and Accelerating Therapeutic Development for Nervous System Disorders: Workshop Summary*. National Academies Press (US), Feb. 6, 2014.

[143] D. Institute of Medicine (US) Forum on Drug Discovery. “Challenges in Clinical Research”. In: *Transforming Clinical Research in the United States: Challenges and Opportunities: Workshop Summary*. National Academies Press (US), 2010.

[144] C. H. Emmerich et al. “Improving target assessment in biomedical research: the GOT-IT recommendations”. In: *Nature Reviews Drug Discovery* 20.1 (Jan. 2021). Publisher: Nature Publishing Group, pp. 64–81. DOI: 10.1038/s41573-020-0087-3.

[145] D. Sun, W. Gao, H. Hu, and S. Zhou. “Why 90% of clinical drug development fails and how to improve it?” In: *Acta Pharmaceutica Sinica B* 12.7 (July 1, 2022), pp. 3049–3062. DOI: 10.1016/j.apsb.2022.02.002.

[146] K. Takashima, Y. Maru, S. Mori, H. Mano, T. Noda, and K. Muto. “Ethical concerns on sharing genomic data including patients’ family members”. In: *BMC Medical Ethics* 19.1 (June 18, 2018), p. 61. DOI: 10.1186/s12910-018-0310-5.

[147] S. S.-J. Lee. “The Ethics of Consent in a Shifting Genomic Ecosystem”. In: *Annual Review of Biomedical Data Science* 4 (Volume 4, 2021 July 20, 2021). Publisher: Annual Reviews, pp. 145–164. DOI: 10.1146/annurev-biodatasci-030221-125715.

[148] K. B. Brothers and M. A. Rothstein. “Ethical, Legal and Social Implications of Incorporating Personalized Medicine into Healthcare”. In: *Personalized Medicine* 12.1 (Jan. 1, 2015). Publisher: Taylor & Francis _eprint: https://doi.org/10.2217/pme.14.65, pp. 43–51. DOI: 10.2217/pme.14.65.

[149] L. Chen, H. Deng, H. Cui, J. Fang, Z. Zuo, J. Deng, Y. Li, X. Wang, and L. Zhao. “Inflammatory responses and inflammation-associated diseases in organs”. In: *Oncotarget* 9.6 (Dec. 14, 2017), p. 7204. DOI: 10.18632/oncotarget.23208.

[150] D. Furman et al. “Chronic inflammation in the etiology of disease across the life span”. In: *Nature medicine* 25.12 (Dec. 2019), pp. 1822–1832. DOI: 10.1038/s41591-019-0675-0.

[151] H. Zhao, L. Wu, G. Yan, Y. Chen, M. Zhou, Y. Wu, and Y. Li. “Inflammation and tumor progression: signaling pathways and targeted intervention”. In: *Signal Transduction and Targeted Therapy* 6.1 (July 12, 2021). Publisher: Nature Publishing Group, pp. 1–46. DOI: 10.1038/s41392-021-00658-5.

[152] L. Yu, M.-M. Zhang, and J.-G. Hou. “Molecular and cellular pathways in colorectal cancer: apoptosis, autophagy and inflammation as key players”. In: *Scandinavian Journal of Gastroenterology* 57.11 (Nov. 2, 2022). Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00365521.2022.2088247, pp. 1279–1290. DOI: 10.1080/00365521.2022.2088247.

[153] A. Derakhshani, D. Javadrashid, N. Hemmat, A. Dufour, A. G. Solimando, M. Abdoli Shadbad, P. H. G. Duijf, O. Brunetti, N. Silvestris, and B. Baradaran. “Identification of Common and Distinct Pathways in Inflammatory Bowel Disease and Colorectal Cancer: A Hypothesis Based on Weighted Gene Co-Expression Network Analysis”. In: *Frontiers in Genetics* 13 (Mar. 31, 2022). Publisher: Frontiers. DOI: 10.3389/fgene.2022.848646.

[154] F. Mariani, P. Sena, and L. Roncucci. “Inflammatory pathways in the early steps of colorectal cancer development”. In: *World Journal of Gastroenterology* 20.29 (Aug. 7, 2014). Publisher: Baishideng Publishing Group Inc., pp. 9716–9731. DOI: 10.3748/wjg.v20.i29.9716.

[155] M. W. Dougherty and C. Jobin. “Intestinal bacteria and colorectal cancer: etiology and treatment”. In: *Gut Microbes* 15.1 (Dec. 31, 2023). Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/19490976.2023.2185028, p. 2185028. DOI: 10.1080/19490976.2023.2185028.

[156] H. Feizi, M. A. Rezaee, R. Ghotaslou, M. Sadrkabir, F. Jadidi-Niaragh, P. Gholizadeh, S. Taghizadeh, K. Ghanbarov, M. Yousefi, and H. S. Kafil. “Gut Microbiota and Colorectal Cancer Risk Factors”. In: *Current Pharmaceutical Biotechnology* 24.8 (2022), pp. 1018–1034. DOI: 10.2174/1389201023666221005103340.

[157] Q. Ni, M. Li, and S. Yu. “Research Progress of Epithelial-mesenchymal Transition Treatment and Drug Resistance in Colorectal Cancer”. In: *Technology in Cancer Research & Treatment* 21 (2022), p. 15330338221081219. DOI: 10.1177/15330338221081219.

[158] G. Calviello, S. Serini, and E. Piccioni. “n-3 Polyunsaturated Fatty Acids and the Prevention of Colorectal Cancer: Molecular Mechanisms Involved”. In: *Current Medicinal Chemistry* 14.29 (), pp. 3059–3069. DOI: 10.2174/092986707782793934.

[159] A. A. Khan, M. T. Ashraf, F. M. Aldakheel, A. S. Yazgan, and R. Zaidi. “Deciphering the involvement of iron targets in colorectal cancer: a network biology approach”. In: *American Journal of Translational Research* 14.1 (Jan. 15, 2022), p. 440.

[160] A. S. Gheorghe et al. “Biochemical and Metabolical Pathways Associated with Microbiota-Derived Butyrate in Colorectal Cancer and Omega-3 Fatty Acids Implications: A Narrative Review”. In: *Nutrients* 14.6 (Jan. 2022). Number: 6 Publisher: Multidisciplinary Digital Publishing Institute, p. 1152. DOI: 10.3390/nu14061152.

[161] E. O’Rourke, H. Tang, A. Chin, A. Long, S. Sindher, and R. S. Chinthrajah. “Current insights: a systemic review of therapeutic options for peanut allergy”. In: *Current Opinion in Allergy and Clinical Immunology* 22.3 (June 2022), p. 188. DOI: 10.1097/ACI.0000000000000824.

[162] M. H. Shamji and R. J. Boyle. “New innovations in allergy treatment and phenotyping”. In: *Clinical & Experimental Allergy* 51.4 (2021). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cea.13630, pp. 514–517. DOI: 10.1111/cea.13630.

[163] G. A. Mueller, S. J. Maleki, and L. C. Pedersen. “The Molecular Basis of Peanut Allergy”. In: *Current Allergy and Asthma Reports* 14.5 (Mar. 16, 2014), p. 429. DOI: 10.1007/s11882-014-0429-5.

[164] T. Zhang, Y. Shi, Y. Zhao, J. Wang, M. Wang, B. Niu, and Q. Chen. "Different thermal processing effects on peanut allergenicity". In: *Journal of the Science of Food and Agriculture* 99.5 (2019). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jsfa.9430, pp. 2321–2328. DOI: 10.1002/jsfa.9430.

[165] Z. Shakoor, F. A. Al, A. B. Al, T. L. N. Al, and O. R. Y. Al. "Prevalence of IgG-mediated food intolerance among patients with allergic symptoms". In: *Annals of Saudi Medicine* 36.6 (Nov. 2016). Publisher: King Faisal Specialist Hospital & Research Centre, pp. 386–390. DOI: 10.5144/0256-4947.2016.386.

[166] G. den Besten, K. van Eunen, A. K. Groen, K. Venema, D.-J. Reijngoud, and B. M. Bakker. "The role of short-chain fatty acids in the interplay between diet, gut microbiota, and host energy metabolism". In: *Journal of Lipid Research* 54.9 (Sept. 1, 2013), pp. 2325–2340. DOI: 10.1194/jlr.R036012.

[167] J. Wang, N. Zhu, X. Su, Y. Gao, and R. Yang. "Gut-Microbiota-Derived Metabolites Maintain Gut and Systemic Immune Homeostasis". In: *Cells* 12.5 (Jan. 2023). Number: 5 Publisher: Multidisciplinary Digital Publishing Institute, p. 793. DOI: 10.3390/cells12050793.

[168] L.-M. Ney, M. Wipplinger, M. Grossmann, N. Engert, V. D. Wegner, and A. S. Mosig. "Short chain fatty acids: key regulators of the local and systemic immune response in inflammatory diseases and infections". In: *Open Biology* 13.3 (Mar. 29, 2023). Publisher: Royal Society, p. 230014. DOI: 10.1098/rsob.230014.

[169] M. A. R. Vinolo, H. G. Rodrigues, R. T. Nachbar, and R. Curi. "Regulation of Inflammation by Short Chain Fatty Acids". In: *Nutrients* 3.10 (Oct. 2011). Number: 10 Publisher: Molecular Diversity Preservation International, pp. 858–876. DOI: 10.3390/nu3100858.

[170] M. Asarat, V. Apostolopoulos, T. Vasiljevic, and O. Donkor. "Short-Chain Fatty Acids Regulate Cytokines and Th17/Treg Cells in Human Peripheral Blood Mononuclear Cells in vitro". In: *Immunological Investigations* 45.3 (Apr. 2, 2016). Publisher: Taylor & Francis _eprint: https://doi.org/10.3109/08820139.2015.1122613, pp. 205–222. DOI: 10.3109/08820139.2015.1122613.

[171] Q. Yan, S. Jia, D. Li, and J. Yang. "The role and mechanism of action of microbiota-derived short-chain fatty acids in neutrophils: From the activation to becoming potential biomarkers". In: *Biomedicine & Pharmacotherapy* 169 (Dec. 31, 2023), p. 115821.

DOI: 10.1016/j.biopha.2023.115821.

[172] Z. Zhang, H. Zhang, T. Chen, L. Shi, D. Wang, and D. Tang. "Regulatory role of short-chain fatty acids in inflammatory bowel disease". In: *Cell Communication and Signaling* 20.1 (May 11, 2022), p. 64. DOI: 10.1186/s12964-022-00869-5.

[173] C. H. Kim. "Control of lymphocyte functions by gut microbiota-derived short-chain fatty acids". In: *Cellular & Molecular Immunology* 18.5 (May 2021). Publisher: Nature Publishing Group, pp. 1161–1171. DOI: 10.1038/s41423-020-00625-0.

[174] S. Bunyavanich, S. L. Rifas-Shiman, T. A. Platts-Mills, L. Workman, J. E. Sordillo, M. W. Gillman, D. R. Gold, and A. A. Litonjua. "Peanut allergy prevalence among school-age children in a US cohort not selected for any disease". In: *The Journal of allergy and clinical immunology* 134.3 (Sept. 2014), pp. 753–755. DOI: 10.1016/j.jaci.2014.05.050.

[175] J. A. Lieberman, R. S. Gupta, R. C. Knibb, T. Haselkorn, S. Tilles, D. P. Mack, and G. Pouessel. "The global burden of illness of peanut allergy: A comprehensive literature review". In: *Allergy* 76.5 (May 2021), pp. 1367–1384. DOI: 10.1111/all.14666.

[176] R.-X. Foong, J. A. Dantzer, R. A. Wood, and A. F. Santos. "Improving Diagnostic Accuracy in Food Allergy". In: *The Journal of Allergy and Clinical Immunology. in Practice* 9.1 (Jan. 2021), pp. 71–80. DOI: 10.1016/j.jaip.2020.09.037.

[177] S. Al-Muhsen, A. E. Clarke, and R. S. Kagan. "Peanut allergy: an overview". In: *CMAJ: Canadian Medical Association Journal* 168.10 (May 13, 2003), pp. 1279–1285.

[178] M. Calvani, A. Bianchi, C. Reginelli, M. Peresso, and A. Testa. "Oral Food Challenge". In: *Medicina* 55.10 (Sept. 27, 2019), p. 651. DOI: 10.3390/medicina55100651.

[179] N. Koutlas, A. Stallings, G. Hall, C. Zhou, J. Kim-Chang, and T. Mousallem. "Pediatric oral food challenges in the outpatient setting: A single-center experience". In: *The Journal of Allergy and Clinical Immunology: Global* 3.1 (Nov. 8, 2023), p. 100187. DOI: 10.1016/j.jacig.2023.100187.

[180] G. Pouessel et al. "Food-Induced Anaphylaxis Reactions at School: A Room for Improvement". In: *Allergy* n/a (n/a Dec. 11, 2024). Publisher: John Wiley & Sons, Ltd. DOI: 10.1111/all.16427.

[181] J. Klueber et al. “High-dimensional immune profiles correlate with phenotypes of peanut allergy during food-allergic reactions”. In: *Allergy* 78.4 (2023). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/all.15408, pp. 1020–1035. DOI: 10.1111/all.15408.

[182] V. S. Wacleche, C. L. Tremblay, J.-P. Routy, and P. Ancuta. “The Biology of Monocytes and Dendritic Cells: Contribution to HIV Pathogenesis”. In: *Viruses* 10.2 (Feb. 6, 2018), p. 65. DOI: 10.3390/v10020065.

[183] S. Gupta and A. Agrawal. “Dendritic cells in inborn errors of immunity”. In: *Frontiers in Immunology* 14 (Jan. 23, 2023). Publisher: Frontiers. DOI: 10.3389/fimmu.2023.1080129.

[184] J. Gong, H. Zhan, Y. Liang, Q. He, and D. Cui. “Role of Th22 Cells in Human Viral Diseases”. In: *Frontiers in Medicine* 8 (Aug. 9, 2021). Publisher: Frontiers. DOI: 10.3389/fmed.2021.708140.

[185] M. P. d. Aguiar and J. H. Vieira. “Entrance to the multifaceted world of CD4+ T cell subsets”. In: *Exploration of Immunology* 4.2 (Mar. 5, 2024). Number: 2 Publisher: Open Exploration, pp. 152–168. DOI: 10.37349/ei.2024.00134.

[186] W. A. Rose, A. J. Okragly, N. N. Hu, M. R. Daniels, A. P. Martin, Y. T. Koh, K. Kikly, and R. J. Benschop. “Interleukin-33 Contributes Toward Loss of Tolerance by Promoting B-Cell-Activating Factor of the Tumor-Necrosis-Factor Family (BAFF)-Dependent Autoantibody Production”. In: *Frontiers in Immunology* 9 (Dec. 6, 2018). Publisher: Frontiers. DOI: 10.3389/fimmu.2018.02871.

[187] S. Chirumbolo, G. Bjørklund, A. Sboarina, and A. Vella. “The role of basophils as innate immune regulatory cells in allergy and immunotherapy”. In: *Human Vaccines & Immunotherapeutics* 14.4 (Jan. 18, 2018), pp. 815–831. DOI: 10.1080/21645515.2017.1417711.

[188] H. Jiang and J. Jiang. “Balancing act: the complex role of NK cells in immune regulation”. In: *Frontiers in Immunology* 14 (Nov. 2, 2023), p. 1275028. DOI: 10.3389/fimmu.2023.1275028.

[189] S. Chapoval, P. Dasgupta, N. J. Dorsey, and A. D. Keegan. “Regulation of the T helper cell type 2 (Th2)/T regulatory cell (Treg) balance by IL-4 and STAT6”. In: *Journal of Leukocyte Biology* 87.6 (June 2010), pp. 1011–1018. DOI: 10.1189/jlb.1209772.

[190] M. J. Butcher and J. Zhu. “Recent advances in understanding the Th1/Th2 effector choice”. In: *Faculty Reviews* 10 (Mar. 15, 2021), p. 30. DOI: 10.12703/r/10-30.

[191] P. Burt, M. Peine, C. Peine, Z. Borek, S. Serve, M. Floßdorf, A. N. Hegazy, T. Höfer, M. Löhning, and K. Thurley. “Dissecting the dynamic transcriptional landscape of early T helper cell differentiation into Th1, Th2, and Th1/2 hybrid cells”. In: *Frontiers in Immunology* 13 (Aug. 16, 2022). Publisher: Frontiers. DOI: 10.3389/fimmu.2022.928018.

[192] L. D. Faustino, J. W. Griffith, R. A. Rahimi, K. Nepal, D. L. Hamilos, J. L. Cho, B. D. Medoff, J. J. Moon, D. A. Vignali, and A. D. Luster. “Interleukin-33 activates regulatory T cells to suppress innate γδ T cell responses in the lung”. In: *Nature immunology* 21.11 (Nov. 2020), pp. 1371–1383. DOI: 10.1038/s41590-020-0785-3.

[193] H. Mühl. “Pro-Inflammatory Signaling by IL-10 and IL-22: Bad Habit Stirred Up by Interferons?” In: *Frontiers in Immunology* 4 (Feb. 4, 2013), p. 18. DOI: 10.3389/fimmu.2013.00018.

[194] V. Carlini, D. M. Noonan, E. Abdalalem, D. Goletti, C. Sansone, L. Calabrone, and A. Albini. “The multifaceted nature of IL-10: regulation, role in immunological homeostasis and its relevance to cancer, COVID-19 and post-COVID conditions”. In: *Frontiers in Immunology* 14 (June 8, 2023). Publisher: Frontiers. DOI: 10.3389/fimmu.2023.1161067.

[195] K. Littringer et al. “Common Features of Regulatory T Cell Specialization During Th1 Responses”. In: *Frontiers in Immunology* 9 (June 13, 2018). Publisher: Frontiers. DOI: 10.3389/fimmu.2018.01344.

[196] S. L. Clay, A. Bravo-Blas, D. M. Wall, M. K. L. MacLeod, and S. W. F. Milling. “Regulatory T cells control the dynamic and site-specific polarization of total CD4 T cells following Salmonella infection”. In: *Mucosal Immunology* 13.6 (Nov. 2020). Publisher: Nature Publishing Group, pp. 946–957. DOI: 10.1038/s41385-020-0299-1.

[197] J. S. Dolina, N. Van Braeckel-Budimir, G. D. Thomas, and S. Salek-Ardakani. “CD8+ T Cell Exhaustion in Cancer”. In: *Frontiers in Immunology* 12 (July 20, 2021). Publisher: Frontiers. DOI: 10.3389/fimmu.2021.715234.

[198] C. T. Weaver, C. O. Elson, L. A. Fouser, and J. K. Kolls. “The Th17 Pathway and Inflammatory Diseases of the Intestines, Lungs and Skin”. In: *Annual review of pathology* 8 (Jan. 24, 2013), pp. 477–512. DOI: 10.1146/annurev-pathol-011110-130318.

[199] V. Abbasova, A. Gül, and G. Saruhan-Direskeneli. “IL-17A and IFN-ɣ are Up-regulated in CD4 and γδ T Cells in Active Behcet’s Disease Patients”. In: *Immunology Letters* 242 (Feb. 2022), pp. 37–45. DOI: 10.1016/j.imlet.2021.11.004.

[200] A. Neunkirchner, K. G. Schmetterer, and W. F. Pickl. “Lymphocyte-Based Model Systems for Allergy Research: A Historic Overview”. In: *International archives of allergy and immunology* 163.4 (Jan. 1, 2014), pp. 259–91. DOI: 10.1159/000360163.

[201] C. Xie, J. Yang, A. Gul, Y. Li, R. Zhang, M. Yalikun, X. Lv, Y. Lin, Q. Luo, and H. Gao. “Immunologic aspects of asthma: from molecular mechanisms to disease pathophysiology and clinical translation”. In: *Frontiers in Immunology* 15 (Oct. 8, 2024). Publisher: Frontiers. DOI: 10.3389/fimmu.2024.1478624.

[202] J. Wei, C. L. Mayberry, X. Lv, F. Hu, T. Khan, N. A. Logan, J. J. Wilson, J. D. Sears, D. Chaussabel, and C.-H. Chang. “IL3-Driven T Cell–Basophil Crosstalk Enhances Antitumor Immunity”. In: *Cancer Immunology Research* 12.7 (July 2, 2024), pp. 822–839. DOI: 10.1158/2326-6066.CIR-23-0851.

[203] T. M. Corpuz, R. Vazquez-Lombardi, J. K. Luong, J. Warren, J. Stolp, D. Christ, C. King, R. Brink, J. Sprent, and K. E. Webster. “IL-2 Shapes the Survival and Plasticity of IL-17-Producing γδ T Cells”. In: *Journal of Immunology (Baltimore, Md.: 1950)* 199.7 (Oct. 1, 2017), pp. 2366–2376. DOI: 10.4049/jimmunol.1700335.

[204] K. Wakahara et al. “Human basophils interact with memory T cells to augment Th17 responses”. In: *Blood* 120.24 (Dec. 6, 2012), pp. 4761–4771. DOI: 10.1182/blood-2012-04-424226.

[205] I. S. Junttila. “Tuning the Cytokine Responses: An Update on Interleukin (IL)-4 and IL-13 Receptor Complexes”. In: *Frontiers in Immunology* 9 (June 7, 2018). Publisher: Frontiers. DOI: 10.3389/fimmu.2018.00888.

[206] G. J. McKenzie, P. G. Fallon, C. L. Emson, R. K. Grencis, and A. N. McKenzie. “Simultaneous Disruption of Interleukin (IL)-4 and IL-13 Defines Individual Roles in T Helper Cell Type 2–mediated Responses”. In: *Journal of Experimental Medicine* 189.10 (May 17, 1999), pp. 1565–1572. DOI: 10.1084/jem.189.10.1565.

[207] T. Van Meerhaeghe, A. Néel, S. Brouard, and N. Degauque. “Regulation of CD8 T cell by B-cells: A narrative review”. In: *Frontiers in Immunology* 14 (Mar. 8, 2023), p. 1125605. DOI: 10.3389/fimmu.2023.1125605.

[208] N. P. Rogozynski and B. Dixon. “The Th1/Th2 paradigm: A misrepresentation of helper T cell plasticity”. In: *Immunology Letters* 268 (Aug. 2024), p. 106870. DOI: 10.1016/j.imlet.2024.106870.

[209] F. A. Haggar and R. P. Boushey. “Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors”. In: *Clinics in colon and rectal surgery* 22.4 (Nov. 2009). tex.pmcid: PMC2796096, pp. 191–197. DOI: 10.1055/s-0029-1242458.

[210] C. M. Johnson, C. Wei, J. E. Ensor, D. J. Smolenski, C. I. Amos, B. Levin, and D. A. Berry. “Meta-analyses of colorectal cancer risk factors”. In: *Cancer causes & control : CCC* 24.6 (June 2013). tex.pmcid: PMC4161278, pp. 1207–1222. DOI: 10.1007/s10552-013-0201-5.

[211] P. M. Calvert and H. Frucht. “The genetics of colorectal cancer”. In: *Annals of internal medicine* 137.7 (Oct. 2002), pp. 603–612. DOI: 10.7326/0003-4819-137-7-200210010-00012.

[212] K. Tutlewska, J. Lubinski, and G. Kurzawski. “Germline deletions in the EPCAM gene as a cause of Lynch syndrome – literature review”. In: *Hereditary Cancer in Clinical Practice* 11.1 (Aug. 2013), p. 9. DOI: 10.1186/1897-4287-11-9.

[213] S. Carr and A. Kasi. “Familial Adenomatous Polyposis”. In: *StatPearls*. tex.copyright: Copyright © 2022, StatPearls Publishing LLC. Treasure Island (FL): StatPearls Publishing, 2022.

[214] H. E. Vogelsang. “Prophylactic Surgery and Extended Oncologic Radicality in Gastric and Colorectal Hereditary Cancer Syndromes”. In: *Visceral medicine* 35.4 (Aug. 2019). tex.pmcid: PMC6738259, pp. 231–239. DOI: 10.1159/000501919.

[215] J. Li, X. Ma, D. Chakravarti, S. Shalapour, and R. A. DePinho. “Genetic and biological hallmarks of colorectal cancer”. In: *Genes & development* 35.11 (June 2021). tex.pmcid: PMC8168558, pp. 787–820. DOI: 10.1101/gad.348226.120.

[216] J. Li, D. Chen, and M. Shen. “Tumor Microenvironment Shapes Colorectal Cancer Progression, Metastasis, and Treatment Responses”. In: *Frontiers in Medicine* 9 (2022).

[217] K. J. Cavagnero and R. L. Gallo. “Essential immune functions of fibroblasts in innate host defense”. In: *Frontiers in Immunology* 13 (2022).

[218] Z. Y. Wong, E. Nee, M. Coles, and C. D. Buckley. "Why does understanding the biology of fibroblasts in immunity really matter?" In: *PLOS Biology* 21.2 (2023), e3001954. DOI: 10.1371/journal.pbio.3001954.

[219] M. B. Buechler and S. J. Turley. "A short field guide to fibroblast function in immunity". In: *Seminars in Immunology*. The mesenchymal and myeloid regulation of immunity 35 (Feb. 2018), pp. 48–58. DOI: 10.1016/j.smim.2017.11.001.

[220] L. A. Bautista-Hernández, J. L. Gómez-Olivares, B. Buentello-Volante, and V. M. Bautista-de Lucio. "Fibroblasts: The Unknown Sentinels Eliciting Immune Responses Against Microorganisms". In: *European journal of microbiology & immunology* 7.3 (Aug. 2017). tex.pmcid: PMC5632742, pp. 151–157. DOI: 10.1556/1886.2017.00009.

[221] R. Peiffer, Y. Boumahd, C. Gullo, R. Crake, E. Letellier, A. Bellahcène, and O. Peulen. "Cancer-Associated Fibroblast Diversity Shapes Tumor Metabolism in Pancreatic Cancer". In: *Cancers (Basel)* 15.1 (Dec. 2022). tex.pmcid: PMC9817728, p. 61. DOI: 10.3390/cancers15010061.

[222] U. Vaish, T. Jain, A. C. Are, and V. Dudeja. "Cancer-Associated Fibroblasts in Pancreatic Ductal Adenocarcinoma: An Update on Heterogeneity and Therapeutic Targeting". In: *International Journal of Molecular Sciences* 22.24 (Jan. 2021), p. 13408. DOI: 10.3390/ijms222413408.

[223] D. Öhlund et al. "Distinct populations of inflammatory fibroblasts and myofibroblasts in pancreatic cancer". In: *The Journal of experimental medicine* 214.3 (Mar. 2017). tex.pmcid: PMC5339682, pp. 579–596. DOI: 10.1084/jem.20162024.

[224] I. Belhabib, S. Zaghdoudi, C. Lac, C. Bousquet, and C. Jean. "Extracellular Matrices and Cancer-Associated Fibroblasts: Targets for Cancer Diagnosis and Therapy?" In: *Cancers (Basel)* 13.14 (July 2021). tex.pmcid: PMC8303391, p. 3466. DOI: 10.3390/cancers13143466.

[225] E. Sahai et al. "A framework for advancing our understanding of cancer-associated fibroblasts". In: *Nature reviews. Cancer* 20.3 (Mar. 2020). tex.copyright: 2020 The Author(s), pp. 174–186. DOI: 10.1038/s41568-019-0238-1.

[226] P.-Y. Chen, W.-F. Wei, H.-Z. Wu, L.-S. Fan, and W. Wang. "Cancer-Associated Fibroblast Heterogeneity: A Factor That Cannot Be Ignored in Immune Microenvironment Remodeling". In: *Frontiers in Immunology* 12 (2021).

[227] E. M. Schatoff, B. I. Leach, and L. E. Dow. “Wnt Signaling and Colorectal Cancer”. In: *Current colorectal cancer reports* 13.2 (Apr. 2017). tex.pmcid: PMC5391049, pp. 101–110. DOI: 10.1007/s11888-017-0354-9.

[228] Y.-S. Jung and J.-I. Park. “Wnt signaling in cancer: therapeutic targeting of Wnt signaling beyond β-catenin and the destruction complex”. In: *Experimental & molecular medicine* 52.2 (Feb. 2020). tex.copyright: 2020 The Author(s), pp. 183–191. DOI: 10.1038/s12276-020-0380-6.

[229] L. Novellasdemunt, P. Antas, and V. S. W. Li. “Targeting Wnt signaling in colorectal cancer. A Review in the Theme: Cell Signaling: Proteins, Pathways and Mechanisms”. In: *American Journal of Physiology-Cell Physiology* 309.8 (Oct. 2015), pp. C511–C521. DOI: 10.1152/ajpcell.00117.2015.

[230] A. Seeber et al. “Molecular landscape of colorectal cancers harboring R-spondin fusions.” In: *JCO* 37.15 (May 2019), pp. 3588–3588. DOI: 10.1200/JCO.2019.37.15_suppl.3588.

[231] J. Bian, M. Dannappel, C. Wan, and R. Firestein. “Transcriptional Regulation of Wnt/β-Catenin Pathway in Colorectal Cancer”. In: *Cells* 9.9 (Sept. 2020). tex.pmcid: PMC7564852, p. 2125. DOI: 10.3390/cells9092125.

[232] H. Zhao, T. Ming, S. Tang, S. Ren, H. Yang, M. Liu, Q. Tao, and H. Xu. “Wnt signaling in colorectal cancer: pathogenic role and therapeutic target”. In: *Molecular cancer* 21.1 (July 2022). tex.pmcid: PMC9281132, p. 144. DOI: 10.1186/s12943-022-01616-7.

[233] T. Hirashima et al. “Wnt5a in cancer-associated fibroblasts promotes colorectal cancer progression”. In: *Biochemical and Biophysical Research Communications* 568 (Sept. 2021), pp. 37–42. DOI: 10.1016/j.bbrc.2021.06.062.

[234] D. Qian, C. Jones, A. Rzadzinska, S. Mark, X. Zhang, K. P. Steel, X. Dai, and P. Chen. “Wnt5a functions in planar cell polarity regulation in mice”. In: *Developmental biology* 306.1 (June 2007). tex.pmcid: PMC1978180, pp. 121–133. DOI: 10.1016/j.ydbio.2007.03.011.

[235] L. Topol, X. Jiang, H. Choi, L. Garrett-Beal, P. J. Carolan, and Y. Yang. “Wnt-5a inhibits the canonical Wnt pathway by promoting GSK-3–independent β-catenin degradation”. In: *The Journal of cell biology* 162.5 (Sept. 2003). tex.pmcid: PMC2172823, pp. 899–908. DOI: 10.1083/jcb.200303158.

[236] R. van Amerongen, A. N. Bowman, and R. Nusse. "Developmental Stage and Time Dictate the Fate of Wnt/β-Catenin-Responsive Stem Cells in the Mammary Gland". In: *Cell Stem Cell* 11.3 (Sept. 2012), pp. 387–400. DOI: 10.1016/j.stem.2012.05.023.

[237] H.-Y. H. Ho, M. W. Susman, J. B. Bikoff, Y. K. Ryu, A. M. Jonas, L. Hu, R. Kuruvilla, and M. E. Greenberg. "Wnt5a-Ror-Dishevelled signaling constitutes a core developmental pathway that controls tissue morphogenesis". In: *Proceedings of the National Academy of Sciences of the United States of America* 109.11 (Mar. 2012). tex.pmcid: PMC3306699, pp. 4044–4051. DOI: 10.1073/pnas.1200421109.

[238] K. Kumawat and R. Gosens. "WNT-5A: signaling and functions in health and disease". In: *Cellular and molecular life sciences : CMLS* 73 (2016). tex.pmcid: PMC4713724, pp. 567–587. DOI: 10.1007/s00018-015-2076-y.

[239] M. L. P. Bueno, S. T. O. Saad, and F. M. Roversi. "WNT5A in tumor development and progression: A comprehensive review". In: *Biomedicine & Pharmacotherapy* 155 (Nov. 2022), p. 113599. DOI: 10.1016/j.biopha.2022.113599.

[240] J. B. Rawson et al. "Promoter methylation of Wnt5a is associated with microsatellite instability and BRAF V600E mutation in two large populations of colorectal cancer patients". In: *British journal of cancer* 104.12 (June 2011). tex.pmcid: PMC3111198, pp. 1906–1912. DOI: 10.1038/bjc.2011.165.

[241] P. D. Da Forno, J. H. Pringle, P. Hutchinson, J. Osborn, Q. Huang, L. Potter, R. A. Hancox, A. Fletcher, and G. S. Saldanha. "WNT5A expression increases during melanoma progression and correlates with outcome". In: *Clinical cancer research : an official journal of the American Association for Cancer Research* 14.18 (Sept. 2008), pp. 5825–5832. DOI: 10.1158/1078-0432.CCR-07-5104.

[242] R.-Q. Wang, W. Zhao, H.-K. Yang, J.-M. Dong, W.-J. Lin, F.-Z. He, M. Cui, and Z.-L. Zhou. "Single-Cell RNA Sequencing Analysis of the Heterogeneity in Gene Regulatory Networks in Colorectal Cancer". In: *Frontiers in Cell and Developmental Biology* 9 (2021).

[243] H. Lin, L. Xia, J. Lian, Y. Chen, Y. Zhang, Z. Zhuang, H. Cai, J. You, and G. Guan. "Delineation of colorectal cancer ligand-receptor interactions and their roles in the tumor microenvironment and prognosis". In: *Journal of Translational Medicine* 19.1 (Dec. 2021), p. 497. DOI: 10.1186/s12967-021-03162-0.

[244] F. Ma, S. Zhang, L. Song, B. Wang, L. Wei, and F. Zhang. "Applications and analytical tools of cell communication based on ligand-receptor interactions at single cell level". In: *Cell & Bioscience* 11.1 (July 2021), p. 121. DOI: 10.1186/s13578-021-00635-z.

[245] K. Bridges and K. Miller-Jensen. "Mapping and Validation of scRNA-Seq-Derived Cell-Cell Communication Networks in the Tumor Microenvironment". In: *Frontiers in Immunology* 13 (2022).

[246] D. Dimitrov et al. "Comparison of methods and resources for cell-cell communication inference from single-cell RNA-Seq data". In: *Nature communications* 13.1 (June 2022). tex.copyright: 2022 The Author(s), p. 3224. DOI: 10.1038/s41467-022-30755-0.

[247] J. Qian et al. "A pan-cancer blueprint of the heterogeneous tumor microenvironment revealed by single-cell profiling". In: *Cell research* 30.9 (Sept. 2020). tex.copyright: 2020 Center for Excellence in Molecular Cell Science, CAS, pp. 745–762. DOI: 10.1038/s41422-020-0355-0.

[248] F. Yu, C. Yu, F. Li, Y. Zuo, Y. Wang, L. Yao, C. Wu, C. Wang, and L. Ye. "Wnt/β-catenin signaling in cancers and targeted therapies". In: *Sig Transduct Target Ther* 6.1 (Aug. 2021). tex.copyright: 2021 The Author(s), pp. 1–24. DOI: 10.1038/s41392-021-00701-5.

[249] A. H. Nile et al. "A selective peptide inhibitor of Frizzled 7 receptors disrupts intestinal stem cells". In: *Nature chemical biology* 14.6 (June 2018). tex.copyright: 2018 The Author(s), pp. 582–590. DOI: 10.1038/s41589-018-0035-2.

[250] E. Palomer et al. "Epigenetic repression of Wnt receptors in AD: a role for Sirtuin2-induced H4K16ac deacetylation of Frizzled1 and Frizzled7 promoters". In: *Molecular psychiatry* 27.7 (2022). tex.pmcid: PMC9205772, pp. 3024–3033. DOI: 10.1038/s41380-022-01492-z.

[251] K. Sompel, A. Elango, A. J. Smith, and M. A. Tennis. "Cancer chemoprevention through Frizzled receptors and EMT". In: *Discov Onc* 12.1 (Sept. 2021), p. 32. DOI: 10.1007/s12672-021-00429-2.

[252] K. Ueno, M. Hiura, Y. Suehiro, S. Hazama, H. Hirata, M. Oka, K. Imai, R. Dahiya, and Y. Hinoda. "Frizzled-7 as a potential therapeutic target in colorectal cancer". In: *Neoplasia (New York, N.Y.)* 10.7 (July 2008). tex.pmcid: PMC2435005, pp. 697–705. DOI: 10.1593/neo.08320.

[253] J. Lyu, V. Yamamoto, and W. Lu. “Cleavage of the Wnt receptor Ryk regulates neuronal differentiation during cortical neurogenesis”. In: *Developmental cell* 15.5 (Nov. 2008), pp. 773–780. DOI: 10.1016/j.devcel.2008.10.004.

[254] Y. Fu, Y. Chen, J. Huang, Z. Cai, and Y. Wang. “RYK, a receptor of noncanonical Wnt ligand Wnt5a, is positively correlated with gastric cancer tumorigenesis and potential of liver metastasis”. In: *American journal of physiology. Gastrointestinal and liver physiology* 318.2 (Feb. 2020), G352–G360. DOI: 10.1152/ajpgi.00228.2019.

[255] Q. Ren, J. Chen, and Y. Liu. “LRP5 and LRP6 in Wnt Signaling: Similarity and Divergence”. In: *Frontiers in cell and developmental biology* 9 (2021). tex.pmcid: PMC8134664, p. 670960. DOI: 10.3389/fcell.2021.670960.

[256] G. Lettini, V. Condelli, M. Pietrafesa, F. Crispo, P. Zoppoli, F. Maddalena, I. Laurenzana, A. Sgambato, F. Esposito, and M. Landriscina. “TRAP1 Regulates Wnt/β-Catenin Pathway through LRP5/6 Receptors Expression Modulation”. In: *International journal of molecular sciences* 21.20 (Oct. 2020). tex.pmcid: PMC7589514, p. 7526. DOI: 10.3390/ijms21207526.

[257] X. Nie et al. “LRP5 promotes cancer stem cell traits and chemoresistance in colorectal cancer”. In: *Journal of cellular and molecular medicine* 26.4 (Feb. 2022). tex.pmcid: PMC8831954, pp. 1095–1112. DOI: 10.1111/jcmm.17164.

[258] A. Shishido et al. “The Relationship between LRP6 and Wnt/β-Catenin Pathway in Colorectal and Esophageal Cancer”. In: *Life (Basel, Switzerland)* 13.3 (Feb. 2023). tex.pmcid: PMC10057833, p. 615. DOI: 10.3390/life13030615.

[259] N. Fenderico et al. “Anti-LRP5/6 VHHs promote differentiation of Wnt-hypersensitive intestinal stem cells”. In: *Nature communications* 10.1 (Jan. 2019). tex.copyright: 2019 The Author(s), p. 365. DOI: 10.1038/s41467-018-08172-z.

[260] L.-S. Chang, M. Kim, A. Glinka, C. Reinhard, and C. Niehrs. “The tumor suppressor PTPRK promotes ZNRF3 internalization and is required for Wnt inhibition in the Spemann organizer”. In: *eLife* 9 (Jan. 2020). Ed. by R. Nusse and M. B. Eisen, e51248. DOI: 10.7554/eLife.51248.

[261] M. Kim, C. Reinhard, and C. Niehrs. “A MET-PTPRK kinase-phosphatase rheostat controls ZNRF3 and Wnt signaling”. In: *eLife* 10 (Sept. 2021). Ed. by R. Nusse, M. B. Eisen, and R. Nusse, e70885. DOI: 10.7554/eLife.70885.

[262] J. H. Jeong, J. W. Yun, H. Y. Kim, C. Y. Heo, and S. Lee. “Investigation of cell signalings and therapeutic targets in PTPRK-RSPO3 fusion-positive colorectal cancer”. In: *PLoS One* 17.9 (2022). tex.pmcid: PMC9491571, e0274555. DOI: 10.1371/journal.pone.0274555.

[263] M. M. Fischer et al. “RSPO3 antagonism inhibits growth and tumorigenicity in colorectal tumors harboring common Wnt pathway mutations”. In: *Scientific reports* 7.1 (Nov. 2017). tex.copyright: 2017 The Author(s), p. 15270. DOI: 10.1038/s41598-017-15704-y.

[264] T. Jin et al. “ANTXR1 Regulates Erythroid Cell Proliferation and Differentiation through wnt/β-Catenin Signaling Pathway *In Vitro* and in Hematopoietic Stem Cell”. In: *Disease Markers* 2022 (Aug. 2022), e1226697. DOI: 10.1155/2022/1226697.

[265] T. Jin et al. “Transmembrane Protein ANTXR1 Regulates γ-Globin Expression by Targeting the Wnt/β-Catenin Signaling Pathway”. In: *Journal of Immunological Research* 2022 (2022). tex.pmcid: PMC9356848, p. 8440422. DOI: 10.1155/2022/8440422.

[266] Ł. Pietrzyk, A. Korolczuk, M. Matysek, M. B. Arciszewski, and K. Torres. “Clinical Value of Detecting Tumor Endothelial Marker 8 (ANTXR1) as a Biomarker in the Diagnosis and Prognosis of Colorectal Cancer”. In: *Cancer management and research* 13 (2021). tex.pmcid: PMC8043785, pp. 3113–3122. DOI: 10.2147/CMAR.S298165.

[267] S. Vijayakumar, G. Liu, H.-C. Wen, Y. Abu, R. Chong, H. Nastri, G. G. Bornstein, Z.-Q. Pan, and S. A. Aaronson. “Extracellular LDLR repeats modulate Wnt signaling activity by promoting LRP6 receptor endocytosis mediated by the Itch E3 ubiquitin ligase”. In: *Genes & cancer* 8.7 (July 2017). tex.pmcid: PMC5620007, pp. 613–627. DOI: 10.18632/genesandcancer.146.

[268] K. Tamai, M. Semenov, Y. Kato, R. Spokony, C. Liu, Y. Katsuyama, F. Hess, J. P. Saint-Jeannet, and X. He. “LDL-receptor-related proteins in Wnt signal transduction”. In: *Nature* 407.6803 (Sept. 2000), pp. 530–535. DOI: 10.1038/35035117.

[269] A. Zilberberg, A. Yaniv, and A. Gazit. “The low density lipoprotein receptor-1, LRP1, interacts with the human frizzled-1 (HFz1) and down-regulates the canonical Wnt signaling pathway”. In: *The Journal of biological chemistry* 279.17 (Apr. 2004), pp. 17535–17542. DOI: 10.1074/jbc.M311292200.

[270] C. Hafemeister and R. Satija. “Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression”. In: *Genome Biology* 20.1 (Dec. 23, 2019), p. 296. DOI: 10.1186/s13059-019-1874-1.

[271] T. Tian, J. Wan, Q. Song, and Z. Wei. “Clustering single-cell RNA-seq data with a model-based deep learning approach”. In: *Nature machine intelligence* 1.4 (Apr. 2019). tex.copyright: 2019 The Author(s), under exclusive licence to Springer Nature Limited, pp. 191–198. DOI: 10.1038/s42256-019-0037-0.

[272] L. Yu, Y. Cao, J. Y. H. Yang, and P. Yang. “Benchmarking clustering algorithms on estimating the number of cell types from single-cell RNA-sequencing data”. In: *Genome Biology* 23.1 (Feb. 2022), p. 49. DOI: 10.1186/s13059-022-02622-0.

[273] G. Pasquini, J. E. Rojo Arias, P. Schäfer, and V. Busskamp. “Automated methods for cell type annotation on scRNA-seq data”. In: *Computational and Structural Biotechnology Journal* 19 (Jan. 2021), pp. 961–969. DOI: 10.1016/j.csbj.2021.01.015.

[274] X. Ji, D. Tsao, K. Bai, M. Tsao, L. Xing, and X. Zhang. “scAnnotate: an automated cell-type annotation tool for single-cell RNA-sequencing data”. In: *Bioinformatics Advances* 3.1 (Jan. 2023), vbad030. DOI: 10.1093/bioadv/vbad030.

[275] A. Ianevski, A. K. Giri, and T. Aittokallio. “Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data”. In: *Nature communications* 13.1 (Mar. 2022). tex.copyright: 2022 The Author(s), p. 1246. DOI: 10.1038/s41467-022-28803-w.

[276] R. Browaeys, W. Saelens, and Y. Saeys. “NicheNet: modeling intercellular communication by linking ligands to target genes”. In: *Nature methods* 17.2 (Feb. 2020). tex.copyright: 2019 The Author(s), under exclusive licence to Springer Nature America, Inc., pp. 159–162. DOI: 10.1038/s41592-019-0667-5.

[277] J. V. de Assis, L. A. Coutinho, I. T. Oyeyemi, O. T. Oyeyemi, and R. F. e. Q. Grenfell. “Diagnostic and therapeutic biomarkers in colorectal cancer: a review”. In: *American journal of cancer research* 12.2 (Feb. 2022). tex.pmcid: PMC8900002, pp. 661–680.

[278] B. A. Alves Martins, G. F. de Bulhões, I. N. Cavalcanti, M. M. Martins, P. G. de Oliveira, and A. M. A. Martins. “Biomarkers in Colorectal Cancer: The Role of Translational Proteomics Research”. In: *Frontiers in Oncology* 9 (2019).

[279] J. L. Slavin. “Carbohydrates, Dietary Fiber, and Resistant Starch in White Vegetables: Links to Health Outcomes”. In: *Advances in Nutrition* 4.3 (May 1, 2013), 351S–355S. DOI: 10.3945/an.112.003491.

[280] A. Bojarczuk, S. Skąpska, A. Mousavi Khaneghah, and K. Marszałek. “Health benefits of resistant starch: A review of the literature”. In: *Journal of Functional Foods* 93 (June 1, 2022), p. 105094. DOI: 10.1016/j.jff.2022.105094.

[281] W. Sheng, G. Ji, and L. Zhang. “Immunomodulatory effects of inulin and its intestinal metabolites”. In: *Frontiers in Immunology* 14 (2023), p. 1224092. DOI: 10.3389/fimmu.2023.1224092.

[282] J. L. S. Donadio, J. P. Fabi, M. B. Sztein, and R. Salerno-Gonçalves. “Dietary fiber pectin: challenges and potential anti-inflammatory benefits for preterms and newborns”. In: *Frontiers in Nutrition* 10 (Jan. 12, 2024). Publisher: Frontiers. DOI: 10.3389/fnut.2023.1286138.

[283] F. Fischer et al. “Dietary cellulose induces anti-inflammatory immunity and transcriptional programs via maturation of the intestinal microbiota”. In: *Gut Microbes* 12.1 (Nov. 9, 2020), p. 1829962. DOI: 10.1080/19490976.2020.1829962.

[284] V. Tosti, B. Bertozzi, and L. Fontana. “Health Benefits of the Mediterranean Diet: Metabolic and Molecular Mechanisms”. In: *The Journals of Gerontology: Series A* 73.3 (Mar. 2, 2018), pp. 318–326. DOI: 10.1093/gerona/glx227.

[285] C. M. Phillips. “Nutrigenetics and Metabolic Disease: Current Status and Implications for Personalised Nutrition”. In: *Nutrients* 5.1 (Jan. 10, 2013), p. 32. DOI: 10.3390/nu5010032.

[286] M. Aldana and R. Robeva. “New Challenges in Systems Biology: Understanding the Holobiont”. In: *Frontiers in Physiology* 12 (Mar. 26, 2021), p. 662878. DOI: 10.3389/fphys.2021.662878.

[287] S. R. Bordenstein and K. R. Theis. “Host Biology in Light of the Microbiome: Ten Principles of Holobionts and Hologenomes”. In: *PLoS Biology* 13.8 (Aug. 18, 2015), e1002226. DOI: 10.1371/journal.pbio.1002226.

[288] J. L. Matthews. “Editorial: Holobiont interactions”. In: *Frontiers in Ecology and Evolution* 12 (Mar. 7, 2024). Publisher: Frontiers. DOI: 10.3389/fevo.2024.1382169.

[289] W. Wang et al. “Decoupling of strain- and intrastrain-level interactions of microbiomes in a sponge holobiont”. In: *Nature Communications* 15.1 (Sept. 18, 2024). Publisher: Nature Publishing Group, p. 8205. DOI: 10.1038/s41467-024-52464-6.

[290] R. Greer, X. Dong, A. Morgun, and N. Shulzhenko. “Investigating a holobiont: Microbiota perturbations and transkingdom networks”. In: *Gut Microbes* 7.2 (Mar. 16, 2016), pp. 126–135. DOI: 10.1080/19490976.2015.1128625.

[291] P. T. Santana, S. L. B. Rosas, B. E. Ribeiro, Y. Marinho, and H. S. P. de Souza. “Dysbiosis in Inflammatory Bowel Disease: Pathogenic Role and Potential Therapeutic Targets”. In: *International Journal of Molecular Sciences* 23.7 (Mar. 23, 2022), p. 3464. DOI: 10.3390/ijms23073464.

[292] N. M. Sahasrabudhe et al. “Dietary Fiber Pectin Directly Blocks Toll-Like Receptor 2–1 and Prevents Doxorubicin-Induced Ileitis”. In: *Frontiers in Immunology* 9 (Mar. 1, 2018). Publisher: Frontiers. DOI: 10.3389/fimmu.2018.00383.

[293] S. Deleu, K. Machiels, J. Raes, K. Verbeke, and S. Vermeire. “Short chain fatty acids and its producing organisms: An overlooked therapy for IBD?” In: *eBioMedicine* 66 (Apr. 1, 2021), p. 103293. DOI: 10.1016/j.ebiom.2021.103293.

[294] D. Zhang, Y.-P. Jian, Y.-N. Zhang, Y. Li, L.-T. Gu, H.-H. Sun, M.-D. Liu, H.-L. Zhou, Y.-S. Wang, and Z.-X. Xu. “Short-chain fatty acids in diseases”. In: *Cell Communication and Signaling* 21.1 (Aug. 18, 2023), p. 212. DOI: 10.1186/s12964-023-01219-9.

[295] Y. P. Silva, A. Bernardi, and R. L. Frozza. “The Role of Short-Chain Fatty Acids From Gut Microbiota in Gut-Brain Communication”. In: *Frontiers in Endocrinology* 11 (Jan. 31, 2020). Publisher: Frontiers. DOI: 10.3389/fendo.2020.00025.

[296] J.-M. Zhang and J. An. “Cytokines, Inflammation, and Pain”. In: *International Anesthesiology Clinics* 45.2 (2007), p. 27. DOI: 10.1097/AIA.0b013e318034194e.

[297] M. J. Cameron and D. J. Kelvin. “Cytokines, Chemokines and Their Receptors”. In: *Madame Curie Bioscience Database [Internet]*. Landes Bioscience, 2013.

[298] O. G. Swann, M. Breslin, M. Kilpatrick, T. A. O’Sullivan, T. A. Mori, L. J. Beilin, and W. H. Oddy. “Dietary fibre intake and its association with inflammatory markers in adolescents”. In: *British Journal of Nutrition* 125.3 (Feb. 2021), pp. 329–336. DOI: 10.1017/S0007114520001609.

[299] Y. Ma et al. “Association between dietary fiber and markers of systemic inflammation in the Women’s Health Initiative Observational Study”. In: *Nutrition* 24.10 (Oct. 1, 2008), pp. 941–949. DOI: 10.1016/j.nut.2008.04.005.

[300] H. H. M. Hermsdorff, K. B. F. Barbosa, A. C. P. Volp, B. Puchau, J. Bressan, M. Á. Zulet, and J. A. Martínez. “Vitamin C and fibre consumption from fruits and vegetables improves oxidative stress markers in healthy young adults”. In: *British Journal of Nutrition* 107.8 (Apr. 2012), pp. 1119–1127. DOI: 10.1017/S0007114511004235.

[301] H. K. Krishnamurthy, I. Rajavelu, M. Pereira, V. Jayaraman, K. Krishna, T. Wang, K. Bei, and J. J. Rajasekaran. “Inside the genome: understanding genetic influences on oxidative stress”. In: *Frontiers in Genetics* 15 (June 25, 2024). Publisher: Frontiers. DOI: 10.3389/fgene.2024.1397352.

[302] L. A. Da Costa, A. Badawi, and A. El-Sohemy. “Nutrigenetics and Modulation of Oxidative Stress”. In: *Annals of Nutrition and Metabolism* 60 (Suppl. 3 May 15, 2012), pp. 27–36. DOI: 10.1159/000337311.

[303] U. M. KUJALA et al. “Polygenic Risk Scores and Physical Activity”. In: *Medicine and Science in Sports and Exercise* 52.7 (July 2020), pp. 1518–1524. DOI: 10.1249/MSS.0000000000002290.

[304] T. Konuma and Y. Okada. “Statistical genetics and polygenic risk score for precision medicine”. In: *Inflammation and Regeneration* 41.1 (June 17, 2021), p. 18. DOI: 10.1186/s41232-021-00172-9.

[305] L. Sun et al. “Polygenic risk scores in cardiovascular risk prediction: A cohort study and modelling analyses”. In: *PLoS Medicine* 18.1 (Jan. 14, 2021), e1003498. DOI: 10.1371/journal.pmed.1003498.

[306] N. M. McKeown, G. C. Fahey, J. Slavin, and J.-W. van der Kamp. “Fibre intake for optimal health: how can healthcare professionals support people to reach dietary recommendations?” In: *The BMJ* 378 (July 20, 2022), e054370. DOI: 10.1136/bmj-2020-054370.

[307] M. Franzago, D. Santurbano, E. Vitacolonna, and L. Stuppia. “Genes and Diet in the Prevention of Chronic Diseases in Future Generations”. In: *International Journal of Molecular Sciences* 21.7 (Apr. 10, 2020), p. 2633. DOI: 10.3390/ijms21072633.

[308] H. T. Nguyen, T. G. Pavey, P. F. Collins, N. V. Nguyen, T. D. Pham, and D. Gallegos. "Effectiveness of Tailored Dietary Counseling in Treating Malnourished Outpatients with Chronic Obstructive Pulmonary Disease: A Randomized Controlled Trial". In: *Journal of the Academy of Nutrition and Dietetics* 120.5 (May 2020), 778–791.e1. DOI: 10.1016/j.jand.2019.09.013.

[309] A. Ayoub, J. McHugh, J. Hayward, I. Rafi, and N. Qureshi. "Polygenic risk scores: improving the prediction of future disease or added complexity?" In: *The British Journal of General Practice* 72.721 (July 29, 2022), pp. 396–398. DOI: 10.3399/bjgp22X720437.

[310] A. C. Lewis, R. C. Green, and J. L. Vassy. "Polygenic risk scores in the clinic: Translating risk into action". In: *Human Genetics and Genomics Advances* 2.4 (July 28, 2021), p. 100047. DOI: 10.1016/j.xhgg.2021.100047.

[311] Y. Wang, W. Qi, X. Guo, G. Song, S. Pang, W. Fang, and Z. Peng. "Effects of Oats, Tartary Buckwheat, and Foxtail Millet Supplementation on Lipid Metabolism, Oxido-Inflammatory Responses, Gut Microbiota, and Colonic SCFA Composition in High-Fat Diet Fed Rats". In: *Nutrients* 14.13 (Jan. 2022). Number: 13 Publisher: Multidisciplinary Digital Publishing Institute, p. 2760. DOI: 10.3390/nu14132760.

[312] X. Qi, Y. Li, C. Fang, Y. Jia, M. Chen, X. Chen, and J. Jia. "The associations between dietary fibers intake and systemic immune and inflammatory biomarkers, a multi-cycle study of NHANES 2015–2020". In: *Frontiers in Nutrition* 10 (Aug. 31, 2023), p. 1216445. DOI: 10.3389/fnut.2023.1242115.

[313] *Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data*. URL: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (visited on 01/31/2025).

[314] M. Martin. "Cutadapt removes adapter sequences from high-throughput sequencing reads". In: *EMBnet.journal* 17.1 (May 2, 2011). Number: 1, pp. 10–12. DOI: 10.14806/ej.17.1.200.

[315] B. Langmead and S. L. Salzberg. "Fast gapped-read alignment with Bowtie 2". In: *Nature Methods* 9.4 (Mar. 4, 2012), pp. 357–359. DOI: 10.1038/nmeth.1923.

[316] E. Garrison and G. Marth. *Haplotype-based variant detection from short-read sequencing*. July 20, 2012. DOI: 10.48550/arXiv.1207.3907.

[317] P. Cingolani, A. Platts, L. L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, and D. M. Ruden. "A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3". In: *Fly* 6.2 (2012), pp. 80–92. DOI: 10.4161/fly.19695.

[318] A. A. Alzahrani and F. S. Alharithi. "Machine learning approaches for advanced detection of rare genetic disorders in whole-genome sequencing". In: *Alexandria Engineering Journal* 106 (Nov. 1, 2024), pp. 582–593. DOI: 10.1016/j.aej.2024.08.056.

[319] S. W. Choi, T. S.-H. Mak, and P. F. O'Reilly. "Tutorial: a guide to performing polygenic risk score analyses". In: *Nature Protocols* 15.9 (Sept. 2020). Publisher: Nature Publishing Group, pp. 2759–2772. DOI: 10.1038/s41596-020-0353-1.

[320] P. Waldmann, G. Mészáros, B. Gredler, C. Fuerst, and J. Sölkner. "Evaluation of the lasso and the elastic net in genome-wide association studies". In: *Frontiers in Genetics* 4 (Dec. 4, 2013), p. 270. DOI: 10.3389/fgene.2013.00270.

[321] J. Stephan, O. Stegle, and A. Beyer. "A random forest approach to capture genetic effects in the presence of population structure". In: *Nature Communications* 6.1 (June 25, 2015). Publisher: Nature Publishing Group, p. 7432. DOI: 10.1038/ncomms8432.

[322] A. Singh, C. P. Shannon, B. Gautier, F. Rohart, M. Vacher, S. J. Tebbutt, and K.-A. Lê Cao. "DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays". In: *Bioinformatics (Oxford, England)* 35.17 (Sept. 1, 2019), pp. 3055–3062. DOI: 10.1093/bioinformatics/bty1054.

[323] B. Wang, A. M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains, and A. Goldenberg. "Similarity network fusion for aggregating data types on a genomic scale". In: *Nature Methods* 11.3 (Mar. 2014). Publisher: Nature Publishing Group, pp. 333–337. DOI: 10.1038/nmeth.2810.

[324] J. M. Eales, J. W. Pinney, R. D. Stevens, and D. L. Robertson. "Methodology capture: discriminating between the "best" and the rest of community practice". In: *BMC Bioinformatics* 9.1 (Sept. 1, 2008), p. 359. DOI: 10.1186/1471-2105-9-359.

[325] Y. Hao et al. "Dictionary learning for integrative, multimodal and scalable single-cell analysis". In: *Nature Biotechnology* 42.2 (Feb. 2024). Publisher: Nature Publishing Group, pp. 293–304. DOI: 10.1038/s41587-023-01767-y.

[326] A. M. Kaizer, H. M. Belli, Z. Ma, A. G. Nicklawsky, S. C. Roberts, J. Wild, A. F. Wogu, M. Xiao, and R. T. Sabo. "Recent innovations in adaptive trial designs: A review of design opportunities in translational research". In: *Journal of Clinical and Translational Science* 7.1 (Apr. 28, 2023), e125. DOI: 10.1017/cts.2023.537.

[327] "Good Clinical Practice Research Guidelines Reviewed, Emphasis Given to Responsibilities of Investigators: Second Article in a Series". In: *Journal of Oncology Practice* 4.5 (Sept. 2008), p. 233. DOI: 10.1200/JOP.0854601.

[328] A. Vijayananthan and O. Nawawi. "The importance of Good Clinical Practice guidelines and its role in clinical trials". In: *Biomedical Imaging and Intervention Journal* 4.1 (Jan. 1, 2008), e5. DOI: 10.2349/biij.4.1.e5.

[329] M. D. Wilkinson et al. "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific Data* 3.1 (Mar. 15, 2016). Publisher: Nature Publishing Group, p. 160018. DOI: 10.1038/sdata.2016.18.

[330] Y. Ohno. "ICH Guidelines—Implementation of the 3Rs (Refinement, Reduction, and Replacement): Incorporating Best Scientific Practices into the Regulatory Process". In: *ILAR Journal* 43 (Suppl_1 Jan. 1, 2002), S95–S98. DOI: 10.1093/ilar.43.Suppl_1.S95.

[331] S. Koppad, A. B, G. V. Gkoutos, and A. Acharjee. "Cloud Computing Enabled Big Multi-Omics Data Analytics". In: *Bioinformatics and Biology Insights* 15 (July 28, 2021), p. 11779322211035921. DOI: 10.1177/11779322211035921.

[332] M. Martyn, E. Forbes, L. Lee, A. Kanga-Parabia, R. Weerasuriya, E. Lynch, P. Gleeson, and C. Gaff. "Secondary use of genomic data: patients' decisions at point of testing and perspectives to inform international data sharing". In: *European Journal of Human Genetics* 32.6 (June 2024). Publisher: Nature Publishing Group, pp. 717–724. DOI: 10.1038/s41431-023-01531-5.

[333] M. Calnan, S. Kirchin, D. L. Roberts, M. N. Wass, and M. Michaelis. "Understanding and tackling the reproducibility crisis – Why we need to study scientists' trust in data". In: *Pharmacological Research* 199 (Jan. 1, 2024), p. 107043. DOI: 10.1016/j.phrs.2023.107043.

[334] M. Baker. "1,500 scientists lift the lid on reproducibility". In: *Nature* 533.7604 (May 1, 2016). Publisher: Nature Publishing Group, pp. 452–454. DOI: 10.1038/533452a.

[335] K. Neves and O. B. Amaral. “Addressing selective reporting of experiments through predefined exclusion criteria”. In: *eLife* 9 (May 22, 2020), e56626. DOI: 10.7554/eLife.56626.

[336] T. Cernava et al. “Metadata harmonization–Standards are the key for a better usage of omics data for integrative microbiome analysis”. In: *Environmental Microbiome* 17 (June 24, 2022), p. 33. DOI: 10.1186/s40793-022-00425-1.

[337] E. Oldoni et al. “Tackling the translational challenges of multi-omics research in the realm of European personalised medicine: A workshop report”. In: *Frontiers in Molecular Biosciences* 9 (Oct. 13, 2022), p. 974799. DOI: 10.3389/fmolb.2022.974799.

[338] K. H. Jones, H. Daniels, E. Squires, and D. V. Ford. “Public Views on Models for Accessing Genomic and Health Data for Research: Mixed Methods Study”. In: *Journal of Medical Internet Research* 21.8 (Aug. 21, 2019). Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada, e14384. DOI: 10.2196/14384.

[339] B. Institute of Medicine (US) Committee on Building Bridges in the Brain, T. C. Pellmar, and L. Eisenberg. “Interdisciplinary Training Programs”. In: *Bridging Disciplines in the Brain, Behavioral, and Clinical Sciences*. National Academies Press (US), 2000.

[340] D. Lachmann, T. Martius, J. Eberle, M. Landmann, L. v. Kotzebue, B. Neuhaus, and S. Herzig. “Regulations and practices of structured doctoral education in the life sciences in Germany”. In: *PLoS ONE* 15.7 (July 30, 2020), e0233415. DOI: 10.1371/journal.pone.0233415.

[341] C. P. Austin. “Opportunities and challenges in translational science”. In: *Clinical and Translational Science* 14.5 (2021), pp. 1629–1647. DOI: 10.1111/cts.13055.

[342] E. F. S. Authority (EFSA). “Dietary Reference Values for nutrients Summary report”. In: *EFSA Supporting Publications* 14.12 (2017). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.2903/sp.efsa.2017.e15121, e15121E. DOI: 10.2903/sp.efsa.2017.e15121.