Systems-level modelling of the cancer and immune metabolome to improve immunotherapeutic outcomes

Sara Manso de Sousa Cardoso

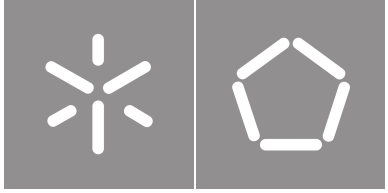
UMinho | 2022

Universidade do Minho
Escola de Engenharia

Sara Manso de Sousa Cardoso

Systems-level modelling of the cancer and immune metabolome to improve immunotherapeutic outcomes

October 2022



Universidade do Minho
Escola de Engenharia

Sara Manso de Sousa Cardoso

Systems-level modelling of the cancer and immune metabolome to improve immunotherapeutic outcomes

Doctorate Thesis
Doctorate in Biomedical Engeneering

Work developed under the supervision of:
Miguel Rocha
Noel de Miranda
Dina Ruano

October, 2022

This document was created with the (pdf/Xe/Lua)LaTeX processor and the NOVAthesis template (v6.7.4) [1].

Acknowledgements

I would like to express my gratitude to everyone who, directly or indirectly, contributed to the completion of this work.

First of all, I would like to thank Fundação para a Ciência e Tecnologia for the PhD scholarship I was awarded (SFRH/BD/138951/2018), without which this work would not be possible. I would also like to thank my main host institution, the Centre of Biological Engineering at the University of Minho, and the Leiden University Medical Center (LUMC), my secondary institution.

I would also like to thank my supervisors, for offering me this opportunity and all the support and guidance. Miguel Rocha, my main supervisor, for not just the past four years of working together, but also those that came before. Noel de Miranda and Dina Ruano, my co-supervisors, for their feedback on this work and for welcoming me in the group at Leiden.

I would also like to thank everyone that has passed through the BISBII research group at the University of Minho while I was here. To everyone at the LUMC, but more specifically those from the pathology group, thank you all.

I would like to thank all my friends for supporting me all these years.

Last, and by no means least, my family. No words would be enough to thank them for all the years that lead to now.

Once again, thank you all.

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the Universidade do Minho.

_____, _____

(Place) (Date)

_____

(Sara Manso de Sousa Cardoso)

Resumo

Modelação do metaboloma do cancro e do sistema imunitário de forma a melhorar resultados imunoterapêuticos

A medicina de precisão busca fornecer terapias para diversas doenças que sejam adequadas para (grupos de) pacientes específicos. O cancro é uma doença causada por células anormais que se multiplicam descontroladamente e, dada a sua heterogeneidade e base genética, é um dos desafios mais relevantes para a medicina de precisão. As imunoterapias podem ser adaptadas para indivíduos específicos, fornecendo formas interessantes de combater o cancro, induzindo ou aprimorando as respostas naturais do sistema imunológico dos pacientes, o que pode traduzir-se em terapias com menos efeitos colaterais.

Neste trabalho, o objectivo foi o de desenvolver abordagens computacionais baseadas em mineração de dados ómicos e modelação metabólica que ajudem os esforços personalizados de descoberta de medicamentos com o mínimo de efeitos secundários. Para isso, foram desenvolvidos modelos metabólicos de células T de tumores e tecidos saudáveis com base em dados ómicos de *single-cell*. O *single-cell RNAseq* (*scRNAseq*) é uma ótima ferramenta para a reconstrução de modelos metabólicos específicos para cada paciente e tipo de célula. No entanto, esta abordagem não foi ainda aproveitada no campo da imunoterapia tumoral.

Numa fase inicial, construímos um atlas de dados de scRNAseq para cancro colorretal, que foi usado para reconstruir 196 modelos de vários sub-tipos de células T do micro-ambiente desse tumor. Além disso, realizou-se uma análise do desempenho de vários métodos de deconvolução do tumor, para permitir que os dados de *bulk RNAseq* amplamente disponíveis sejam usados na modelação metabólica de tipos de células presentes no micro-ambiente do cancro colorretal.

Palavras-chave: Modelos metabólicos, *Single-cell RNAseq*, Cancro colorectal, Deconvolução de tumores

Abstract

Systems-level modelling of the cancer and immune metabolome to improve immunotherapeutic outcomes

Precision medicine seeks to provide therapies for different diseases that are adequate for specific (groups of) patients. Cancer is a a disease caused by abnormal cells that multiply uncontrollably and given its heterogeneity and genetic basis, is one of the most relevant challenges for precision medicine. Immunotherapies can be tailored for specific subjects, providing attractive ways to fight cancer by inducing or enhancing patients' natural immune system responses, which can lead to therapies with less side effects.

In this work, we aimed to develop computational approaches based on omics data mining and metabolic modelling that could help, in the future, personalised drug discovery efforts with minimum off-target effects. For this, metabolic models of T-cells from tumour and healthy tissues based on single-cell omics data were developed. Single-cell RNAseq is a great tool for the reconstruction of metabolic models with patient- and cell-type- scpecificity. However, this approach has not been exploited in the field of tumour immunotherapy.

We constructed an atlas of scRNAseq data for colorectal cancer, which was used to reconstruct 196 models of various T-cell subtypes from the micro-environment of this tumour. Furthermore, the benchmarking of several tumour deconvolution methods was performed to allow the extensively available bulk RNAseq data to be optimally used in cell-type specific modeling of the colorectal cancer.

Keywords: Metabolic modeling, Single-cell RNAseq, Colorectal cancer, Tumour deconvolution

Contents

List of Figures

List of Tables

Acronyms

APC	Antigen-presenting cell
ATP	Adenosine triphosphate
CAF	Cancer-associated fibroblasts
CBM	Constraint-based modeling
CMS	Consensus molecular subtype
CNV	Copy number variation
CPM	Counts per million
CRC	Colorectal cancer
EFM	Elementary flux mode
EMT	Epithelial-mesenchymal transition
ETC	Electron transport chain
FA	Fatty acid
FADH2	Flavin adenine dinucleotide
FAO	Fatty acid oxidation
FAS	Fatty acid synthesis
FBA	Flux balance analysis
FVA	Flux variability analysis
GAS	Gene activity score
GPR	Gene-protein-reaction
GSMM	Genome-scale metabolic model
IDO	Indoleamine-2,3-dioxygenase

ILC	Innate lymphoid cell
MCC	Mathews correlation coefficient
MDSC	Myeloid-derived suppressor cell
MHC	Major histocompatibility complex
MSC	Mesenchymal stem cell
NADH	Nicotinamide adenine dinucleotide
NO	Nitric oxide
OXPHOS	Oxidative phosphorilation
PBMC	Peripheral blood mononuclear cell
pFBA	Parsimonious flux balance analysis
PPP	Pentose phosphate pathway
RAS	Reaction activity score
RMSE	Root mean square error
RNAseq	RNA sequencing
ROS	Reactive oxygen species
SBML	Systems biology markup language
scRNAseq	single-cell RNA sequencing
SRC	Spare respiratory capacity
TAM	Tumour-associated macrophage
TCA	Tricarboxylic acid cycle
TCR	T-cell receptor
TIL	Tumour-infiltrating leukocyte
TME	Tumour micro-environment
UMAP	Uniform manifold approximation and projection

1 Introduction

1.1 Context and Motivation

Cancer is a disease caused by abnormal cells that multiply uncontrollably, which may generate a solid mass called tumour, and possibly invade the organism. Although tumours are usually portrayed as homogeneous cell populations, they actually contain a diverse amount of cells beyond cancer cells, including stromal and immune cells. In fact, the cells in these tumour micro-environments may be shaped by tumour cells or even by each other to aid tumour proliferation and metastasis, as well as affect responsiveness to therapeutics.

Although many characteristics span all or most tumours, the heterogeneity in how, when and what leads these characteristics to manifest are specific to each patient or group of patients. As such, precision medicine seeks to provide therapies for different diseases, including cancer, that are adequate for specific (groups of) patients. Besides finding ways of killing only the malignant cells with minimum impact on healthy cells, natural immune system responses can be induced or enhanced to fight cancer (immunotherapy), which can lead to less side effects.

Besides the large number of genomic changes that promote uncontrollable proliferation, cancer cells must undergo metabolic changes, when compared to normal cells, to support the acquisition and maintenance of malignant properties. For instance, changes in energy metabolism such as the shift to aerobic glycolysis in cancer cells even under normal oxygen conditions, as observed firstly by Warburg *et al* [2] in 1958, allows these cells to rapidly proliferate and survive under stressful conditions. Interestingly, many of the changes observed in the tumour micro-environment are also caused at the metabolic level, providing an interesting target for cancer and immune therapies.

The discovery and characterisation of the reprogrammed metabolism of cancer cells and of those in the tumour micro-environment may hence help to study tumour tissue non-invasively, predict tumour behaviour, and prevent tumour progression.

The recent major advances in high-throughput omics data, such as genomics, transcriptomics and metabolomics, have allowed modelling the metabolism of several species at a large scale through the reconstruction of Genome-Scale Metabolic Models (GSMMs). Human metabolic models, when applied

together with data specific to (groups of) patients, can be important contributors in gaining knowledge towards a better comprehension of cancer mechanisms. These models can even be essential, in the future, to find therapies specific for each patient, or group of patients.

Thus, this work focused on developing system-level computational approaches based on omics data mining and metabolic modelling, so that efficient immunotherapies, complemented with personalised drug discovery efforts, can be developed in the future with minimum off-target effects. For this, models of T-cells from tumour and healthy tissues based on patient- and cell-type- specific omics data were developed. To accomplish this, an atlas of single-cell RNAseq data for colorectal cancer was constructed. Furthermore, benchmarking of tumour deconvolution methods was performed to allow the extensively available bulk RNAseq data to be optimally used in cell-type specific modeling of the colorectal cancer.

This will enable the study of the effects of metabolic targets and off-targets on each individual cell-type/tissue, as well as interactions between the different cell-types in the tumour micro-environment.

1.2 Research Objectives

The main focus of this work is to develop system-level approaches that allow personalised immunotherapies, which will be materialised by the development of computational tools based on constraint-based modelling, omics data mining and immunoinformatics.

This work provides innovative contributions along the following axes of research and development:

- Reviewing the state of the art on the metabolism of the immune system, cancer, and the tumour micro-environment, as well as on constraint-based modelling and its applications in human and cancer metabolism.
- Development of a single-cell RNAseq atlas of colorectal cancer micro-environment, using publicly available datasets;
- Development of metabolic models of a set of T-cells, based on the atlas constructed;
- Benchmark tumour deconvolution methods using bulk RNAseq data provided by the Leiden University Medical Centre (LUMC);
- Implement the code in open-source software, allowing the full reproducibility of the studies conducted, as well as bringing important resources for the scientific community.
- Write articles with the results of the work in selected international journals and conferences.

1.3 Thesis Outline

The present document is divided into six chapters. The first chapter gives a brief introduction to the subject of the work, by providing a context, motivation and main objectives.

The following chapter gives a background for this work. It starts by introducing the immune system and cancer, and how these two are connected. This is followed by a description of the metabolism in T- and cancer cells, and how it can be taken advantage of for immunotherapy. Lastly, we detail on what constraint-based modelling is and how it can be used in human cancer.

The third chapter describes the creation of a single-cell RNAseq (scRNAseq) atlas of colorectal cancer (CRC) tumour-microenvironment. It starts by explaining what datasets were collected and how they were properly integrated. This is followed by a detailed description of how the cells were annotated, with emphasis on T-cells. Lastly, an overview of the annotated cell-types is given.

Chapter four details the creation of genome-scale metabolic models of different T-cell subtypes from the micro-environment of CRC and normal matched colon. The data collected in chapter 3 was used to create these models. This is followed by a thorough analysis of the models, including model structure, flux predictions and gene essentiality.

Chapter five details the benchmarking of several tumour deconvolution methods that use single-cell RNAseq data as reference to do so. This chapter includes a summary of each method tested and how they were compared, followed by in-depth analysis of what are the best methods to use in different situations.

In the final chapter, the present work is reviewed, followed by a discussion on work to do in the future.

2

Background

This chapter gives first an overview of the immune system and cancer. The role that the immune system plays in cancer and how the different cells in a tumour micro-environment interact with each other will also be discussed, with emphasis on metabolism and how this can be taken advantage of for immunotherapy. Having the biological background of the work explained, we detail on what constraint-based modeling is and how it can be used in human cancer is overviewed.

2.1 Overview of the Immune System

Immunity is the state of protection from a disease and can be differentiated into two components, innate and adaptive immunity.

2.1.1 From innate to adaptive immune system

Innate immunity is the first to respond to the exposure of an antigen. The defense mechanisms generated by this type of immunity are not specific to a particular pathogen but can recognise classes of molecules characteristic of pathogens. This type of immunity is carried out by myeloid cells like neutrophils, basophils, eosinophils and mast cells.

The other cells from the myeloid lineage are phagocytic cells that have professional antigen-presenting cell (APC) function. These cells play a role in connecting the innate and adaptive immune systems, by secreting proteins that attract and activate lymphocytes once they make contact with a pathogen at a site of infection [3].

Adaptive immunity, even though it responds only within 5 to 6 days after the initial exposure to the antigen, has a high degree of specificity. This type of immunity can exhibit immunologic memory, where a subsequent exposure to the same antigen results in a quicker, stronger and more effective response [4]. Lymphocytes play an important role in effective adaptive immunity. They are a type of white blood cells that display antigen-binding receptors with specificity, diversity, memory and self/non-self recognition capacity. There are two main types of lymphocyte populations: B lymphocytes (B-cells) and T lymphocytes

(T-cells). B- and T- cells recognise and bind to discrete sites on the antigen, called antigen determinants or epitopes [5].

The adaptive immunity can be divided into humoral, carried out by B-cells, and cell-mediated, by T-cells.

B-cells mature within the bone marrow and each one expresses a unique antigen-binding receptor on its membrane, the antibody. When a naïve B cell finds an antigen that binds to its membrane-bound antibody, the cell starts dividing rapidly into a clone of cells that differentiate into memory B cells and effector B cells called plasma cells. These cells have the same antigenic specificity as the original parent cell (clonal selection). Plasma cells produce enormous amounts of antibodies that can be secreted into circulation. After binding to the antigen, these antibodies can form clusters that are readily ingested by phagocytic cells, or cause the lysis of the foreign organism where the antigen is [4].

Like B-cells, T-cells arise in the bone marrow but then migrate to the thymus, where they mature. Each T-cell acquires a unique antigen-binding molecule, known as T-cell receptor (TCR). There are two well-defined subpopulations of T-cells: helper (T_H) and cytotoxic (T_C) T-cells [3]. However, T-cells can only recognise epitopes when bound to cell-membrane glycoproteins called major histocompatibility complex (MHC) molecules [5]. Class I MHC molecules are expressed on the surface of nearly all nucleated cells, while class II MHC molecules are primarily expressed on APCs like monocytes, macrophages, dendritic cells and B cells.

When a naïve helper T-cell encounters and recognises an antigen loaded into a MHC class II molecule, their interaction causes production of a costimulatory signal by the APC, causing activation of the helper T-cell. This activated cell proliferates and differentiates into memory and effector cells, all with the same antigen specificity (clonal expansion). The effector cells then start secreting various growth factors, known collectively as cytokines. These cytokines play a role in activating various cells that participate in the immune response, such as B-cells, cytotoxic T-cells and macrophages. Differences in the patterns of cytokines result in different types of immune response [4].

Under the influence of these cytokines, the cytotoxic T-cell that recognises the antigen proliferates and differentiates into effector cells called activated cytotoxic T lymphocytes, which then monitor the body and eliminate any cell that display that same antigen [4].

2.1.2 T-cells

The TCR is an heterodimer, with each chain containing two extracellular domains connected by a dissulfide bond. The transmembrane region anchors each chain in the plasma membrane and interact with CD3 chains, forming the TCR-CD3 complex. CD3 is not only closely related to TCR, but its expression is also required for membrane expression of the TCRs [6].

The chains that compose a TCR are predominantly α and β chains ($\alpha\beta$TCR). $\alpha\beta$T-cells comprise 90% to 95% of all T-cells [3].

$\alpha\beta$T Lymphocytes These cells are divided into two main groups, based on whether they express CD4 or CD8 in their membranes. Cytotoxic T-cells ($_{CD8^+}$T-cells) are normally those that display the glycoprotein CD8 in their membranes and recognize the complex antigen-MHC class I. The cells that normally display the glycoprotein CD4 in their membranes are the helper T-cells ($_{CD4^+}$T-cells) and recognize the complex antigen-MHC class II [3].

The naïve form of $_{CD8^+}$T-cells browse the surfaces of antigen-presenting cells with their TCRs and the co-receptors CD8. If bound to an MHC-peptide complex, they become activated, proliferate and differentiate into either an effector cell, the activated cytotoxic T-cell, that will monitor the body and eliminate any cell with the foreign antigen complexed with MHC class II, or a memory cell. To proliferate and differentiate optimally, naïve $_{CD8^+}$T-cells need help from mature $_{CD4^+}$T-cells [7].

The naïve form of $_{CD4^+}$T-cells also browse the surfaces of antigen-presenting cells with their TCRs and the co-receptors CD4. If bound to a MHC-peptide complex, they become activated, proliferate and differentiate into either an effector or regulatory $_{CD4^+}$T-cell subtype or into a memory cell. Which subtype dominates the immune response depends on the type of pathogen or malign cell, as the different subtypes produce different sets of cytokines that enable the activation of cytotoxic T-cells and other cells [7, 8]. Some of the subtypes are:

- *T Helper Type 1 (T_{H1})*: Regulates the immune response to intracellular pathogens and it is characterized by the secretion of IFN-γ and TNF-β;
- *T Helper Type 2 (T_{H2})*: Characterized by the secretion of IL-4, IL-5 and IL-13, it regulates the response to many of the extracellular pathogens;
- *T Helper Type 17 (T_{H17})*: Named after their secretion of the cytokine IL-17, they play an important role in cell-mediated immunity and may help the defence against fungi. They also secrete IL-21 and IL-22;
- *Follicular Helper T-Cell (T_{FH})*: Has a role in humoral immunity and regulates B-cell development. It secretes IL-21;
- *Regulatory T-cell (T_{REG})*: This type of cell is able to inhibit an immune response, helping with maintenance of immune tolerance [9]. Regulatory T-cells are distinguished from other cells by their presence of CD25 on their surfaces, and by the expression of the internal transcription factor FoxP3.

Memory T-cells, regardless of expressing CD4 or CD8, are often divided into two main subsets [8, 10]. While central memory T-cells express the CCR7 receptor and are mostly present in secondary lymphoid organs, effector memory T-cells do not express CCR7 and exhibit rapid effector function *ex vivo*, while mostly residing in peripheral lymphoid organs or recently infected tissues. Central memory were shown to generate effector memory T-cells *in vitro* [11]. In fact, memory CCR7$^-$CD62L$^-$CD28$^+$ T-cells have been found in peripheral blood of healthy individuals and are seen as 'transitional' memory cells, as they appear

to be more differentiated than central memory T-cells but not as fully as effector memory T-cells in terms of phenotype and magnitude of expansion in response to IL-15 *in vivo* [10].

Other subsets of memory T-cells have been found, such as the tissue-resident memory T-cells, which is emerging as pivotal in the protection of mucosal surfaces and epithelial from invading pathogens [10].

Innate Lymphoid cells (ILCs) Lacking antigen-specific receptors like those of B- and T- cells, ILCs are a diverse family of lymphocytes that comprise natural killer (NK) cells, lymphoid tissue inducer (LTi) cells and helper-like ILCs. ILCs are often considered as a subset of T-cells.

NK cells can be distinguished by the expression of specific surface markers and presence of cytotoxic granules. NK cells are efficient cell killers that attack abnormal cells, by killing any cell that does not have receptors for self MHC class I. The binding of these receptors to NK cells inhibits their killing ability [3]. MHC class I is expressed by almost all normal cells but often down-regulated in tumour cells [12].

Helper-like ILCs mirror the T-helper polarization of conventional $_{CD4^+}$T-cells and can be divided into three subtypes: ILC1, ILC2 and ILC3. These cells produce not only cytokines to orchestrate and amplify anti-microbial denfenses, but also soluble factors that promote tissue maintenance [13].

Unconventional T-cells $\gamma\delta$T-cells are scarce in lymphoid tissues but abundant at mucosal sites such as skin, tongue, intestine and reproductive organs [14, 15]. Other unconventional T-cells include CD8$\alpha\alpha$ T-cells, which express the CD8$\alpha\alpha$ heterodimer instead of the CD8$\alpha\beta$ characteristic of conventional cytotoxic T-cells, and NKT cells, which have characteristics of both conventional T-cells and NK cells.

2.2 Cancer

Although most cells in the adult body are quiescent, certain cell populations retain the ability to proliferate throughout the adult life, which is essential for proper tissue homeostasis. Cell division is influenced by exogenous signals like nutrients, growth and inhibitory factors, as well as interaction with neighbour cells and extracellular matrix [16].

During cell division, a series of surveillance pathways, known as 'cell cycle checkpoints' monitor for potential problems during the cell cycle. Once a problem is detected, cell cycle checkpoints activate signalling pathways that induce a temporary cell cycle arrest so that it can be fixed. Depending on the cell type and degree of damage, checkpoints can even induce permanent cell cycle arrest (senescence) or apoptosis [16].

Cancer is a disease caused by abnormal cells that escaped the checkpoint mechanisms and multiply uncontrollably. Interestingly, cancer cells are well known for carrying genetic alterations that often affect genes of the cell cycle machinery and checkpoint pathways. Cancer cells might generate a solid mass called tumour (or neoplasm/ neoplasia), and possibly invade the organism. Those tumours that invade the organism are mentioned as malignant tumours, while those that do not are referred to as benign.

Based on their origin, cancers can be divided into four types. Carcinomas, the most common type, occur at internal organs and usually form solid tumours. These include breast cancer, lung cancer and colorectal cancer. Sarcomas begin in tissues that connect and support the body, like nerves, muscles, blood and lymph vessels, and bone. Leukemias, on the other hand, consist on uncontrollable growth of the blood cells, while lymphomas begin in the lymphatic system.

2.2.1 Cancer hallmarks

Cancer cells display many other biochemical and biological features common to most cancers. While some are particular to distinct tumour types, most human tumours share six of these biochemical and biological features. They are known as hallmarks of cancer and are essential to the multistep process of tumourigenesis, which reflects the progressive transformation of normal human cells into highly malignant derivatives [17].

Underlying these hallmarks are genome instability, which generates the genetic diversity that expedites their acquisition. Cancer cells often show increased number of mutations, which can happen due to increasing sensitivity to mutagenic agents, a breakdown in one or several components of the genomic maintenance machinery, or both. Inflammation also fosters multiple hallmark functions [18].

Sustaining proliferative signalling Normal cells require mitogenic growth signals to move from a quiescence state to an active proliferative state. These signals are transmitted into the cell by transmembrane receptors, which then regulate progression through the cell cycle as well as cell growth [18]. Tumour cells, however, show a greatly reduced dependence on exogenous growth stimulation. They can synthesize and secrete the necessary growth factors, or even over-express growth factor receptors to make the cells hyper-responsive to low levels of growth factor [17, 18].

Evading growth suppressors Within a normal tissue, multiple anti-proliferative signals operate to maintain cellular quiescence and tissue homeostasis [17]. Disruptions in the pathways of these growth suppressors help cancer cells evading anti-proliferative signals.

Resisting cell death Alterations on the apoptotic machinery can dramatically affect the dynamics of tumour progression. In fact, apoptosis is attenuated in those tumours that succeed in progressing to high-grade malignancy and resistance to therapy [19].

Replicative immortality The previous hallmarks do not ensure, on their own, expansive tumour growth, as normal cells carry a cell-autonomous program that limits their multiplication. Telomeres are multiple tandem hexanucleotide repeats that protect the ends of chromosomal DNA and suffer progressive erosion through successive cycles of replication, loosing the ability to protect the ends of chromosomal DNA and thus cell viability. Malignant cells are able to maintain telomeres at a length above a critical threshold allows unlimited multiplication of descent cells. Most do so by up-regulating expression of the telomerase enzyme, which adds hexanucleotide repeats onto the ends of telomeric DNA [20].

Inducing Angiogenesis Oxygen and nutrients are crucial for cell function and survival. For that, cells need blood vessels nearby. The growth of new blood vessels, called angiogenesis, is transitory and carefully regulated. In the adult, angiogenesis only happens as part of physiologic processes as wound healing and female reproductive cycle [18]. Many tumours show increased expression of angiogenesis inducing factors, while others down-regulate expression of endogenous inhibitors [21].

Tissue invasion and metastasis Often, tumour masses have cells that move out, invade adjacent tissues, and travel to distant sites where they may establish new colonies [17]. These distant settlements are called metastasis and their occurrence is mostly regulated by the developmental regulatory program called *epithelial-mesenchymal transition* (EMT). After developing tissue-colonising ability, the cells in metastatic colonies may disseminate to new sites or back to the primary tumours. This can significantly modify the phenotype and underlying gene expression program of the cancer cells within the primary tumour [18].

Reprogramming of energy metabolism The uncontrolled cell proliferation characteristic of cancer cells has also to be enabled by adjustments to energy metabolism in order to fuel growth and division [17]. For example, Warburg *et al* [2] observed that cancer cells can reprogram their energy production by limiting it largely to glycolysis, leading to a state termed aerobic glycolysis. Additionaly, oxygenation fluctuates temporally and regionally, as a result of the instability and chaotic organisation of the tumour-associated neovasculature. Because of this, some tumours have been found to contain two sub-populations of cancer cells that differ in their energy generating pathways. One is glucose-dependent and secretes lactate, while the other, better oxygenated due to its position in the tumour's periphery, preferentially imports and utilises this lactate as energy source [22].

Evading immune destruction Highly immunogenic cancer cells may evade immune destruction by disabling components of the immune system that could eliminate them [18]. For example, cancer cells may paralyse infiltrating cytotoxic T and NK cells, by secreting immunosuppressive factors [23], or recruit immunosuppressive cells like regulatory T-cells [24].

The cancer hallmarks are not necessarily expressed equally nor continuously in all the tumour, as there is cell-to-cell phenotypic variability due to several constrains such as location, size, and history.

The initial causes of oncogenic events include prolonged hormone stimulation, viruses, chemical carcinogens, radiation, inflammation, and acquired or inborn genetic defects. Necessary but not sufficient, causal events simply increase the probability of cancer development [25]. The oncogenic events can either be genetic, like mutations, recombinations and copy number changes, or epigenetic, transcriptional or post-transcriptional. The final cancerous phenotype may result from one or a successive addition of oncogenic events [25].

These oncogenic events affect molecules that take part on signalling pathways or gene regulatory networks. One event can induce several hallmarks. For example, the disruption of p53 can induce angiogenesis and proliferation, and suppress apoptosis [17, 21]. Furthermore, while a specific genetic event

may only contribute to the acquisition of a single hallmark in certain tumours, the same event may lead to simultaneous acquisition of several distinct hallmarks in other tumours [17].

The cell's paths for becoming malignant vary greatly. Mutations in certain oncogenes and tumour suppressor genes can occur early in some tumours and late in others, which causes the particular sequence in which hallmarks are acquired to vary widely. Nonetheless, the hallmarks that are ultimately reached are shared by virtually all types of tumours [17].

2.2.2 Role of the Immune System in Cancer

The human organism is able to undergo a series of stepwise events called the cancer-immunity cycle so that a productive anti-tumour immunity response can occur. Firstly, neoantigens released from tumours are passively transported or captured and delivered by dendritic cells to regional lymph nodes via afferent lymphatic vessels. In tumour-draining lymph nodes, dendritic cells present the captured cancer-specific antigens to T-cells, activating their responses. At this stage, the balance between T effector cells and T regulatory cells is an important key to the final outcome [26].

The newly activated cytotoxic T-cells exit the lymph node and circulate throughout the body via the bloodstream, whose chemokine gradients and adhesion molecules direct circulating T-cells to extravasate through the blood vessels and migrate into the tumour. Once here, they scan the cancer cells and kill them, which releases additional tumour-associated antigens to increase the range and depth of the response in subsequent steps [26, 27]. While cytotoxic T-cells, and NK cells, engage in tumour killing, T_{H1} and, sometimes, T_{H17} cells provide important help that boosts cytotoxic immunity [28]. Under ideal circumstances, this cycle leads to the eradication of malignant cells by the cytotoxic cells, establishes tumour-specific immunological memory, and prevents further tumour progression [27].

In the vast majority of established tumours, however, the leukocytes that infiltrate the tumour are insufficient to stop tumour growth [28]. For instance, tumour antigens may not be detected, or dendritic cells and T-cells may treat antigens as self rather than foreign. T-cells may not even properly infiltrate tumours due to factors present in the tumour micro-environment [26]. Also, if T regulatory cells exceed effector responses, they will be able to inhibit these responses against tumour cells.

It is clear now that tumours are not an homogeneous cell population of only cancer cells, they are also composed by several types of non-malignant cells. This environment, referred to as tumour microenvironment (TME), includes blood and lymphatic endothelial cells, tumour-infiltrating leukocytes (TILs), mesenchymal stem cells (MSCs) and their differentiated progeny, such as cancer-associated fibroblasts (CAFs) and pericytes, accompanied with the extracellular matrix.

The different constituents of the TME can interact closely with each other and the tumour cells. These interactions control and shape tumour cell survival, invasiveness and metastatic dissemination, as well as access and responsiveness to therapeutics. This dynamic relationship is set early on in malignant growth and evolves throughout the life history of a tumour [27, 28]. Besides direct contact and contact through cytokine and chemokine production, metabolism is also an important part in this interaction.

2.3 Metabolism in T-cells and Cancer

Glycolysis, oxidative phosphorilation (OXPHOS), fatty acid oxidation (FAO), the tricarboxylic acid (TCA) cycle and fatty acid synthesis (FAS) are core cooperative metabolic pathways in any type of cell.

Glycolysis is the pathway by which glucose is broken down into pyruvate, which can then be used as substrate in the TCA cycle in the presence of oxygen, or for production of lactate when the mitochondria are damaged or in the absence of oxygen. However, in 1958, Warburg *et al* [2] found that activated leukocytes and tumour cells, two types of cells that need to rapidly divide, produce lactate from pyruvate even in the presence of oxygen, a process termed aerobic glycolysis. Although it only produces a net of 2 ATP, aerobic glycolysis provides metabolic intermediates necessary for many other metabolic pathways, allowing these cells to rapidly proliferate.

In mitochondria, the TCA cycle incorporates Acetyl-CoA from pyruvate produced by glycolysis or FAO, to generate reducing equivalents (NADH and $FADH_2$), which donate electrons to the electron transport chain (ETC). This leads to the generation of a proton gradient across the inner mitochondrial membrane and ultimately to the generation of up to 36 ATP in a process named OXPHOS. NADH and $FADH_2$ can also be produced during FAO. During this process, reactive oxygen species (ROS) are also produced [29].

2.3.1 T-cells

Throughout the life of an immune cell, energy and substrate requirements change considerably, as certain metabolic pathways must be engaged, or suppressed, to facilitate development and activation [29].

Naïve T-cells The naïve $CD8^+$ and $CD4^+$ T-cells formed in the development stage leave the thymus and enter the circulation as resting cells. Traveling throughout the organism on immune surveillance requires constant cytoskeletal rearrangements, a process that is ATP expensive but requires only basal replacement biosynthesis [30]. These cells need a metabolic balance that favours energy production over biosynthesis to move through tissues and prevent cell death, without leaving a quiescence state [29].

To do this, naïve T-cells rely greatly on the high-energy yielding processes of FAO, and pyruvate and glutamine oxidation via the TCA cycle (figure 1) [30].

Naïve T-cells Activation Activated T-cells undergo high production of metabolite precursors and ATP to support the cell growth phase prior to the first T-cell division that will enter the cell in the clonal expansion phase (figure 1).

Activated T-cells engage into mitochondrial one-carbon metabolism associated with induction of the folate cycle, which coincides with the increased TCA cycle and OXPHOS [31]. Also, lipid synthesis is suppressed and instead lipid oxidation is promoted [32].

ROS are generated in the mitochondria as a consequence of the TCA cycle and OXPHOS [33]. High levels of ROS lead to uncontrolled proliferation of T-cells. Glutathione, generated through one-carbon units derived from serine in the mitochondria, titrates ROS, maintaining moderate levels of ROS [34].

Rapid cell division requires biosynthesis of intracellular constituents including lipid membranes, nucleic acids and proteins, which need increased glucose and glutamine to satisfy the metabolic requirements, while decreasing lipid oxidation (figure 1) [30].

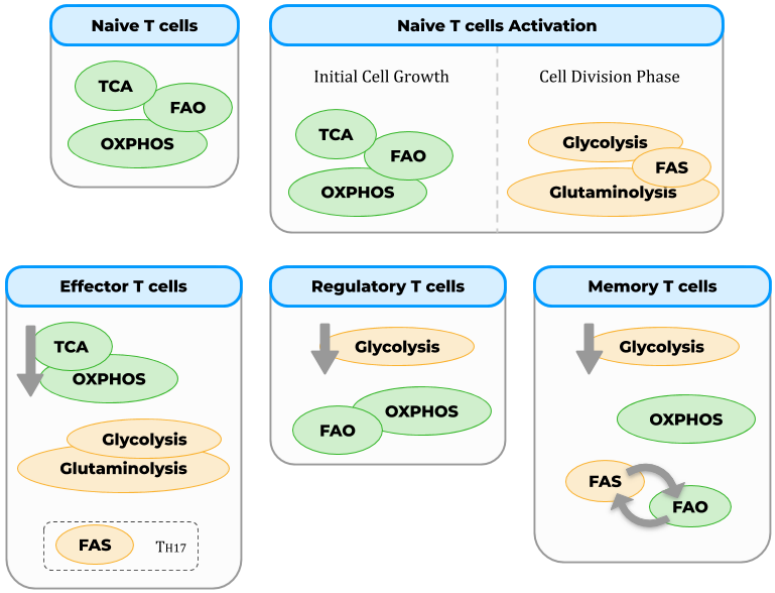


Figure 1: Overview of the characteristic metabolic phenotype throughout T-cells life cycle. In green are the metabolic pathways characteristic of quiescent/immunosuppressive cells and in yellow those of proliferating cells. Naïve T-cells and their activated form in the initial growing phase are characterised by high activity of fatty acid oxidation (FAO), tricarboxylic acid (TCA) cycle and oxidative phosphorilation (OXPHOS). Activated naïve T-cells in the cell division phase highly express glycolysis, glutaminolysis and fatty acid synthesis (FAS). Effector T-cells have low TCA and OXPHOS activity, and high glycolysis and glutaminolysis. T_{H17} cells are further characterised by high levels of FAS. Regulatory T-cells have low glycolysis but high OXPHOS and FAO. Finally, memory T-cells are characterised by low glycolysis, high OXPHOS, and the futile cycle between FAS and FAO.

Activated T cells also increase methionine uptake. This amino acid is highly important for proliferating cells, as it is the predominant 'start' amino acid in protein synthesis. The methionine cycle further generates methyl donors required for RNA and histone methylation[35].

Finally, activated T-cells go through asymmetric cell division. Associated with different distribution of metabolic mediators, asymmetric division drives activated T-cells toward either an effector or memory phenotype [36].

Those cells that inherit higher levels of amino acids go through a regulatory process that makes them more glycolytic and more likely to develop into effector cells [37]. They exhibit increased expression of effector molecules and increased expression of the large neutral amino acid transporter CD98, critical for clonal expansion and effector cell differentiation [38].

The other cells have enhanced lipid metabolism, spare respiratory capacity (SRC, the extra capacity that cells have available to produce energy in response to increased stress or work) and survival. All features of the memory phenotype, these cells are primed to become long-lived memory cells [37].

Effector T-cells Activated effector T-cells use glycolysis and glutaminolysis for ATP generation and redox balance. In fact, these cells upregulate Glut1, which mediates the increased glucose uptake [39]. Cells that starve of glucose have increased amounts of glutamine-derived glutamate and pyruvate to allow them maintain TCA activity (figure 1) [40]. This metabolite allocation allows lipids and other amino acids to be redirected to generate biomass, such as nucleotides for DNA and lipids for membranes, to support cell division and effector cell functions.

Effector T cells have increased fission and more punctate mitochondria with looser cristae, which leads to a physical dissociation of ETC supercomplexes that may cause electrons to linger in the complexes and imbalance redox reactions. This imbalance can cause an increase in NADH levels that slow the TCA cycle. To restore the redox balance, cells augment glycolysis and shunt pyruvate as excreted lactate, known as aerobic glycolysis. This allows regeneration of NAD+ from cytosolic NADH [41].

Memory T-cells Besides the initial asymmetric division during T-cell activation, some effector CD4+ and CD8+ T-cells are capable of differentiating into long-lived quiescent memory cells after pathogen clearance instead of suffering apoptosis. In this phase, T-cells no longer undergo rapid growth that requires high rates of biosynthesis. Instead, they require efficient energy generation to support basic cellular functions and prevent cell death [30]. Thus, aerobic glycolysis is reduced, while FAO and mitochondrial metabolism is favoured (figure 1). This also allows memory T-cells to increase their capacity to undergo oxidative metabolism under metabolic stress, due to higher SRC [29, 30].

Fatty acids are synthesized by memory T-cells from glucose in internal lysossomal stores, rather than acquiring them from an extracellular source. The fatty acids formed are then broken down by lisossomal-acid-lipase-mediated lipolysis to liberate free fatty acids from storage to be used as substrates in FAO. This futile cycle may be engaged to ensure a continuous lipid supply for FAO, regardless of the extracellular lipid content, and maintain enzyme expression to keep cells primed and ready for rapid recall in the event of pathogen re-encounter [42].

Although SRC and this futile cycle are important for the function of central memory T-cells in the lymph nodes, effector memory T-cells that reside in tissues rely on the import of extracellular fatty acids and on glycolysis [43]. Nevertheless, and although not yet clear, the substrates used for FAO in different memory T-cell populations may be due to substrate availability in different tissue environments [29].

Memory T-cells have increased expression of phosphoenolpyruvate carboxykinase (PCK1), which mediates glycogen biosynthesis and subsequent glutathione production through the pentose phosphate pathway, which maintains memory T-cells by reducing ROS levels [44].

Effector $_{CD4^+}$ T-cells subsets The distinct metabolic pathways engaged by the different $_{CD4^+}$T-cell subsets (figure 1) are not only crucial for their differentiation and survival, but also support important cell-specific functions.

Although all effector $_{CD4^+}$T-cell subsets have higher rates of glycolysis than naïve T-cells, T_{H1}, T_{H2} and T_{H17} cells have higher levels of Glut1 compared with regulatory T-cells [45]. Increased glucose uptake is not only sufficient to selectively enhance effector T function, but the inhibition of glucose metabolism

is capable of selectively inhibit effector T-cells, specially T_{H17} [46]. T_{H17} cells display lower OXPHOS than other helper T-cells. Regulatory T-cells are only partially dependent on glycolysis, resulting from the simultaneous increase in OXPHOS [47]. While high levels of glycolysis induces their proliferation, it limits their suppressive capacity [48]. Once glycolysis-related genes are inhibited and lipid and oxidative metabolism genes are promoted, regulatory T-cells reach their maximal suppressive ability.

T_{H17} cells rely on *de novo* FAS, rather than acquisition of extracellular fatty acids, to meet lipid requirements [49]. Cholesterol biosynthesis is required for suppressive function in regulatory T-cells [50].

In spite of exhibiting high FAO levels, Cluxton D. *et al* [47] observed that regulatory T-cells do not entirely rely on FAO. Glycolysis can serve as an alternative energy source. Raud B. *et al* [51] even found that deletion of the mitochondrial long-chain fatty acid transporter CPTI, essential for long-chain FAO, did not affect regular T-cell development and function.

2.3.2 Tumour cells

Cancer cells go through metabolic changes relative to normal cells that support the acquisition and maintenance of malignant properties. Many of these changes are observed among most types of cancer cells, which supports reprogrammed metabolism as a hallmark of cancer [52].

Nutrients Uptakes The uptake of nutrients from the environment must be increased to fulfil the biosynthetic demands associated with the high proliferation rate of cancer cells. Glucose and glutamine are the two most important components. Their catabolism provides various carbon intermediates that are used for the assembly of various macromolecules. Furthermore, controlled oxidation of carbon skeletons of glucose and glutamine allows generation of NADH, $FADH_2$ and NADPH. Glutamine further contributes with reduced nitrogen for the *de novo* biosynthesis of nitrogen-containing compounds [52].

Normal cells do not import nutrients in a constitutive manner, as this process is strictly regulated by growth factor signalling and interactions with the extracellular matrix. However, accumulated oncogenic mutations in cancer cells allows tumour cells to constantly scavenge for glucose, glutamine and essential amino acids from the extracellular environment [18]. The glucose transporter *Glut1* [53] is up-regulated, just like the glutamine transporters ASCT2 and SN2.

Glutaminase expression is also promoted [54], which converts glutamine into glutamate, whose accumulation promotes TCA cycle. Increased cysteine uptake is promoted by the glutamate accumulation, as the xCT transporter imports cysteine in exchange of glutamate [55]. Cysteine, a sulfur containing amino acid, can be involved in the biosynthesis of glutathione, iron-sulfur clusters, and hydrogen sulfide (H_2S). H_2S is involved in the protection from oxidative stress, increased mitochondrial respiration, protection from apoptosis, and facilitation of angiogenesis [56].

Although the majority of normal proliferating cells require exogenous supply of glutamine despite the existence of a glutamine biosynthetic pathway, some cancer cells over-express glutamine synthetase and are able to produce glutamine *de novo* [57].

Tumour cells are also able to use opportunistic modes of nutrient acquisition to access normally inaccessible nutrient sources, as well as to recover pre-made molecules when their synthesis within the cell is compromised[52]. The deficit of unsaturated fatty acids in cancer cells can be overcome by the import of ready-made fatty acids. Cancer cells can even induce the neighbouring normal cells to release stored lipids [58].

Bioenergetics Cancer cells exhibit aerobic glycolysis, a robust provider of precursors and reducing equivalents necessary for biosynthesis of macromolecules essential for cell proliferation [52].

The first metabolite produced in the glycolysis pathway, glucose-6-phosphate, enters the pentose phosphate pathway (PPP), generating NADPH and ribose-5-phosphate, a structural component of nucleotides. In fact, PPP utilisation is frequently elevated in tumour cells [59].

3-Phosphoglycerate can be used as a precursor for the biosynthesis of serine and glycine, and as means to generate methyl donor groups and NADPH. Serine, for example, is a major substrate of the folate cycle, an essential source of precursors for the biosynthesis of purines and thymidine. *Methylene tetrahydrofolate dehydrogenase 2* (MTHFD2), a component of this cycle, has been found to be one of the most frequently over-expressed metabolic enzymes in cancer [60]. Furthermore, the final reaction of glycolysis is catalysed by *pyruvate kinase* (PK) in the form PKM2 in most tissues, including tumours [61]. PKM2 is activated by serine [62].

Nevertheless, most cancer cells still generate the majority of ATP through mitochondrial function, despite the high glycolytic rates [52]. There is no actual shift between TCA and glycolysis, like initially proposed by Warburg *et al* [2], but rather a considerable reduction of the TCA cycle activity to a state sufficient to maintain mitochondrial integrity and ATP production.

In addition to pyruvate derived from glycolysis, fatty acids and amino acids can supply substrates to the TCA cycle. In fact, glutamine can provide acetyl-CoA as a precursor when pyruvate oxidation to acetyl-CoA is compromised by hipoxia or ETC impairment. Also, most proliferating cells depend on a continuous supply of glutamine to maintain TCA cycle intermediates [52].

Biosynthesis of macromolecules The production of biosynthetic intermediates through metabolic pathways such as glycolysis, PPP, TCA cycle and non-essential amino acid synthesis allows the assembly of larger and more complex molecules, required for replicative cell division and tumour growth. Among these, the most commonly studied in cancer metabolism are proteins, lipids and nucleic acids [52].

While glutamine-derived glutamate works as a nitrogen donor for the production of several non-essential amino acids via transamination, the amide nitrogen of glutamine is used by *asparagine synthetase* (ASNS) to produce asparagine from aspartate. Notably, ASNS is frequently up-regulated in tumours and is associated with poor prognosis [63, 64].

Essential amino acids, in turn, are acquired from the extracellular space through surface transporters under the influence of growth factor signalling [65].

Arginine is a non-essential amino acid that can become conditionally essential in some tumourigenic contexts. For example, *arginino succinate synthase* (ASS1), essential to the *de novo* biosynthesis of arginine, is frequently epigenetically silenced in pancreatic and renal cancers [66, 67]. Inactivation of this enzyme causes cancer cells to accumulate ornithine, which is then used in the production of polyamines. These compounds have been shown to inhibit apoptosis and promote tumour growth invasion [52]. Proline can be produced from glutamate or from arginine-derived ornithine. Notably, the principal enzyme in proline production, *pyrroline-5-carboxylate reductase* (PYCR1), is one of the most commonly overexpressed enzymes in tumours [60]. The suppression of ASS1-driven argininosuccinate production can also cause accumulation of its substrate, aspartate, required for nucleotide production [52].

Purine and pyrimidine nucleotides are required for synthesis of RNA and DNA. The expression of *phosphoribosyl pyrophosphate synthetase 2* (PRPS2) [68] and *carbamoyl phosphate synthetase II* (CAD) [69] are up-regulated by c-Myc in tumour cells. These enzymes are involved in purine and pyrimidine biosynthesis, respectively.

The capacity to rapidly produce lipids in cancer cells facilitates the formation of membranes, the alteration of membrane composition in favour of oxidative damage-resistant saturated fatty acids, lipidation reactions, and cellular signalling. The activity of several enzymes involved in lipid synthesis, even in lipid-replete conditions, are up-regulated in cancer cells.

2.3.3 Tumour Micro-Environment

The most frequently found tumour infiltrating lymphocytes (TILs) within the TME are tumour-associated macrophages (TAMs) and T-cells [28]. TAMs are alternatively activated macrophages reprogrammed to display various tumour-promoting functions [28, 70]. Polymorphonuclear leukocytes are rarely seen in human TMEs [71].

High numbers of cytotoxic T cells and T_{H1} cells are correlated with better survival in some cancers, including invasive colon cancer, melanoma, multiple myeloma, and pancreatic cancer [28]. For example, T_{H1} cells maximize the killing efficiency of macrophages and proliferation of $CD8^{+}$T-cells [72]. However, during *de novo* carcinogenesis, anti-tumour T-cells cannot control tumour growth, due to tumour-induced tolerance mechanisms in most cancers [73].

In fact, tumour cells have the ability to actively downregulate all phases of anti-tumour immune responses through metabolism, affecting the recruitment and function of immune cells. The different constituents of the TME can also interact closely with each other to control and shape tumour cell survival, invasiveness and metastatic dissemination. These interactions dictate the ability of the immune system to fight cancer and even responsiveness to therapeutics. This dynamic relationship is set early on in malignant growth and evolves throughout the life history of a tumour [27, 28].

Figure 2 gives an overview of some of the metabolic interactions within the TME, discussed below. When tumour cells are exposed to hypoxic conditions, the production of the hypoxia-inducible factor 1α (HIF-1α) is up-regulated, which promotes glycolysis and leads to activation of angiogenesis-promoting

factors. In the presence of oxygen, HIF-1α is degraded. The acidic microenvironment caused by tumours, that up-regulates glycolysis and increases production of lactic acid, can 'simulate' the effects of hypoxia, even in the presence of oxygen. Thus, HIF-1α may not be suppressed even in normoxia conditions [72].

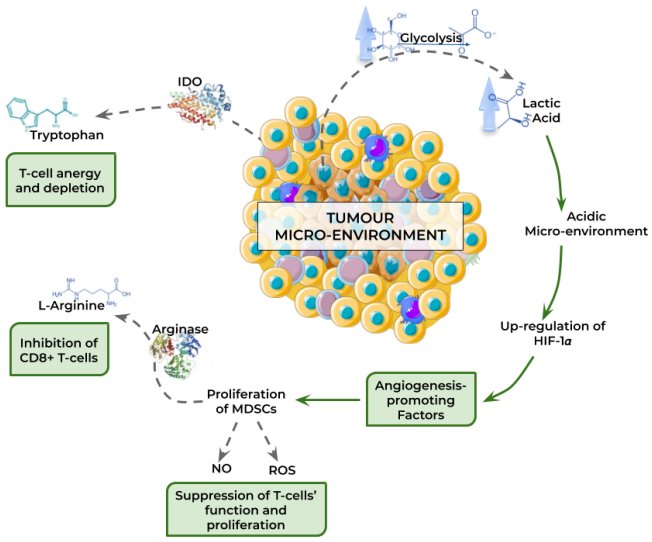


Figure 2: Summary of some of the metabolic interactions within the tumour micro-environment (TME). The up-regulation of glycolysis in tumour cells increases the secretion of lactic acid into the micro-environment, causing an acidic microenvironment that 'simulates' the effects of hypoxia. This leads to an up-regulation of HIF-1α, activating angiogenesis-promoting factors. These factors can stimulate the proliferation of MDSCs, which liberate arginase into the micro-environment that will consume L-arginine. Shortage of this metabolite can inhibit $CD8^+$T-cell function. Constitutively expressed in most human tumours, IDO is involved in the catabolism of tryptophan, which induces immunosuppression through T-cell anergy and depletion. | IDO: Indoleamine-2,3-dioxygenase; HIF-1α: hypoxia-inducible factor 1α; MDSCs: myeloid-derived suppressor cells; NO: nitric oxide; ROS: reactive oxygen species

Angiogenic factors also stimulate the proliferation of myeloid-derived suppressor cells (MDSCs), which include immature dendritic cells, neutrophils, monocytes, and early myeloid progenitors. When stimulated, these cells up-regulate and liberate arginase into the micro-environment, which consumes L-arginine. Shortage of L-arginine in the environment can inhibit $CD8^+$T-cell function. Stimulated MDSCs also increase production of nitric oxide (NO) and ROS. NO can suppress T-cell function through inhibition of MHC-II expression and T-cell proliferation and apoptosis [72].

TAMs are capable of blocking $CD8^+$T-cells proliferation or infiltration by releasing factors with immunosuppressive potential like ROS [70]. TAMs can further suppress surface proteins on infiltrating T-cells through nitrosylation, inhibiting T-cells' anti-tumour functions [74].

ROS released by tumour cells can induce cancer-associated fibroblasts to up-regulate aerobic glycolysis and secrete lactate and pyruvate. Tumour cells then consume these two metabolites [75]. Alternatively, the lactate secreted by cancer cells is taken up by cancer-associated fibroblasts and used as fuel to drive tumour-promoting functional activities [76]. Cancer-associated fibroblasts are the most predominant non-hematopoietic stromal cell type in the TME [27].

Indoleamine-2,3-dioxygenase (IDO) is constitutively expressed in most human tumours [77]. IDO is involved in the catabolism of tryptophan, an essential amino acid for T-cell proliferation and differentiation [71]. The catabolism of tryptophan into kynurenine by this enzyme induces immunosuppression through T-cell anergy and depletion [77]. Plasmacytoid dendritic cells have shown expression of indoleamine-2,3-dioxygenase (IDO), as well as defective production of type I interferon [74].

Presence of TNF-α and IFN-γ can dramatically increase MSCs' expression of *inducible nitric oxide synthase* (iNOS) [78] and production of IDO [71]. iNOS consumes arginase, producing ROS and NO.

Mevalonate pathway intermediates produced by tumour cells were shown to activate and promote $\gamma\delta$T-cells' anti-tumour responses [15].

2.3.4 Targeting metabolism for therapy

The most common therapies applied to cancer patients are surgery, radiation therapy and/or chemotherapy. Surgery is mostly used for non-invasive solid tumours and coupled with other treatments. While radiation therapy uses high doses of radiation to kill or slow the growth of tumour cells, by damaging their DNA beyond repair, chemotherapy uses drugs to this end. However, these two treatments do not only kill tumour cells, they also affect healthy cells. For example, the most common side effects comprise fatigue, hair loss, nausea and vomiting.

Through more recent years, other alternatives for cancer therapy have been studied to generate as few harmfull side effects as possible to the patient. These can be achieved by searching for specific targets in tumours cells that do not affect healthy cells, or even specific ways of enhancing the immune response against cancer. As cancer cells go through metabolic changes relative to normal cells that support the acquisition and maintenance of malignant properties, the metabolism of cancer cells has been studied for immunotherapy.

Naturally, inhibition of glycolytic and glutaminolytic enzymes has been extensively studied. Diclofenac has been reported to reduce tumour growth, the quantity of regulatory T-cells and lactate in the microenvironment in a glioma model [79]. Neutralisation of the TME's acidic environment with bicabornate or esomeprazole, for example, improves cytotoxic T-cell and NK cell anti-cancer immune responses [80]. Hexokinase (HK), a glycolytic protein, is overexpressed in many tumour cells and its inhibition was shown to delay tumour progression in pre-clinical mouse models [81]. However, 1-deoxyglucose (2DG), an inhibitor of HK, also leads to impairment of T-cells' metabolism [82].

Bis-2-(5-phenylacetamido-1,2,4-thiadiazol-2-yl) ethyl sulfide (BPTES) is a glutaminase inhibitor that showed anti-cancer immunity in several tumour models with elevated activity [83].

OXPHOS is also a great potential target to eliminate tumour cells.

The anti-diabetic drug metformin can act as an anti-cancer agent by inhibiting the complex I from ETC. This causes a decrease in ATP levels that lead to cancer cell death [80]. However, this drug's uptake occurs through the *organic cation transporters* (OCTs), only present in a few tissues, such as liver and

kidney, and in certain tumour cells [65]. Regarding effects on the immune system, this drug can further enhance memory T-cells and regulatory T-cell expansion [80].

Dichloroacetate (DCA) induces a shift from glycolysis to OXPHOS, thus inhibiting tumour cells growth *in vitro* and in mouse models. However, it also affects T-cells, favouring regulatory T-cell formation [84].

Down-modulation of IDO has been shown to improve anti-tumour responses [74, 80]. Imatinib, for instance, activates effector T-cells and suppresses regulatory T-cells in an IDO-dependet manner [85].

Inhibition of the rate-limiting enzyme in FAO, CPTI, has anticancer effects *in vitro* and *in vivo*. However, etomoxir showed hepatotoxicity in patients with congestive heart failure, and other inhibitors are still to be approved for cancer therapy [86].

Resistance to cancer therapies may result from not taking into consideration the TME, as briefly noted above in few examples. Furthermore, CSCs are typically therapy-resistant due to decreased oxidative stress response, increased genomic stability, and expression of multiple drug resistance transporters [78].

TME cells are not subject to mutational and epigenetic changes that result in drug resistance. Thus, targeting the TME cells, specially immune cells, along with tumour cells can be advantageous.

Regarding tumour stroma, targeting the tumour extracellular matrix can boost natural anti-tumour immunity and improve immune-therapeutics efficacy. However, it may also enhance regulatory T-cell infiltration and increase angiogenesis [27].

Despite the progresses, clinical responses may be transitory and have limited benefits in long term [74], mostly due to drug resistance caused by the existence of similar pathways that are alternatively upregulated by the cell. Furthermore, investigation of a potential metabolic target is normally only performed in tumour cells, without counting with the possible negative side effects on other cells in the TME and outside.

These problems show the value in creating *in silico* metabolic models of the whole metabolome of immune and tumour cells to study the effects of metabolic targets and off-targets on each individual cell, as well as interactions between the different cells in the TME upon disruption with approved drugs. This could be further fine-tuned to patient-specific cases, personalising each patient's therapy. Metabolomics can aid in such a way that cancer therapy and cancer immunotherapy act specifically on malignant cells, remarkably reducing the side effects to the patient.

2.4 Constraint-based Modeling and Human Cancer

Two popular, but very different, approaches to model a cell's metabolism are the kinetic (dynamic) and stoichiometric modeling. The kinetic modeling, as the name suggests, relies on the enzyme kinetics information to model the metabolite concentrations and reaction fluxes through time [87, 88]. However, kinetic models require a lot of details for their construction, which must be obtained through experiments that are difficult to perform. Because of this, they usually end up covering only a few pathways or models from small organisms [88, 89].

Stoichiometric modeling in turn disregards this dynamic intracellular behaviour, by assuming the pseudo-steady state for internal metabolites, due to the intracellular dynamics being much faster than extracellular dynamics. With this assumption, the variation in concentration of each metabolite throughout time is considered to be zero [89]. As such, they are better fitted to be applied at the genome-scale, as a smaller amount of information per reaction allows the development of larger models. However, the simulation of any changes in time and metabolite concentrations is not possible [88].

For a reaction $S_1 + S_2 \longleftrightarrow 2P$, the reaction rate ($v_t$) is the difference between the rate of the forward reaction (V_f) and the rate of the backward reaction (v_b), leading to: $v_t = v_f - v_b$. As the molecularity is 1 for each substrate and 2 for the product, the variation in concentration through time of S_1 and S_2 is $\frac{d[S_1]}{dt} = \frac{d[S_2]}{dt} = -v_t$, and the variation in concentration of P through time is $\frac{d[P]}{dt} = 2.v_t$.

To model all the metabolic interactions within a cell, the structure of metabolic models is represented by a matrix S, where each row represents a metabolite from a metabolic network and each column a reaction in the same network.

2.4.1 Stoichiometric modeling

For stoichiometric modeling, the matrix values represent the stoichiometric coefficients of the metabolites in the reactions. A positive value denotes that the metabolite in question is a product of the reaction, while a negative value means that the metabolite is a substrate. If the coefficient value is zero, the metabolite does not participate in that reaction.

The structure of this matrix is often complemented by information gathered about gene-protein-reaction (GPR) associations, which list the set of metabolic reactions encoded in the genome. These associations can go from the simplest one, where one gene encodes one protein that catalyses one reaction, to more complex associations, where several genes encode part of one enzyme (enzyme complexes). It is also possible that multiple enzymes, encoded by the respective genes, perform the same function (isozymes), while one protein can catalyse several reactions.

With this, the stoichiometric model of a network can be defined by:

$$S.v = 0, \tag{2.1}$$

where *S* represents the stoichiometric matrix and *v* the flux vector. The *i* element of *v* represents the rate of the reaction *i*, present in the column *i* of the stoichiometric matrix. Figure 3 represents an example of how a stoichiometric model can be obtained from a metabolic network. The reactions v_4, v_5 and v_6 in that figure represent exchange reactions, used to explain the flow of metabolites in and out of the cell.

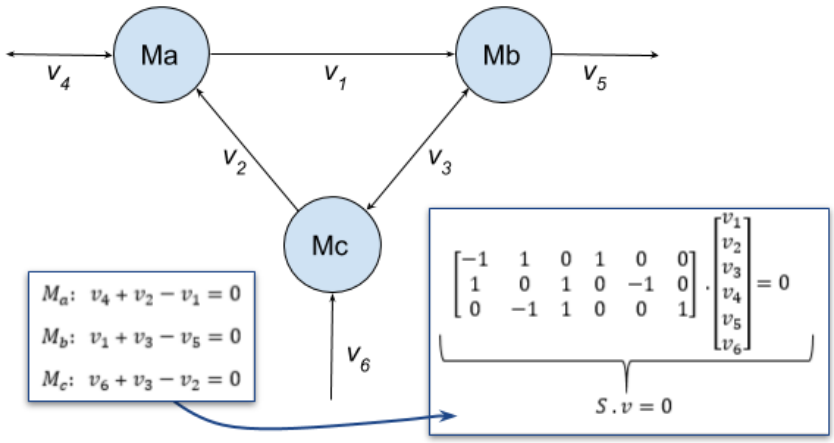


Figure 3: Example of a stoichiometric representation of a metabolic network.

Furthermore, cell growth is normally also taken into consideration in the stoichiometric matrix, by adding a column, where the metabolites that are consumed during production of biomass, such as nucleic acids, proteins and lipids, have a negative stoichiometry value. If not, they have a null stoichiometry value. The corresponding flux rate in v corresponds to the growth rate of the organism.

Normally, there are more reactions than metabolites in a metabolic network, i.e., there are more unknown variables than equations. This leaves the stoichiometric model underdetermined, as there is not a unique solution to this system.

Since cells are subject to constraints that limit their behaviour, defining these in the stoichiometric model leads to a space of flux distributions that can actually be achieved by a cell. As the metabolic phenotype can be defined in terms of flux distributions, this space represents, or contains, all feasible phenotypes. This establishes the link between stoichiometric modeling and constraint-based modeling (CBM), as stoichiometric modeling can be viewed as a particular case of constraint-based modeling that only has stoichiometric constraints [89].

There are two main types of constraints in CBM [89]. The non-adjustable, invariant, constraints are time-invariant restrictions of possible cell behaviour. The invariant constraints are the ones that compose the general assumptions of every stoichiometric model [88]: the mass conservation principle, applied to limit metabolic network behaviour in models; energy balance, derived from the law of conservation of energy in an isolated system; the steady-state, where metabolites' concentrations do not vary over time; and thermodynamic constraints, which limit the direction and capacity of reactions.

Adjustable constraints depend on environmental conditions and vary from one individual cell to another. These include regulation and experimental measurements [89].

A graphical representation of this is provided in figure 4, where each axis represents the flux through a reaction in the network. With the assumption of steady state ($S.v = 0$), the space of possible flux distributions is firstly restricted to an hyperplane, which is converted into a convex polyhedral cone when the irreversibility of fluxes is taken into consideration, further constraining the model. By convention, the natural flow direction of a reaction is assigned the positive direction. Therefore, every irreversible reaction

is constrained to only non-negative values: $v \geq 0$. If a reaction is reversible, its lower bound is normally set to minus infinity or a large negative value. After this, if the capacity of all reactions is set, which refers to the maximum flux values of enzyme or transport capacities ($v \leq v_{max}$), also called upper bounds, the space of possible solutions is bounded, having no longer infinite possibilities.

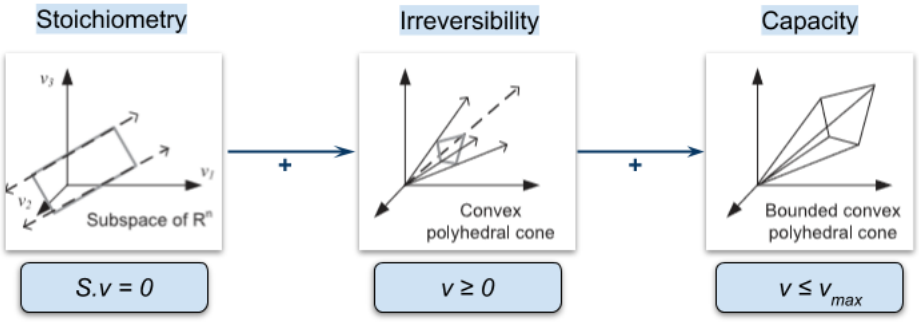


Figure 4: Graphical representation of the space of feasible flux distributions of a stoichiometric model. Adapted from [89].

The more non- and adjustable constraints are incorporated, the more restrict the space of possible flux distributions is [89].

There are two major types of analysis that can be performed to extract information from CBMs: network-based pathway analysis and determination of flux distributions under certain conditions.

Network-based Pathway Analysis Network-Based Pathway Analysis attempts to study the systemic properties of the network and the capabilities of the cell metabolism.

Elementary flux modes (EFMs) represent unique routes of a metabolic network that may connect inputs to outputs. Each EFM is non-decomposable, i.e., the removal of a reaction prevents the occurrence of a non-zero steady flux distribution by means of the remaining reactions [90]. Thus, each metabolic network is represented by its own unique set of EMs [91].

As most reactions in a metabolic network are catalysed by enzymes, these reactions show the minimal sets of enzymes that must be expressed for a proper functioning of the metabolic routes. If one of the reactions is blocked, the route(s) containing this reaction do not occur any more [92].

Indeed, EFMs can be used in the detection of minimal cut sets (MCSs) [90]. These are irreducible sets of reactions in the network whose inactivation prevents a feasible flux distribution of a certain objective reaction or set of objective reactions in the network.

Determination of flux distributions Experimental measurements of extracellular fluxes (uptake or excretion of metabolites) can be used to estimate the current flux of a metabolic network, which may be accomplished by methods from the field of Metabolic Flux Analysis (MFA) [89]. In some cases, these experimental measurements are provided as intervals, taking into account the uncertainty of the measurements. Therefore, instead of having a measured flux v_m, it will be given as an interval $\left[v_{(mmin)}, v_{(mmax)}\right]$.

If enough fluxes are measured, this may enable the calculation of intervals of possible values for each non-measured flux [93].

On the other hand, the flux balance analysis (FBA) approach predicts a phenotype (flux distribution) based on the main objective to be accomplished by the cell and the restrictions that this metabolic system has to obey. This approach assumes that cells evolve to achieve an optimal behaviour owing to evolutionary pressure, i.e., cells regulate their fluxes toward optimal flux distribution [89], thus aiming at achieving a biologically meaningful description of the metabolic state of a system.

The mathematical representation is as follows [94]:

$$\begin{aligned} \text{Maximize (or Minimize)} \quad & Z = c^T.v \\ \text{Subject to:} \quad & S.v = 0 \\ & lb_j \leq v_j \geq ub_j \quad j = 1, ..., N \end{aligned} \tag{2.2}$$

In the objective function Z, c is a vector of weights indicating how much each reaction in the network (v) contributes to the objective to be accomplished by the cell. So, c contains values of 1 at the positions of the respective reactions in v that enter the objective function, and 0 for those that do not; lb and ub are the lower and upper bounds, respectively, of each reaction v_j.

The most commonly used objective function is the maximization of biomass. However, there are many other widely used objective functions: minimize ATP production, minimize nutrient uptake, or even maximize a given metabolite's production [89]. Different outcomes can be achieved by not only playing with the objective function, but also by altering the constraints of the model [94]. One example is the simulation of gene knockouts, by forcing the unwanted reactions to zero flux. Another one is limiting substrates uptake to zero to mimic substrates not available to the cell.

FBA has its limitations, however, and various alternative methods based on FBA have been put forward throughout the past years [95–99]. Parsimonious enzyme usage FBA (pFBA) [95], for example, assumes that there is a selection for the fastest growing strains, like FBA, but that they also require the least amount of enzymes.

There are many situations where different flux distributions reach the same exact quantitative objective value. FVA (Flux Variability Analysis) [96] allows to determine the maximum and minimum values that each flux can have so that the constraints are satisfied and the same optimal objective values is reached, identifying as much alternative optimal solutions as possible.

Minimization of Metabolic Adjustment (MOMA) [100] is used when predicting flux distributions resulting from gene knockouts or other genetic perturbations. By assuming that a mutant is likely to initially display a sub-optimal flux distribution that is intermediate between the wild-type optimum and the mutant optimum, it calculates the flux distribution that minimizes the differences to a reference distribution, which may be provided or calculated using FBA or pFBA.

The MOMA approach tends to favour numerous small changes in fluxes over a few large changes with an equal total sum of distance to the original steady state. Thus, it is said that MOMA is more suitable to

predict the transient metabolic states that occur right after the genetic perturbations, instead of predicting well the metabolic state after the adaptation of the organism to the perturbation has occurred [101].

The regulatory on/off minimization (ROOM) approach [101] tries to minimise the total number of significant fluxes that have to change compared to the wild-type version, with the assumption that the cell's objective is to only perform the regulatory changes that minimise the adaptation cost.

2.4.2 Metabolic Models in Humans

Genome-scale metabolic models (GSMMs) are stoichiometric models that contain all the known reactions that occur in a type of cell. The construction of these models is an iterative process that may require cycling several times between different steps. It starts with a draft reconstruction that combines information on genes that encode enzymes or membrane transporters, from several sources, to construct a preliminary set of reactions and constraints. This set is analysed to detect potential faults and correct them, in a step normally called reconstruction refinement. The reconstruction is then converted to a mathematical model, where the matrix S and respective constraints are defined. The stoichiometric model is evaluated by comparing its predictions with experimental or literature data and revised if necessary.

Generic Human Models Human genome-scale metabolic networks first appeared after the Human Genome Project, allowing for the full sequence and annotation of the human genome. With this extensive information on the human genes, along with knowledge gathered through decades of research on numerous metabolic genes and enzymes, reaction mechanisms and interactions, large-scale modelling of human metabolism has been constantly progressing.

The first attempts of a genome-scale metabolic network for a generic human cell culminated in two projects: Recon 1 [102] and EHMN (Edinburgh Human Metabolic Network) [103], by 2007. Although EHMN network contained more genes and unique reactions than Recon 1, it lacked subcellular compartments that exist in the human cells, thus not considering transport or exchange reactions. By 2010, EHMN was extended to integrate subcellular location information for the reactions [104], based on Recon 1 and other sources.

Recon 2 was released in 2012 [105] by adding metabolic information present in different sources, from EHMN and literature to Recon 1. This model had 1789 genes, 7440 reactions, 2626 unique metabolites, and eight compartments (cytoplasm, mitochondria, nucleus, endoplasmic reticulum, Golgi apparatus, lysosome, peroxisome and extracellular space). Another model was released around the same time, known as HMR (Human Metabolic Reaction Database) [106, 107]. HMR was constructed by grouping information from Recon 1 and EHMN, as well as from HumanCyc [108] and KEGG databases [109]. In 2013, HMR was further expanded into a new version, HMR 2.0 [110], to include an extensive lipid metabolism. HMR was extended by merging data from previously published hepatocyte models, Recon 2, literature, HumanCyc [108] and KEGG [109]. HMR 2.0 contained 3765 genes, 3160 unique metabolites and 8181 reactions.

Since the release of Recon 2, several updates by different groups led to different versions of Recon 2. Recon 2.2. [111] integrated all these alternative versions in 2016, alongside with additional manual corrections and updates. By redefining the representation of oxidative phosphorylation, a new compartment was defined, the mitochondrial intramembrane.

Shortly after this release, Recon 3D [112] expanded Recon 2 by integrating sources such as HMR 2.0. This model contained 3 288 genes, 13 543 reactions and 4 140 unique metabolites. Two gender-specific whole-body metabolism reconstructions based on the Recon 3D model were created [113]. These models, named Harvey (male model) and Harvetta (female), were organized into 22 and 24 organs, respectively, due to sex-specific organs, and 6 types of blood cells.

More recently, a new human generic model, Human-GEM [114], was created by integrating and extensively curating several sources, including Human Protein Atlas (HPA) [115], HMR 2.0, Recon 3D and others. The novelty compared to previous models is the availability of a version-controlled open source repository, which enables community-driven curation and refinement that allow rapid and trackable updates. This model contains 13 802 reactions, 8 378 metabolites and 3 625 genes.

Context-specific Models Many reactions in a generic human GSMM are not active under certain conditions or cell/tissue types. Although a generic human model contains all the metabolic genes and reactions in the organism, not all cells/tissues express all the genes encoded in the genome or at the same levels. An example of this is the metabolism of cancer cells, which changes considerably in relation to healthy cells, as mentioned before. These differences lead to marked discrepancies in the metabolic model. Thus, reducing generic models into context-specific ones aids the *in silico* prediction of particular situations.

Various tissue-specific reconstruction algorithms have been developed throughout the years. With a general GSMM, literature and experimental data, as a starting point, these algorithms output a specific model with a subset of the reactions (and metabolites) of the original GSMM.

Tissue-specific reconstruction algorithms can be mainly divided in two major groups. The flux-dependent methods can either try to find an optimal flux distribution for the given experimental data, like GIMME [116], or use experimental data to get the subset of reactions that are active based on a defined threshold, like iMAT [117, 118], INIT [107] and its extension tINIT [119], and PRIME [120].

For example, INIT (Integrative Network Inference for Tissues) [107] is mainly focused on scoring each protein based on how strong the experimental evidence on its presence, or absence, is. With this, it finds the subnetwork whose sum of proteins' evidence scores is maximized, while the ability of all reactions in the resulting model to carry flux is tested. For this, the steady state condition is not imposed, but instead, a small positive net accumulation or secretion rate for all metabolites is allowed. The tINIT (task-driven Integrative Network Inference for Tissues) algorithm extension [119] allows the user to define metabolic tasks that the resulting model should be able to perform. It starts by identifying the set of reactions in the generic model without which at least one of the tasks fails, and goes on performing like the INIT algorithm with the additional constraint that these reactions must be in the solution.

Pruning methods, in turn, start with a set of core reactions, known to be present in the desired tissue by going through literature or experimental data, and remove the other reactions of the generic GSMM whose removal does not cause loss of reaction functionality in the core set. To achieve this, these methods establish a trade-off between maintaining the model as concise as possible and including all core reactions in the final model, allowing a core reaction to be removed if it requires too many undesirable reactions to be active [121].

The main advantages of this last type of methods relies on making it possible for the user to define the set of core reactions using multiple different sources and know that reactions with high evidence of being present in a tissue are always included in the final model, besides generating a flexible and functional metabolic model. Examples of such algorithms are MBA [122], mCADRE [123], fastCORE [124], and CORDA [121].

However, pruning algorithms are not deprived of disadvantages. The order in which reactions are removed can affect the outcome of the final model, as well as the possible removal of fundamental reactions to achieve a concise tissue-specific reconstruction can cause physiologically unlikely flux distributions [121].

The fastCORE [124] algorithm, for instance, takes a set of reactions that have strong evidence to be active in the context of interest and searches for a subnetwork that contains all reactions from the core set and a minimal set of additional reactions, by assessing the flux consistency through FVA. This flux consistency is characterized by each reaction in the subnetwork being active in at least one feasible flux distribution.

Several human cancer metabolic models have been constructed in the past years. Among these are models of glioblastoma [125], ovarian cancer [126], hepatocellular carcinoma [126], melanoma [127], breast, urothelial, lung and renal cancers [128], and colorectal cancer [129]. Several other studies have focused on constructing models specific for cancer cell lines [130–132], or even from patient-specific samples [119, 133–136].

Regarding the study of the immune system, tissue-specific reconstruction algorithms have also been used to generate metabolic models of immune cells, including CD4 T [113], naive CD4 T [137–139], T_{H1} [138, 139], T_{H2} [138, 139] and T_{H17} [138] cells, as well as macrophages [140, 141], monocytes [113], B cells [113] and NK cells [113]. However, none were applied in the context of cancer.

2.4.3 Omics Data

The advent of omics technologies have revolutionised the way biological research is conducted. They allow the analysis of a global set of molecules and their interactions simultaneously. Integrating data from different omics helps creating an overall 'snapshot' of a cell's metabolism and evaluate how different mutations affect metabolism.

Omics data can be used to find the reactions that should be present in a cell-type or tissue in order to reconstruct a context-specific metabolic model from a generic model. In fact, developing large-scale

metabolic models became possible due to the advance of omics technologies.

Genomics It consists on the study of the genomic material of an organism. In cancer patients, genomic tests are often performed before treatment. Genetic mutations are identified by comparing the tumour genome with the patient's normal tissue or a reference genome at a single base pair resolution [142].

The most used techniques have been Sanger sequencing, microarrays, and Next-Generation Sequencing (NGS) [143].

Sanger sequencing consists on a base-by-base sequencing, but it can only capture up to one thousand bases per run. Microarrays, in turn, involve the binding of the different cDNA sequences obtained in a solution to the respective probe in the array, allowing the measurement of the relative concentrations of those sequences [144]. NGS enables the identification and quantification of transcripts without prior knowledge of a particular gene sequence, and can provide information regarding alternative splicing and sequence variation. Although NGS allows whole genome sequencing (WGS), allowing identification of all coding and non-coding variants, it is also possible to only screen variants in the coding region (whole exome sequencing, WES) [143].

The great technical advances achieved throughout the years to obtain genomic data culminated in the sequence of the whole human genome, the Human Genome Project, in 2003. Since then, Sanger sequencing and microarrays have been completely substituted by NGS in sequencing the human cells' genome.

Transcriptomics This technology studies the transcriptome, which represents all RNA transcripts in a cell, both coding (mRNA) and non-coding (ribosomal, transfer, etc) RNAs. Normally, mRNAs are the main focus, as the quantification of each transcript provides and insight into the gene expression levels. This allows a better understanding of the dynamics of metabolism [143].

To assess transcriptomics data, similar techniques to those in genomics are used, namely microarrays and RNA-sequencing (RNA-seq). While microarrays only allow measurement of relative concentrations of the different transcripts, quantification in RNA-seq is done by counting the number of sequence reads assigned to different transcripts [145].

More recently, single-cell techniques like single-cell RNAseq (scRNAseq) have been developed. This technique allows quantitative characterisation of each cell's transcriptome in a sample, giving valuable information about cell-types and states at high resolution [146]. Especially in cancer, where samples from bulk transcriptomics are often analysed as if they were an homogenous population of tumour cells, not accounting for the diversity of stromal and immune cells present that can be crucial in the understanding of cancer.

Proteomics This omics studies the entire set of proteins in a cell, at a precise developmental or cellular phase. Techniques used in proteomics, mainly mass spectrometry (MS), are less scalable than those used to study nucleic acids, like NGS. Still, considering that protein quantities and activities of malignant cells

are affected by distinct replication and metabolic processes, proteomics has the potential to uniquely characterise the malignant cell or tissue and to discover diagnostic markers of a disease [142, 143].

Although MS allows identification and quantification of proteins in a sample, the eventual function of an existing protein in a sample depends on how the protein is folded, i.e., its 3D structure. To derive the proteins' structures in a sample, techniques such as X-Ray, Nuclear Magnetic Resonance (NMR) and cryo-electron microscopy come into play. These techniques allow visualization of protein domains, deduce protein function, study structural changes following disease associated mutations, and discover and develop drugs [143].

Proteomics can also be used to study protein-protein interactions (PPIs), which consist on either two proteins that physically interact in a complex, or simply share the same location, as interacting proteins are likely to share common tasks or functions [143].

Because the proteome is extremely dynamic, there is an elevated sample heterogeneity, even with the same types of samples and conditions, that complicates the development of a universal and comprehensive human proteome reference and the comparison of different studies [143].

Epigenomics Epigenomic changes include DNA methylation and chromatin modifications like histone acetylation, methylation, phosphorylation and others. These heritable changes have a great impact in the expression patterns of genes, and can be affected by mutations in enzymes in charge of DNA methylation, demethylation and chromatin modification. This may allow tumour progression if the suppressed or activated gene hinders or promotes tumour progression. Epigenomic changes are often observed in many cancers.

Identification of DNA methylation status is performed by using bisulfite treatment, which only modifies unmethylated cytosines to uracils, followed by sequencing and comparison between bisulfite-treated and untreated samples [142].

Regarding histone modifications, these are identified by high-throughput DNA sequencing technologies coupled with *chromatin immunoprecipitation* (ChIP), where modification-specific antibodies immunoisolate DNA-histone complexes with the desired histone modifications. The selected DNA sequences are then identified via microarrays (ChIP-chip) or sequencing (ChIP-seq) [142].

Metabolomics This technique comprises the analysis of the metabolites produced during biochemical reactions, which depend on gene expression to be produced. This omics is thus closely related to the ones described above, as the production of metabolites can reflect particular combinations of individuals' genetics and environmental exposures. Metabolomics can be used to develop diagnostics and understand relevant molecular pathways under specific conditions [143].

The most used techniques in this field are Mass Spectrometry, coupled with liquid or gas chromatography (LC/GC-MS), and Nuclear Magnetic Resonance (NMR). These methods, despite being able to give important information, are far less common than those used in the previous omics, and there is not a single one that allows the analysis of the whole metabolome.

2.4.4 Applications of Modeling in Human Cancer

At first, the reconstructed models are used to simulate cellular responses under certain conditions and compare the predictions to experimental data or expected behaviour from literature. This model evaluation helps to further refine the model and assess its accuracy.

As cancer cells are expected to have reduced TCA cycle activity and increase glycolysis, even in aerobic conditions, many authors start out by evaluating if the cancer models show reduced fluxes in the reactions that are part of the TCA cycle and increased fluxes in those part of glycolysis, in aerobic conditions [125–127, 132], as opposed to normal cell models [126, 127]. Other functions are tested to assess if the models simulate what has been experimentally found. Özcan E. et al [125] glioblastoma metabolic models showed active flux for the pyruvate dehydrogenase reaction, where glucose is metabolised through pyruvate dehydrogenase rather than pyruvate carboxylase in glioblastoma cells. A metastatic melanoma model [127], when compared to the primary melanoma model, showed increased fluxes in purine and pyrimidine synthesis and glutamate, arginine and proline metabolism, while reaction fluxes in coenzyme-A metabolism, fatty acid synthesis and OXPHOS decreased.

As for T-cells, Puniya et al [138] showed that T_{H1}, T_{H2} and T_{H17} models had more flux through fatty acid biosynthesis and less flux through fatty acid β oxidation than the naïve CD4 T-cell model. Limiting glucose from the environment resulted in decreased growth rate in all the 4 models. However, there was not a significant effect in growth rate when glutamine was removed from the medium of T_{H1}, T_{H2} and T_{H17} models, and growth rate of naïve CD4 T-cell model was more dependent on glucose and glutamine uptake than the other models. Still, gene deletion analysis revealed that more than 70% of the gene essentiality predictions agreed with previous experimental tests.

There are several methodologies that make use of a GSMM to gain insights on the capabilities of a cell's metabolism. This could be of great value in better understanding the underlying metabolic mechanisms of cancer and the tumour micro-environment.

Metabolite Biomarkers A challenge in cancer diagnosis is the identification of metabolite biomarkers that are present in biofluids such as plasma, urine and feces. This allows measurement of biomarkers in a non-invasive, cost-effective way for early diagnosis and monitoring treatment efficiency [147]. These biomarkers can distinguish not only between healthy and disease cases, but also between clinical groups such as subtypes of cancer. The study by Nam et al [148] is an example of potential biomarkers that were found using metabolic models. While succinate and fumarate were shown to be biomarkers for gastric cancer, as seen in literature, due to loss of function in sub-unit B of succinate dehydrogenase (SDHB), palmitate, D-glucose, and adenosine are potential biomarkers of leukemia due to fatty acid synthase's (FASN) loss of function.

Pathway Analysis Pathway analysis allows to understand different important aspects of the metabolic network. The more a reaction appears in different EFMs, the more important it is for the structure of the network [149]; when different sets of connected reactions lead to the same output from the same input,

the network has pathway redundancy. A high degree of redundancy might translate into the network being better at tolerating knockouts or drug inhibition than one of low redundancy, thus showing high robustness [91]. Minimal cut sets (MCSs) [90] represent a set of potential drug targets to prevent the functioning of an objective reaction.

Drug Targets As mentioned before, the metabolism of cancerous cells is modified during tumour development to meet requirements of fast cellular proliferation or due to genetic alterations. There might be certain reactions whose enzymes are essential for a tumour cell's viability but not for healthy cells. These enzymes can become interesting drug-targets. Also, as metabolism is evolutionarily more conserved than other biological processes, it is less probable for cancer cells to evolve resistance to these drugs by developing alternative pathways [150]. While the discovery of new drugs presents itself as a challenging task that requires a very long period of research and development before any new compound can be commercialised, exploiting the properties of already available drugs, whose information about the therapeutic and toxicity effects are already known, frequently becomes the focus when using metabolic models [121, 127, 131, 133, 151].

The identification of drug targets comes from studying the essentiality of reactions and metabolites. Ghaffari et al [131] tested cancer cell line models for metabolite essentiality and, after testing the essential metabolites in normal tissue models, 85 metabolites were found to be essential only in the cancer models. One of these metabolites was L-carnitine. Knowing that perhexiline malate salt inhibits CPT1 (responsible for the translocation of conjugated L-carnitine and long chain fatty acids from cytosol to the mitochondria), and partly CPT2, the authors treated two different cancer cell lines with this compound. Significant reduction in both cell lines viability was observed.

Yizhak *et al* [132] predicted 17 metabolic enzymes as targets to mitigate cancer cell migration, as their individual knockout from the models reduced the glycolytic to oxidative ATP flux ratio, positively associated with this event. Indeed, most have been found to have significantly higher expression levels in metastatic breast cancer patients than those non-metastatic and lower expression of 9 of those enzymes are associated wit improved long-term survival.

Immune cell models were used to find potential drug targets for T cells in rheumatoid arthritis, multiple sclerosis or primary biliary cholangitis [138], and to study metabolic changes during infection of macrophages by a pathogen [141]. However, none were applied in the context of cancer.

3

A Colorectal Cancer Atlas of scRNAseq Data

The type of cancer focused throughout this work was colorectal cancer (CRC). Also known as colorectal adenocarcinoma, this cancer usually emerges from the glandular, epithelial cells of the colon or rectum [152, 153]. Colon or rectal cancers are often merged due to the many biological and clinical common features [152].

Colorectal cancer is often asymptomatic until it substantially grows and spreads, hindering the prognosis and subsequent survival. In fact, the 5-year survival rate is close to 90% when it is diagnosed at an early stage, but of only 13% when the diagnosis is delayed. The limited number of tests that can be used for timely and efficient screening or diagnosis also affects the diagnosis, with up to 90% of the cases being diagnosed after symptoms onset [152].

According to the statistics provided by the International Agency for Research and Cancer (IARC) of the World Health Organization (WHO) in 2020 [154], colorectal cancer is the third most frequent malignant disease around the world, comprising 10% of total malignancies. The number of deaths in 2020 for colorectal cancer was approximately 935 000, representing 9.4% of cancer-related deaths, only preceded by lung cancer.

As CRC is a very heterogenous malignancy with different pathological and genetic signatures, it is very difficult to find one single molecular therapy to treat this type of cancer. In fact, surgery remains the primary course of treatment for early diagnosis cases, while in advanced cases cytotoxic therapies are met with rapid evolution of drug resistance and cancer recurrence [153]. However, therapeutic stratification based on the pathological and gene signatures may ultimately lead to improved therapeutic outcomes. The most widely used stratification approach is the Consensus Molecular Subtypes (CMS) of CRC [155].

Briefly, there are 4 CMS subtypes and an additional mixed phenotype when a clear CMS subtype cannot be assigned [156]. CMS1 is characterised by high microsatellite instability (MSI), CpG island methylator phenotype (CIMP+), hypermutation, frequent mutations of BRAF gene, and immune infiltration and activation. CMS2 is known for high somatic copy number alteration (SCNA) and activation of the WNT and MYC signalling pathways. CMS3, in turn, has frequent KRAS mutations and is characterised by metabolic deregulation. Finally, CMS4 is characterised by high SCNA, stromal infiltration, TGF-β activation, and angiogenesis.

To construct a wide number of T-cell genome-scale metabolic models that characterised different subtypes of T-cells, we wanted to make use of scRNAseq data, as this type of data allows quantitative characterisation of each cell's transcriptome in a sample. This lets us construct models of different T-cell types from the same tumour micro-environment. For this, we first constructed a colorectal cancer (CRC) atlas of scRNAseq data from different studies.

The following tools were used to read, process, analyse, and save scRNAseq data: R packages *Seurat* (v.4.0.3) [157], *SeuratDisk* (v.0.0.0.9019) [158], *SeuratObject* (v.4.0.2) [159], *clustree* (v. 0.4.3) [160], *SIGMA* (v.0.0.0.1) [161], *infercnv* (v.1.8.0) [162], *CMScaller* (v.0.99.2) [163]. The scripts were run with the R version 4.1.0. More detailed information, including scripts, is available in the GitHub project `https://github.com/saracardoso/CRC_ATLAS`.

3.1 Datasets Collected

Raw counts from four publicly available datasets, named *CRC_Qian* [164], *GSE132465* [165], *GSE144735* [165] and *Colon_smillie* [166], were used to construct the atlas. Three correspond to CRC studies, while one (*Colon_smillie*) has data from colon of patients with ulcerative colitis and healthy individuals. As such, the samples related to the healthy individuals were used. Table 1 shows an overview of the datasets collected.

Table 1: Overview of the datasets used to construct the CRC atlas

	CRC_Qian	***GSE132465***	***GSE144735***	***Colon_smillie***
Nº Patients	7	23	6	12
Nº Samples	21	33	18	48
Nº Cells (before QC)	44 684	63 689	27 414	51 705
Nº Cells (after QC)	29 793	57 804	24 510	50 811
Technology	3' 10xGenomics	3' 10xGenomics	3' 10xGenomics	3' 10xGenomics

Studies *CRC_Qian* and *GSE144735* have tumour (border and core) and normal matched mucosa samples for each patient. For study *GSE132465*, 10 of the 23 patients have tumour and normal matched mucosa samples, while the other 13 only have tumour samples. Regarding *Colon_smillie* dataset, each donor has a sample extracted from two different locations of the colon.

3.2 Quality Control

Before merging the datasets together, they were individually checked for quality control. This allowed us to filter low quality cells and lowly expressed genes. Cells were kept if they followed all the following criteria: *(1)* At least 1000 UMIs in each cell; *(2)* Between, including, 250 and 6 000 of detected genes (i.e., genes with non-zero UMIs) in each cell; *(3)* A complexity metric ($\frac{\log_{10} Number\_Genes}{\log_{10} Number\_UMIs}$) of at least 0.8; *(4)* Percentage

of mitochondrial RNA not greater than 20%. Genes were considered lowly expressed, and thus removed, if they were not detected in 10 or more cells. This quality control was performed using the R language. At this point in the pipeline, datasets were already loaded into a *Seurat* object and stored as a *h5Seurat* format file, using the R packages *Seurat* (v.4.0.3) [157] and *SeuratDisk* (v.0.0.0.9019) [158].

3.3 Dataset Integration

The different datasets were integrated using *Seurat* (R package, v.4.0.3) [157], so that proper cell-type annotation could be performed.

First, datasets were processed. Each dataset was normalised individually by dividing the counts from each cell by the total of the respective cell and multiplied by the scale factor 10000, followed by natural logarithmic transformation. Then, the top 2000 variables were calculated for each dataset and integration features calculated. Each dataset was then scaled with regression of the difference between the S and G2M phases (so that we could still distinguish proliferative cells from those not proliferating) and the percentage of mitochondrial RNA. PCA was finally performed to reduce dimensionality of datasets, in order to perform a faster and less computationally heavy integration.

After this, the integration anchors were found and the data integrated.

3.4 Cell Annotation

We started by separating the cells into 5 big groups of cell-types: Epithelial, Stromal, Myeloid, B-cells and T-cells. After this, we further annotated the cells for each of these groups individually, so that we could obtain more detailed and better annotations.

To perform the initial separation of the cells into the mentioned groups, we used *Seurat* (R package, v.4.0.3) [157] to find the clusters at a resolution of 0.1. After inspecting the quality of the clustering by mapping into UMAP plot metrics, such as number of UMIs and genes per cell, cell-cycle scores for the S and G2M phases, and percentage of mitochondrial RNA, we explored the expression of gene markers to annotate the cells (figure 5). The following genes were used: *EPCAM* (epithelial cells); *S100B*, *COL1A1*, *VWF* and *ENG* (stromal cells); *CD14*, *FCGR3A*, *FCER1A*, *GZMB*, *TPSAB1* and *TPSB2* (myeloid cells); *CD79A* and *MZB1* (B-cells); and *CD3D* (T-cells).

For each group of cell-types, we found the clusters using a set of 10 different resolutions, ranging from 0.1 to 1 at a step of 0.1. After assessing the quality of the clustering, we found the resolution that best separates the cells into clusters with interesting biological information using *clustree* (R package, v.0.4.3) [160] and the expression of genes that were markers for cell-types that we wanted to find. The gene expression was evaluated by visualizing their expression mapped into UMAP and violin plots. Regarding T- and epithelial cells, more steps were performed to be able to fully annotate the cells correctly. These steps are mentioned next in their respective sub-sections.

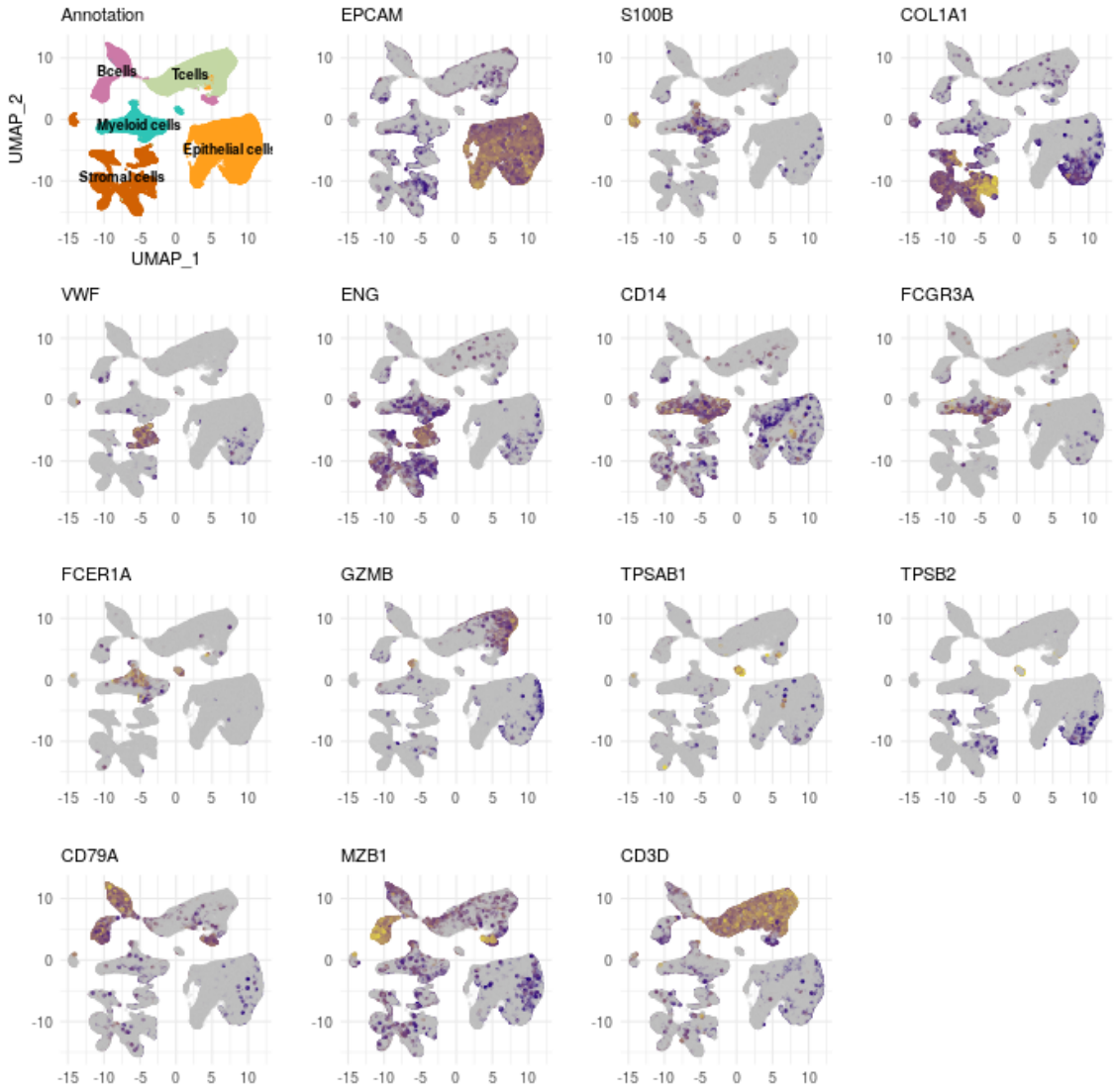


Figure 5: Annotation of the big groups of cells. Top-left UMAP is coloured by the annotated cell-types. All other UMAPs are coloured by the RNA expression of the genes used to identify the cells, which goes from blue (low expression) to yellow (high expression). Grey cells have no expression of the respective gene.

3.4.1 Stromal cells

Regarding stromal cells, the following genes were used to annotate the cell-types (figure 6): *VWF*, *PLVAP*, *CDH5* (endothelial cells); *RGCC*, *RAMP3* (tip-like vascular endothelial cells); *ACKR1*, *SELP* (stalk-like vascular endothelial cells); *LYVE1*, *PROX1* (lymphatic endothelial cells); *S100B, PLP1* (enteric glia); *SYNPO2*, *CNN1*, *PDGFRB* (vascular smooth muscle-cells); *RGS5*, *ABCC9*, *KCNJ8* (pericytes); *TAGLN*, *ACTA2*, *ACTG2*, *MYH11*, *MYLK*, *DES* (myofibroblasts); *COL1A1*, *COL1A2*, *COL6A1*, *COL6A2*, *COL3A1*, *DCN* (fibroblasts); *THY1*, *FAP* (cancer associated fibroblasts).

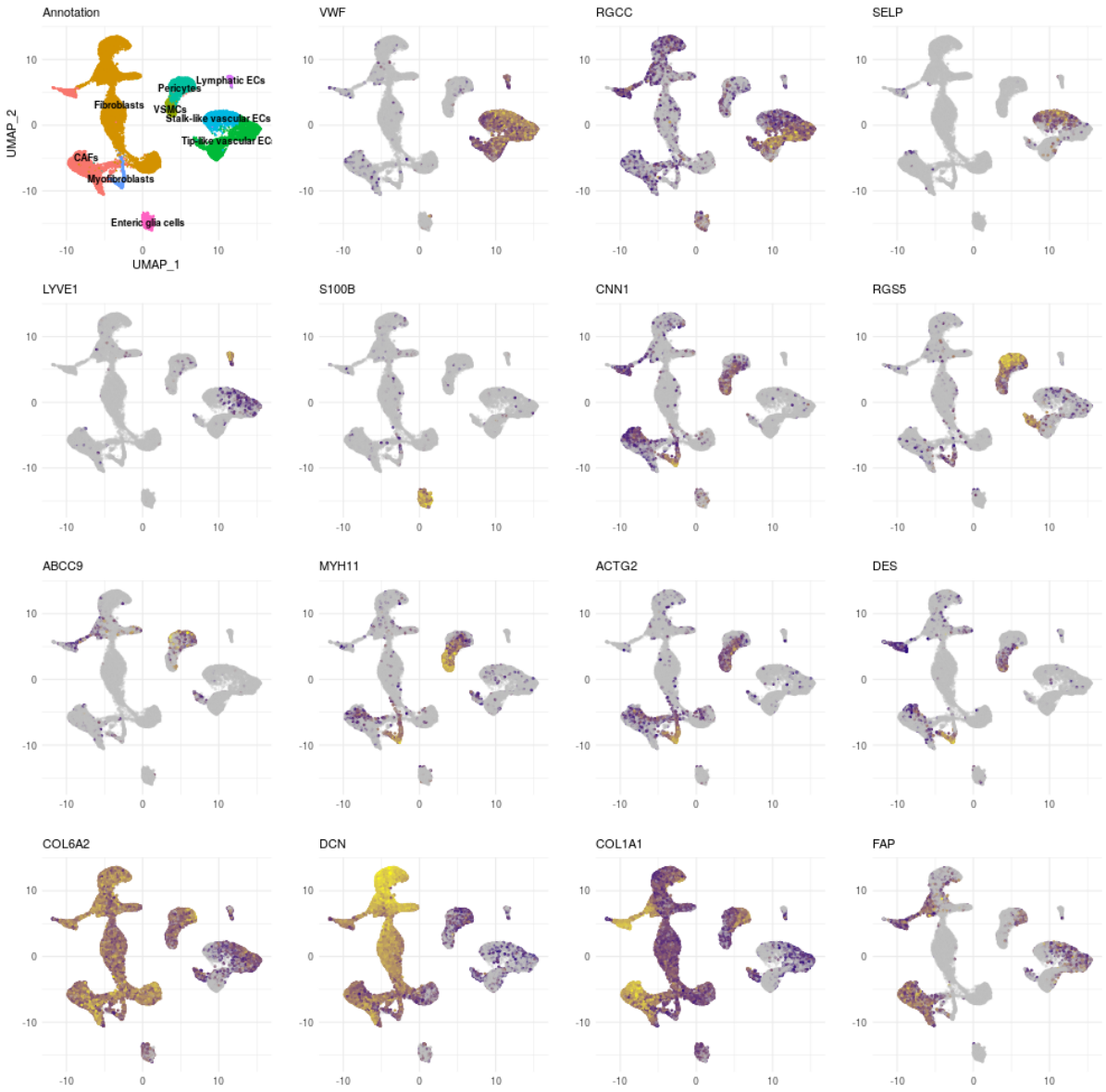


Figure 6: Annotation of the stromal cells. Top-left UMAP is coloured by the annotated cell-types. All other UMAPs are coloured by the RNA expression of some of the genes used to identify the cells, which goes from blue (low expression) to yellow (high expression). Grey cells have no expression of the respective gene.

3.4.2 Myeloid cells

Regarding myeloid cells, the following genes were used to annotate the cell-types (figure 7): *CD14*, *FCGR3A*, *MARCO*, *ITGAM* (monocyte/ macrophage lineage); *FCER1A*, *CST3* (conventional dendritic cells); *IL3RA*, *SERPINF1*, *GZMB*, *ITM2C* (plasmacytoid dendritic cells); *KIT*, *TPSB2*, *TPSAB1* (mast cells). In our data, we were not able to separate monocytes from macrophages. However, we further separated the monocyte/ macrophage lineage into further groups: *IL1B*, *IL6*, *S100A8*, *S100A9* (pro-inflammatory macro/mono); *CD163*, *SEPP1*, *APOE*, *MAF* (anti-inflammatory macro/mono); *SPP1* (SPP1+ macro/mono);

FCN1 (FCN1+ macro/mono).

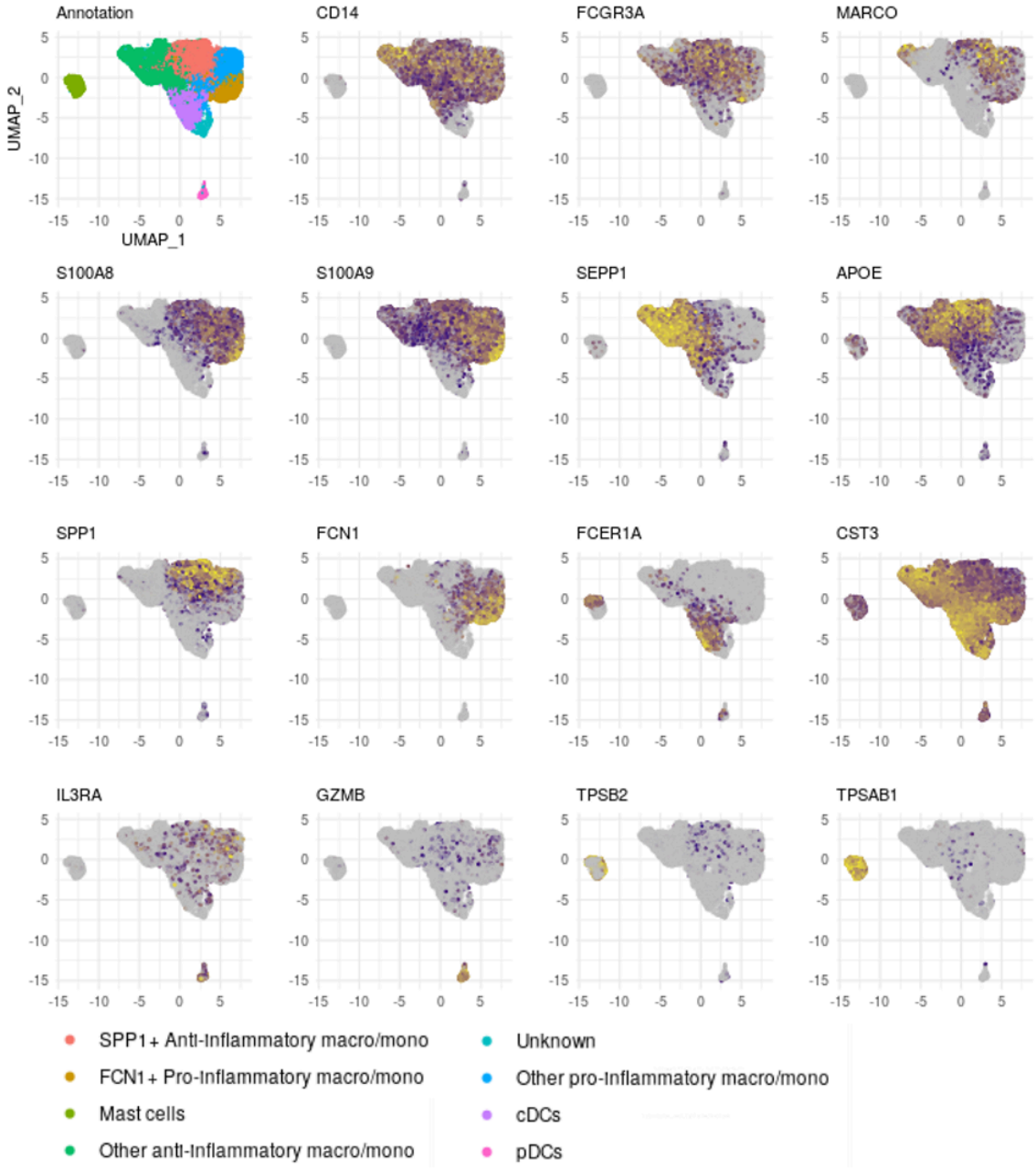


Figure 7: Annotation of the myeloid cells. Top-left UMAP is coloured by the annotated cell-types. All other UMAPs are coloured by the RNA expression of some of the genes used to identify the cells, which goes from blue (low expression) to yellow (high expression). Grey cells have no expression of the respective gene.

3.4.3 B-cells

For B-cells, the following genes were used to annotate the cell-types (figure 8): *IGHD*, *MS4A1*, *CXCR4*, *NR4A2* (naïve B-cells); *CD27*, *MS4A1*, *CXCR4*, *NR4A2* (memory B-cells); *STMN1*, *ACTB*, *RGS13*, *MKI67*,

PCNA, *MARCKSL1*, *HMGN1*, *HMGN2* (proliferative cells); *MZB1*, *CD27* (plasma cells); *IGHA1*, *IGHA2* (IgA+); *IGHM* (IgM+); *IGHG1*, *IGHG2*, *IGHG3*, *IGHG4*, *IGHGP* (IgG+).

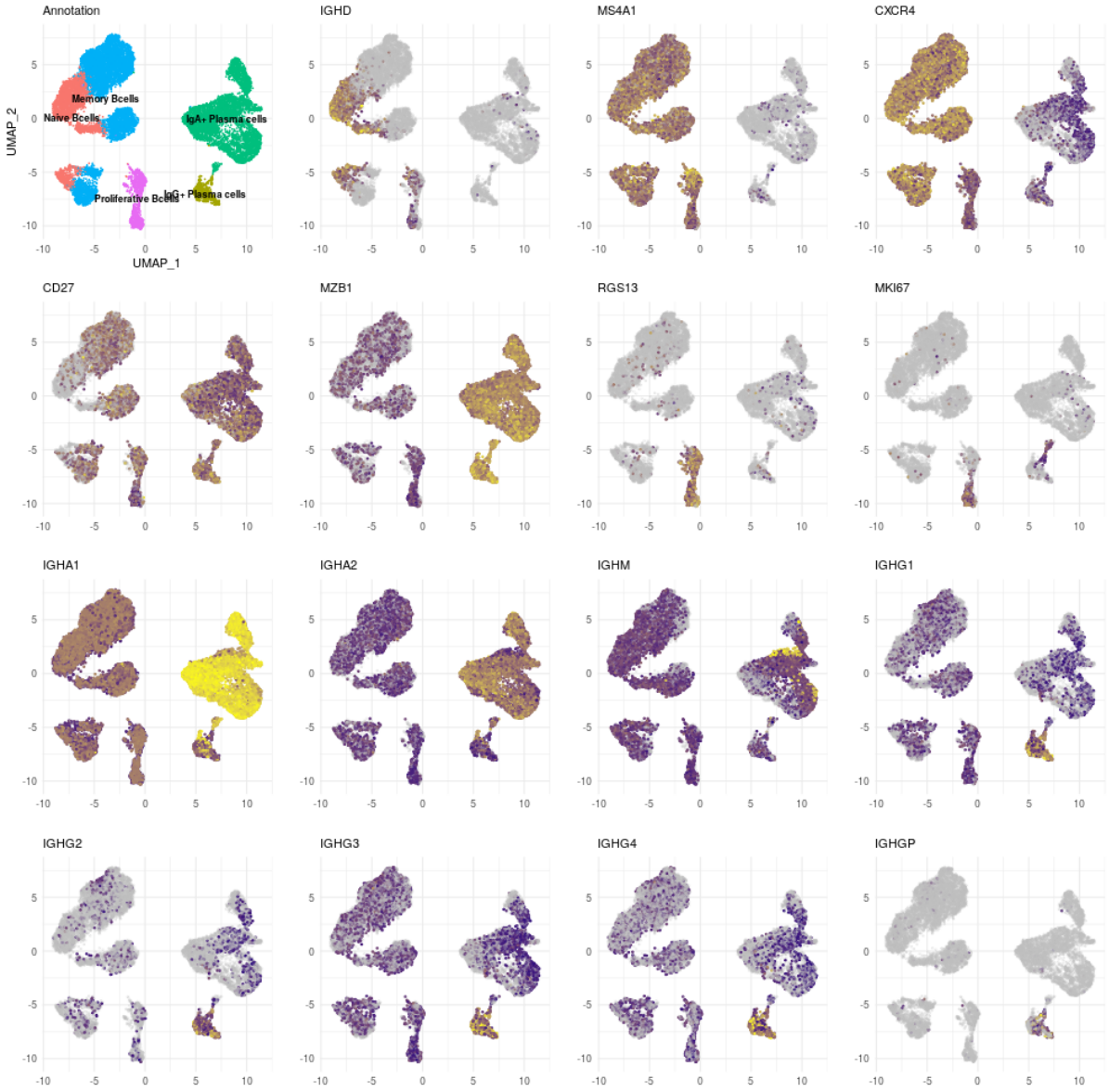


Figure 8: Annotation of the B-cells. Top-left UMAP is coloured by the annotated cell-types. All other UMAPs are coloured by the RNA expression of some of the genes used to identify the cells, which goes from blue (low expression) to yellow (high expression). Grey cells have no expression of the respective gene.

3.4.4 T-cells

To annotate the group of T-cells, we first clustered them with a low resolution (0.2) where we were able to obtain 8 clusters: ***(0+7)*** CD4 T-cells; ***(1)*** CD8 T-cells and more; ***(2)*** regulatory T-cells; ***(3)*** $\gamma\delta$T-cells + NK-like cells + other ILCs; ***(4)*** CD8+ and CD4+ T-cells expressing *CXCL13*; ***(5)*** CD4 T-cells mostly expressing *IL17*; and ***(6)*** proliferative T-cells. These annotations were obtained based on the following

genes (figure 9): *CD3E*, *CD3D*, *CD3G* (T-cells); *TRAC*, *TRBC1*, *TRBC2* ($\alpha\beta$T-cells); *CD8A* (CD8+ T-cells); *CD4* (CD4+ T-cells); *TRGC1*, *TRGC2*, *TRDC* ($\gamma\delta$T-cells); *KLRB1* (NK-like); *CXCL13* (CXCL13+ cells); *IL2RA* (regulatory T-cells); *IL17A* (IL17+ cells); and *MKI67*, *PCNA* (proliferative cells).

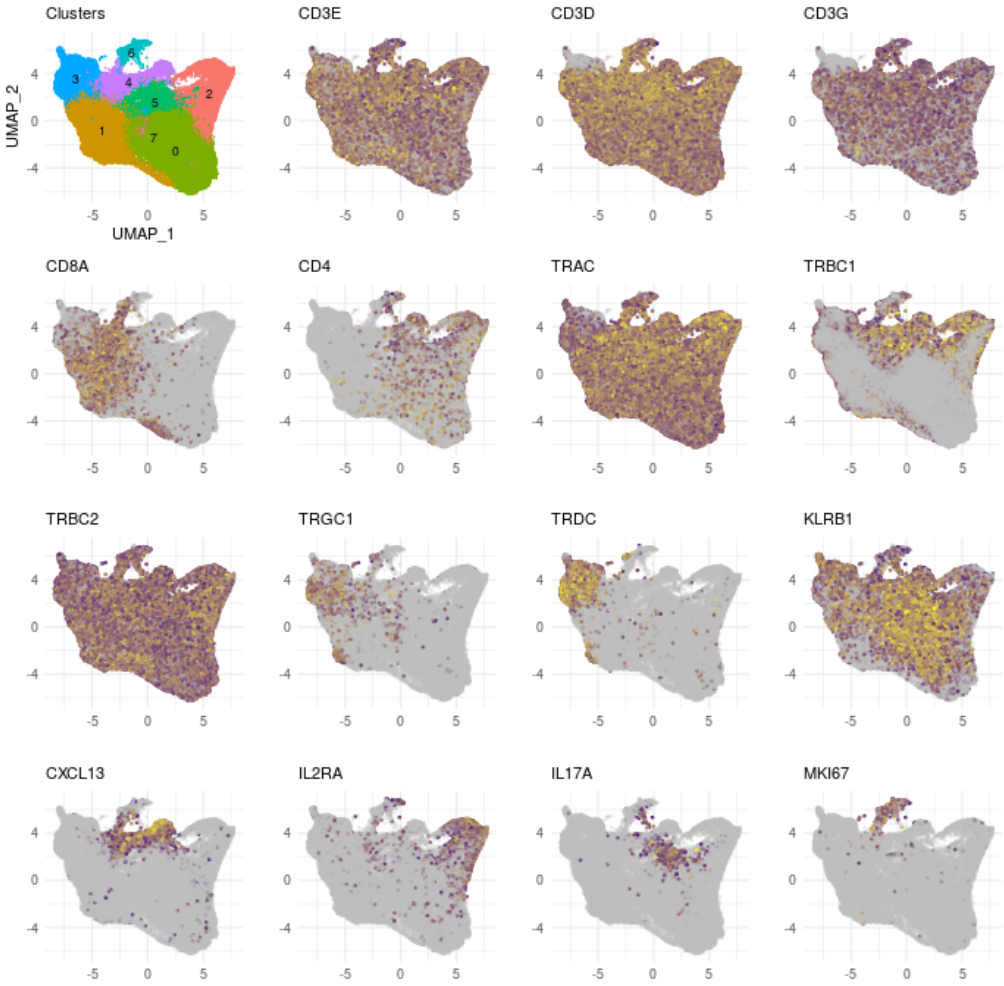


Figure 9: Gene expression over the group of cells identified as T-cells. Top-left UMAP is coloured by the clusters obtained under resolution 0.2. All other UMAPs are coloured by the RNA expression of the respective genes, which goes from blue (low expression) to yellow (high expression). Grey cells have no expression of the respective gene.

As we wanted to have more granularity in the annotation of T-cells and increasing the resolution was not originating satisfactory results, we used *SIGMA* (R package, v.0.0.0.1) to evaluate the clusterability of each cluster from resolution 0.2, and if that was due to meaningful variability (for example, we do not want to end up clustering cells according to dataset of origin).

Having a good clusterability among all clusters except cluster 7 (figure 10B), we decided to merge clusters 0 and 7, as they represented the same cell-type(s) (figures 9 and 10). We then separately sub-clustered each of the 7 clusters for further annotation, after visually checking that the clusterability was not due to dataset of origin (supplementary figure 51).

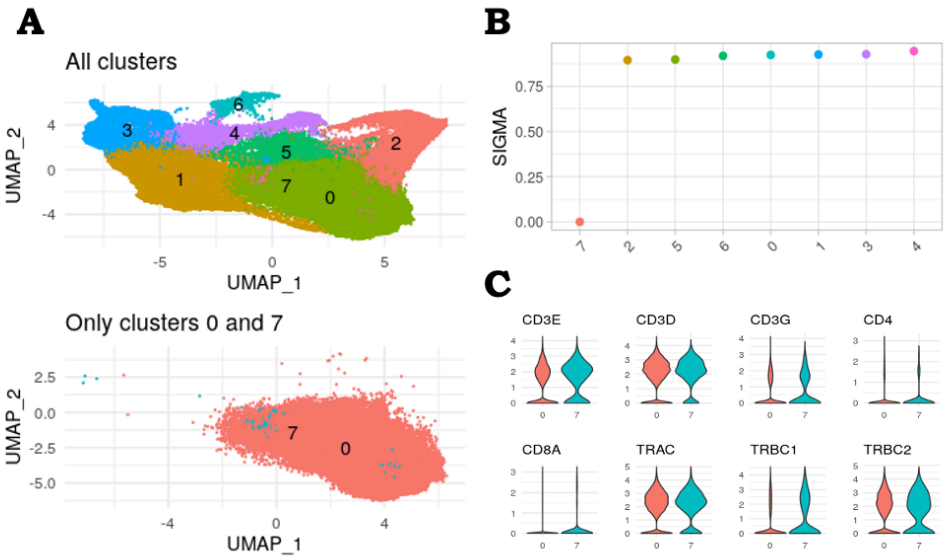


Figure 10: (A) Same UMAP, one (top) with all Tcells coloured using resolution 0.2, and the other (bottom) with only the clusters 0 and 7. (B) Clusters' SIGMA clusterability. This metric goes from 0 to 1. The closer to 1, the more clusterable. (C) Violin plots of the gene expression distribution in clusters 0 and 7.

When finding the sub-clusters for cluster ***(0+7)***, we assessed if the cells were being separated according to the original clusters 0 and 7 (figure 11C), which did not happen. Thus, we could further trust that these two clusters could be merged and evaluated together for further sub-clustering.

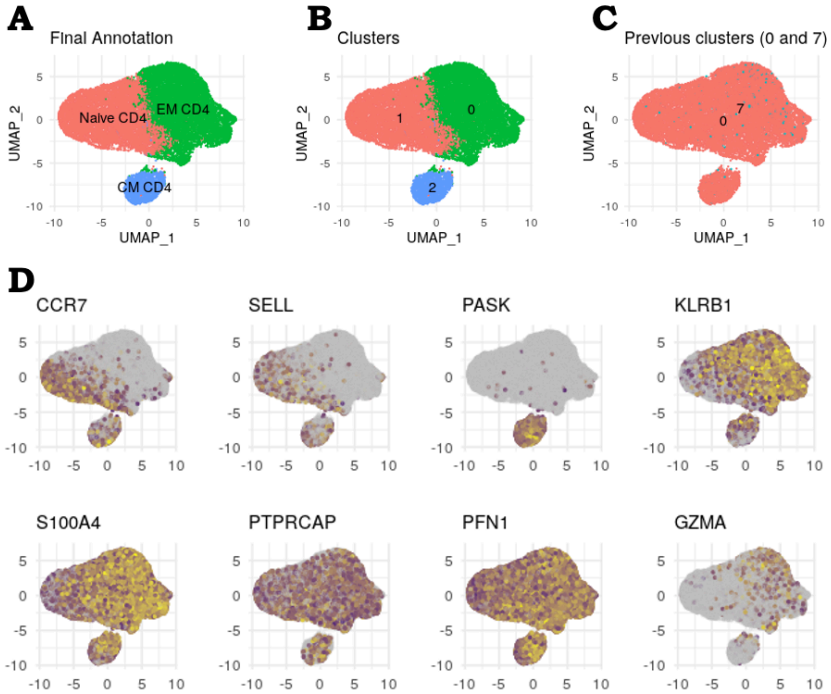


Figure 11: Cluster *(0+7)*. UMAPS with (A) Final annotation, (B) clusters before the final annotation, (C) orignal clusters, and (D) expression of relevant genes (expression goes from blue (low) to yellow (high), while grey cells is the absence of expression).

For cluster ***(0+7)***, as observed in figure 11, we were able to annotate naïve (*CCR7*, *SELL*), central memory (*CCR7*, *SELL*, *PASK*) and effector memory (*KLRB1*, *S100A4*, *PTPRCAP*, *PFN1*, *GZMA*) cells. The cluster of regulatory T-cells ***(2)*** was not further sub-clustered, as we did not found any interesting biological information between the sub-clusters.

The cluster constituted by mostly CD4 T-cells expressing IL17 ***(5)*** was separated into IL17A+, IL17A+ IL17F+, IL17F+, IL17A+ IL22+ and IL22+ CD4 T-cells, and effector memory CD4 T-cells (figure 12A), as they expressed effector memory related genes (*KLRB1*, *S100A4*, *PTPRCAP*, *PFN1*, *GZMA*) but not IL17 or IL22 genes (figure 12C), or any other markers that could indicate these could be another type of CD4 T-cells (data not shown).

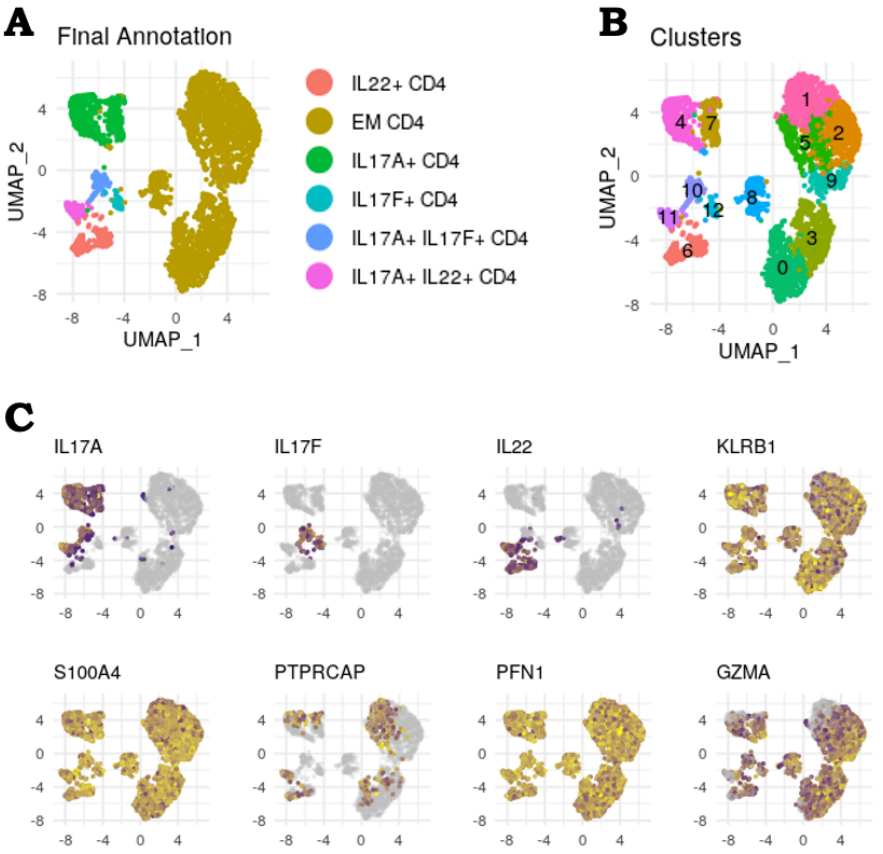


Figure 12: Cluster *(5)*. UMAPS with (A) Final annotation, (B) clusters before the final annotation, and (C) expression of relevant genes (expression goes from blue (low) to yellow (high), while grey cells is the absence of expression).

In the cluster with mostly CD8 T-cells ***(1)*** (figure 13), we found naïve CD8 T-cells (*CCR7*, *SELL*); tissue resident memory CD8 T-cells (*ZNF683*); effector memory CD8 T-cells (*KLRB1*, *S100A4*, *PTPRCAP*, *PFN1*, *IL7R*, *ANKRD28*); memory CD8 T-cells expressing *CD160*, cytotoxic CD8 T-cells (*NKG7*, *GZMA*, *GZMB*, *GZMH*, *GZMK*, *PRF1*, *IFNG*); double-negative T-cells (expressing $\alpha\beta$TCR and *CD3* genes, but not *CD8* or *CD4* genes); CD8$\alpha\alpha$ IELs (expressing $\alpha\beta$TCR genes and *CD8A*, but no *CD8B*); and a cluster with an NK-like signature (*KLRD1*, *KLRB1*, *XCL1*, *XCL2*) that also expressed some T-cell markers, which was

annotated as NKT cells.

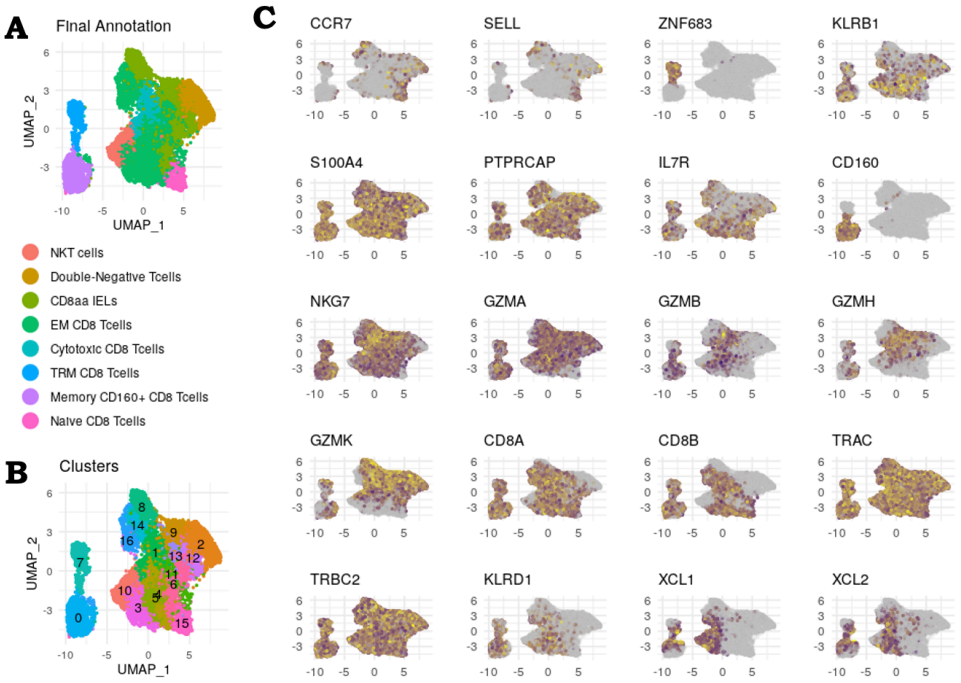


Figure 13: Cluster *(1)*. UMAPS with (A) Final annotation, (B) clusters before the final annotation, and (C) expression of relevant genes (expression goes from blue (low) to yellow (high), while grey cells is the absence of expression).

In the cluster with cells expressing *CXCL13* ***(4)*** (figure 14A), we found follicular (CXCL13+) CD4 T-cells, CXCL13+ CD8 T-cells, and tissue resident memory CD8 T-cells (expressing *ZNF683* but no *CXCL13*). The cluster of proliferative cells ***(6)*** was divided into CD4 and CD8 proliferative T-cells (figure 14B).

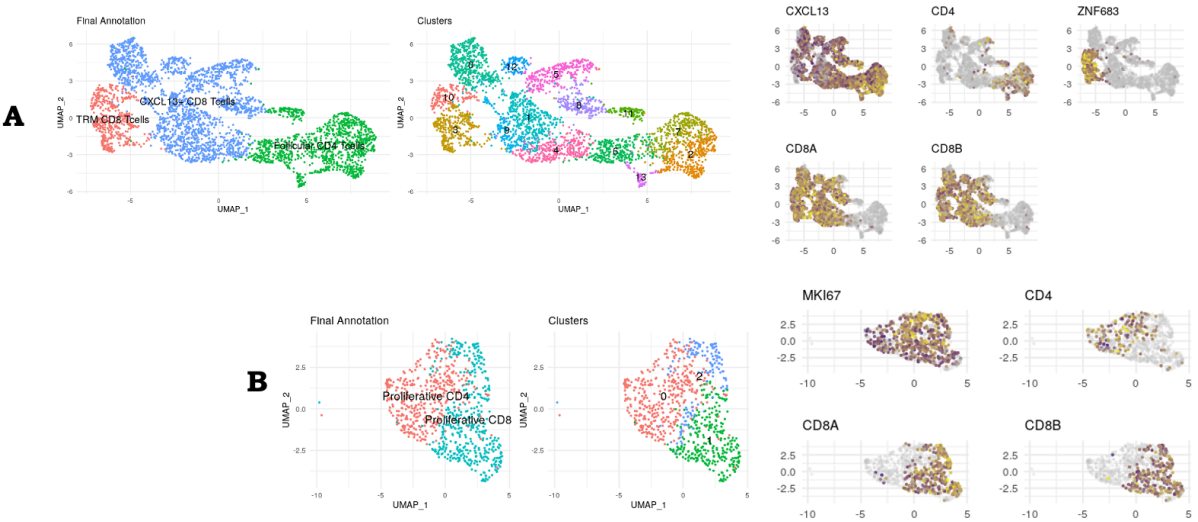


Figure 14: UMAPS with final annotation, clusters before the final annotation, and expression of relevant genes (expression goes from blue (low) to yellow (high), while grey cells is the absence of expression). (A) Cluster *(4)*; (B) Cluster *(6)*.

Finally, the cluster with $\gamma\delta$T-cells and other non-conventional T-cells ***(3)*** (figure 15) contained $\gamma\delta$T-cells (*TRDC*, *TRGC1*, *TRGC2*), NK cells (*KLRD1*, *KLRB1*, *XCL1*, *XCL2*), lymphoid tissue-inducer cells (*RORC*, *LTA*, *LTB*), NKT cells (NK-like signature genes and some T-cell markers). In this cluster, a sub-cluster of cells with high expression of heat-shock proteins (*HSPA6*, *HSPA1B*, *HSPA1A*, *HSPB1*, *HSP90AA1*, *HSPD1*, *HSPH1*) was removed from the dataset.

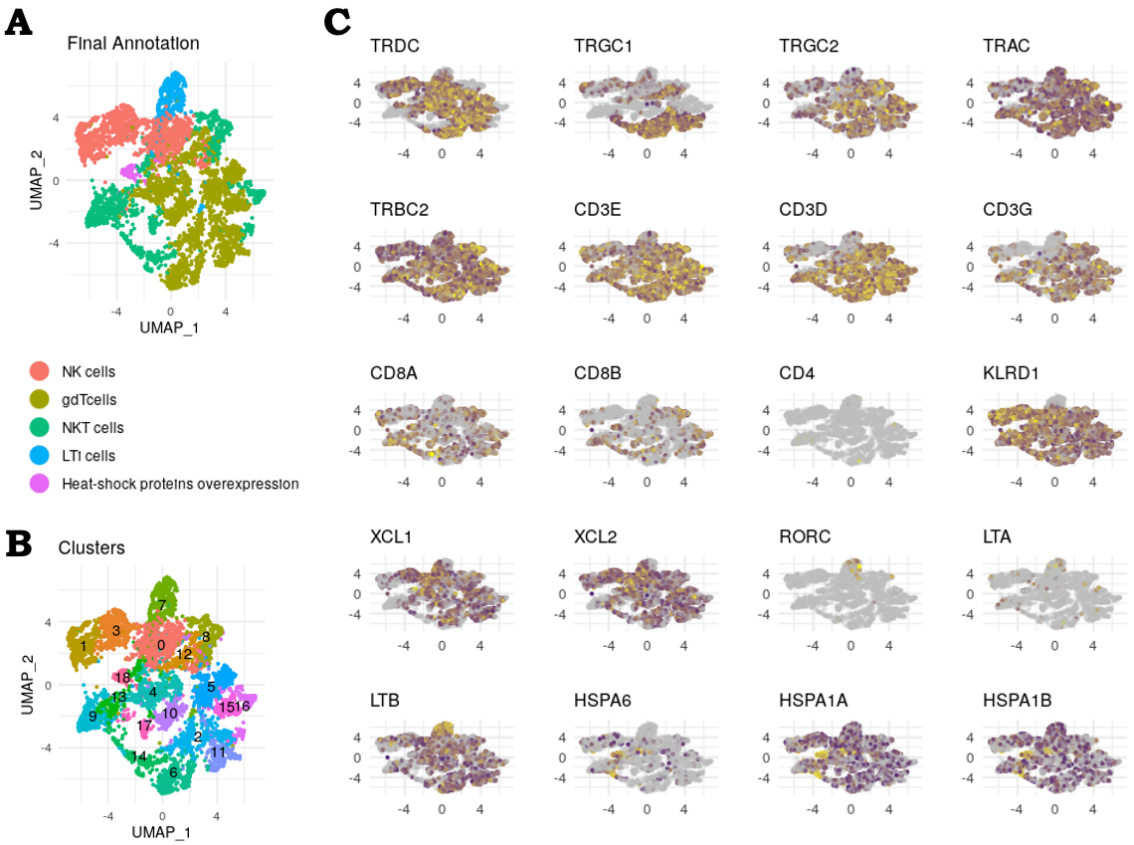


Figure 15: Cluster *(3)*. UMAPS with (A) Final annotation, (B) clusters before the final annotation, and (C) expression of relevant genes (expression goes from blue (low) to yellow (high), while grey cells is the absence of expression).

3.4.5 Epithelial cells

First, we wanted to separate tumour from normal cells in tumour samples. For that, we used the R package *infercnv* (v.1.8.0) [162] to identify somatic large-scale chromosomal copy number variations (CNVs) by comparing the gene expression of the epithelial cells from tumour samples with the gene expression of the epithelial cells from normal matched samples. After processing the raw counts data accordingly, we used the six-state HMM-based CNV prediction method (i6 HMM) to predict, for each gene in each cell, the CNV level. There are 6 different levels: *(1)* complete loss; *(2)* loss of one copy; *(3)* neutral (i.e., no change); *(4)* addition of one copy; *(5)* addition of two copies; *(6)* addition of more than two copies. From these predictions, we considered the existence of a chromosome arm loss if more than a third of the genes from that chromosome arm were predicted to have levels 1 or 2. If more than a third of a

chromosome arm's genes were predicted with levels 4 or greater, we considered to have a gain in that chromosome arm. Finally, an epithelial cell was considered a tumour cell if it had at least one chromosome arm alteration (gain or loss), while those with no large-scale alterations were classified as possibly normal. CNV predictions for some of the patients can be observed in the heatmap figures 53 to 55.

For those tumour samples having too many epithelial cells classified as normal (figure 55 is an example of that), we further analysed the expression of genes known to be differentially expressed in CRC, to account for those cases where tumour do not have copy number variation. These genes are: *MYC*, *RNF43*, *AXIN2*, *CTNNB1*, *CD44*, *MLH1* [167–172]. While *MLH1* is under-expressed in tumours, all others are over-expressed, which happens in our atlas (figure 16). For each of those samples, we compared the expression distribution of these genes between normal, putative normal and putative tumour cells. The genes that had similar distribution between putative normal and putative tumour and different distribution between normal and putative normal were used to re-annotate the putative normal cells. For over expressed genes, the cells were re-annotated as tumour if the gene expression was higher than the median gene expression from normal cells. For the under expressed gene, the cells were re-annotated as tumour if the gene expression was smaller than the median gene expression from normal cells.

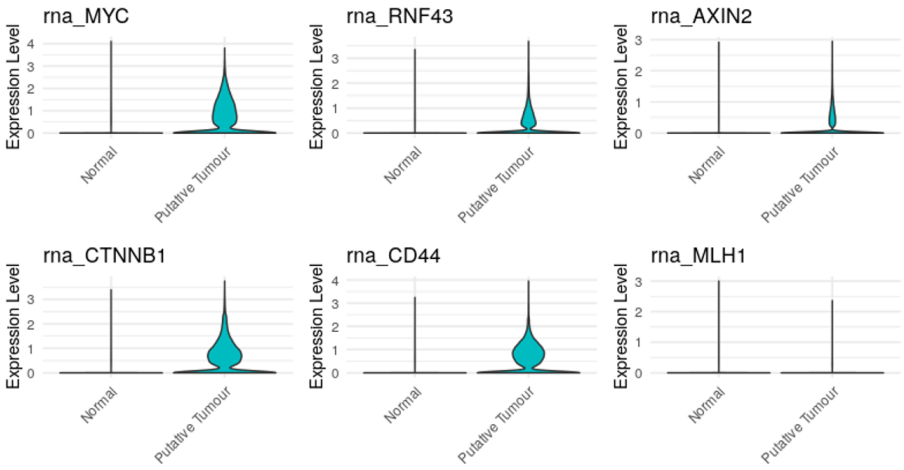


Figure 16: Distribution of the expression of genes *MYC*, *RNF43*, *AXIN2*, *CTNNB1*, *CD44*, *MLH1* in normal epithelial cells vs cells classified as tumour after CNV predictions.

The following genes were used to annotate the group of normal epithelial cells (figure 17): *LGR5*, *SMOC2*, *ASCL2* (stem cells); *SOX9*, *CDK6*, *MUC4*, *FABP5*, *PLA2G2A*, *LCN2* (progenitor cells); *MUC2*, *ITLN1*, *CLCA1* (secretory progenitors); *TOP2A*, *CCNA2*, *MCM5*, *OLFM4*, *SLC12A2* (transit-amplifying cells); *POU2F3*, *TRPM5*, *SPIB*, *IL17RB*, *HTR3E* (tuft cells); *CHGA*, *CHGB*, *CPE*, *NEUROD1*, *PYY* (enteroendocrine cells); *MUC2*, *CLCA* (goblet cells); *FABP1*, *SLC26A3*, *TMEM37*, *BEST4* (enterocyte/colonocyte cells); *LYZ*, *CA7*, *CA4*, *SPIB*, *FKBP1A* (paneth-like cells).

Tumour cells were annotated according to the consensus molecular subtypes (CMS), using the R package *CMScaller* (v.0.99.2) [163]. Gene expression of cells was aggregated by sample before prediction. As such, a cell was annotated with the CMS type that was predicted to the respective sample. If the tool

was not able to confidently classify a sample, that sample was classified with a *Mixed* type.

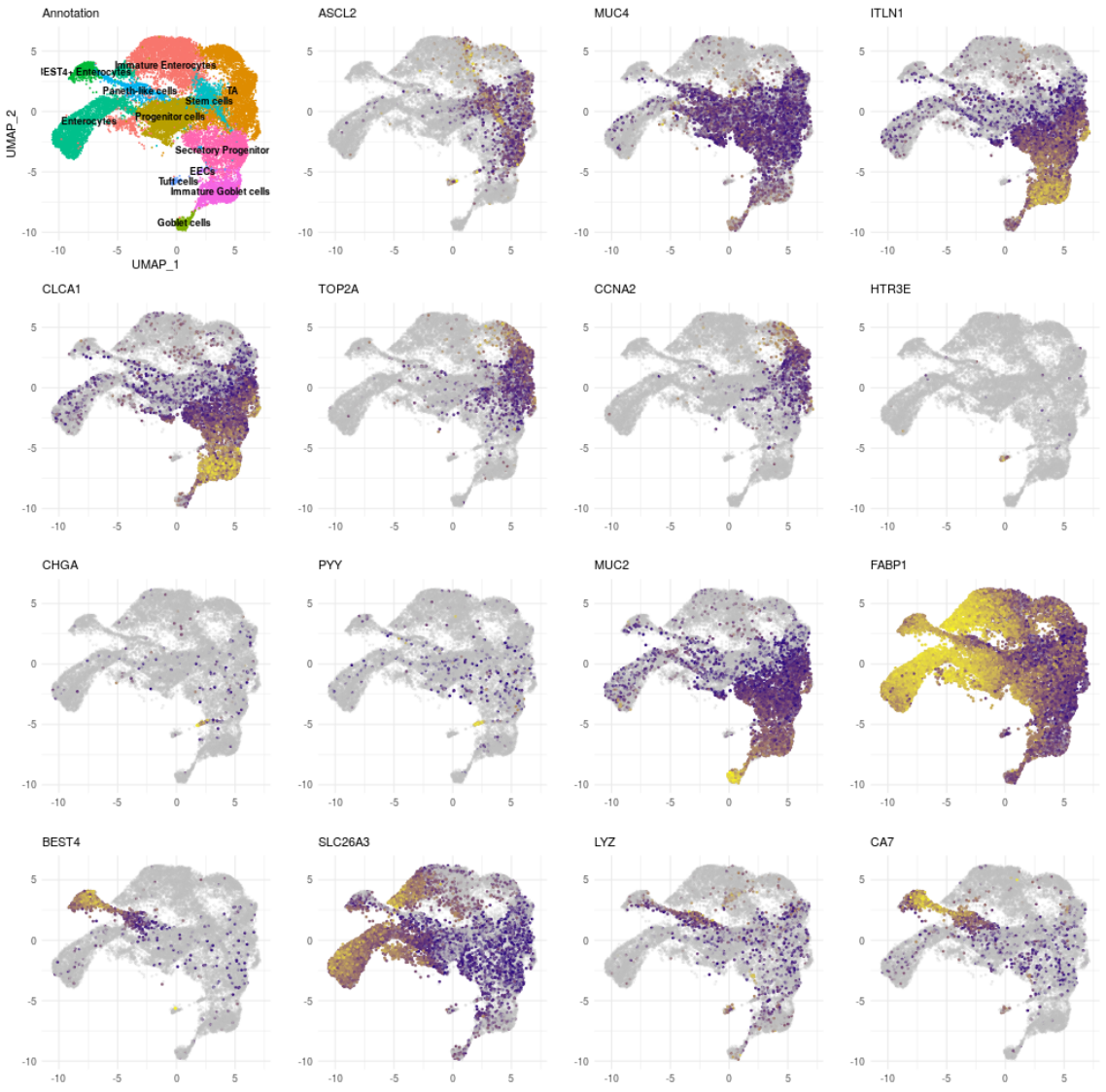


Figure 17: Annotation of the normal epithelial cells. Top-left UMAP is coloured by the annotated cell-types. All other UMAPs are coloured by the RNA expression of some of the genes used to identify the cells, which goes from blue (low expression) to yellow (high expression). Grey cells have no expression of the respective gene.

3.5 Overview of the atlas

Our atlas has a total of 163 810 cells, separated into 51 044 T-cells, 47 462 epithelial cells, 30 187 stromal cells, 17 674 B-cells, and 17 443 myeloid cells. The number of cells for each cell-subtype identified is summarised in table 2.

Table 2: Number of cells in each cell-subtype present in the CRC atlas. *TA*: transit-amplifying cells; *CAFs*: Cancer-associated fibroblasts; *VSMCs*: Vascular smooth muscle cells; *DCs*: Dendritic cells; *LTi*: Lymphoid tissue-inducer cells.

Cancer cells			28 208	
Normal Epithelial cells	Progenitor cells		2 818	19 254
	Secretory progenitors		2 094	
	TA cells		3 862	
	Tuft cells		191	
	Enteroendocrine cells		74	
	Goblet cells	Not immature	391	
		Immature	1 505	
	Enterocyte/colonocyte cells	Immature	3 926	
		BEST4+	536	
		Other	3 201	
	Paneth-like cells		656	
Stromal cells	Fibroblasts		14 007	30 187
	CAFS		4 914	
	Myofibroblasts		693	
	Pericytes		2 014	
	Enteric glia cells		1 649	
	VSMCs		791	
	Endothelial cells	tip-like vascular	3 934	
		stalk-like vascular	1 792	
		lymphatic	393	
Myeloid cells	Anti-inflammatory macro/mono lineage	Other	4 963	17 443
		SPP1+	4 180	
	Pro-inflammatory macro/mono lineage	Other	2 548	
		FCN1+	1 639	
	conventional DCs		2 072	
	plasmacytoid DCs		335	
	Mast cells		1 222	
	Unkown		484	
B-cells	Naive		2 872	17 674
	Memory		7 323	
	Plasma cells	IgA+	5 624	
		IgG+	685	
	Proliferative		1 170	
T-cells	CD4+	Poliferative	377	51 044
		Naive	8 484	
		Memory	13 484	
		Regulatory	6 801	
		Follicular	1 156	
		IL22+	201	
		IL17+	755	
	CD8+	Poliferative	439	
		Naive	414	
		Memory	6 534	
		CXCL13+	1 977	
		Cytotoxic	1 105	
	Unconventional	$\gamma\delta$	2 317	
		NKT	1 759	
		NK	1 382	

Table 2 (cont.)

		CD8$\alpha\alpha$	2 488
		LTi cells	327
	Double-Negative		1 044
TOTAL	163 810		

The consensus molecular subtype (CMS) of colorectal cancer most present in our atlas is CMS2, with 22 samples, followed by CMS1 (10), CMS4 (8) and CMS3 (4). Two samples were not confidently classified with any CMS and were thus labeled with a Mixed type. Gene set enrichment analysis (GSEA) with the assigned groups confirms the CMS classification (figure 18A). The group of samples with a CMS2 type are characterised by a high MSS/MSI ratio and activated MYC and E2F target gene sets. The CMS1 group, as expected, is MSI-like, but also revealed overexpression of genes related with MTORC1 signaling and E2F targets. CMS3 samples are MSS-like and have up-regulation of metabolic processes, namely fatty acid metabolism and oxidative phosphorilation. CMS4, in turn, have the characteristic strong activation of epithelial mesenchymal transition (EMT), TGFβ and angiogenesis. GSEA was not performed with the Mixed samples.

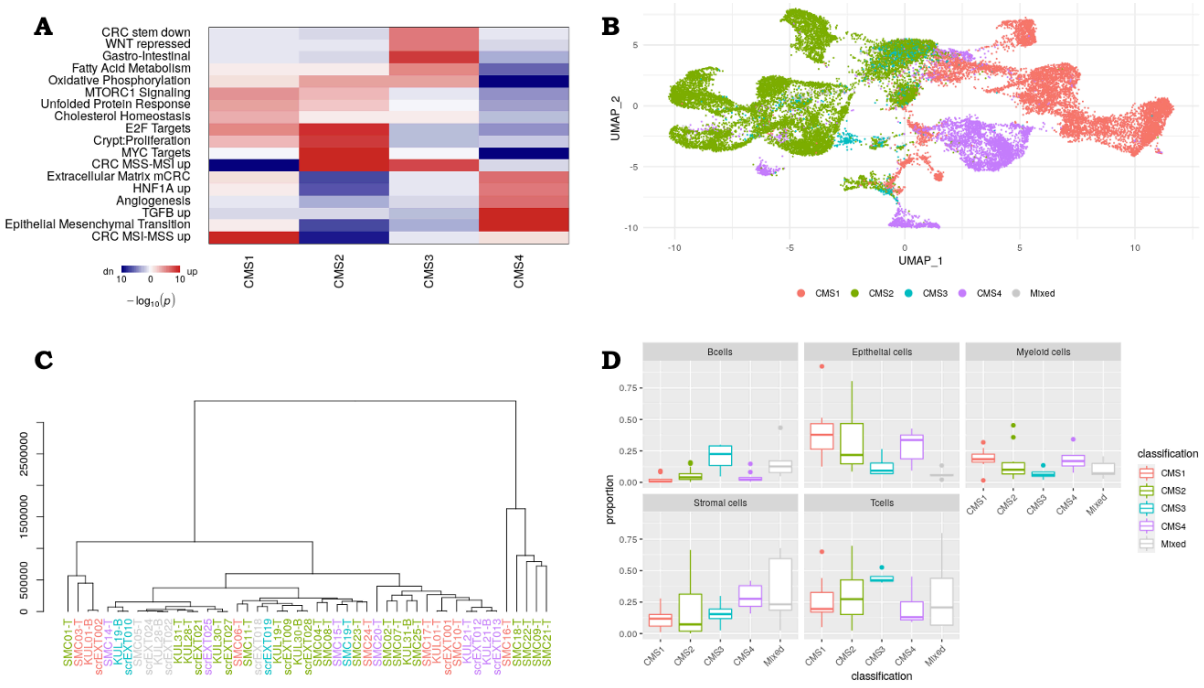


Figure 18: (A) Heatmap of the results of gene set enrichment analysis using the CMS predictions. Blue and red mean under- and over- representation, respectively, of the gene set. (B) UMAP visualization of tumour epithelial cells, coloured by CMS type. (C) Hierarchical cluster of samples, coloured by CMS type. (D) Distribution of the proportions of B-, epithelial, myeloid, stromal and T- cells by CMS type.

The UMAP visualization of the tumour cells coloured by the respective sample's CMS (figure 18B) shows that the cells from different samples but same CMS tend to group together. There is a clear separation of the CMS1, CMS2 and CMS4 cells, while CMS3 seems to overlap with the other CMS types. In fact, when performing hierarchical clustering of the samples (by aggregating the genes counts across

all cells of each sample, figure 18C), CMS3 samples clustered closer to the Mixed type. The remaining types tended to group with other samples of the same CMS type.

We further assessed the proportion of the major cell-types across each CMS type (figure 18D). CMS4 showed a higher proportion in stromal cells than other types, as expected, but we did not observe a higher immune infiltration in CMS1 samples than other CMS types. CMS3 samples showed the highest proportions of B- and T- cells, although it should be kept in mind that only 4 samples were classified as CMS3.

The construction of this CRC atlas of scRNAseq data with cells spanning various T-cell subtypes from several different colorectal patients with tumour and normal matched samples allowed us to construct T-cell metabolic models that characterised not only the T-cells present in the tumour micro-environment, but also those in an unaffected part of the colon (or rectum) of the patient.

4

Modeling T-cells from the Colorectal Cancer Environment

This chapter discusses the construction of a wide number of genome-scale metabolic models for different subtypes of T-cells from the micro-environment of colorectal cancer (CRC) and normal matched colon. We made use of the scRNAseq atlas for colorectal cancer (CRC) constructed in the previous chapter.

A total of 196 models were constructed. Even though the structure of these models showed a lot of common pathways across cell-types and tissues of origin, the models for regulatory T-cells showed interesting differences between tissue of origin, for example. The prediction of models' fluxes was very close to expected results regarding biomass and energy production. Cell-type prediction showed good results not only on datasets of gene expression but also on datasets of models' structure (reaction presence) and of flux predictions. Models were further tested for gene essentiality and for the effect of different media conditions.

All code was run with the R version 4.1.0 or the python version 3.6.9. More detailed information, including scripts, is available in the GitHub project `https://github.com/saracardoso/Metabolic_Models`.

4.1 Methods

Starting from a generic human metabolic model, we reconstructed cell-type specific models with the aid of the scRNAseq data from the colorectal cancer (CRC) atlas constructed in the previous chapter. Not all samples from this atlas were used tough. We only used CRC patients with samples from both tumour and normal matched tissues, after filtering the samples with less than 1000 cells. Within the total of 14 patients with normal matched tissue samples, 7 had samples from tumour border and core. The cell-types considered for the construction of the models were: naïve CD8, memory CD8, proliferative CD8, cytotoxic CD8, naïve CD4, memory CD4, proliferative CD4, IL17+ CD4, follicular CD4, and regulatory CD4 T-cells.

4.1.1 Generic human model

The generic human model used to reconstruct context-specific models was Human-GEM [173], v.1.8.0, retrieved from the GitHub repository `https://github.com/SysBioChalmers/Human-GEM` in June 2021 in the SBML format. Human-GEM is a consensus genome-scale metabolic model (GSMM) created by integration of the preceding models HMR2.0 and Recon3D, and curation of different databases. This model contains a total of 13 802 reactions, 8 378 metabolites and 3 625 genes. Reactions span 9 compartments: extracellular, cytosol, inner mitochondria, mitochondria, endoplasmic reticulum, Golgi apparatus, lysosome, peroxisome, and nucleus.

Before model reconstruction, we ensured consistency of the generic model chosen by identifying and removing blocked reactions. These are reactions whose maximum and minimum fluxes are null with open medium exchanges when performing flux variability analysis (FVA), meaning that they are not able to carry flux on any condition.

Furthermore, we tested the consistent model for its capability to perform metabolic tasks [174] known to occur in human cells, as well as for biomass production under a specific medium (see section 4.1.3 for a detailed information on the medium used).

The python module *COBRApy* [175] was used to handle the SBML file of the Human-GEM model, as well as to process this model into a consistent one. The evaluation of the consistent model regarding metabolic tasks and biomass production was performed through the python module *troppo* [176].

4.1.2 Model Reconstruction

To construct cell-type specific models from the human generic model, we used the fastCORE algorithm [177], which was available through the python module *troppo* [176]. fastCORE takes a set of reactions that have strong evidence to be active in the context of interest and searches for a subnetwork that contains all reactions from the core set and a minimal set of additional reactions to allow flux consistency through FVA. The choice of this algorithm was based on Vieira *et al* [178], where different combinations of data processing and reconstruction algorithms were tested for the reconstruction of tissue-specific models.

The models returned by fastCORE were further gap-filled to ensure that flux through the biomass reaction when running the models with a specific medium (see section 4.1.3 for a detailed information on the medium used) is possible. To do so, the gap-filling was performed assuming no limitation (i.e., high upper bounds) of the metabolites that compose the medium. All models whose biomass flux was still null were excluded from the analysis. As such, in downstream analysis, whenever a model is predicted with null biomass, it can only be due to specific conditions (e.g., concentration of metabolites in medium, inhibited internal reactions, etc.).

From expression data to core reaction set To construct the models using scRNAseq data, the raw gene counts were aggregated across cells of the same cell-type, generating a pseudo-bulk expression data for each cell-type in each sample. A cell-type in a sample was only considered for aggregation if more

than 5 cells representing that cell-type were available. A table summarising the cell-types represented in each scRNAseq sample and respective number of cells is present in the supplementary table 13. After aggregation, a gene expression matrix was created for each sample, where the rows corresponded to the genes and the columns to the cell-types considered for that sample. After this, the gene expression matrices were normalised into CPM counts. This was performed with the aid of the R packages Seurat [157] and SeuratDisk [158].

To get the set of core reactions for each cell-type model reconstruction, we used a similar approach to that of Richelle *et al* [179]. A gene is active in a sample's cell-type if its expression is greater than a global maximum threshold, defined by the 75^{th} percentile of the expression distribution of all genes in all cell-types of that sample. On the other hand, a gene is inactive in a sample's cell-type if its expression is smaller than a global minimum threshold, defined by the 10^{th} percentile of the expression distribution of all genes in all cell-types of that sample. Genes whose expression falls between these two thresholds, are considered active, or inactive, based on a local threshold. This local threshold is defined by 25^{th} percentile of the expression distribution of that gene in all cell-types of that sample. The percentiles were defined based on the best combinations found by Vieira *et al* [178] for the reconstruction of metabolic models.

To achieve this, and for further downstream analysis, we first transformed the normalised pseudo-bulk data into gene activity scores (GASs), calculated using the above thresholds (table 3).

Table 3: Calculation of the gene activity scores (GAS) from a sample. x_g: expression, in CPMs, of gene g; $global\_max$: 75th percentile of the expression distribution of all genes in all cell-types of a sample; $global\_min$: 10th percentile of the expression distribution of all genes in all cell-types of a sample; $local\_threshold$: 35th percentile of the expression distribution of a gene in all cell-types of a sample.

Condition	**State**	**GAS**
$x_g \geq global\_max$	Active	$\log_2 \frac{x_g}{global\_max}$
$x_g \leq global\_min$	Inactive	$\log_2 \frac{x_g}{global\_min}$
$global\_min \leq x_g \geq global\_max$	Moderate (Active or Inactive)	$\log_2 \frac{x_g}{local\_threshold}$

To obtain the set of core reactions to give as input to the reconstruction algorithm, we then need to translate the GASs into reaction activity scores (RASs). To do this, the gene-protein-reaction (GPR) rules present in the generic human model are used. These rules describe which combinations of genes are involved in the production of the enzyme(s) that catalyse the reactions. If two genes are necessary together (e.g., enzyme complex) for a reaction to occur, its GPR rule will be *Gene A AND Gene B*. On the other hand, if only one of two genes is necessary for a reaction (e.g., enzyme isoforms), the GPR rule is *Gene A OR Gene B*. To translate these GPR rules into the continuous RASs, the *AND* operators are replaced with a *minimum* function, assuming that the enzyme/reaction is limited by the lowest expressed gene. The *OR* operators, on the other hand, are replaced by a *maximum* function, assuming that the reaction activity is the activity of the highest expressed gene. A few examples on how to get the RASs from the GASs are present in table 4.

Table 4: Examples for calculating RASs. x_a, x_b, x_c: expression, in CPMs, of genes a, b and c, respectively; max: maximum; min: minimum.

GPR	**Replacement**	**RAS** (if $x_a < x_b < x_c$)
$x_a\ OR\ x_b$	$max(x_a,\ x_b)$	x_b
$x_a\ AND\ x_b$	$min(x_a,\ x_b)$	x_a
$(x_a\ AND\ x_b)\ OR\ x_c$	$max(min(x_a,\ x_b),\ x_c)$	x_c
$(x_a\ OR\ x_b)\ AND\ x_c$	$min(max(x_a,\ x_b),\ x_c)$	x_b

Finally, we only need to obtain the set of core reactions that will be used in the reconstruction model. By the way that the GASs and RASs were calculated (base 2 logarithm of a ratio), active reactions are simply those reactions whose RAS is positive.

All these calculations were made using the python language.

4.1.3 Media used in the experiments

To obtain flux predictions as close as possible to the *in vivo* reality of the T-cells, we sought to create a medium that best represented the metabolites in healthy human blood, instead of using a cell culture medium. We also wanted to recreate the metabolic medium in the tumour micro-environment, as it would be even closer to reality. However, there isn't enough information on the concentration of metabolites in this situation, when compared to that found for metabolites under normal conditions. Nevertheless, several studies focused on how much the presence of metabolites change between the blood of normal individuals and CRC patients were found.

Normal Medium The concentrations for the metabolites in the external compartment of our models were gathered from the Serum Metabolome Database (SMDB), which is integrated in the Human Metabolome Database (HMDB) [180]. For those model metabolites with more than one 'normal' concentration in the database, the final concentration was averaged. Metabolites with no information in the database regarding 'normal' concentrations were not included in the medium. Three metabolites (water, oxygen, and H^+) were considered as always available, and thus their bounds were opened (maximum uptake set to 1 000). A total of 537 metabolites composed the final *normal medium*.

Metabolic models work with reaction fluxes (mmol/gDW/h) instead of metabolite concentrations. Regarding the metabolites in a medium, they are represented by fluxes that characterise the entry rate of the metabolites in the cell. As such, the concentrations gathered from the database were transformed into fluxes in the following manner [181]:

$$Flux_{Ma} = \frac{[Ma]}{[cells]\ .\ cell_{weight}\ .\ time} \tag{4.1}$$

The flux of a metabolite Ma in a model of a specific cell-type is thus obtained by dividing its concentration in the medium ($[Ma]$, in mM) by the concentration of viable cells in the medium ($[cells]$, in n°cells/L) after an experiment that lasted a certain amount of time, dry weight of the cell ($cell_{weight}$, in gDW) and said time ($time$, in h).

All these values, apart from the metabolite concentrations, were not known. However, we followed Aurich M. *et al* [181] on how to calculate these values. Knowing that osteosarcoma (U2OS) cells have a cell dry weight of approximately 60 pg [182] and a cell volume of 4000 um3 [183], we can get the weight of T-cells knowing that their volume is approximately 176 um3 [184]. Thus, the calculated T-cell weight is, approximately, 2.640×10e-12 gDW. The optimal concentration of T-cells in a medium was considered 2.5×10e8 cells/L, assuming that the culture media is normally changed every two days (time = 48h) [185].

The supplementary table 11 summarises the metabolites present in the normal medium.

Tumour Medium We found several studies [186–191] that performed NMR- or MS- based metabolomics on blood of normal individuals and CRC patients and assessed how much the peak intensities of the respective metabolites changed between conditions. As such, we used the information available regarding the fold changes between these two conditions to modify the normal human blood medium into a tumour one.

We found information for 84 of the metabolites present in our normal human blood medium. Thus, we changed the concentrations of these metabolites using the fold change averaged across the studies, while maintaining the remaining metabolites' concentrations. A table with the information on the fold changes calculated by the different studies and respective average is present in supplementary table 12.

4.1.4 Flux prediction

To predict the reaction fluxes in our models, we resorted to the Parsimonious Flux Balance Analysis (pFBA) approach.

pFBA maximises an objective to be accomplished by the cell and, from the possible solutions that lead to that maximum, gives the solution that leads to the lowest overall flux through the metabolic network. The objective was set to maximise the biomass reaction for proliferative T-cells. Biomass production is not the main objective of non-proliferative T-cells, with the production of energy being as important. In line with other studies that constructed models for normal, non-proliferative, cells [129, 137, 148], the objective of the remaining T-cells was the maximisation of biomass and ATP production in equal weights.

It is in the lower and upper bounds restrictions of the exchange reactions that the medium fluxes come into play. The upper bound (i.e., maximum flux possible) of medium metabolites' exchange reactions are set to the fluxes calculated in section 4.1.3.

pFBA was run using the python module COBRApy [175].

4.1.5 Model Evaluation

After reconstructing the models, we carried a series of evaluations to assess how good the models represented the cell-types and to extract additional valuable information.

We first assessed the ability to distinguish the different cell-types using transcriptomics data versus the models reconstructed. We also analysed the fluxes of certain reactions and compared them between models, and assessed how the models were affected when specific metabolites were removed from the medium. We further tested the models for gene essentiality and compared the flux predictions between using a normal and tumour human blood medium.

Cell-type Prediction We compared the ability to predict the type of T-cells from the different types of data used, namely the pseudo-bulk RNAseq dataset, in CPMs, used to construct the respective models, the absent and present reactions of each reconstructed model, and the pFBA flux predictions using the normal or tumour human blood media. For the reaction presence and pFBA prediction datasets, only the reactions with GPRs were used.

For this, a random forest classifier with repeated 10-fold cross-validation was trained using the R package *caret* [192]. Each dataset used was divided into the same 70% samples used for training and 30% for testing. Prior to training, genes/reactions with zero variance across the training samples were removed.

The predictions obtained from the test samples were evaluated using the metric *Mathews correlation coefficient* (MCC). This metric is appropriate to multi-class problems, perfectly symmetric (no class is more important than the other), and not sensitive to class-imbalance (i.e., when the different classes are not evenly represented, which happens in these datasets with some T-cell subsets having far more models than others). The general *MCC* formula is the following:

$$MCC = \frac{TP.TN - FP.FN}{\sqrt{(TP + FP).(TP + FN).(TN + FP).(TN + FN)}} \tag{4.2}$$

MCC is calculated through the true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). This metric varies from -1 to 1. An *MCC* value 1 means that all cell-types were correctly classified, while a value of -1 means that all cell-types were not well classified. An *MCC* value of 0 means that the classifier is no better than random guessing.

Gene Essentiality As mentioned before, the objective was set to maximise the biomass reaction for proliferative T-cells, while both the biomass and ATP production were maximised in the other T-cell subsets. For this simulation, however, the objective of all models was set to maximise only biomass, as we wanted to assess which genes could be essential in the production of biomass.

After excluding house-keeping genes from our analysis, only 932 genes were potentially essential for our models. These were the only ones whose deletion led to the deletion of one or more reactions. For example, a reaction whose GPR rule is *Gene A OR Gene B* will not be deleted if only one of the genes is. On the other hand, a reaction with a GPR rule of *Gene A AND Gene B* will be deleted if one of the genes is.

The 932 genes were deleted individually, and the biomass maximised using the FBA approach (see section 2.4.1). A gene was considered essential for a model if its objective decreased to 0. Furthermore, a gene was considered essential for a cell-type if it was essential for more than 50% of the models of that cell-type.

Gene essentiality was carried using python and the module *COBRApy* [175], while the analysis of its results was carried in the R language.

4.2 The different representations of the transcriptomics dataset

We used the dimension reduction technique *Uniform Manifold Approximation and Projection* (UMAP) to have a visual representation of how well the different cell-types can be separated when using the different representations of the transcriptomics dataset used prior to model reconstruction. Those representations are the scRNAseq dataset of T-cells with all genes (figure 19A); the scRNAseq dataset of T-cells with only the metabolic genes present in the generic model Human-GEM (figure 19B); and the pseudo-bulk dataset created (figure 19C).

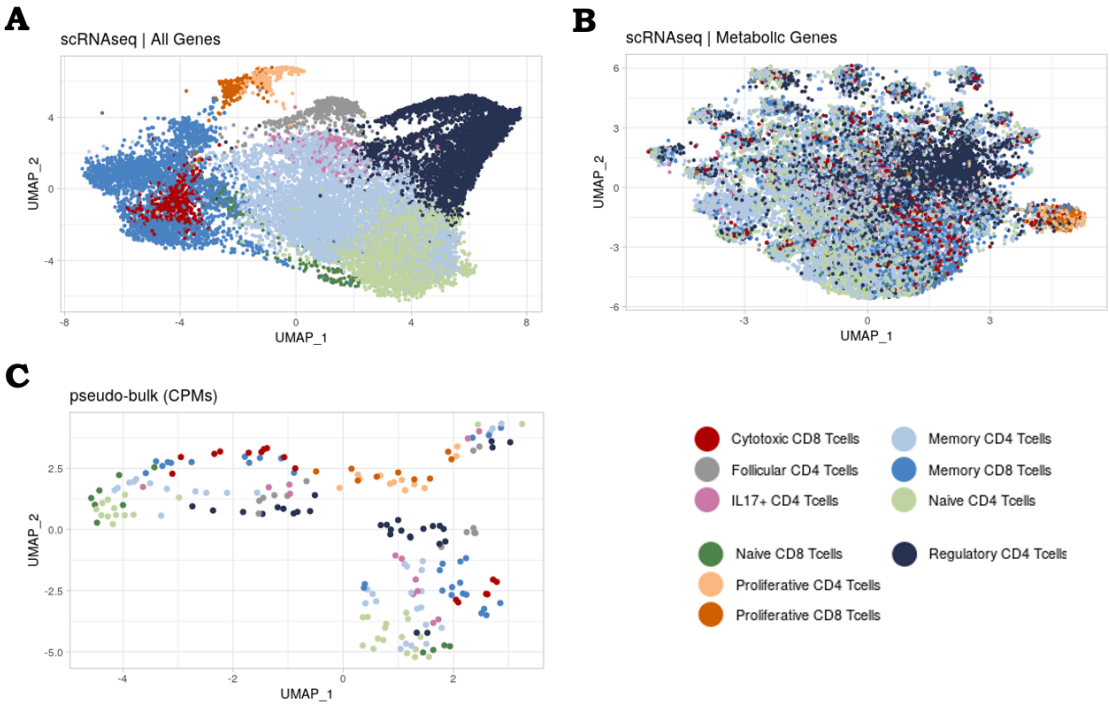


Figure 19: UMAP plots for the (A) scRNASeq data with all genes, (B) scRNAseq data with only the metabolic genes, and (C) pseudo-bulk RNAseq data. Each dot corresponds to a cell in case of the scRNAseq datasets, or a cell-type in a sample in case of the pseudo-bulk data. The plots were coloured according to the cell-types.

Using all genes to plot the T-cell subtypes from the CRC atlas that were modelled (figure 19A) shows

that most cells group together with others from the same cell-type. However, a few of the cell-types tend to overlap. That is the case of cytotoxic CD8 T-cells and memory CD8 T-cells, IL17+ CD4 T-cells and memory CD4 T-cells, and a small subset of naïve CD8 T-cells with memory CD8 T-cells.

Filtering the dataset for metabolic genes relevant for the model reconstruction (figure 19B) shows a marked overlap. All cell-types overlap, except for proliferative CD4 and CD8 T-cells, which group apart from the rest. This makes sense, as metabolically T-cells are very closely related, with the proliferative T-cells being the most distinct ones due to the known metabolic shifts all cells go through to acquire the ability to proliferate.

The pseudo-bulk data (figure 19C), however, no longer shows such a strong overlap. Proliferative CD4 and CD8 T-cells still group together and there is no distinct separation between the other cell-types.

4.3 Models Reconstructed

A total of 196 models were reconstructed (table 5 and supplementary table 13): 16 cytotoxic CD8 T-cells; 13 follicular CD4 T-cells; 13 IL17+ CD4 T-cells; 33 memory CD4 T-cells; 30 memory CD8 T-cells; 30 naïve CD4 T-cells; 11 naïve CD8 T-cells; 12 proliferative CD4 T-cells; 9 proliferative CD8 T-cells; and 29 regulatory CD4 T-cells.

Table 5: For each cell-type (*Cell Type)*, number of reconstructed models *(Number of Models)* and their distribution regarding tissue of origin *(State Distribution)* and CMS classification *(CMS Distribution)*.

Cell Type	Number of Models	State Distribution	CMS Distribution
Cytotoxic CD8	16	Tumour: 10 Normal Matched: 6	CMS1: 2 CMS2: 6 CMS3: 2 CMS4: 0 Mixed: 0
Follicular CD4	13	Tumour: 13 Normal Matched: 0	CMS1: 3 CMS2: 6 CMS3: 3 CMS4: 0 Mixed: 1
IL17+ CD4	13	Tumour: 8 Normal Matched: 5	CMS1: 2 CMS2: 4 CMS3: 1 CMS4: 0 Mixed: 1
Memory CD4	33	Tumour: 20 Normal Matched: 13	CMS1: 6 CMS2: 6 CMS3: 3 CMS4: 4 Mixed: 1
Memory CD8	30	Tumour: 18 Normal Matched: 12	CMS1: 6 CMS2: 6 CMS3: 3 CMS4: 2 Mixed: 1
Naive CD4	30	Tumour: 19 Normal Matched: 11	CMS1: 6 CMS2: 6 CMS3: 2 CMS4: 4 Mixed: 1
Naive CD8	11	Tumour: 6 Normal Matched: 5	CMS1: 0 CMS2: 4 CMS3: 2 CMS4: 0 Mixed: 0
Proliferative CD4	12	Tumour: 12 Normal Matched: 0	CMS1: 3 CMS2: 4 CMS3: 2 CMS4: 2 Mixed: 1
Proliferative CD8	9	Tumour: 9 Normal Matched: 0	CMS1: 3 CMS2: 3 CMS3: 2 CMS4: 0 Mixed: 1
Regulatory CD4	29	Tumour: 20 Normal Matched: 9	CMS1: 6 CMS2: 6 CMS3: 3 CMS4: 4 Mixed: 1

The cell-types follicular and proliferative CD4 T-cells, and proliferative CD8 T-cells do not have models from normal matched tissue.

Regarding the number of reactions that are present in the different models (figure 20A), there is a visible difference in the number of reactions present between the models from normal matched tissue and those from tumour tissue for the cell-types cytotoxic CD8 T-cells, IL17+ CD4 T-cells and regulatory CD4 T-cells. On the other hand, there is not much difference for the models from memory and naïve cell-types.

The models from naïve CD8 T-cells, normal cytotoxic CD8 T-cells, normal IL17+ T-cells, and normal regulatory CD4 T-cells have the least number of reactions. Excluding normal regulatory CD4 T-cells, all these models have less than 5 000 reactions.

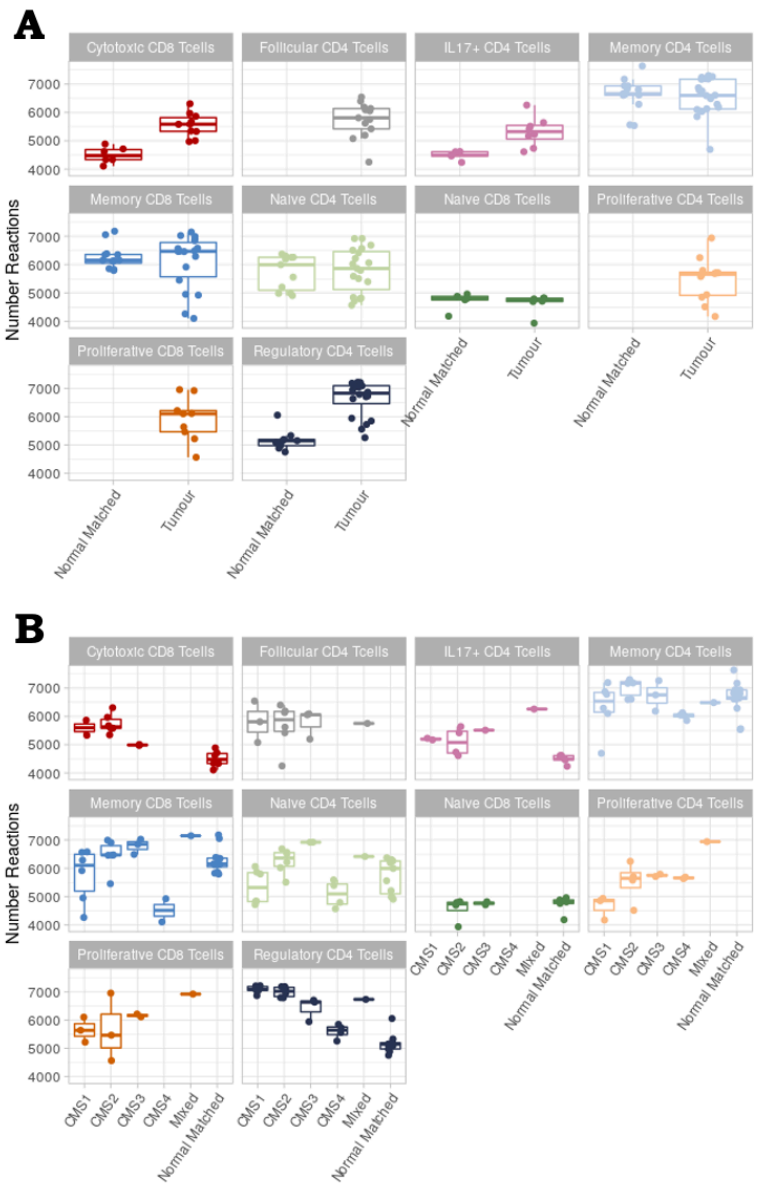


Figure 20: Distribution of the number of reactions per model, (A) separated by tissue of origin, and (B) separated by the CMS subtype classification.

When considering the CMS subtypes of the respective samples (figure 20B), we can see some interesting differences. The number of reactions in the regulatory CD4 T-cell models tends to decrease in the following manner: CMS1>CMS2>CMS3>CMS4>Normal Matched. Also, while most memory CD8 T-cell models from tumour tissue have a similar number of reactions to those from normal matched mucosa, the models from CMS4 and 2 of the CMS1 have a lot less reactions than other ones.

4.4 Pathway Coverage

To analyse pathway coverage, we calculated the percentage of reactions in each metabolic pathway that were present in each model.

4.4.1 Pathways covered in all models

The most and least covered pathways across all models are depicted in the heatmap from figure 21. The heatmap with all metabolic pathways is in supplementary figure 57.

There are two pathways that clearly have little to no presence across the great majority of the models, which are *peptide metabolism* and *dietary fibre binding*. Both pathways only have reactions that are not catalysed by any enzyme, meaning that none of the reactions have a GPR rule associated. It is more difficult for these reactions to be added in the reconstruction process, as there is no omics data supporting their presence or absence. They will only be added if they are included manually in the core reaction set if, for example, literature supports their existence in the respective cell-type; or if during the reconstruction process, they are necessary for flux consistency.

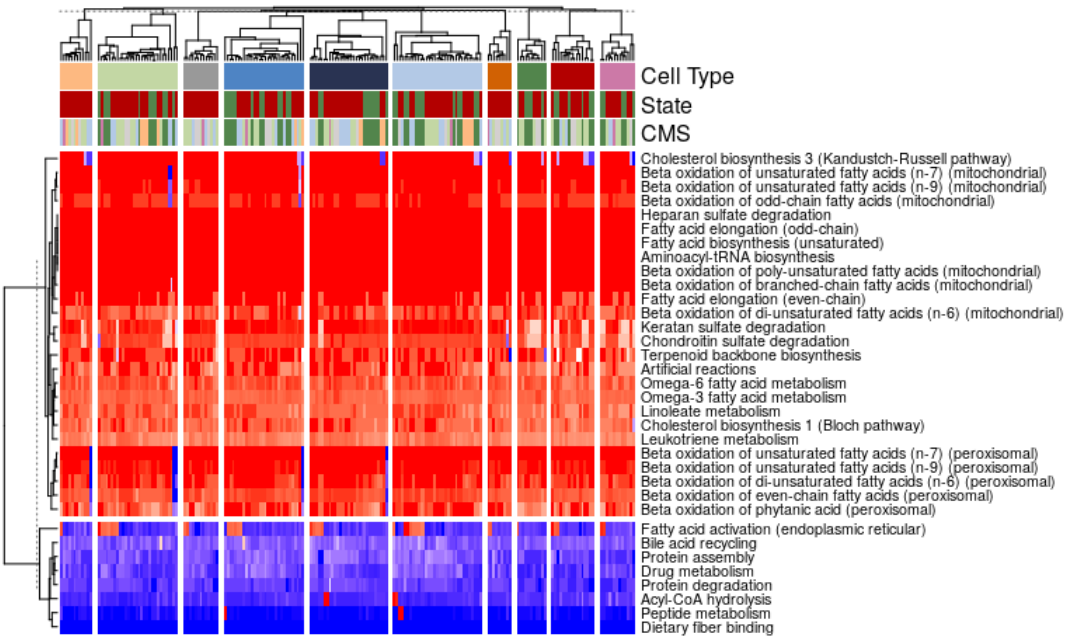


Figure 21: Most and least covered pathways (%) across all reconstructed models.

The other pathways that also have little percentage of reactions present in the models are *fatty acid recycling in the endoplasmic reticulum*, *bile acid recycling*, *protein assembly*, *drug metabolism*, *protein degradation*, and *acyl-CoA hydrolysis*. *Bile acid recycling*, *acyl-CoA hydrolysis* and *drug metabolism*, for example, are pathways mostly associated with the liver [193, 194].

Two clear sets of pathways that are very present in all models are the pathways for β-oxidation of fatty acids in both the mitochondria and peroxisome, and those related to fatty acid metabolism (omega-3 and -6, linoleate), elongation and biosynthesis.

Cholesterol biosynthesis (both *Kandustch-Russel* and *Bloch* pathways) is also present. Cholesterol has a critical role in T-cell function and signalling [195, 196] because T-cells rely greatly on motility and membrane-membrane interactions with other cells and cholesterol has been shown to be important in maintaining cell membrane stiffness [197]. Inhibiting enzymes from the cholesterol metabolism and transporters leads to changes in T-cell function, activation, and reprogramming.

For example, treating a population of naive T-cells with an inhibitor of the enzyme that catalyses the rate-limiting step of cholesterol biosynthesis suppresses progression of their differentiation and cell cycle [197]. Both CD4 and CD8 naive T-cells reprogram their cholesterol metabolism upon activation, promoting both import and biosynthesis [198–200]. SREBP proteins, required for cholesterol biosynthesis, are essential for cytotoxic CD8 T-cells to acquire sufficient cholesterol levels that allow proliferation and acquisition of an effector phenotype [199]. IL17+ CD4 T-cells' differentiation induced by RORγ was preceded by enhanced cholesterol biosynthesis [200]. In fact, increased cholesterol content in the plasma membrane has been associated with a pro-inflammatory phenotype, even though membrane cholesterol enrichment in FOXP3+ CD4 regulatory T-cells did not alter their suppressogenic function [201].

Other pathways highly present across models include heparan sulfate degradation, keratan sulfate degradation, chondroitin sulfate degradation, aminoacyl-tRNA biosynthesis, terpenoid backbone biosynthesis, and leukotriene metabolism.

The interaction of chemokine, integrins and selectins with glycosaminoglycans (GAGs) regulates the recruitment, adhesion, and migration of leukocytes from the circulation to the site of inflammation. GAGs are divided into four main groups: Heparan sulfate, chondroitin sulfate, keratan sulfate, and hyaluronic acid. Apart from hyaluronic acid, GAGs are attached to the core protein of proteoglycans (PGs)[202].

For instance, heparanase, the only mammalian enzyme that directly cleaves heparan sulfate, has also been described to be expressed in T-cells and up-regulated upon activation [203]. Furthermore, lymphocytes from peripheral blood mononuclear cells (PBMCs) of breast cancer patients were found to display higher heparanase expression than those from healthy patients [204]. Even though its expression has been shown to promote leukocytes migration and penetration of the basement membrane and blood vessel entry [203], it has been suggested that it can either be pro- or anti- tumorigenic [205].

Leukotrienes are one of the types of eicosanoids that derive from arachidonate. They have been linked to proliferation, apoptosis, cytokine production, differentiation, and chemotaxis in T-cells [206], and recognised as possibly having both pro- and anti- inflammatory roles in the immune response of T-cells [207].

ALOX5, an enzyme that catalyses the first reaction from this pathway (conversion of arachidonic acid into leukotriene A4, LTA4), was shown to occur in human T cell lines as well as in purified peripheral blood T cells, including naive and memory CD4 T-cells and cytotoxic CD8 T-cells [208]. The production of leukotriene B4 (LTB4) and cysteinyl leukotrienes (LTC4, LTD4 and LTE4) was also detected in various T-cell lines and primary cells [209, 210].

4.4.2 Models' structure differs between normal and tumour tissue

We wanted to assess if any cell-type was affected by the tumour micro-environment, i.e., if the reaction presence differed between normal and tumour models. For that, we performed pathway differential analysis based only on comparing the pathway coverage between normal- and tumour- derived models of each cell-type. From all the cell-types, regulatory CD4 T-cells and cytotoxic CD8 T-cells showed the most interesting results.

To calculate the most differently covered pathways for each cell-type, we first calculated the fold change on pathway coverage between normal- and tumour- derived models using the R package *gtools* [211]. Only those with absolute fold changes higher than 1.5 were subsequently tested using the non-parameteric test Mann-Whitney using the R package *stats* [212]. p-values were adjusted using the false discovery rate (FDR) method by Benjamin & Hochberg [213]. Pathways with an adjusted p-value smaller than 0.05 were considered differentially covered.

Regulatory CD4 T-cells Thirty-two pathways are differentially covered between normal and tumour regulatory T-cell models (supplementary table 15). When plotting these pathways in a heatmap (figure 22), it is possible to see that, even though most models from normal matched tissue group together, there is not a complete separation between normal and tumour derived models. In fact, three main groups can be distinguished: (1) only models from normal matched tissues; (2) all models from tumour tissue classified as CMS4, two from normal tissue and one from tumour tissue classified as CMS3; (3) remaining tumour derived models.

This goes in line with the previously shown results (figure 20) regarding the distribution of number of reactions in the models for each CMS type, where it was visible a decreasing trend: CMS1 > CMS2 > CMS3 > CMS4 > Normal Matched. From all the tumour derived models, the number of reactions in the models from CMS4 samples is the closest to that in the normal matched models. Also, one of the models from CMS3 has clearly less reactions than the other two CMS3 models. These 3 groups are still visible when visualizing all metabolic pathways' coverage in a heatmap (supplementary figure 57).

The pathways *fatty acid biosynthesis (even-chain)*, *glycerolipid metabolism*, *keratan sulfate biosynthesis*, *fatty acid biosynthesis (odd-chain)* give the clearest separation between tumour (heatmap groups (2) and (3)) and normal (heatmap group (1)) derived models. The TCA cycle (*Tricarboxylic acid cycle and glyoxylate/dicarboxylate metabolism*) is also more significantly present in tumour-derived models than normal matched-derived ones.

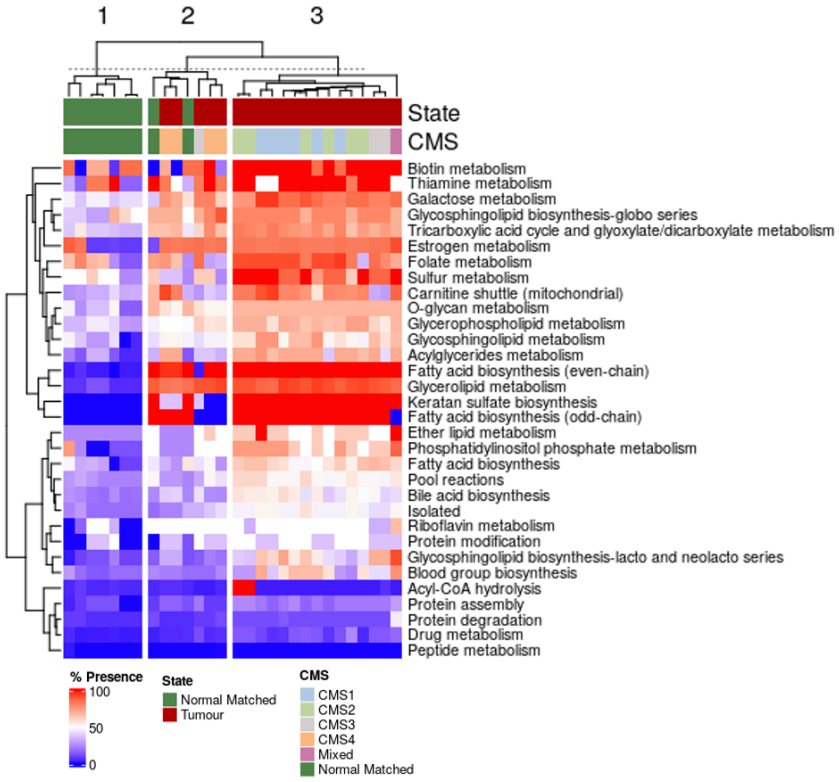


Figure 22: Differentially covered pathways between normal- and tumour- derived regulatory CD4 T-cell models.

When comparing mice regulatory T-cells from a tumour site and the spleen, Pacella *et al* [214] found that those from the tumour site tended to acquire less FAs, despite the up-regulation of *CD36*, a transporter of FAs into the cell. Considering that *CPT1A*, which moves FAs into the mitochondria, was significantly more expressed in tumour regulatory CD4 T-cells, and that these cells accumulated significantly more FAs, FA synthesis might be more present in tumour regulatory CD4 T-cells. Pacella *et al* [214] further showed higher consumption of the first intermediates of the TCA cycle by the tumour-infiltrating regulatory CD4 T-cells.

There are a number of pathways whose distinction between groups (1) and (2) is not as clear. These pathways include: *biotin metabolism*, *folate metabolism*, *sulfur metabolism*, *carnithine shuttle (mitochondrial)*, *ether lipid metabolism*, *phosphatidylinositol phosphate metabolism*, *fatty acid biosynthesis*, *pool reactions*, *bile acid biosynthesis* and *glycosphingolipid biosynthesis-lacto and neolacto series*.

This shows that regulatory CD4 T-cell models from CMS4 tumour seem to be less anti-inflammatory than the other tumour-derived models.

Biotin deficiency decreases differentiation toward anti-inflammatory regulatory T-cells [215]. It thus makes sense that this pathway is very present in regulatory CD4 T-cells, even more so in tumour regulatory CD4 T-cells. Although most regulatory CD4 T-cell models have relatively high presence of *biotin metabolism*, it is visibly higher in group (3).

Folate metabolism coverage is lower in groups (1) and (2). Yamaguchi *et al* [216] reported that folate

receptor 4 (FR4) is crucial for murine regulatory CD4 T-cell expansion *in vivo* and its blockade enhanced anti-tumour immunity.

High levels of hydrogen sulfide (H2S), produced in the *sulfur metabolism* pathway, are known to limit the release of pro-inflammatory molecules and promote the secretion of anti-inflammatory cytokines [217]. Knockout of genes that encode for H2S-producing enzymes decreases regulatory CD4 T-cell proliferation [218].

The *carnithine shuttle (mitochondrial)* is in charge of transporting FAs in and out of the mitochondria, crucial for FAO to occur.

Cytotoxic CD8 T-cells Twelve pathways are differentially covered between normal and tumour cytotoxic CD8 T-cell models (supplementary table 16). When plotting these pathways in a heatmap (figure 23), it is possible to see a clear separation between tumour- and normal- derived models. The pathways that are more clearly different between the two groups are *oxidative phosphorylation* (OXPHOS), *biopterin metabolism*, *pantothenate and CoA biosynthesis*, *carnitine shuttle (peroxisomal)*, *folate metabolism*, and *estrogen metabolism*. There are sphingolipid-related pathways (*glycosphingolipid biosynthesis-ganglio series*, *sphingolipid metabolism* and *glycosphingolipid metabolism*). There is one model from a CMS2 tumour, however, that seems closer to the normal-derived models, as pathway coverage is very similar apart from *folate metabolism* and *estrogen metabolism*. Also, from the 10 tumour-derived models, only 3 have relatively low coverage of *estrogen metabolism*.

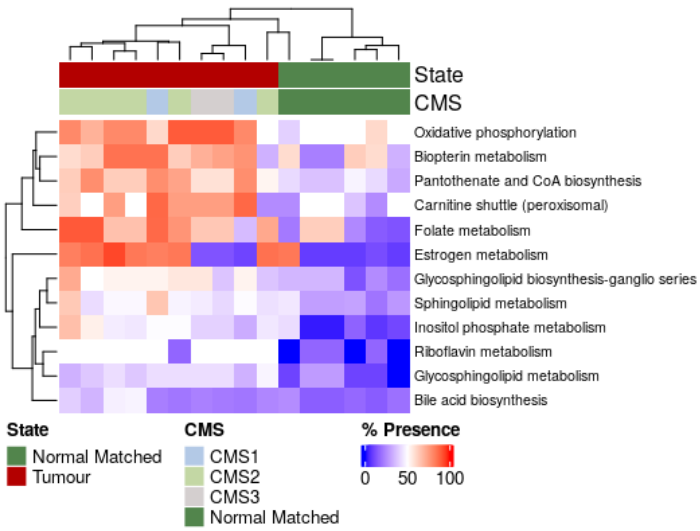


Figure 23: Differentially covered pathways between normal- and tumour- derived cytotoxic CD8 T-cell models.

Folate deficiency was shown to reduce CD8+ T-cells capacity to proliferate in response to activation. This sensitivity to the lack of folate is higher in CD8+ T-cells than CD4+ T-cells [219]. We showed previously that the *folate metabolism* in regulatory CD4+ T-cell models from CMS4 tumours, and normal-matched tissues, was less covered than the other tumour-derived models. Regarding *estrogen metabolism*, estrogens are usually correlated with an immunoenhancement effect on the immune system, with CD8+ T-cells

showing a high response [220]. Navarro *et al* [221] showed that the female human CD8+ T-cells incubated with higher concentrations of the estrogen 17β-estradiol (E2) were significantly more cytotoxic, while the male human CD8+ T-cells showed an increase but not significant. Of note, the 3 tumour-derived models with low *estrogen metabolism* are not all from male patients, and the other tumour-derived models are not all from female patients. The same occurs in the normal-derived models.

Increased levels tetrahydrobiopterin (BH4), by either providing high concentrations of it in the medium or by overexpression of *GCH1*, a gene from the *biopterin metabolism*, were shown to enhance proliferation of stimulated mice CD4 and CD8 T-cells. Cronin *et al* [222] further showed that stimulated BH4-deficient T-cells hold decreased mitochondrial respiration and oxygen consumption. Indeed, *OXPHOS* and *pantothenate and CoA biosynthesis* (essential for the TCA cycle) are two pathways that are also more present in the tumour-derived cytotoxic CD8 T-cell models than those derived from normal-matched tissue.

4.4.3 Differences in models' structure between cell-types

We also wanted to assess if any pair of cell-types' metabolism was significantly different, i.e., if the pathway presence differed between them. For that, we performed pathway differential analysis based only on comparing the pathway coverage. From all the pairs of cell-types, regulatory vs IL17+ CD4 T-cells, and naïve vs proliferative CD8 T-cells showed the most interesting results.

To calculate the most differentiated pathways for each pair of cell-types, we first calculated the fold change on pathway coverage between two cell-types using the R package *gtools* [211]. Only those with absolute fold changes higher than 1.5 were subsequently tested using the non-parameteric test Mann-Whitney using the R package *stats* [212]. p-values were adjusted using the Benjamin & Hochberg (FDR) method [213]. Pathways with an adjusted p-value smaller than 0.05 were considered differentially covered.

IL17+ CD4 vs regulatory CD4 T-cells Nine pathways are differentially covered between IL17+ and regulatory CD4 T-cell models (supplementary table 18). When plotting these pathways in a heatmap (figure 24), it is possible to see a separation between these two cell-types. The pathways that are more clearly different between the two cell-types are *biotin metabolism*, *fatty acid biosynthesis*, *glycerolipid metabolism*, and *keratan sulfate biosynthesis*.

We previously showed that biotin metabolism was more present in tumour-derived regulatory CD4 T-cell models than those derived from normal matched mucosa and that biotin deficiency revealed a decreased differentiation towards anti-inflammatory regulatory T-cells [215]. Now, when comparing IL17+ and regulatory CD4 T-cells, we were also able to capture the difference in biotin metabolism presence between these two cell-types, where this pathway is less present in IL17+ CD4 T-cell models. Indeed, biotin deficiency did not only lead to a decreased differentiation towards anti-inflammatory regulatory T-cells, but also induced Th1 and Th17-mediated pro-inflammatory responses [215]. This shows that, like corroborated in the literature, that biotin metabolism is important for an anti-inflammatory function of regulatory T-cells, while it is not necessary for IL17+ CD4 T-cell function.

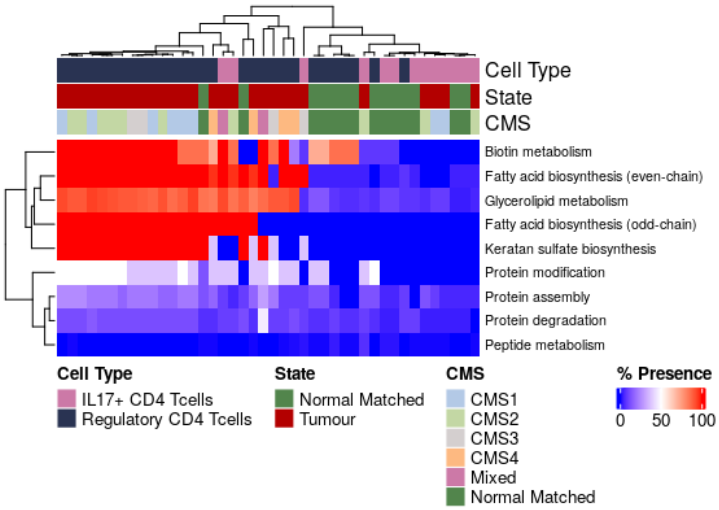


Figure 24: Differentially covered pathways between IL17+ and regulatory CD4 T-cell models.

The normal-derived regulatory CD4 T-cell models tend to cluster closer to IL17+ T-cell models than the other regulatory CD4 T-cell models. Interestingly, the pathways that are clearly different between these two cell-types were also considered differentially present between normal- and tumour- derived regulatory CD4 T-cell models. These pathways are *biotin metabolism*, *fatty acid biosynthesis (even-chain)*, *fatty acid biosynthesis (odd-chain)*, *glycerolipid metabolism*, and *keratan sulphate biosynthesis*.

Protein related pathways (protein modification, assembly and degradation, and peptide metabolism) were also differentially more present in regulatory CD4 T-cells but were not considered differentially present between normal- and tumour- derived regulatory CD4 T-cell models.

Naive CD8 vs proliferative CD8 T-cells Ten pathways are differentially covered between naïve and proliferative CD8 T-cell models (supplementary table 17). When plotting these pathways in a heatmap (figure 25), it is possible to see a clear separation between these two cell-types. The pathways that are more clearly different between the two cell-types are *carnithine shuttle* (endoplasmic reticular and mitochondrial), *thiamine metabolism*, *sulfur metabolism*, and *O-glycan metabolism*.

The *carnithine shuttle (mitochondrial)* is in charge of transporting FAs in and out of the mitochondria. This is known to be crucial for fatty acid oxidation, process in which naïve T-cells greatly rely on [30]. However, our models show higher presence of carnithine shuttle (endoplasmic reticular and mitochondrial) in proliferative CD8 T-cell models than in naïve CD8 T-cell models, except for a few naïve models that also have a high presence of *carnithine shuttle (endoplasmic reticular)*. Still, Cong-Hui *et al* [223] showed that transport of fatty acids into the mitochondria by *CPT1* may be required for anabolic processes that support healthy mitochondrial function and proliferation independent of fatty acid oxidation in cancer cells. It could be the same case for proliferative T-cells. Endoplasmic reticulum's carnithine shuttle has been shown to function as an antioxidant to inhibit endoplasmic reticulum stress in proliferative cells other than T-cells [224].

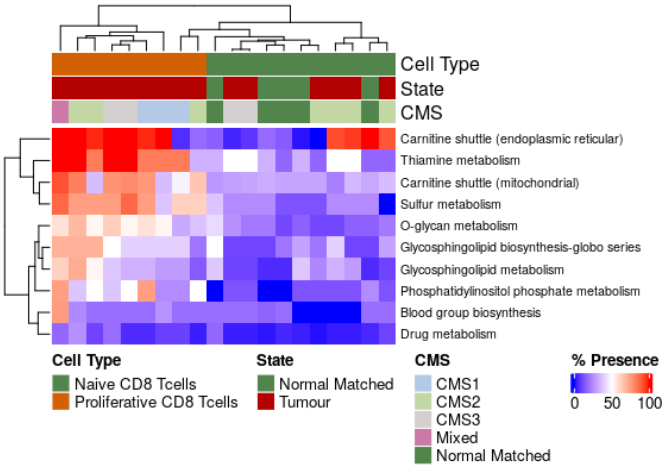


Figure 25: Differentially covered pathways between naive and proliferative CD8 T-cell models.

Glycans are essential for signaling and cell-cell interactions, and they have been shown to be important in T-cell development, activity, differentiation and proliferation. Although O-glycans are *de novo* synthesized by cells, naïve T-cells cannot synthesize core 2 O-glycans. Following TCR stimulation, T-cells increase expression of 2 O-glycan synthesis, which allows proliferative T-cells to extravasate into non-lymphoid tissues [225].

4.5 Predicting cell-types in the different stages of model reconstruction

We sought to assess how well the data from different stages of model reconstruction predict the different cell-types. Using pseudo-bulk transcriptomics data leads to the best results out of all stages (figure 26), with an MCC of 0.793. Nevertheless, the model structure (presence/absence of reactions) is still a good classifier for cell-type, with an MCC of 0.526. If we filter the reactions without a GPR, this value increases to 0.583.

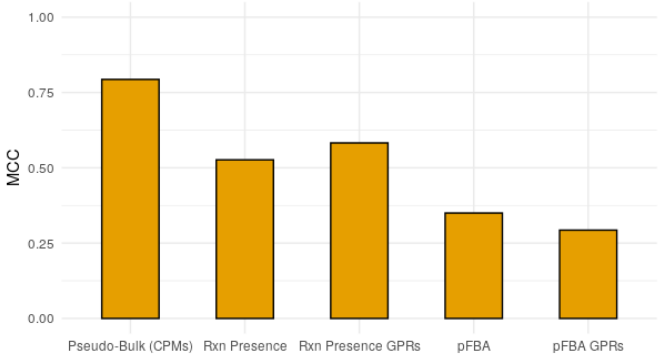


Figure 26: MCC results when predicting cell-type using test samples from the pseudo-bulk RNAseq (in CPMs), reactions presence, or pFBA predicted fluxes datasets.

The fluxes predicted using the normal human blood medium lead to the worst results, with an MCC of 0.350. This suggests that the structure of a model is a better indicator of cell-type than predicted fluxes. Also, a good structure that correctly resembles the experimental data does not necessarily mean that predicted fluxes will be good, as the model's objective and constraints like the medium used affects predictions. Nevertheless, the fluxes still have some predictive power, as the MCC is clearly higher than 0. This time, the pFBA fluxes without filtration of the reactions without a GPR give better predictions than those with (MCC 0.350 > 0.293).

Even though different, we can thus expect high flux similarities between the models of the different T-cell subtypes. We calculated the Euclidean distances of the structure (supplementary figure 56A) and of the predicted fluxes under normal medium (supplementary figure 56B) between the different models. Indeed, all models were structurally very different from each other, while leading to very similar flux predictions.

4.6 Flux Predictions

4.6.1 Biomass and ATP production

We chose different objectives for the models of proliferative T-cells and the remaining, non-proliferating, T-cells. As mentioned, biomass production is not the main objective of non-proliferative T-cells, with the production of energy being as important. With this, the proliferative T-cells' models were optimised for biomass production while the remaining were optimised for both biomass and ATP production, using pFBA.

Having this in mind, prediction results do show that the biomass of proliferative T-cells is higher than their naïve counterparts (figure 27), as expected. Cytotoxic CD8 T-cell follicular CD4 T-cells and IL17+ CD4 T-cell models also have relatively low biomass.

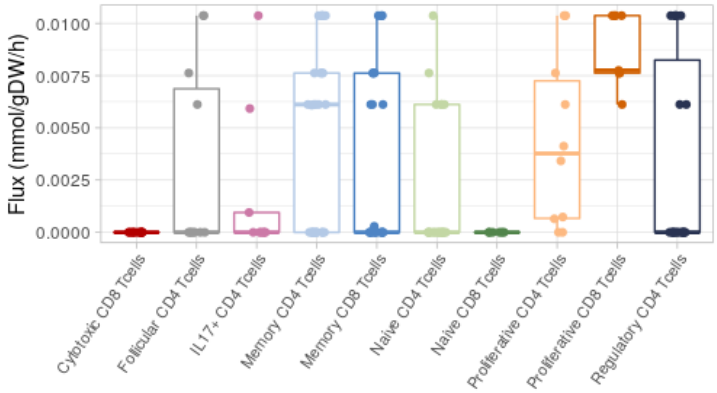


Figure 27: Biomass flux prediction using normal human blood.

Some models from naïve CD4, follicular CD4, memory CD4, memory CD8 and regulatory CD4 T-cells showed biomasses as big as those observed in the proliferative T-cells' models, even though the objectives

are different. While for memory T-cells the flux going through the biomass reaction varies in both tumour- and normal matched- derived models (figure 28A), all naive CD4 T-cell models, except one, that have high biomass are from tumour tissue samples. In line with the high biomass fluxes, all naïve CD4 T-cell models, except one, that have low ATP production levels are from tumour tissue samples (figure 28B).

As for regulatory CD4 T-cell models, the models with high biomass are from tumour samples of CMS1 and CMS2 types (figure 28A). Interestingly, regulatory CD4 T-cell models from CMS4 and one from CMS3 clustered closer to those from normal tissue when looking into the pathway coverage in section 4.4.2, while models from CMS1 and CMS2 showed higher coverage of pathways connected to pro-inflammatory functions. While the models from tumour tissues of type CMS1 and CMS2 have high biomass flux, unlike the remaining regulatory CD4 T-cell models, only those from CMS1 have low ATP production (figure 28B).

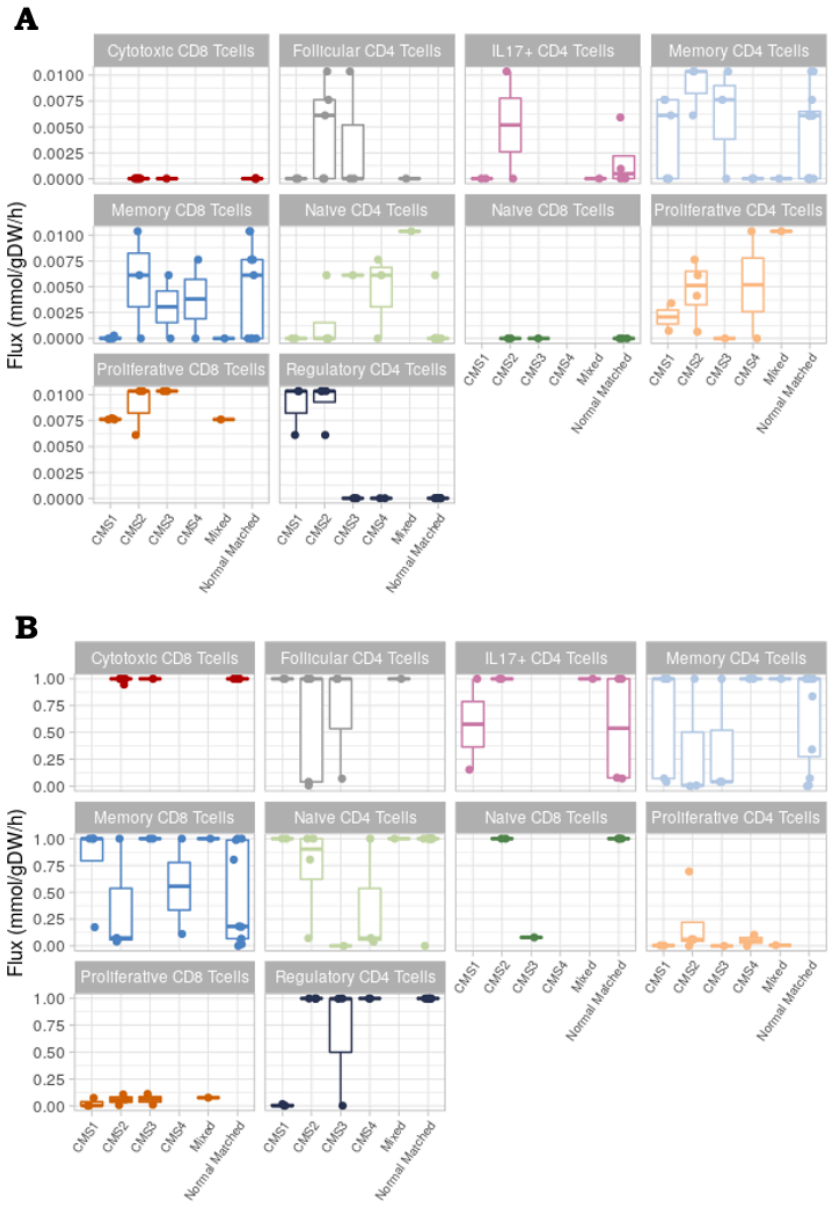


Figure 28: Biomass flux (A) and ATP production (B) predictions using normal human blood, separated by CMS type.

Regarding ATP production, it was expected that it would be higher in non-proliferative T-cells, due to the objectives chosen for pFBA. It is visible (figure 28A) that proliferative T-cells do have less ATP production than the remaining T-cell types. However, apart from cytotoxic CD8 T-cells, some of the models of these non-proliferating T-cell types have relatively lower or equal ATP production to those of proliferative T-cells.

We further checked what would happen to the biomass and ATP production if all models, irrespective of cell-type, were optimised for biomass production only (supplementary figure 58). Even though we can see some difference in biomass fluxes between naive and proliferative CD8 T-cells, all cell-types show relatively the same biomass and ATP production flux. This shows the importance of choosing the right objective to predict a model's fluxes that can depict the real fluxes of a cell-type. Previous studies [129, 137, 148], including one for T-cells, have also applied an objective that would combine biomass and ATP production for models characterising normal, non-proliferative, cells, while proliferative cell models (tumour and normal) would only have the maximisation of biomass as their objective.

Performing prediction of the cell-types based on the pFBA fluxes predicted using biomass as the only objective for all models (figure 29) lead to worse results, especially when comparing those obtained using all reactions (MCC of 0.158 vs 0.350).

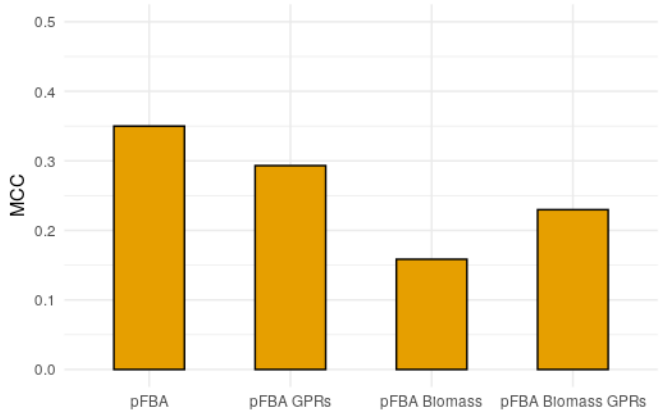


Figure 29: MCC results when predicting cell-type using test samples from the pFBA predicted fluxes datasets. *pFBA*: pFBA predictions with the different objectives and all reactions; *pFBA GPRS*: pFBA predictions with the different objectives and only reactions with GPRs; *pFBA Biomass*: pFBA predictions with biomass as the only objective and all reactions; *pFBA Biomass GPRS*: pFBA predictions with biomass as the only objective and only reactions with GPRs.

Ploting the biomass against the ATP production (figure 30) shows that most proliferative models have relatively low ATP production flux and a variable biomass flux that is mostly high (13 out of 19 models). Regarding the remaining models, which are non-proliferative and thus were maximised for both ATP and biomass production, most either have high biomass and low ATP production (28 out of 123 models), or low biomass and high ATP production (73 out of 123 models). This shows that even when maximising for both biomass and ATP production, the tendency of models with high biomass to have low ATP production, and vice versa, is still captured. Furthermore, most models where this is not observed are from tumour

tissue samples: 9 of the 12 models with low ATP production and biomass and 10 out of the 15 models with high ATP production and biomass.

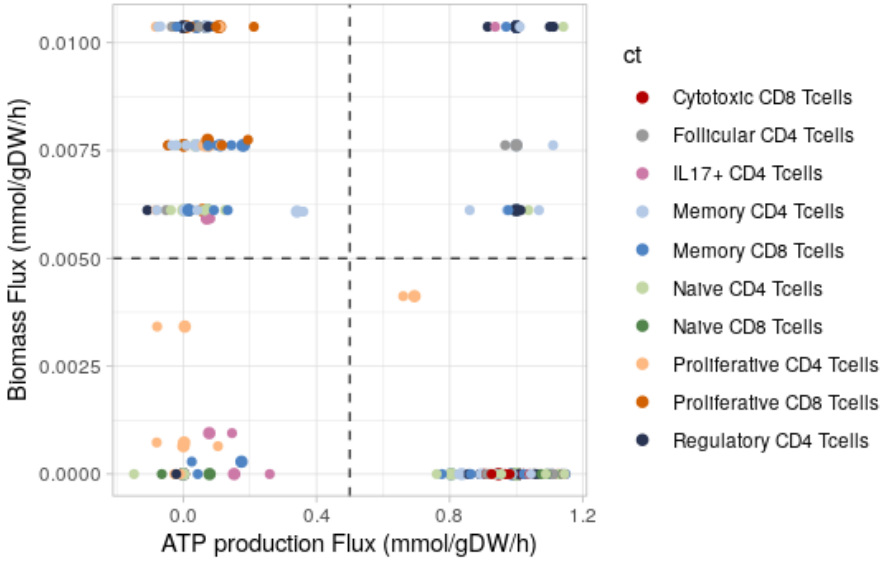


Figure 30: Models' fluxes of the ATP production and biomass. *Upper left*: high biomass and low ATP; *upper right*: high biomass and high ATP; *bottom left*: low biomass and low ATP; *bottom right*: low biomass and high ATP.

4.6.2 Sources of FADH2 and NADH and fatty acid (FA) uptake

In general, all cell-types are obtaining most of their FADH2 (figure 31A) and NADH (figure 31B) from fatty acid oxidation (FAO) pathway. This was assessed by calculating for each pathway the total flux going through the reactions that produce FADH2 or NADH.

Regarding the pathways that work as NADH sources (figure 31B), glutaminolysis is the one that contributes the least, followed by glycolysis. After FAO, the TCA cycle is the pathway that most contributes to the production of NADH in the models. Regarding proliferative CD4 and CD8 T-cells, however, their models show that the biggest source of NADH is gycolysis, followed by TCA cycle (supplementary figure 59).

The models for naive T-cells are in line with literature [30], as they do rely greatly on FAO, followed by the TCA cycle (figure 31). Proliferating T-cells also show an expected higher dependency on glycolysis than other pathways. The flux going through NADH producing reactions from glycolysis is not higher in proliferative versus naive T-cell models, however.

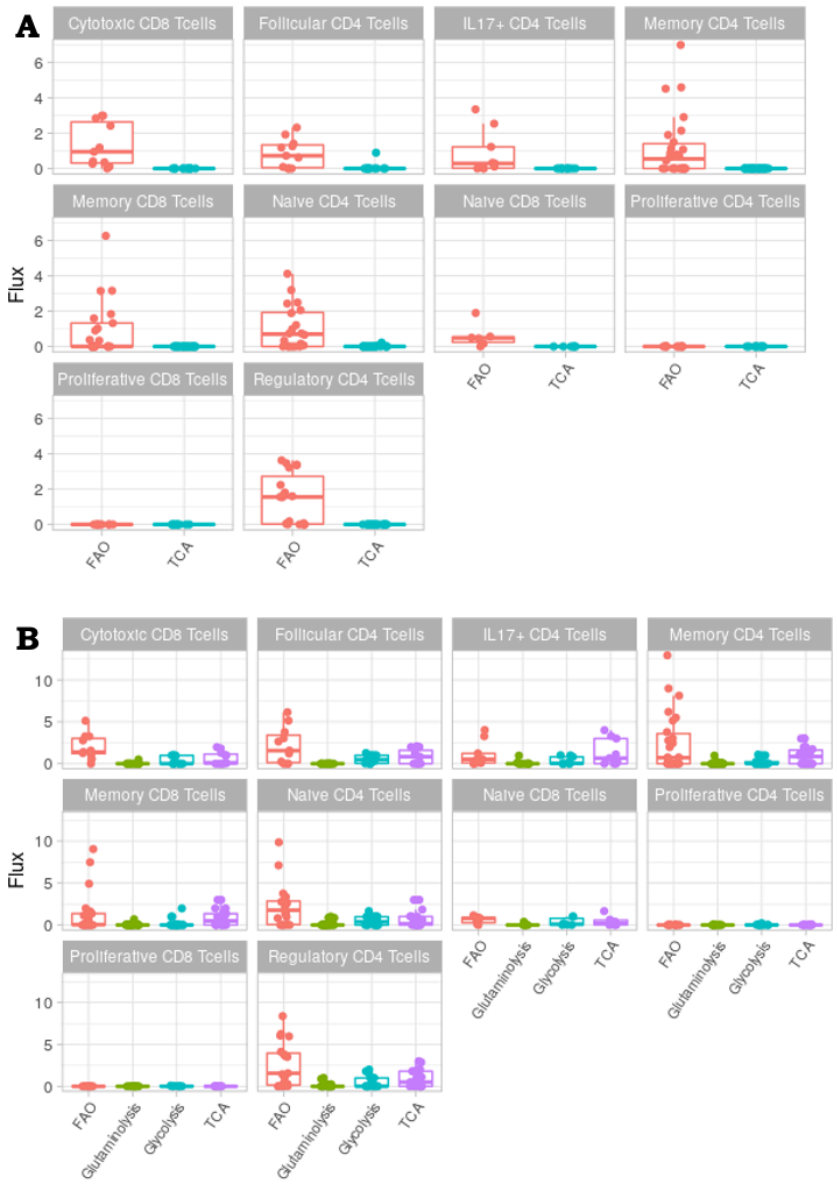


Figure 31: Cumulative fluxes (mmol/gDW/h) of the reactions that produce (A) FADH2 or (B) NADH, for each source pathway.

Although memory T-cells synthesize FAs and aerobic glycolysis is reduced, effector memory T-cells that reside in tissues rely on the import of extracellular FAs and glycolysis [43]. This might be why there are memory CD4 and CD8 models that either have relatively high flux of FAs uptake or no/very close to no uptake (figure 32A). Also, memory CD4 T-cells have a higher median of fluxes producing NADH and FADH2 from FAO and glycolysis than memory CD8 T-cells (figure 31).

Proliferative T-cells are characterised by relying in fatty acid synthesis instead of uptake (figure 1), which shows in our results (figure 32A), with a null or very close to null uptake of fatty acids.

Regarding regulatory CD4 T-cells, Pacella *et al*'s study [214] found that regulatory CD4 T-cells from the tumour site tended to acquire less FAs, despite the up-regulation of *CD36*, a transporter of FAs into the cell. We not only showed in section 4.4.2 that fatty acid synthesis was more present in tumour regulatory

CD4 T-cell models, but also now pFBA predictions show smaller fatty acid uptake fluxes in a lot of the tumour-derived models (figure 32B), when compared to the normal-derived ones.

Finally, although IL17+ CD4 T-cells were found to rely on *de novo* FA synthesis rather than acquisition of extracellular fatty acids to meet lipid requirements [49], a lot of the IL17+ CD4 T-cell models show a relatively high flux of FA uptake (figure 32A).

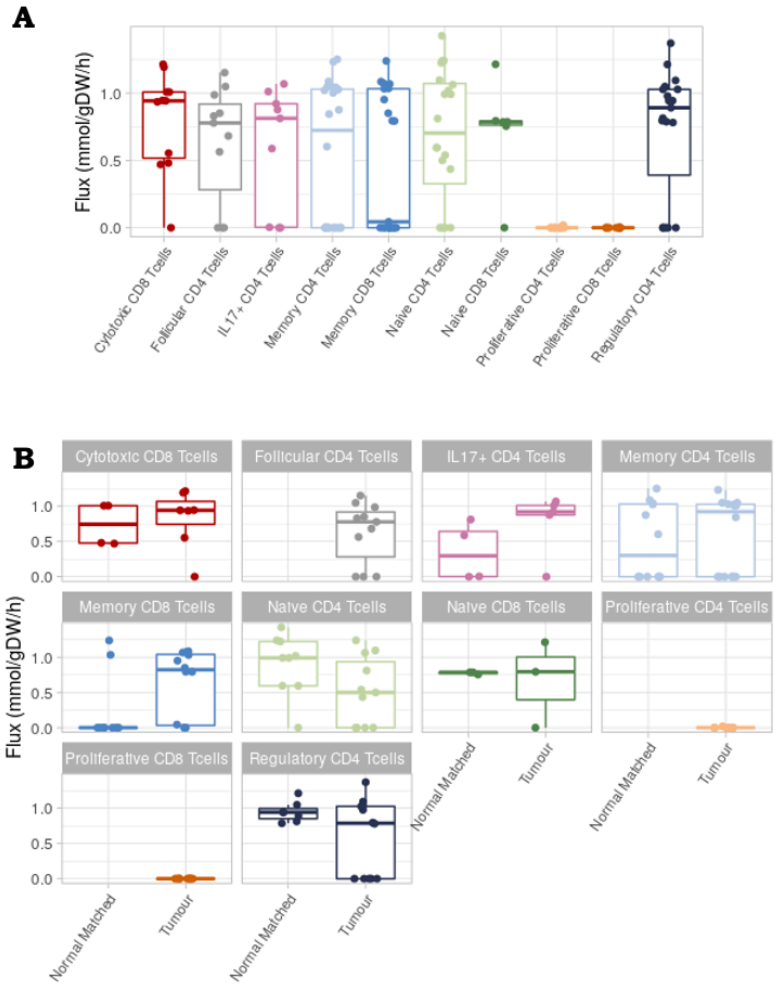


Figure 32: Cumulative fluxes (mmol/gDW/h) of the reactions that uptake fatty acids (FAs) from the medium, (A) per cell-type and (B) separated by tissue of origin.

4.6.3 Effect of metabolite availability on biomass

We next evaluated the effect that the absence of certain metabolites has on the biomass of the T-cell models. Thus, the objective of all models was set to maximise only biomass.

No Tryptopahn It is expected that absence of tryptophan from medium causes T-cells' biomass to decrease, as it has been shown that indoleamine-pyrrole 2,3-dioxygenase (IDO), which catalyses tryptophan metabolism in the kynurenine pathway, inhibits T-cell activation by tryptophan deprivation and by

promoting the expansion of regulatory T-cells [226]. The biomass of all models decreases to zero or very close to zero once tryptophan is removed from medium (figure 33), even regulatory CD4 T-cells.

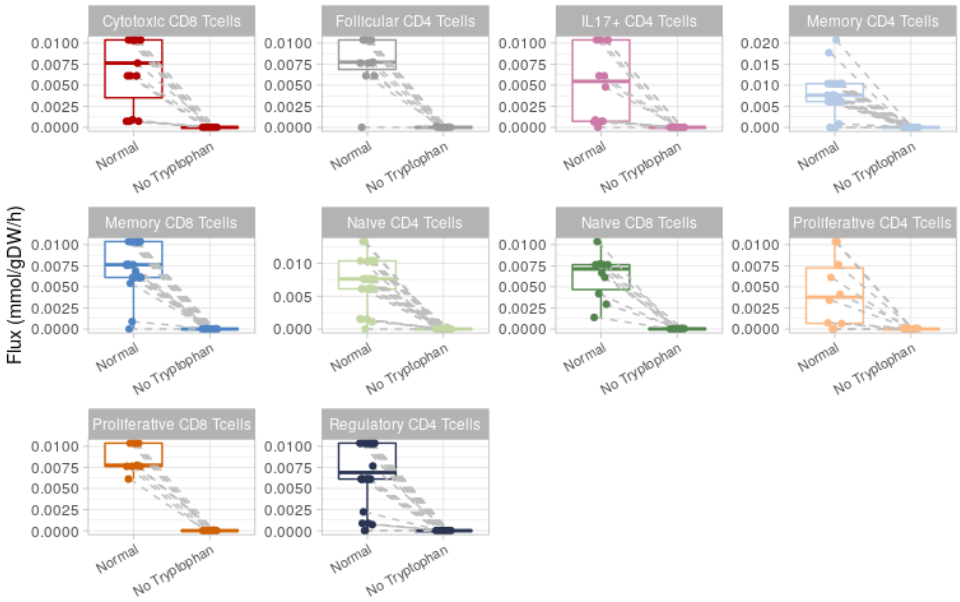


Figure 33: Distribution of the biomass flux of the T-cell types, with and without tryptophan in the medium.

No Oxygen Although a lot of models across most cell-types suffer a decrease in biomass flux, most cell-types, on average, do not suffer a big change (figure 34).

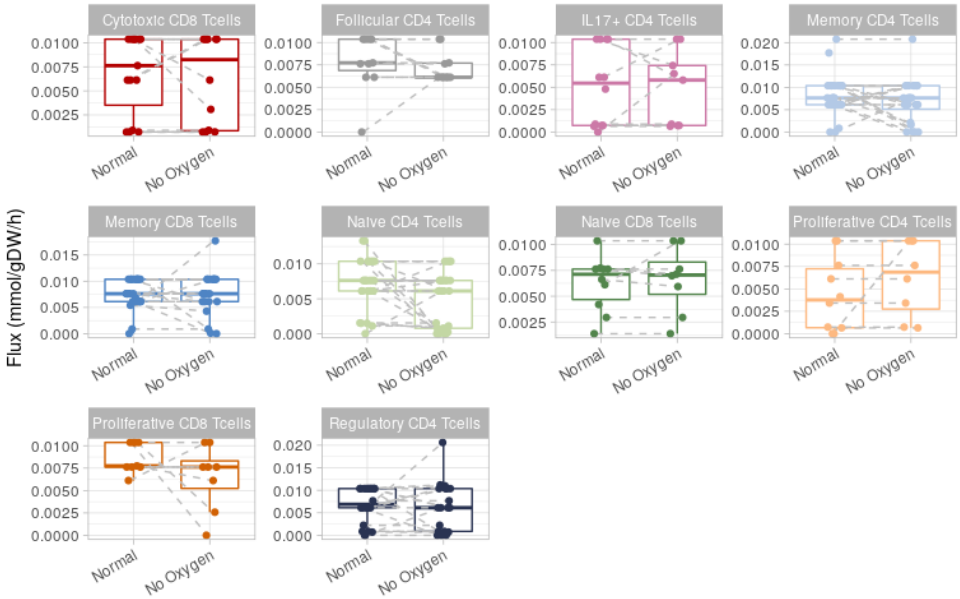


Figure 34: Distribution of the biomass flux of the T-cell types, with and without oxygen in the medium.

In general, it is expected that T-cells suffer reduced proliferation in an environment with no oxygen [227, 228], even though reduced amounts of oxygen up-regulate genes involved in glycolytic ATP production and down-regulates the OXPHOS pathways [229], associated with higher [230, 231] or no effect [229] in proliferation. Still, it has been pointed out that the impact of oxygen in cell viability relies on the type of stimulus that the stimulated cultures received, as two different stimuli revealed different impacts of oxygen levels on T-cells proliferation [232, 233]. Thus, removing oxygen from the metabolic models' medium can result in either decreased, increased or no effect on biomass.

It is well known that the effects of hypoxia are coordinated by the hypoxia-inducible transcription factors (HIFs). As noted before, however, the metabolic models do not capture the signaling and/or regulation pathways. Thus, discrepancies between the effect of hypoxia in our models and literature might be due to this. Furthermore, most studies on the effect of hypoxia test the cells under very low oxygen levels, and not in complete absence of it.

No Glutamine Unavailability of glutamine in medium, with glucose present, decreases biomass flux to or very close to zero across all cell-types (figure 35).

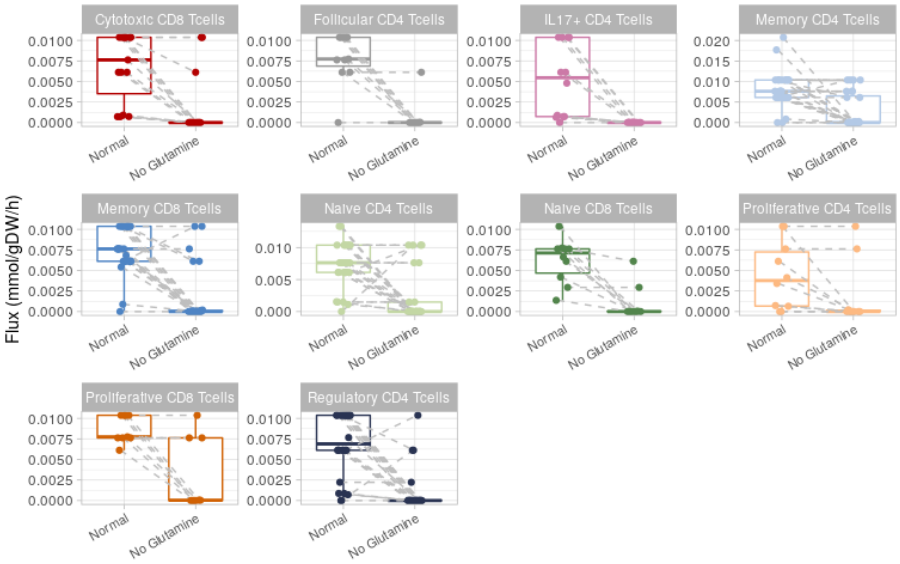


Figure 35: Distribution of the biomass flux of the T-cell types, with and without glutamine in the medium.

As reported in the literature, glutamine seems to be essential for all T-cells' proliferation [226, 227, 234], especially because it acts as a nitrogen donor for DNA and RNA nucleotide production [227, 234]. Indeed, DNA and RNA production decreases to zero, or very close to zero, when no glutamine is available in the models' medium (supplementary figure 60).

No Nucleotides Overall, there is no difference in biomass flux (figure 36) and production of DNA (supplementary figure 61) when no nucleotides are available in the medium. This goes in line with Ma *et al*'s study [235] on *in vitro* vs *in vivo* metabolism of CD8+ T-cells, where the effector cells where shown to almost entirely rely on *de novo* nucleotide biosynthesis.

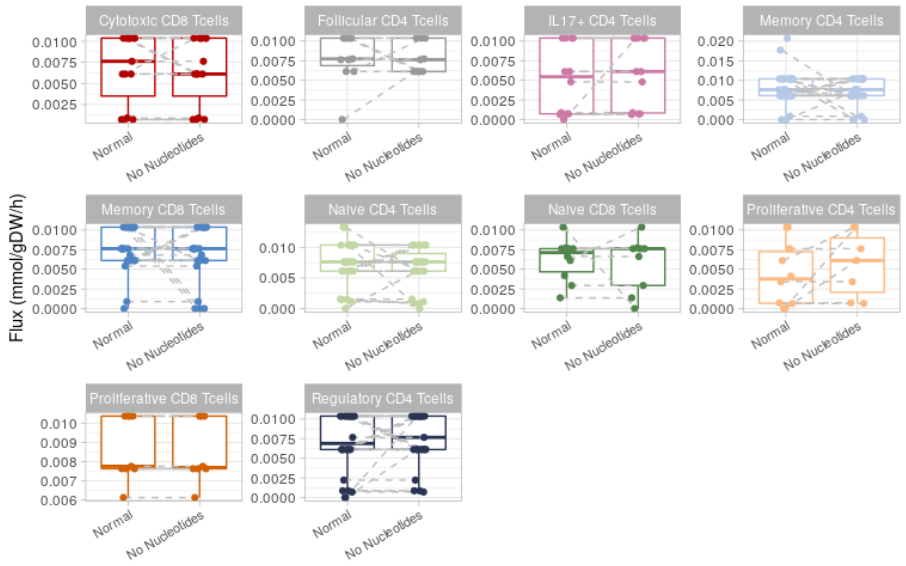


Figure 36: Distribution of the biomass flux of the T-cell types, with and without nucleotides in the medium.

No Glucose When removing glucose from medium, most models do not suffer changes in the biomass flux, with only some showing decreased biomass (supplementary figure 62). This, however, is not corroborated by literature, as it has been reported decreased T-cell proliferation rates in glucose-deficient media [227].

However, these studies are mainly done *in vitro*, where not all metabolites present in the blood are used and glucose concentration is significantly higher than physiological levels. The existence of alternative metabolites in the medium may reduce the dependence on glucose.

Essentiality of genes that encode transporters As detailed in the following section, we performed gene essentiality for all models. From the 932 genes that are potentially essential, 201 catalise transport reactions. From these, 43 genes were predicted as essential in our pipeline for at least one cell-type. Focusing on those that catalise uptake of metabolites, we were left with 14 genes (supplementary figure 14).

Interestingly, some of these genes are involved in co-uptake of sodium and chloride (SLC12A3); sodium- and chloride- dependent transportation of glycine (SLC6A5); sodium- and chloride- dependent transportations of taurine and β-alanine (SLC6A6); sodium-dependent transportation of monocarboxylates and short-chain fatty acids (SLC5A8); and sodium-dependent uptake of the bile acid sulfotaurolithocholate (SLC10A6). Other essential genes catalise uptake of compounds like amino acids (SLC7A11, SLC7A8, SLC7A5); prostaglandines, leukotrines and other eicosanoids (SLCO2A1).

It is noteworthy that no single-deletion of a gene responsible for the uptake of tryptophan or glutamine was reported as essential. This is due to the uptake of these metabolites being done by reactions that are catalised by more than one alternative gene, or even by different reactions catalised by different genes.

Thus, all genes related to the uptake of each of these metabolites would have to be deleted to replicate the results obtained upon their removal from the medium.

4.7 Gene Essentiality

4.7.1 Validation with CRISPR-CAS9 studies

We compared the gene essentiality predictions with two CRISPR-CAS9 studies that evaluated how essential the genes were for in vitro human CD4 [236] and CD8 [237] T-cell proliferation.

Ting *et al* [236] evaluated a total of 2 658 genes, of which only 274 are present in genes tested *in silico* (figure 37A). Shifrut *et al* [237], on the other hand, tested a total of 9 329 genes, of which only 462 are part of the genes tested *in silico* (figure 37A). 324 genes tested *in silico* were not tested in any of the studies (figure 37A). It should be kept in mind that some of the genes could be present in two or all the three datasets, but the gene symbol used could be different. As a simple comparison was made between the symbols present in the three datasets, this might be the case for some genes.

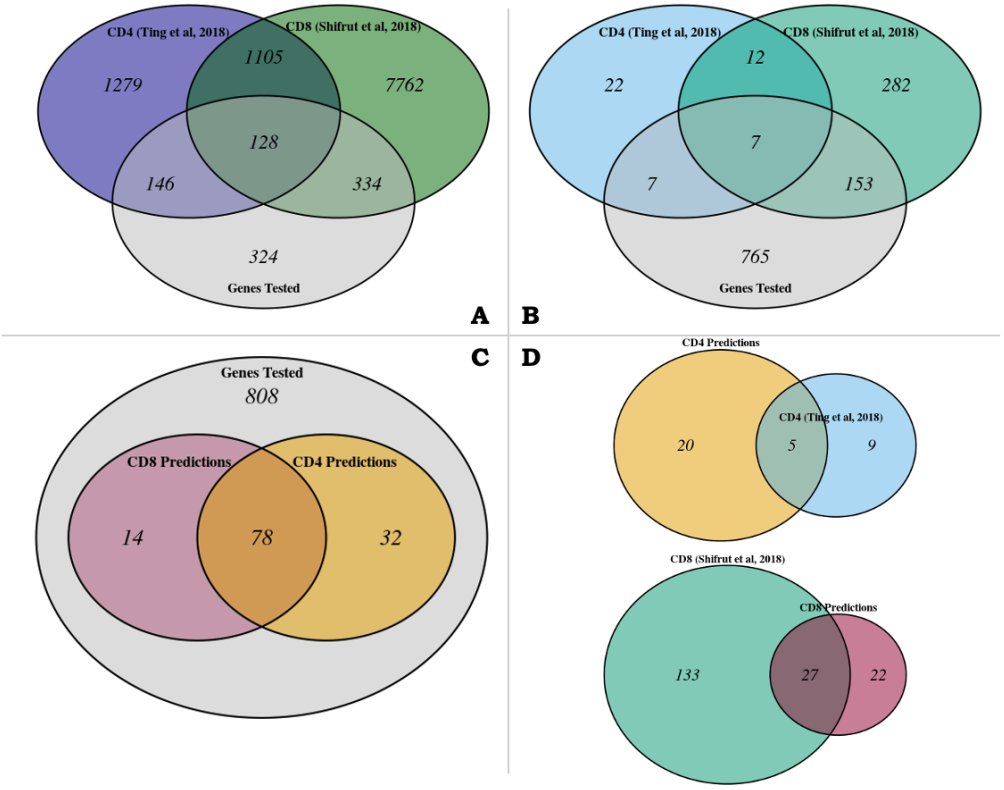


Figure 37: Venn diagrams of (A) the genes that were tested by the 3 different datasets, (B) the genes tested *in silico* and the genes reported as essential by the studies, (C) the *in silico* predictions and the genes tested, and (D) from the common genes between the studies and the pipeline, the predicted essential genes and the essential genes reported by the respective study, for both CD4 T-cells (top diagram) and CD8 T-cells (bottom diagram).

As no CRISPR-CAS9 studies testing cell proliferation were found for any specific CD4 or CD8 T-cell subtype, we joined the genes that were predicted as essential for each CD4 T-cell subtype together and compared them with those from the study of Ting *et al* [236]. The same thing was done for the CD8 T-cell subtypes, for comparison with Shifrut *et al* [237].

From the 932 genes tested in our pipeline, 14 were only essential in CD8 T-cell subtypes, while 32 were only essential in CD4 T-cell subtypes (figure 37C). 78 were essential in both CD4 and CD8 T-cell subtypes.

From the 2 658 genes tested by Ting *et al* [236], 48 were considered essential. Of these 48, only 14 are present in the group of genes tested in silico (figure 37B). For the study of CD8 T-cells [237], 454 genes were found to be essential, in the total of 9 329 genes tested, and only 160 are present in the group of genes tested in our pipeline. The CRISPR-CAS9 studies share 19 essential genes, 7 of which are part of our tested genes.

For a fair comparison between our predictions and the studies' essential genes, we only used the *in silico* genes that were also tested in the studies (figure 37D). For the CD4 T-cells, only 5 were considered essential in both the study [236] and this work. While 9 *in silico* predictions were not essential in the study, 20 study's essential genes were not considered so in our pipeline. For the CD8 T-cells, 27 were considered essential by both our pipeline and Shifrut *et al* [237]. However, 22 *in silico* essential genes were not considered essential by the study, and 133 of study's essential genes were not predicted essential in our pipeline.

There is a marked difference between the *in silico* predictions and the results reported by both studies. However, from the 274 common genes, 240 genes are not considered essential in both CD4 T-cells' datasets, while 280 were not essential in both CD8 T-cells' datasets, in the total of 462 common genes.

It is good to keep in mind, however, three main points: (1) the CRISPR-CAS9 studies were performed with healthy human T-cells *in vitro*, while our models represent T-cells from a tumour patient (some from tumour tissue, others from normal matched tissue); (2) the medium used in the predictions does not resemble a medium usually used *in vitro*, but instead resembles the metabolites present in healthy human blood; and (3) the metabolic models can only predict the essentiality of a gene at the metabolism level, they do not account for regulatory and signalling pathways that can affect the essentiality of a gene.

4.7.2 Pathways affected across cell-types

We wanted to check the pathways with the most essential genes, i.e., the most affected pathways. For that, from all the potential essential genes involved in a pathway, we calculated the percentage of those that were essential to each model. We focused on the most affected pathways whose number of potential essential genes is bigger than one (table 6).

Table 6: Top twenty pathways with the most median percentage of essential genes.

Pathways	**% of essential genes**
Sulfur metabolism	33.3
Butanoate metabolism	25.0
Cholesterol biosynthesis 2	25.0
Glycosphingolipid biosynthesis-globo series	25.0
Phosphatidylinositol phosphate metabolism	25.0
Aminoacyl-tRNA biosynthesis	19.0
Fructose and mannose metabolism	16.7
Propanoate metabolism	16.7
ROS detoxification	16.7
Eicosanoid metabolism	16.7
Glycosphingolipid metabolism	16.7
Metabolism of other amino acids	14.3
Cholesterol metabolism	12.8
Tryptophan metabolism	12.5
Beta-alanine metabolism	12.5
Chondroitin / heparan sulfate biosynthesis	12.5
Amino sugar and nucleotide sugar metabolism	11.8
N-glycan metabolism	11.8
Pyrimidine metabolism	11.5
Pentose and glucuronate interconversions	11.1

Eicosanoid production has been described to normally have low constitutive levels [206], although eicosanoids are recognised as possibly having both pro- and anti- inflammatory roles in the immune response of T-cells [207]. Some eicosanoids are produced from eicosapentaenoic acid (EPA) and dihomo-γ-linolenic acid (DGLA), but most are derived from arachidonate [206]. These arachidonate-derived compounds include hydroxyeicosatetraenoic acids (HETEs), epoxides, hydroperoxyeicosatetraenoic acids (HPETEs), lipoxins (LXs), leukotrienes (LT), and prostanoids.

All the 6 genes that affect at least one reaction in the *Eicosanoid metabolism* once deleted catalyse reactions that produce several arachidonate-derived eicosanoids (table 7).

Table 7: Potential essential genes from the *Eicosanoid metabolism* pathway and respective products.

Gene	**Product(s)**
ALOX12B	12-HPETE
CYP2F1	12-HETE
CYP4F8	18-HETE
CYP4F12	10-HETE
LTC4S	Leukotriene C4
HPGD	Prostaglandin F2α /15-Keto-Prostaglandin F2A

As mentioned earlier, arachidonate-derived prostanoids and leukotrienes have been linked to proliferation, apoptosis, cytokine production, differentiation, and chemotaxis in T-cells [206].

CYP4F8 was tested in both CRISPR-CAS9 studies [236, 237] and LTC4S was tested in the CD8 T-cell related study [237]. Only LTC4S was considered essential.

Glycosphingolipids (GSLs) are present in the cell membranes and have been related to T-cell activation, differentiation and function [238]. A group of GSLs are (iso)globosides, whose synthesis is represented in the metabolic models through the metabolic pathway *Glycosphingolipid biosynthesis-globo series*, one of the most affected pathways by gene essentiality. Indeed, the synthesis of this group of GSLs, as well as other types, has been shown to occur in normal human T-cells [238, 239]. From the 8 genes in this pathway that were tested for gene essentiality, 4 were tested by ther CD8 T-cell related study (*A4GALT*, *B3GALNT1*, *ST8SIA1* and *GLA*) [237], and only 1 by the CD4 T-cell related study (*GLA*) [236]. Only Shifrut et al [237] reported essential genes, namely *A4GALT* and *B3GALNT1*.

Other highly affected pathways include *Butanoate metabolism*, which has been shown to play an important role in the cytotoxic capacity of in vitro mouse CD8+ T-cells when these cells were supplemented with low-dose butyrate [240]. In fact, one of the genes tested from this pathway was considered as essential for the cytotoxic CD8+ T-cell type (the gene was essential for more than 90% of this cell-type's models).

As mentioned before, hydrogen sulfide (H2S) is produced in the *sulfur metabolism* pathway. Knockout of genes that encode for H2S-producing enzymes was shown to decrease regulatory CD4 T-cell proliferation [218].

Some genes catalyse reactions from different metabolic pathways. So, it should be kept in mind that only the inactivity of some of those reactions affected by a gene deletion might have an actual effect on the model's objective. So, a pathway that is very affected upon gene deletions might not be 'causing' the inability of a model to perform its objective when that pathway is not fully working. Evaluating the essentiality of each reaction individually would aid in understanding what metabolic pathways could be seen as more important for a cell-type.

Due to having a lot of genes associated to it, the *Transport reactions* pathway does not come up as one of the most affected pathways when looking into the percentage of genes that affect the models' objective. However, it is the pathway that has the biggest number of genes considered as essential in each model. This was expected, due to the high number of potentially essential genes (201) that represent the transport of metabolites between compartments of a cell or the uptake of metabolites. Thus, we also checked the pathways with the highest median number of essential genes (table 8), as looking into the percentage of essential genes in the pathways can focus the analysis only on pathways with less genes.

The 20 pathways with the most median number of genes that are essential are crucial for cell proliferation. There are several pathways related to amino acids metabolism (*glycine, serine and threonine metabolism*, *valine, leucine and isoleucine metabolism*, *phenylalanine, tyrosine and tryptophan metabolism*, *arginine and proline metabolism*, *alanine, aspartate and glutamate metabolism*, and *cysteine and methionine metabolism*), DNA/RNA production (*nucleotide metabolism* and *pyrimidine metabolism*), and fatty acids (*fatty acid oxidation* and *fatty acid biosynthesis*). Other important pathways include *oxidative*

phosphorylation, *cholesterol metabolism*, and *aminoacyl-tRNA biosynthesis*.

Table 8: Top twenty pathways with the most median number of essential genes.

Pathways	**Number of essential genes**
Transport reactions	40.5
Nucleotide metabolism	11.0
Oxidative phosphorylation	8.0
Sphingolipid metabolism	7.0
Fatty acid oxidation	6.0
Cholesterol metabolism	5.0
Aminoacyl-tRNA biosynthesis	4.0
N-glycan metabolism	4.0
Steroid metabolism	4.0
Glycerophospholipid metabolism	4.0
Pyrimidine metabolism	3.0
Glycine, serine and threonine metabolism	3.0
Valine, leucine, and isoleucine metabolism	3.0
Phenylalanine, tyrosine and tryptophan biosynthesis	3.0
Fatty acid biosynthesis	3.0
Bile acid biosynthesis	3.0
Drug metabolism	3.0
Arginine and proline metabolism	2.5
Propanoate metabolism	2.0
Alanine, aspartate and glutamate metabolism	2.0
Cysteine and methionine metabolism	2.0

4.8 Effect of a tumour blood medium

We compared the pFBA predictions obtained using the normal human blood medium, constructed using information from the *Serum Metabolome DataBase* (SMDB), with those obtained using the tumour human blood medium. This tumour human blood medium was created from information on fold changes between the blood of normal and tumour patients. From the total of 537 metabolites in our normal human blood medium, we only found information for 84 of the metabolites (approximately 15.6%).

We performed a non-parametric paired statistical test, *Wilcoxon signed rank test*, to find which models' reaction fluxes differed significantly between the two types of medium. This was only observed for 53 out of the 116 models that returned feasible pFBA solutions (approximately 46% of the models, figure 38A). More than half of the models of naive CD4 (56.25%), memory CD4 (71.43%), follicular CD4 (55.56%), proliferative CD4 (57.14%), and proliferative CD8 (55.56%) T-cells significantly differed (figure 38B). No naive CD8 T-cell models were affected by the change in medium.

We next checked what pathways changed the most when the medium was changed. For this, we

selected the top 30 pathways with the highest median of ratio of affected reactions and plotted the distribution of this ratio across all models for each pathway (figure 38C). As expected, the most affected pathways are those directly affected by the metabolites whose concentration changed.

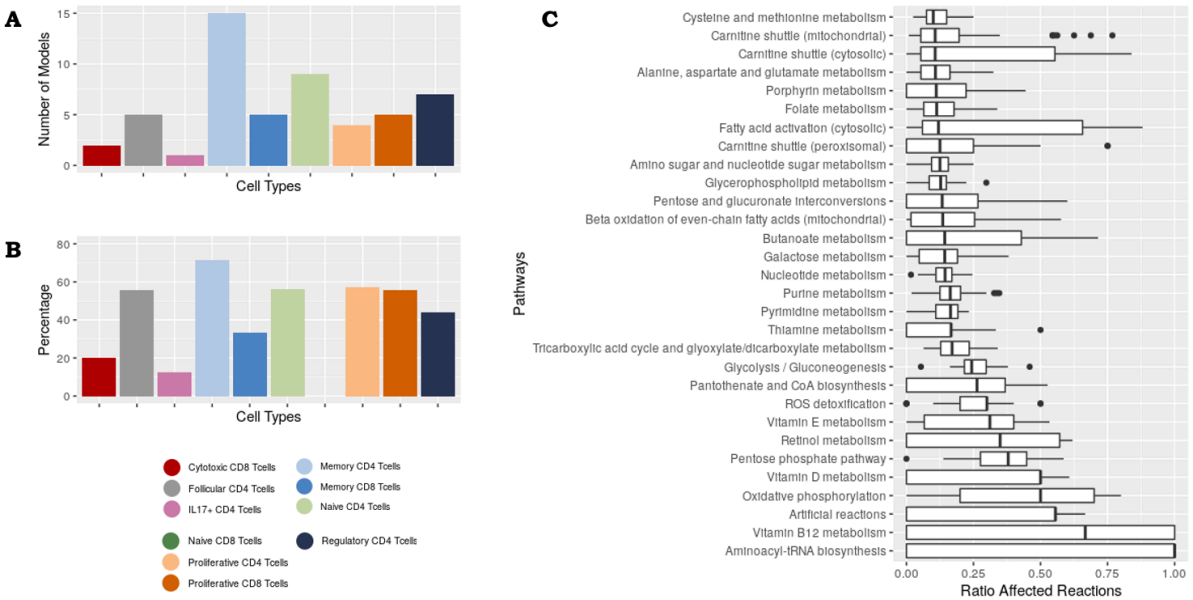


Figure 38: (A) Number and (B) percentage of models, by cell-type, whose metabolism is significantly different when the medium was changed to a tumour blood-like one. (C) Top 30 pathways that changed the most when the medium was changed.

The change in concentration of amino acids (asparagine, aspartate, glutamate, glutamine, glycine, histidine, isoleucine, leucine, lysine, methionine, serine, threonine tryptophan, tyrosine, and valine) affected the flux of a lot of the reactions from two pathways related to the metabolism of this type of metabolites: *alanine, aspartate and glutamate metabolism*; *cysteine and methionine metabolism*.

β-oxidation of even-chain fatty acids (mitochondrial) and *fatty acid activation (cytosolic)* pathways were affected. The concentration of four unsaturated FAs (elaidate, linolenate, nervonic acid, oleate) and two saturated FAs (lauric acid, myristic acid) changed in the tumour medium.

Moreover, *glycolysis / gluconeogenesis* was affected by changes in acetate, glucose, lactate, Pi, PEP and pyruvate levels, while the *TCA cycle* was affected by citrate, fumarate, glycerate, and malate.

Other pathways were affected not because the metabolites that enter those pathways suffered a change in concentration in the medium, but because of the change in other pathways. This is the case of *ROS detoxification*.

We finally checked if the flux of the biomass reaction also changed, considering that pathways like amino acids' metabolism, fatty acid oxidation, glycolysis, TCA cycle, and oxidative phosphorylation were some of the most affected by the change in medium.

Only 45 models (not necessarily part of those that had a significantly different metabolism) had different biomass flux. The biomass flux increased in 26 of these models (supplementary figure 63A). No naïve and cytotoxic CD8 T-cell models' biomass changed (supplementary figure 63D).

4.9 Discussion

The metabolic models of T-cells structurally resemble well the respective gene expression data, with not-so-distant predictive capabilities of the cell-types (figure 26). However, there are still a lot of similarities between the different types of T-cells, whether in gene expression (figure 19) or in the models' structure (figure 21). Pathways like cholesterol biosynthesis, heparan sulfate degradation, keratan sulfate degradation, chondroitin sulfate degradation, leukotriene metabolism, and fatty acid related pathways are all highly present across all T-cell models and known for being important for T-cell function.

Nevertheless, some groups of models show interesting differences in pathway coverage. For example, regulatory CD4 t-cell models from normal tissue samples seem to have less anti-inflammatory function due to less presence of pathways related to anti-inflammatory functions in T-cells (figure 22). Interestingly, we found models from CMS4 tumour samples to have less coverage of anti-inflammatory related pathways than the other tumour models. These regulatory CD4 T-cell CMS4 models also grouped closer to the normal-derived models, suggesting that regulatory CD4 T-cells from CMS4 tumours are not able to be as anti-inflammatory as they could. Further corroborating the less anti-inflammatory function of normal-derived regulatory CD4 T-cells is the close clustering of these models with those of IL17+ CD4 T-cells (figure 24). Normal-derived regulatory CD4 T-cell models also have visibly less reactions than their tumour-counterparts (figure 20A), with models from CMS4 tumour being closer to normal-derived ones than the other tumour-derived models (figure 20B). When performing flux prediction with the pFBA approach, the biomass of normal-derived models, as well as of those from CMS4 tumours, of regulatory T-cells was null or very close to null, unlike those from other tumour-derived models (figure 28). All of this suggests that normal-derived and CMS4 regulatory CD4 T-cell models are also generally less metabolically active than the other regulatory CD4 T-cell models.

A good structure that correctly resembles the experimental data does not necessarily mean that predicted fluxes will be good, however. Especially when knowing that these models do not take into consideration enzymatic constraints and gene / metabolic regulation, which are important aspects of a cell that greatly affect its metabolism. As such, flux predictions are still known to be quite imprecise [178], which is noticeable in our cell-type prediction results using reaction fluxes vs gene expression and reaction presence (figure 26).

Nevertheless, the prediction of major metabolic aspects related to T-cells were correct. It was shown that the median biomass of proliferative T-cell models was higher than the naïve T-cell models, and that most models (114 out of 142 models with feasible pFBA solutions) have relatively high biomass and low ATP production or vice-versa.

Fatty acid oxidation (FAO) is the pathway that contributes the most to produce FADH2 in all models. While models from non-proliferative T-cell models rely mostly in FAO for NADH production, proliferative T-cell models' biggest source of NADH is glycolysis. Proliferative T-cells, characterised by relying in fatty acid synthesis instead of uptake (figure 1), showed to have null or very close to null uptake of fatty acids in the pFBA predictions (figure 32A).

The individual deletion of tryptophan and glutamine from the medium decreases T-cell's biomass to or very close to zero, while the deletion of nucleotides did not affect the biomass, as expected [226, 227, 234]. Even though biomass was expected to decrease, the deletion of glucose did not affect the biomass of the models, which could be explained by existence of alternative metabolites in the medium that are not present in *in vitro* studies and may reduce the dependence on glucose.

All flux predictions mentioned above were made using a normal human blood medium. Thus, we tried to construct a medium as close as possible to what a tumour blood medium could be. As there were no quantitative measurements of a tumour blood medium, unlike the normal one, we constructed the tumour medium looking into studies that compared peak intensities from mass spectrometry between normal and tumour blood samples. We were able to change only 84 of the 537 metabolites we have on the normal medium, which might explain why only 46% of the models differentially changed with the new medium. Another reason for a low percentage of models that differed might be the fact that cells always try to adjust so that the metabolism differs as little as possible and this was captured by the models. Nevertheless, the most affected pathways were those directly affected by the metabolites whose concentration changed. A better quantitative characterisation of a tumour blood medium would be very helpful in the future to better estimate the reaction fluxes of these cell-types in the colorectal cancer environment.

5

Benchmarking of Tumour Deconvolution Methods

Although scRNAseq gives valuable information about cell-types and states at single-cell resolution, this technique is not an accurate representation of cell-type proportions in a sample. Bulk RNA sequencing, on the other hand, is used for diagnostic purposes in routine clinical settings, allowing an unprecedented amount of data describing tumours.

Cell-type profiles can be recovered from bulk RNAseq data and used for cell-type metabolic model reconstruction. The proportions can aid in the construction of community models when modelling the tumour micro-environment.

Having this in mind, we tested and compared several tumour deconvolution methods that were developed to use scRNAseq data as reference to estimate cell-type proportions. Our CRC atlas of scRNAseq data was used.

5.1 Methods

5.1.1 Bulk RNAseq Deconvolution

The mixed gene expression profile of a tumour micro-environment can be seen as the linear combination of the gene expression profiles of each cell type in that micro-environment and the respective proportions of those cell types. This relation can be portrayed by equation 5.1:

$$A.X = B \tag{5.1}$$

Here, B represents a matrix of the gene expression profile for a tumour micro-environment, the bulk data, with an expression value for each gene (rows) in each sample (columns). A is a matrix representing the cell type gene expression profiles, with an expression value for each gene (rows) in each cell type (columns). This matrix is often referred to as reference matrix. Genes in A are the same as those in B. Finally, X represents the proportions of these cells in the micro-environment. This matrix contains the proportion of each cell type (rows) for each sample (columns).

In this chapter, we focus on deconvoluting bulk RNAseq data (B) from CRC samples to obtain the cell-type proportions (X).

5.1.2 Bulk data and ground-truth proportions

We used in-house bulk RNAseq tumour samples from CRC patients, gathered at the Leiden University Medical Center (Leiden cohort) [241] and whose cell counts were calculated by using Hyperion mass cytometry imaging. A total of 21 samples were available.

The ground-truth phenotypes were divided into 9 different cell-types plus a group named 'other cells' (see section 5.1.3). To calculate the proportion of a cell-type in a sample (figure 39B), we aggregated the cell counts of all phenotypes of that cell-type (figure 39A and supplementary table 19) and divided them by the total number of cells in the sample.

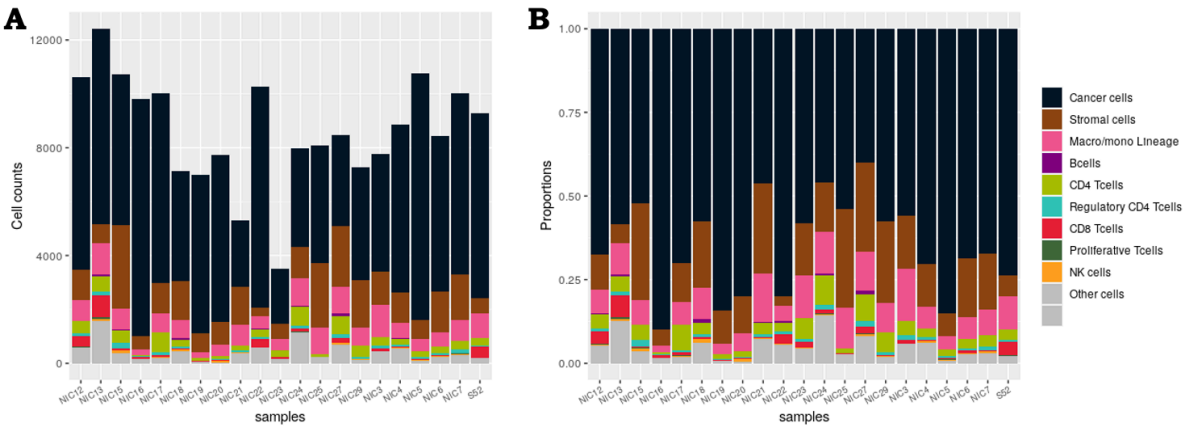


Figure 39: Ground-truth (A) cell counts and (B) proportions of the cell-types used for deconvolution.

5.1.3 CRC atlas as the reference data

The CRC atlas of scRNAseq data constructed was used to build the reference matrix. Usually, scRNAseq data from peripheral blood mononuclear cells (PBMCs) is used as reference for the deconvolution of tumour samples. However, this will always lead to incorrect estimation of immune cell-type proportions: (1) gene expression profile of immune cells from peripheral blood might not represent well the expression of tumour-infiltrating immune cells; (2) several cell-types present in the tumour are not PBMCs and are thus not accounted for when deconvoluting the samples (all methods normalize the proportions estimation to sum to 1).

Mapping cell-types between CRC atlas and ground-truth proportions To test the different deconvolution methods, we had to find the best overlap between the atlas's cell-types and the ground-truth's cell phenotypes. While doing so, we found the best balance between having appropriate cell-types matched and enough subtype detail to deconvolute the most important cell subtypes. All cell-types /

phenotypes that did not have a match between the atlas and the ground-truth were simply grouped in the *Other cells* cell-type.

We grouped the cell-types in the CRC atlas, as well as the ground-truth phenotypes, into 10 different cell-types (more details in supplementary table 19): cancer cells, stromal cells, macro/mono lineage, B-cells, CD4+ T-cells, regulatory T-cells, CD8+ T-cells, proliferative T-cells, NK cells, and other cells.

Subset of original CRC atlas For computational reasons, and because one of the methods tested (*CIBERSORTx*) has a file quota of 1000 MB for each of their website users, a subset of the original atlas was created and used for all methods, including the calculation of cell-type markers (necessary for some of the methods). Also, only the cells from tumour samples were kept, as we only used tumour bulk RNAseq samples for deconvolution.

A maximum of 1 200 cells per cell-type were kept, which reduced the number of cells to a total of 11 187: 904 for NK cells; 683 for proliferative T-cells; and 1 200 for each of the remaining cell-types. The percentage of cells per patient was maintained.

Cell-Type Markers For each of the 10 cell-types, the cells from that cell-type were compared to all others to search for gene markers of that cell-type. We used *Seurat* (R package, v.4.0.3) [157] for this. To reduce the genes to test, only those with a log2 fold-change value, on average, greater than 0.8 between the cell-type's cells and all others were analysed. Also, genes had to be expressed in at least 30% of the cells in the cell-type. The *Wilcoxon Rank Sum* test was used. A gene was considered a marker if the respective p-value was smaller than 0.01. A total of 1169 different genes were obtained.

5.1.4 Methods evaluated

We compared a total of 10 methods. In this section, we give a summary of what these methods can do. Generally, the methods can be separated into two major groups: those that try to solve the deconvolution equation 5.1, and thus are non-negative matrix factorization (NMF) based algorithms, and those that are based on neural networks to predict the proportions (and thus do not necessarily use the deconvolution equation 5.1). Only two of the methods evaluated belong to the last group of methods: *DigitalDLSorter* and *Scaden*.

AutoGeneS (python package, v.1.0.4) [242] starts by selecting the most interesting genes to use in the deconvolution by simultaneously minimizing the correlation and maximizing the distance between cell-types. The signature matrix is generated with only those genes and the proportions are estimated by minimising the regression error E: $A.S.X + E = B$. Three different regression models can be used to perform this prediction: *NuSVR*, *non-negative least squares* and *linear*. This method accounts for the mRNA content bias, where differing cell size, and thus varying per cell RNA content, can affect the correct estimation of cell-types proportions (see section 5.1.5). *AutoGeneS* uses the average number of mRNAs in each cell-type (S in mentioned equation) to try to correct for this bias.

BisqueRNA (R package, v.1.0.5) [243] starts by generating the signature matrix by averaging read counts within each cell-type in the single-cell data. Together with the proportions of the cell-types in the single-cell data, the method then learns gene-specific transformations of the bulk data to account for technical biases between single-cell and bulk technologies. *BisqueRNA* then estimates the bulk RNAseq data proportions using the signature matrix and the transformed bulk data.

BSeqSC (R package, v.1.0) [244] generates a signature matrix to be used by *CIBERSORTx* (method explained below) to estimate the proportions of a bulk RNAseq dataset. When this method was first developed, and due to licensing requirements, the source code for estimating the proportions needed to be downloaded separately from *CIBERSORTx* website. However, it seems that new releases of *CIBERSORTx* no longer have the source code available. As such, we generated the signature matrix through this R package and then uploaded it to *CIBERSORTx* website to perform the estimation. Regarding the generation of the signature matrix from the scRNAseq data, the gene expression is averaged across all cells within each cell type. A list of genes known to be markers of the cell-types in question must be given so that the signature matrix is only calculated with those. Optionally, the single-cell counts can be re-scaled before computing the average gene expression. If so, the data is transformed into counts per million (CPM) and re-scaled using the cell-types' average counts.

CIBERSORTx (Website) [245] first sets to zero the counts of genes whose average expression in a log2 space is low, to then aggregate cells from the same cell-type by summing the counts in non-linear space from a few cells and transforms the summed counts into CPM counts. This aggregation process is repeated several times to generate several transcriptome replicates per cell-type. The genes used in the final signature matrix are those that are differentially expressed in at least one of the cell-types. This method gives the opportunity to optionally handle the technical variation between the signature matrix (single-cell RNAseq) and the bulk mixture (bulk RNAseq) while preserving biological signal. *CIBERSORTx* then estimates the bulk RNAseq data proportions using the signature matrix and the bulk data.

DigitalDLSorter (R package, v.0.1.1) [246] trains Deep Neural Network (DNN) models with bulk RNAseq samples simulated from aggregated and pre-characterised scRNAseq data and whose composition is thus known. The final trained model is then used to estimate the proportions of true bulk RNAseq samples. Furthermore, this method can simulate new single-cell profiles from real ones to increase signal and variability in small datasets or in those with under-represented cell-types.

DWLS (R script) [247] starts by calculating the cell-types' differentially expressed genes using *Seurat* (R package, v.4.0.3) [157] under the *bimod likelihood ratio* test. After that, several candidate signature matrices with 50 to 200 marker genes, where the expression values of the genes are averaged across each cell-type, are tested. The best signature matrix corresponds to the candidate with the lowest condition number. The signature matrix is then used to estimate the cell-types' proportions of the bulk RNAseq data by solving the deconvolution equation 5.1 through a weighted least squares approach.

MOMF (R package, v.0.2.0) [248] aggregates cells from the same cell-type by summing the raw counts. These cell-type profiles are normalised by dividing them by the multiplication of the corresponding cell-type's total counts with a weight factor (proportion of total cell-type counts in the total counts of the scRNAseq data). *MOMF* then estimates the bulk RNAseq data proportions using the signature matrix and the bulk data.

MuSiC (R package, v.0.2.0) [249] starts by choosing genes with low cross-subject variance, critical for transferring cell type-specific gene expression information from one dataset to another. Prediction of proportions for closely related cell-types is difficult, due to their gene expression being closely correlated. Because of this, *MuSiC* optionally employs a tree-guided procedure that recursively zooms in on closely related cell types to estimate the proportions.

Scaden (python package, v.0.9.4) [250] uses a deep neural network ensemble trained on artificial bulk data simulated with a given scRNAseq reference to infer the cellular composition of the samples. The final trained model is then used to estimate the proportions of true bulk RNAseq samples.

SCDC (R package, v.0.0.0.9000) [251] is an ensemble method that combines the deconvolution results from different scRNAseq reference datasets. The references that better recapitulate the true underlying gene expression profiles of the bulk samples are given higher weights when integrating the different results together to produce a final estimation.

5.1.5 RNA content bias correction

Cell-types have differing cell size and thus varying per cell RNA content. For example, a monocyte usually has more mRNA content than a T-cell. This can affect the correct estimation of the cellular proportions of the cell-types, as methods will most likely estimate the RNA proportions of the cell-types instead.

Previous studies have tried to address this matter, by applying RNA content bias correction on the estimated proportions [252] or on the reference matrix, prior to the estimation of proportions [253]. In both cases, this correction is applied using *cell-type factors*, which are values that represent the relative RNA content per cell for the cells in the mixture.

We decided to test each method with and without RNA content bias correction. Both ways of correcting RNA content bias were evaluated. RNA content bias correction was not tested on methods *AutoGeneS* and *MuSiC_woGrouping* (*MuSiC* version where the tree-guided procedure is not performed when estimating the proportions), as these methods were already implemented by taking into consideration the RNA content bias.

When applying the correction to the reference matrix created from the scRNAseq data provided, the following was applied [253]:

$$(r^T.c)^T \tag{5.2}$$

where r is the genes x cell-types reference matrix and c the cell-type factors vector.

To apply correction on the estimated proportions, the following was applied [252]:

$$\frac{c^{-1}.f}{\sum c^{-1}.f'} \tag{5.3}$$

where c is the cell-type factors vector and f the estimated proportions.

To obtain the factor of a cell-type, the average of total gene counts of all cells of the scRNAseq atlas in that cell-type was calculated (table 9).

Table 9: Bias factors used for RNA content bias correction.

Cell-types	Bias factor
Cancer cells	16 865.324
Stromal cells	9 361.539
Macro/mono lineage	6 323.148
Proliferative T-cells	4 863.001
B-cells	3 581.911
Regulatory T-cells	2 979.075
CD4+ T-cells	2 946.983
NK cells	2 738.166
CD8 T-cells	2 514.997

5.1.6 Comparison of methods

The methods were run using default parameters or, when available, the parameters advised by the authors. For *AutoGeneS*, the three different regression models were tested (*NuSVR*, *non-negative least squares* and *linear*), and *MuSiC* was run using the tree-guided procedure (named from here on *MuSiC_wGrouping*) and without it (named from here on *MuSiC_woGrouping*). All methods were run through R language version 4.1.0. For python based methods, they were run with the aid of the R package *reticulate* (v.1.2.0) [254], an R interface to Python modules. More detailed information, including scripts, is available in the GitHub project `https://github.com/saracardoso/Tumour_Deconvolution`.

The pearson correlation and the root mean square error (RMSE) between estimated and ground-truth proportions were the metrics used to assess how good the methods were. We chose both metrics and not just one for three main reasons (figure 40).

If the proportions are perfectly predicted (i.e., estimations equal ground-truth and overlap the y=x line), we have a correlation of 1 and an RMSE of 0 (figure 40A). However, in situations where the predicted proportions vary in parallel to the y=x line (figure 40B), the correlation will still be 1, even though the proportions were not well estimated. In this case, RMSE will be greater than zero. In a last hypothetical case, if all proportions are predicted 0 and the ground-truth proportions are very close to zero, the RMSE will be very close to zero, giving the impression that the proportions are being estimated with very little error (figure 40C). However, the method was not able to detect the cell-type(s) in question. In this case, the correlation cannot be calculated (originating a missing value) due to null standard deviation.

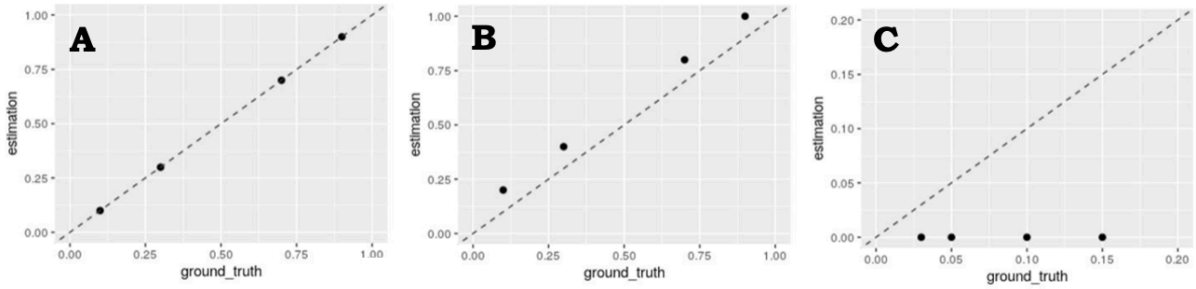


Figure 40: Examples of three different situations of estimated vs ground-truth proportions. (A) correlation is 1, while RMSE is 0. (B) correlation is 1, but RMSE is 0.1. (C) correlation cannot be calculated (as all predictions are 0), but RMSE is 0.09.

5.2 Results

5.2.1 *CIBERSORTx*, *DigitalDLSorter* and *Scaden* are the best methods overall

Overall, three methods standout as the best (figure 41): *CIBERSORTx*, *DigitalDLSorter* and *Scaden*. These methods have a RMSE smaller than 0.12 and a correlation that goes above 0.8. Closely followed by *AutoGeneS_nusvr* and *DWLS*, the remaning methods have a poor correlation (under 0.42) and most have a RMSE greater than 0.2. *AutoGeneS_linear* is by far the worst method. Scatterplots of the estimated proportions against the ground-truth for the best methods and the worst one are present in figure 42, while the remaining methods are in supplementary figure 64.

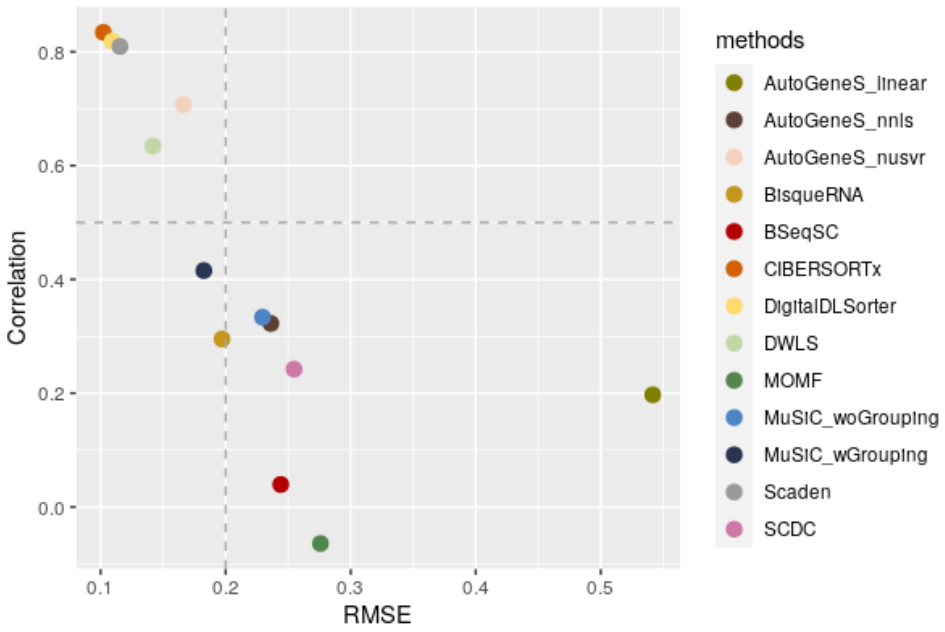


Figure 41: Overall correlation vs RMSE for all methods. A good method has a high correlation and a small RMSE. Grey horizontal and vertical lines mark a correlation of 0.5 and a RMSE of 0.2, respectively.

We imagined that just because a method is, overall, one of the best, does not mean that it is the best at predicting every single cell-type. In fact, a good estimation regarding some of the cells might mask those others that are very poorly estimated. Also, the cell-types that are constantly more present in the samples, like cancer cells, affect more the overall RMSE and correlation than those present in proportions almost close to zero. This is noticeable when plotting the estimated proportions against the ground truth (figures 42 and supplementary figure 64). Thus, it is a good idea to evaluate methods deconvolution ability by looking into each cell-type separately.

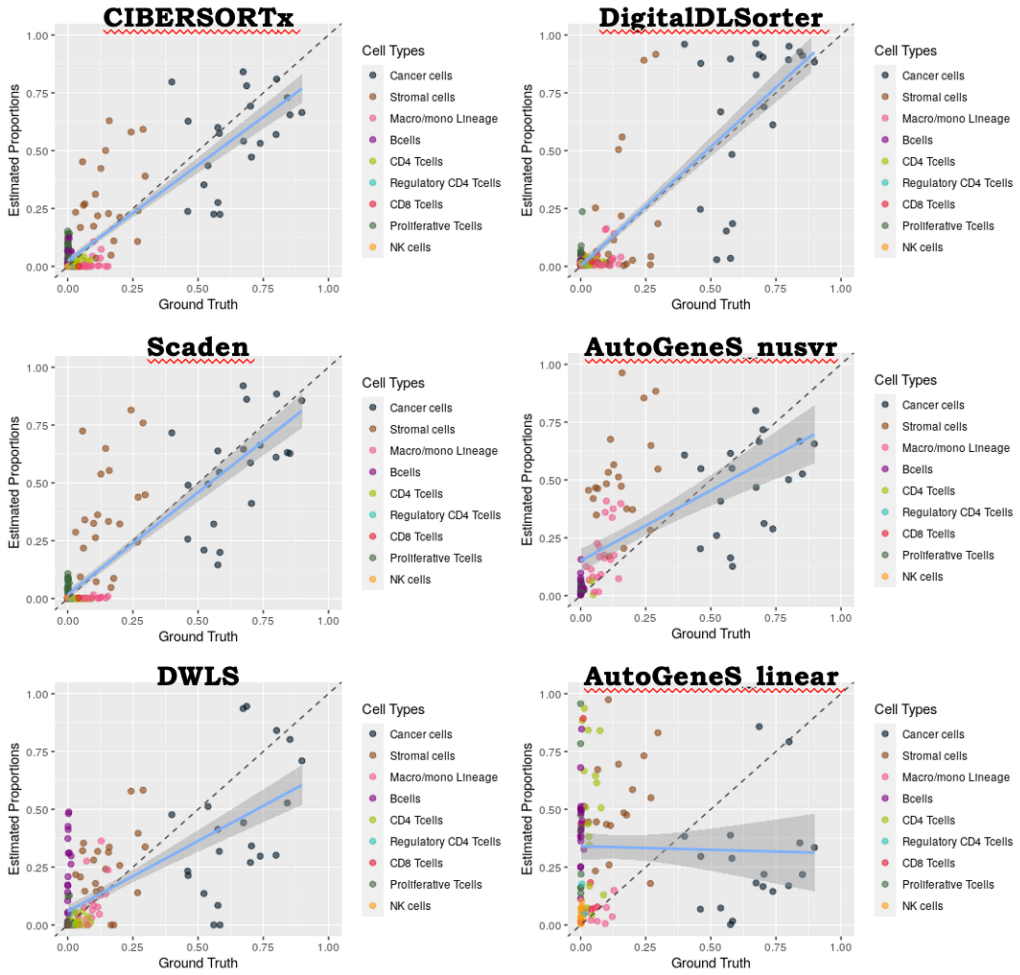


Figure 42: Scatter plots of estimated vs ground-truth proportions for the three clear best methods (*CIBERSORTx*, *DigitalDLSorter* and *Scaden*), the two other good methods (*AutoGeneS_nusvr* and *DWLS*), and the worst method (*AutoGeneS_linear*) overall.

5.2.2 Other methods can be better at predicting a cell-type individually

Per cell-type metrics do show that in overall metrics a good estimation regarding some of the cells masks those others that are very poorly estimated (figure 43). This is especially evident for *AutoGeneS_linear*. This method scores an RMSE of almost 1 for regulatory CD4 T-cells, but for the macro/mono lineage cells it is smaller than 0.2. The remaining cell-types have scores between 0.5 and 0.75, which is closer to the

overall RMSE of this method (0.542). Regarding correlation, it is evident that no method scores higher than 0.8 in any cell-type, although the three best methods overall have a correlation greater than 0.8.

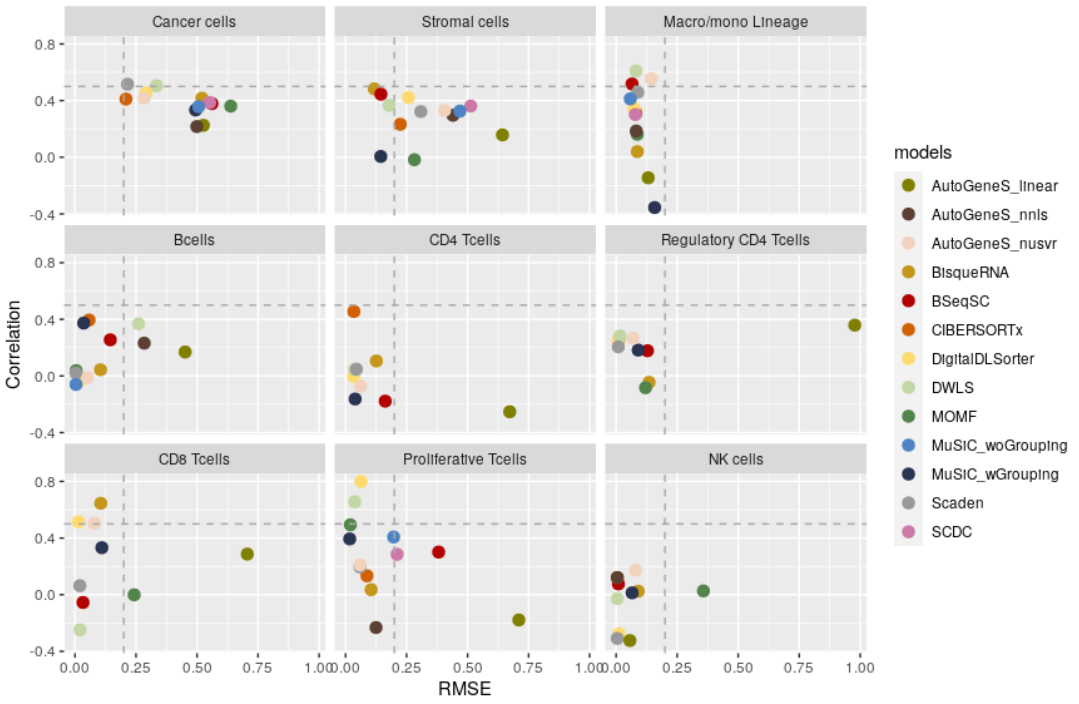


Figure 43: Cell-type correlation vs RMSE for all methods. A good method has a high correlation and a small RMSE. Grey horizontal and vertical lines mark a correlation of 0.5 and a RMSE of 0.2, respectively.

We next evaluated which method is the best for predicting each cell-type individually (table 10). For this, we focused on RMSE and correlations metrics (figure 43), but also on visual inspection of the estimated versus ground-truth proportions (supplementary figures 65 to 73).

Table 10: Best methods for each cell-type.

Cell-type	**Best Method**
Cancer cells	*Scaden*
Stromal cells	*BisqueRNA*
Macro/mono lineage cells	*BSeqSC*
B-cells	*Scaden*
CD4 T-cells	*CIBERSORTx*
Regulatory CD4 T-cells	*Scaden*
CD8 T-cells	*DigitalDLSorter*
Proliferative T-cells	*MuSiC_wGrouping*
NK cells	*Scaden*

The best method to estimate cancer cells is *Scaden*, followed by *CIBERSORTx*, *DigitalDLSorter*, *AutoGeneS_nusvr* and *DWLS* (figure 43 and supplementary figure 65). These are the methods that are the

best overall. As no other cell-type shows these methods as the clear best ones, this shows that the cell-types that are constantly more present in the samples affect more the overall RMSE and correlation of a method.

While stromal cells (figure 43 and supplementary figure 66) are best estimated using *BisqueRNA*, followed by *BseqSC*, *BseqSC* is the best one at estimating the proportions for the cells from the macro/mono lineage (figure 43 and supplementary figure 67).

Even though *MuSiC_woGrouping* and *CIBERSORTx* have better correlation and slight worse RMSE at estimating B-cells (figure 43), the estimations versus the ground-truth plots (supplementary figure 68) show that *Scaden* is the right choice. The very low RMSE is not due to all samples being estimated as zero and the bad correlation is highly influenced by two samples, without which *Scaden*'s correlation would be better and with which *MuSiC_woGrouping* and *CIBERSORTx* have better correlation.

When assessing the methods' RMSE and correlation metrics for regulatory CD4 T-cells (figure 43), *DWLS*, *AutoGeneS_nusvr* and *Scaden* are the three best. However, visually comparing estimated and ground-truth proportions (supplementary figure 70) shows that *AutoGeneS_nusvr* gives negative proportions, while most samples were predicted by *DWLS* to not have regulatory CD4 T-cells. *Scaden* underestimates regulatory CD4 T-cells, but is still able to detect them. Regarding CD4 (figure 43 and supplementary figure 69) and CD8 T-cells (figure 43 and supplementary figure 71), the best methods are *CIBERSORTx* and *DigitalDLSorter*, respectively.

With better RMSE than *DigitalDLSorter*, *DWLS* and *MOMF*, *MuSiC_wGrouping* is the best method to predict proliferative T-cells (figure 43 and supplementary figure 72). Even though they have better correlation than *MuSiC_wGrouping*, *MOMF* only detects the presence of proliferative T-cells in 4 samples, *DWLS* returns negative proportions for 4 of the samples, and *DigitalDLSorter* over-estimates proportions a lot more than the previous methods.

NK cells is by far the most difficult cell-type to estimate, with no method performing particularly well (figure 43 and supplementary figure 73). *Scaden*, followed by *DigitalDLSorter*, seem to be the methods that better capture NK cells proportions, even though the respective correlations are negative.

With these results in mind, we decided to see what would happen if a pipeline using the best method for each cell-type was used (figure 44). Combining the best methods for each cell-type creates better overall estimations, with a correlation that now surpasses 0.9 and a RMSE smaller than 0.1 (figure 44A). We tested how different the results would be between a simple combination of the results from the different methods (*Combined*) and normalising the combined results of each sample to sum to 1 (*Combined_norm*, figures 44C and 44D). Overall, normalising the estimation leads to better results, although the difference is of only 0.013 and -0.001 for the correlation (0.925 - 0.912) and RMSE (0.085 – 0.086) respectively.

We tried to find an independent dataset for colorectal cancer that also provided information on cell-type proportions based on mass cytometry to further corroborate these findings. However, we could not find one. Nevertheless, we checked the overall correlation and RMSE by sample, to check how many samples benefit from a combined pipeline as opposed to the *Scaden* method (figure 44B). Indeed, most samples (76%) were better estimated using the combined approach. The samples that greatly benefited

were NIC13, NIC24 and NIC4, which showed some of the worst estimations throughout most methods (supplementary figure 85) including *Scaden*. The samples that seem to have not benefitted from the combined pipeline were NIC16, NIC20, NIC27, NIC6, and NIC7. Still, the metrics did not change much.

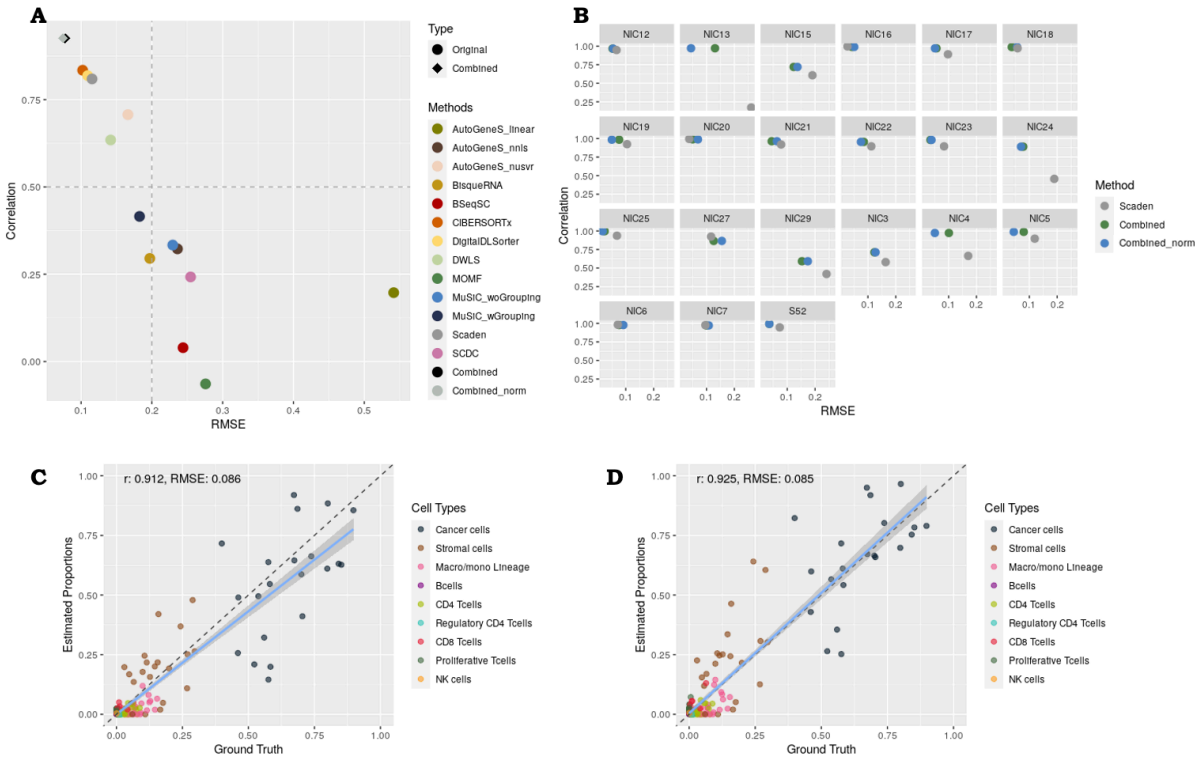


Figure 44: (A) Overall correlation vs RMSE for all methods, including the methods combining the best methods for each cell-type (*Combined* and *Combined_norm*). (B) Sample correlation vs RMSE for *Scaden*, *Combined* and *Combined_norm*. A good method has a high correlation and a small RMSE. Grey horizontal and vertical lines mark a correlation of 0.5 and a RMSE of 0.2, respectively. Scatter plots of estimated vs ground-truth proportions for (C) *Combined* and (D) *Combined_norm*.

5.2.3 RNA content bias correction does not improve predictions

Both RMSE and correlation metrics (figure 45) of predictions with correction for mRNA content bias do not show improvement from those without any correction, even though the decrease in prediction ability is not too steep. There is only one method with improved predictions upon correction on the reference matrix (*Corrected Before*). This method is *Scaden*, but the improvement is not high and, in both cases, (correction and no correction) *Scaden* is one of the three best methods. The improvement observed in the *Scaden* method seems to be mostly due to an improvement in predicting the proportions of the samples NIC3, NIC4, NIC13, NIC15, NIC24 and NIC29 (supplementary figure 86). Furthermore, the methods that already have correction for RNA content (*AutoGeneS_linear*, *AutoGeneS_nnls*, *AutoGeneS_nusvr* and *MuSiC_woGrouping*) have some of the worst predictions (figure 45 and 41), with only *AutoGeneS_nusvr* showing satisfactory results overall.

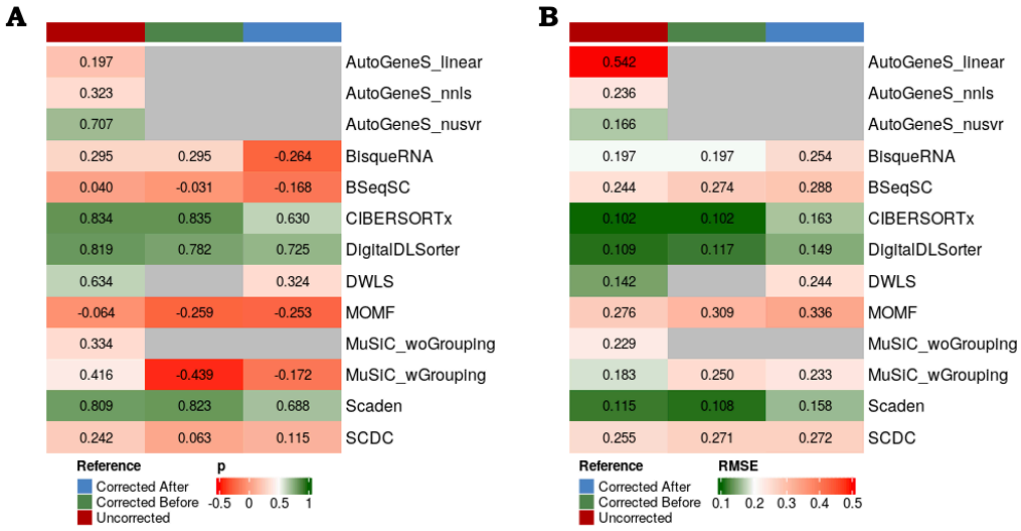


Figure 45: Pearson correlation (A) and (B) RMSE values of the methods without and with correction. Methods were tested for correction on the reference matrix (*Corrected Before*) and on the estimated proportions (*Corrected After*).

This goes against previous findings [252, 253], where correction for mRNA content bias was shown to improve deconvolution. However, it is noteworthy that Zaitsev *et al* [252], which tested the correction for this bias on the predicted proportions, used a dataset created with only two cell-types whose total RNA content is considerably different and was measured by analysing bulk RNA data of pure samples. The small number of cell-types, the big difference in RNA content and its calculation from the exact same cells used for creation of the samples used in deconvolution might be decisive for a good correction. In our case, there are much more than 2 cell-types (but the exact cell-types present in the bulk samples is not known), and the total RNA content is known through a scRNAseq dataset independent from the bulk samples. Beyond this, Zaitsev *et al* [252] developed a complete deconvolution method, i.e., a deconvolution method that does not use prior information on cell-type gene expression to estimate the proportions. Interestingly, the predictions with correction for RNA content on the estimated proportions lead to the worst results overall in our work.

Sosina *et al* [253] tested the correction for this bias on the reference matrix using the *MuSiC* method with default parameters. In that study, the single-cell reference used was obtained using the Fluidigm C1 system, which normalises cDNA libraries to the same concentration prior to sequencing and thus removes potential variability in RNA abundances across cell-types. The datasets collected in this work were constructed using 10XGenomics.

If we explore the results obtained without any correction for RNA content for each cell-type individually, it is possible to see that the biggest cell-type, cancer cells, are not over-estimated in any of the methods that does not use correction (supplementary figure 65). In fact, all methods that already come with RNA content bias correction (*AutoGeneS* and *MuSiC_woGrouping*) under-estimate cancer cells. It is noteworthy, however, that *Scaden* slightly improves the estimation of cancer cells upon correction of the estimated

proportions (figure 46 and supplementary figure 74), whose uncorrected version was the best method at predicting cancer cells.

The only other two cell-types whose RNA content bias correction seems to help is stromal and macro/mono lineage cells.

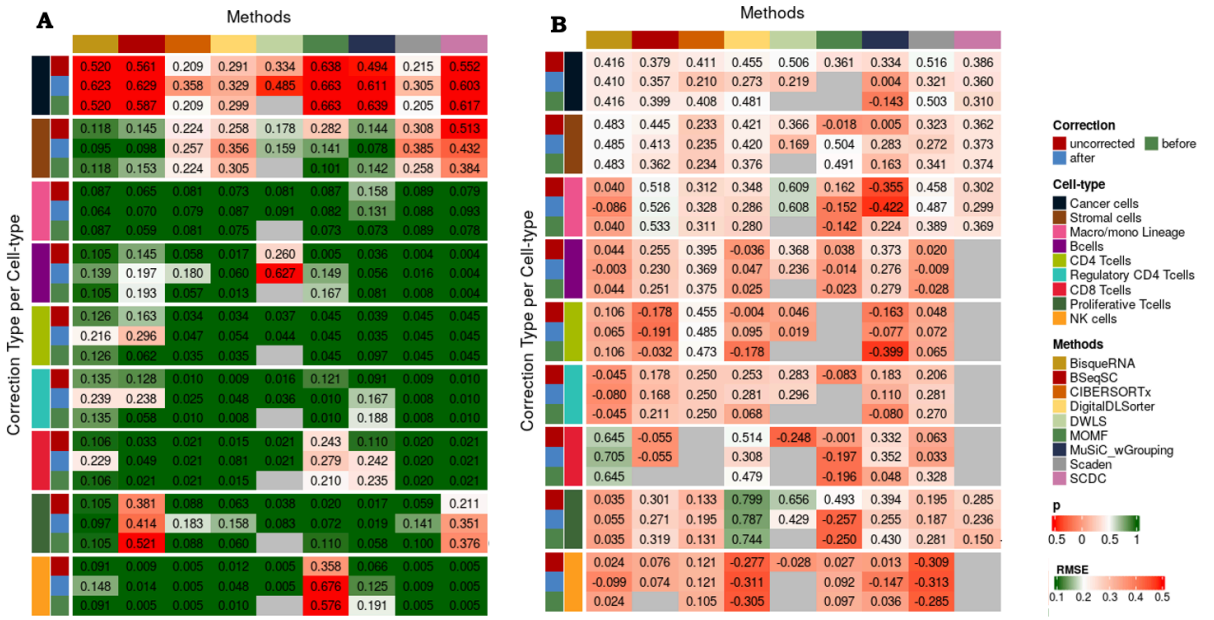


Figure 46: RMSE (A) and (B) pearson correlation values of the methods without and with correction, separated by cell-type. Methods were tested for correction on the reference matrix (*Corrected Before*) and on the estimated proportions (*Corrected After*).

Stromal cells was the cell-type with the most amount of methods where correction for RNA content improved the predictions (7 out of 9 methods – (figure 46A). Of these, only 3 showed better RMSE than the best uncorrected method for this cell-type. Comparing the estimated and ground-truth proportions (supplementary figure 75), however, shows that the uncorrected method (*BisqueRNA*) might still be the best option to predict stromal cells proportions.

For the macro/mono lineage cell-type, *BsqesSC* corrected for RNA content on the reference matrix shows visibly better results than its uncorrected version (figure 46 and supplementary figure 76).

Regarding the remaining cell-types, some methods benefited from this correction, but none showed better results than the uncorrected method that was best for the respective cell-type (figure 46 and supplementary figures 74 to 82).

5.2.4 Effect of real cell-types' proportions in samples estimations

Finally, we were interested in assessing how the amount of cells or RNA content in the samples affected the correct estimation of the samples' proportions.

The total number of cells in a sample seems to not affect much its predictions for most methods (figure 47A). For *AutoGeneS_nnls*, *CIBERSORTx*, *MuSiC_wGrouping*, however, the greater the number of

cells, the worse the RMSE of the sample. The correlation between the RMSE and the total number of cells is bigger than 0.34 in these methods. Even though there is some positive correlation between the total number of cells in a sample and its total number of reads (0.309, supplementary figure 83), the samples' RMSEs from *AutoGeneS_nnls* and *CIBERSORTx* improve with the increase of total read counts (-0.316 and -0.233, respectively). The total number of reads also positively affects other methods (figure 47B), like *AutoGeneS_nusvr* (-0.481), *MuSiC_woGrouping* (-0.478) and *SCDC* (-0.443).

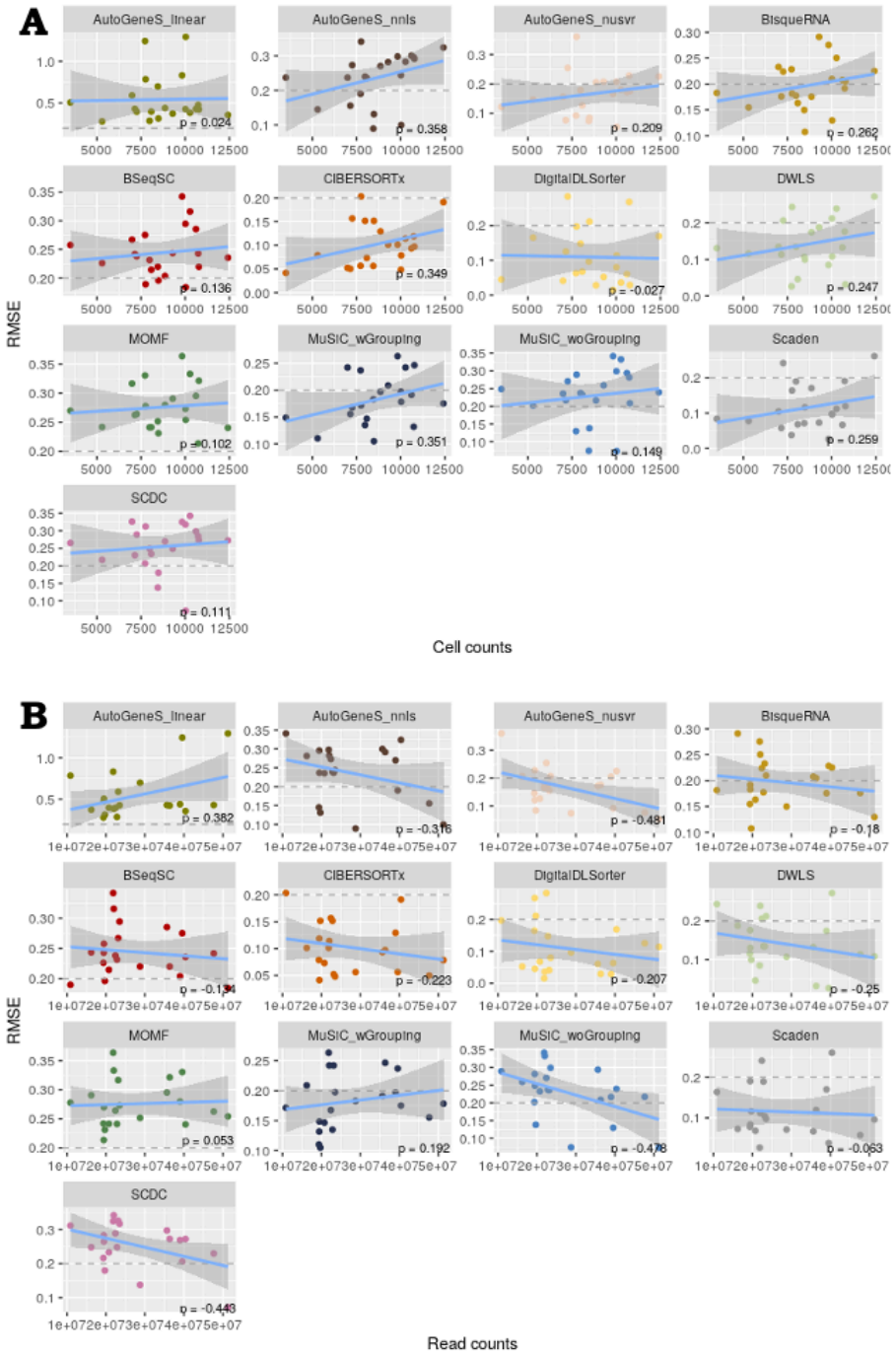


Figure 47: For each method, RMSE of the samples is mapped against their corresponding (A) total number of cells and (B) total number of read counts.

Even though the total number of cells or reads might not seem to affect much the overall prediction of samples' proportions, the proportion of certain cell-types in the samples might affect such predictions. This was hinted before, with cancer and stromal cells being the cell-types that most affect the overall predictions of the methods (figures 42 and supplementary figure 64).

With different levels of correlation, the best methods to predict cancer cells are positively affected by the increase of the proportions the cancer cells in the samples (figure 48). Apart from *AutoGeneS* (-0.034) and *DWLS* (-0.157), these methods have a correlation smaller than -0.3. *DigitalDLSorter* is the most affected one, with a correlation of -0.729. All other models, which did not perform as well, are negatively affected by the increase of cancer cells proportions in the samples, especially *MuSiC_wGrouping* (0.98), *BisqueRNA* (0.661) and *MOMF* (0.885).

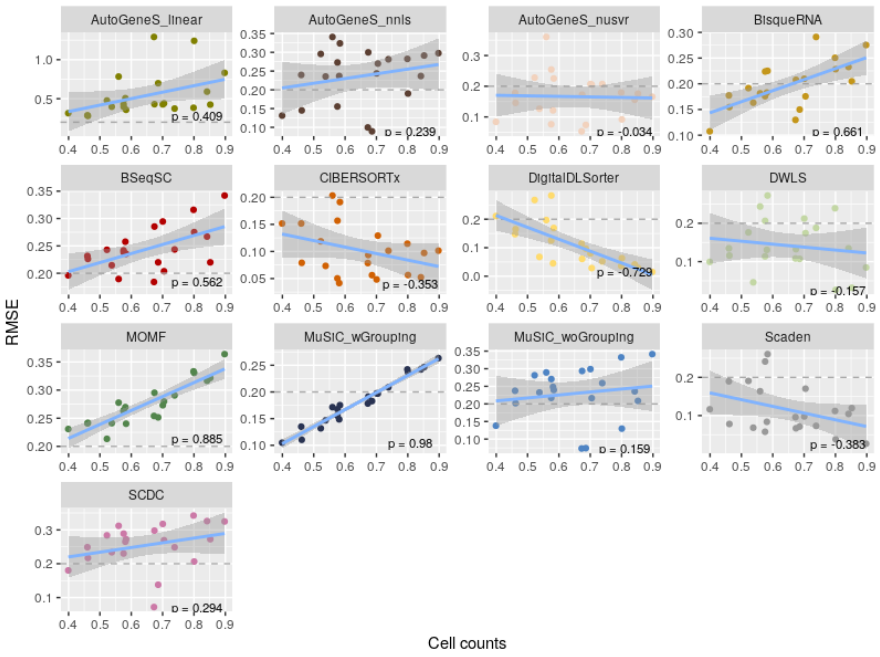


Figure 48: For each method, RMSE of the samples is mapped against their corresponding proportion of cancer cells.

Regarding stromal cells (supplementary figure 84), *CIBERSORTx*, *AutoGeneS_nusvr* and *Scaden* are not too affected by the proportion that stromal cells hold in a sample, while *DigitalDLSorter* is negatively affected, with a correlation of 0.651.

The best methods overall are negatively affected by the increasing proportion that cells not deconvoluted (*Other cells*) have in the samples (figure 49). As expected, increasing the proportion of a group of cells not deconvoluted by the methods hinders estimations because the methods assume that the deconvoluted cell-types are the only cell-types present, by making proportions of estimated cell-types to have to sum up to 1. Despite that, most of these methods still score RMSEs smaller than 0.2 for the samples with higher proportion of *Other cells*. Surprinsingly, some methods benefit from a higher presence of *Other cells*. These methods, however, score an RMSE higher than 0.2 for practically all samples.

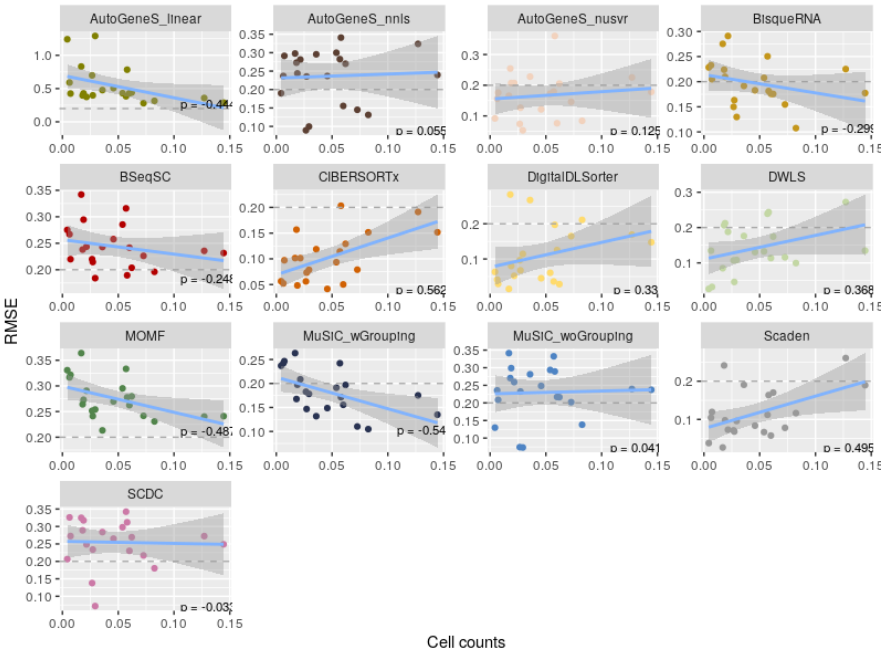


Figure 49: For each method, RMSE of the samples is mapped against their corresponding proportion of other cells.

The same trend seen for the *Other cells* seems to happen for the immune cells (figure 50). The best methods overall are negatively affected by the increasing proportion of immune cells, while the worst methods are either positively affected or not affected.

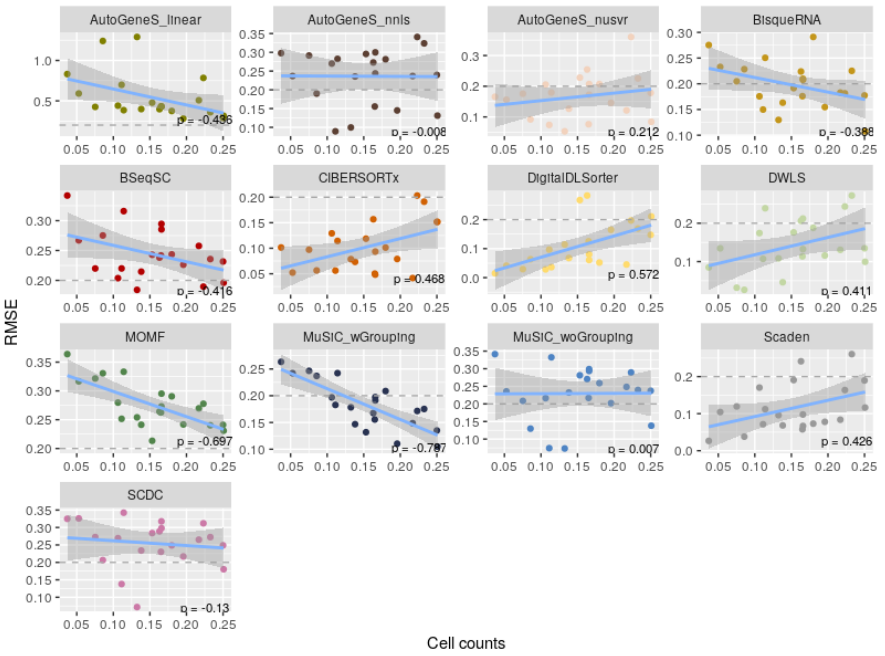


Figure 50: For each method, RMSE of the samples is mapped against their corresponding proportion of immune cells.

5.3 Discussion

In this chapter, we wanted to assess the best methods to use for the deconvolution of colorectal cancer samples of bulk RNAseq data. There are two studies [255, 256] that performed benchmarking of tumour deconvolution methods that include those that use scRNAseq data as reference. However, none were focused on colorectal cancer and methods' performance can change according to the tissue to be deconvoluted. The authors from one of these studies, a DREAM challenge that sought to evaluate a wide range of deconvolution methods and pipelines [256], even concluded that the best method may be problem specific. Although this challenge covered a lot of methods, very few were based on scRNAseq data and those methods were compared using their own default single-cell reference, which does not allow to properly assess if a better performance is due to the reference used or the deconvolution algorithm. Furthermore, they only used correlation metrics (pearson and spearman) to evaluate the results.

We found that the best method overall was *CIBERSORTx*, closely followed by *DigitalDLSorter* and *Scaden*. The authors from the DREAM challenge [256] also reported *CIBERSORTx* as the top-performing method for their datasets. However, when looking into which methods are the best for each cell-type, *CIBERSORTx* is only the best for CD4 T-cells and *DigitalDLSorter* for CD8 T-cells. *Scaden*, on the other hand, was found to be the best method for 4 cell-types (cancer cells, B-cells, regulatory CD4 T-cells, and NK cells). *BisqueRNA*, *BseqSC* and *MuSiC_wGrouping* were some of the worst methods overall but they were the best at predicting stromall cells, macro/mono lineage cells, and proliferative T-cells, respectively. *DWLS*, all *AutoGeneS* methods, *MOMF*, *MuSiC_woGrouping*, and *SCDC* were not the best in any cell-type.

Combining the best methods for each cell-type seems promising for estimating the proportions of all cell-types in the samples. However, further testing with an independent dataset with known cell-type proportions should be pursued, to assure that this is scalable for other CRC datasets and samples and does not work only for our specific case. The authors from the DREAM challenge [256] also combined the outputs of different methods but their ensemble method found only modest improvement relative to the top-scoring individual methods.

We also found that certain cell-types are more easily deconvoluted than others, with NK cells being, by far, the most difficult cell-type to estimate. CD4 and regulatory CD4 T-cells were also rather difficult to estimate.

The total number of read counts seems to influence sample estimations, with those having higher total mRNA content showing better results in most methods. We were also able to show that, as expected, the proportion of cells in the samples that are not part of a cell-type considered for deconvolution hinders the estimation of the deconvoluted cell-types. Notwithstanding, the best methods overall still showed low (< 0.2) RMSE for most samples, in a dataset where the proportion of unknown cells (*Other cells*) in the samples goes up to 0.15.

Finally, assuming that bigger cell-types would tend to be over-estimated, it would be expected that cancer, stromal and macro/mono lineage cells would suffer over-estimation across most methods without RNA content bias correction. This was not the case, however. All methods that already implemented

correction (*AutoGeneS* and *MuSiC_woGrouping*) under-estimated cancer cells and only some of the other methods benefitted (modestly) from this correction. Stromal cells also slightly benefited from RNA content bias correction. Only macro/mono lineage cells showed to visibly benefit from this correction when comparing the best methods with and without correction.

These results give a clear indication of the best methods to use when assessing the proportions of different cell-types in a colorectal cancer sample of bulk RNAseq data. Nevertheless, further work is needed. For example, it is necessary to assess cell-types' *spillover* in CRC samples, i.e., what other cell-types are being attributed signal that belongs to a certain cell-type. This can only be done by estimating proportions of samples purified for each cell-type and could help understand to what detail we could estimate cell subtypes.

6

Conclusion and Future Work

In this work, we have created genome-scale metabolic models of T-cells from tumour and normal-matched samples of different patients with colorectal cancer. This was made possible due to the construction of an atlas of scRNAseq data for CRC patients with normal matched mucosa.

The CRC atlas has a total of 163 810 cells, separated into 51 044 T-cells, 47 462 epithelial cells, 30 187 stromal cells, 17 674 B-cells, and 17 443 myeloid cells. A total of 50 phenotypes were found. We also characterised the tumour cells according to their CMS type.

This atlas allowed us to characterise not only the T-cells present in the tumour micro-environment, but also those in an unaffected part of the colon (or rectum) of the patient. A total of 196 models were reconstructed, spanning various T-cell subtypes: cytotoxic CD8 T-cells, follicular CD4 T-cells, IL17+ CD4 T-cells, memory CD4 and CD8 T-cells, naïve CD4 and CD8 T-cells, proliferative CD4 and CD8 T-cells, and regulatory CD4 T-cells. We have shown that these models do a good job at predicting the different T-cell subtypes, when using reaction presence or pFBA predictions datasets, with MCC values of 0.583 and 0.350 respectively. However, gene expression was still better at predicting the subtypes, with an MCC value of 0.793. Even though these flux predictions lead to quite imprecise results, something that also happened in a previous study of breast cancer models [178], major metabolic aspects of T-cells were observed. These include biomass and ATP production, pathways that most contribute to the production of NADH and FADH2, fatty acid uptake, and the effect of metabolites absence from medium.

Interesting differences in pathway coverage regarding regulatory CD4 T-cell models were found. The models from normal tissue samples seem to have less anti-inflammatory function, due to less presence of pathways related to anti-inflammatory functions in T-cells, and being less metabolically active, due to null or close to null biomass fluxes, than tumour-derived models. Furthermore, models from CMS4 tumours seemed to share these characteristics of normal-derived models.

Another topic we worked on was tumour deconvolution, where we used our CRC atlas of scRNAseq data to test and compare several tumour deconvolution methods that were developed to use scRNAseq data as reference to estimate cell-type proportions. Since bulk RNAseq data is used for diagnostic purposes in routine clinical settings, allowing for an unprecedented amount of data describing tumours, cell-type profiles can be recovered from bulk RNAseq data and used for cell-type metabolic model reconstruction.

Our results gave a clear indication of the best methods to use in colorectal cancer samples. We found that the best method overall was *CIBERSORTx*, closely followed by *DigitalDLSorter* and *Scaden*. However, when looking into each cell-type individually, we found out that other methods were better at predicting some of the cell-types. Combining the best methods for each cell-type seemed promising for estimating the proportions of all cell-types in the samples.

6.1 Contributions

The first major contribution from this work is the CRC atlas of scRNAseq data. The different studies that compose this atlas were integrated and re-annotated so that all would be annotated in the same way and share the same level of detail regarding cell-types annotated. Although this atlas was crucial for the development of the remaining work of this thesis, it can also be used by the community for various other ends outside the scope of this work.

Secondly, we developed a pipeline that can be easily replicated for the reconstruction of metabolic models from any kind of single-cell RNAseq dataset. This pipeline was proven to be very relevant, as the models obtained from this pipeline not only do characterise well the cell-types that they represent, but also revealed interesting aspects of T-cells, especially regulatory CD4 T-cells. Also, the models here constructed and validated can be used by the community.

Lastly, we provided important results regarding the benchmark of tumour deconvolution methods, which can aid the community in choosing the method(s) to use when deconvoluting their bulk RNAseq data from CRC samples. The contributions of this work are not restricted, however, to colorectal cancer. This pipeline can, indeed, be used in other types of samples.

6.2 Publications

At the time of writing, chapters 2 to 5 are being prepared for submissions, including a review of the current state-of-the-art on using metabolic models in cancer and the immune system (chapter 2), a paper on the creation of the atlas and its use in tumour deconvolution (chapters 3 and 5), and one regarding the T-cell metabolic models (chapter 4).

Apart from this, the following publication, although not part of the thesis' goals, was written during the doctoral program: Cardoso, S. *et al.* NMRFinder: a novel method for 1D 1H-NMR metabolite annotation. *Metabolomics*, 17(21), 2021.

6.3 Future Work

Although the main objectives were accomplished, the work done in this thesis can be extended and improved.

The model reconstruction pipeline can be applied to other cell-types from the CRC micro-environment. For example, reconstruction of tumour cell models would allow finding drug target combinations for specific patients, by exploring their effects on both tumour and T-cells (and other immune and non-immune cells). Modeling all cell types in a tumour micro-environment could enable the optimisation of immune cells' response to cancer, by modeling the metabolic competition in the tumour micro-environment through a community of models. The recovered proportions of the different cell-types in the environment from bulk RNAseq data could aid in constructing this community.

Extending the model reconstruction pipeline by complementing the data used for reconstruction with other bulk or single-cell omics data would help on reconstructing more accurate models.

Also, the reconstruction of a metabolic model for each single-cell can also be explored. Two main drawbacks have to be considered, however. The possible computational burden of such task and the low-depth of sequencing per cell that might result in wrongly considering certain genes as not expressed.

Although we gave a clear indication of the best methods to use in colorectal cancer samples and that combining the best methods for each cell-type seemed promising for estimating the proportions of all cell-types in the samples, there is still room for improvement. First, further testing with an independent dataset should be pursued, to assure that combining methods to predict the proportions of the different cell-types is scalable for other CRC samples and does not work only for our specific case. Other future work should include, as discussed before, the assessment of cell-types' spillover in CRC samples. Testing if more detailed subtypes of cells can be recovered in deconvolution is also of interest. In this work, we only deconvoluted, among the T-cells, CD8+, CD4+, regulatory CD4+, and proliferative T-cells. This is not on the same level of detail as those subtypes used to construct the models, due to having to find a balance between the atlas' subtypes and the ground-truth information available about the bulk samples.

Bibliography

[1] J. M. Lourenço. *The NOVAthesis LaTeX Template User's Manual*. NOVA University Lisbon. 2021. url: `https://github.com/joaomlourenco/novathesis/raw/master/template.pdf`.

[2] O. Warburg, K. Gawehn, and A. Geissler. "Metabolism of leukocytes". In: *Zeitschrift fur Naturforschung. Teil B, Chemie, Biochemie, Biophysik, Biologie und verwandte Gebiete* 13.8 (1958), p. 515.

[3] D. Chaplin. "Overview of the immune response". In: *Journal of Allergy and Clinical Immunology* 125.2 (2010), S3–S23.

[4] R. Goldsby et al. "Overview of the Immune System". In: *Immunology*. 5th. WH Freeman, 2002.

[5] D. Chaplin. "Overview of the human immune response". In: *Journal of Allergy and Clinical Immunology* 117.2 (2006), S430–S435.

[6] R. Goldsby et al. "T-Cell Receptor". In: *Immunology*. 5th. WH Freeman, 2002.

[7] J. Owen et al. "Cells, Organs, and Microenvironments of the Immune System". In: *Kuby Immunology*. 7th. WH Freeman, 2013.

[8] V. Golubovskaya and L. Wu. "Different subsets of T cells, memory, effector functions, and CAR-T immunotherapy". In: *Cancers* 8.3 (2016), p. 36.

[9] Y. Chen et al. "Cellular Metabolic Regulation in the Differentiation and Function of Regulatory T Cells". In: *Cells* 8.2 (2019), p. 188.

[10] Y. Mahnke, T. Brodie, F. Sallusto, et al. "The who's who of T-cell differentiation: human memory T-cell subsets". In: *European Journal of Immunology* 43.11 (2013), pp. 2797–2809.

[11] F. Sallusto, D. Lenig, R. Förster, et al. "Two subsets of memory T lymphocytes with distinct homing potentials and effector functions". In: *Nature* 401.6754 (1999), p. 708.

[12] I. Algarra, A. Collado, and F. Garrido. "Altered MHC class I antigens in tumors". In: *International Journal of Clinical and Laboratory Research* 27.2-4 (1997), pp. 95–102.

[13] X. Fan and A. Y. Rudensky. “Hallmarks of tissue-resident lymphocytes”. In: *Cell* 164.6 (2016), pp. 1198–1211.

[14] Y. Zhao, C. Niu, and J. Cui. “Gamma-delta ($\gamma\delta$) T cells: friend or foe in cancer development?” In: *Journal of Translational Medicine* 16.1 (2018), p. 3.

[15] S. Paul and G. Lal. “Regulatory and effector functions of gamma-delta ($\gamma\delta$) T cells and their therapeutic potential in adoptive cellular therapy for cancer”. In: *International Journal of Cancer* 139.5 (2016), pp. 976–985.

[16] M. Malumbres. “Control of the Cell Cycle”. In: *Abeloff’s Clinical Oncology*. Elsevier, 2020, pp. 56–73.

[17] D. Hanahan and R. Weinberg. “The hallmarks of cancer”. In: *Cell* 100.1 (2000), pp. 57–70.

[18] D. Hanahan and R. Weinberg. “Hallmarks of cancer: the next generation”. In: *Cell* 144.5 (2011), pp. 646–674.

[19] S. Cory and J. Adams. “The Bcl2 family: regulators of the cellular life-or-death switch”. In: *Nature Reviews Cancer* 2.9 (2002), p. 647.

[20] M. Jafri et al. “Roles of telomeres and telomerase in cancer, and advances in telomerase-targeted therapies”. In: *Genome Medicine* 8.1 (2016), p. 69.

[21] J. Teodoro, S. Evans, and M. Green. “Inhibition of tumor angiogenesis by p53: a new role for the guardian of the genome”. In: *Journal of Molecular Medicine* 85.11 (2007), pp. 1175–1186.

[22] G. Semenza. “Tumor metabolism: cancer cells give and take lactate”. In: *The Journal of Clinical Investigation* 118.12 (2008), pp. 3835–3837.

[23] L. Yang, Y. Pang, and H. Moses. “TGF-β and immune cells: an important regulatory axis in the tumor microenvironment and progression”. In: *Trends in Immunology* 31.6 (2010), pp. 220–227.

[24] J. Marie et al. “TGF-β1 maintains suppressor function and Foxp3 expression in CD4+ CD25+ regulatory T cells”. In: *Journal of Experimental Medicine* 201.7 (2005), pp. 1061–1067.

[25] S. Floor et al. “Hallmarks of cancer: of all cancer cells, all the time?” In: *Trends in Molecular Medicine* 18.9 (2012), pp. 509–515.

[26] D. Chen and I. Mellman. “Oncology meets immunology: the cancer-immunity cycle”. In: *Immunity* 39.1 (2013), pp. 1–10.

[27] S. Turley, V. Cremasco, and J. Astarita. “Immunological hallmarks of stromal cells in the tumour microenvironment”. In: *Nature Reviews Immunology* 15.11 (2015), p. 669.

[28] S. Grivennikov, F. Greten, and M. Karin. “Immunity, inflammation, and cancer”. In: *Cell* 140.6 (2010), pp. 883–899.

[29] R. Klein Geltink, R. Kyle, and E. Pearce. “Unraveling the complex interplay between T cell metabolism and function”. In: *Annual Review of Immunology* 36 (2018), pp. 461–488.

[30] N. MacIver, R. Michalek, and J. Rathmell. “Metabolic regulation of T lymphocytes”. In: *Annual Review of Immunology* 31 (2013), pp. 259–283.

[31] N. Ron-Harel, D. Santos, J. Ghergurovich, et al. “Mitochondrial biogenesis and proteome remodeling promote one-carbon metabolism for T cell activation”. In: *Cell Metabolism* 24.1 (2016), pp. 104–117.

[32] E. Mills, B. Kelly, and L. O’Neill. “Mitochondria are the powerhouses of immunity”. In: *Nature Immunology* 18.5 (2017), p. 488.

[33] L. Sena, S. Li, A. Jairaman, et al. “Mitochondria are required for antigen-specific T cell activation through reactive oxygen species signaling”. In: *Immunity* 38.2 (2013), pp. 225–236.

[34] T. Mak, M. Grusdat, G. Duncan, et al. “Glutathione primes T cell metabolism for inflammation”. In: *Immunity* 46.4 (2017), pp. 675–689.

[35] L. Sinclair, A. Howden, A. Brenes, et al. “Antigen receptor control of methionine metabolism in T cells”. In: *Elife* 8 (2019), e44210.

[36] J. Chang, V. Palanivel, I. Kinjyo, et al. “Asymmetric T lymphocyte division in the initiation of adaptive immune responses”. In: *Science* 315.5819 (2007), pp. 1687–1691.

[37] K. Pollizzi, I. Sun, C. Patel, et al. “Asymmetric inheritance of mTORC1 kinase activity during division dictates CD8+ T cell differentiation”. In: *Nature immunology* 17.6 (2016), p. 704.

[38] K. Verbist, C. Guy, S. Milasta, et al. “Metabolic maintenance of cell asymmetry following division in activated T lymphocytes”. In: *Nature* 532.7599 (2016), p. 389.

[39] S. Jacobs, C. Herman, N. MacIver, et al. “Glucose uptake is limiting in T cell activation and requires CD28-mediated Akt-dependent and independent pathways”. In: *The Journal of Immunology* 180.7 (2008), pp. 4476–4486.

[40] J. Blagih, F. Coulombe, E. Vincent, et al. “The energy sensor AMPK regulates T cell metabolic adaptation and effector responses in vivo”. In: *Immunity* 42.1 (2015), pp. 41–54.

[41] M. Buck, D. O’Sullivan, R. I. G., et al. “Mitochondrial dynamics controls T cell fate through metabolic programming”. In: *Cell* 166.1 (2016), pp. 63–76.

[42] D. O’Sullivan, G. van der Windt, S. Huang, et al. “Memory CD8+ T cells use cell-intrinsic lipolysis to support the metabolic programming necessary for development”. In: *Immunity* 41.1 (2014), pp. 75–88.

[43] Y. Pan, T. Tian, C. Park, et al. “Survival of tissue-resident memory T cells requires exogenous lipid uptake and metabolism”. In: *Nature* 543.7644 (2017), p. 252.

[44] R. Ma, T. Ji, H. Zhang, et al. “A Pck1-directed glycogen metabolic program regulates formation and maintenance of memory CD8+ T cells”. In: *Nature cell biology* 20.1 (2018), p. 21.

[45] R. Michalek, V. Gerriets, S. Jacobs, et al. “Cutting edge: distinct glycolytic and lipid oxidative metabolic programs are essential for effector and regulatory CD4+ T cell subsets”. In: *The Journal of Immunology* 186.6 (2011), pp. 3299–3303.

[46] L. Shi, R. Wang, G. Huang, et al. “HIF1α–dependent glycolytic pathway orchestrates a metabolic checkpoint for the differentiation of TH17 and Treg cells”. In: *Journal of Experimental Medicine* 208.7 (2011), pp. 1367–1376.

[47] D. Cluxton, B. Moran, and J. Fletcher. “Differential regulation of human Treg and Th17 cells by fatty acid synthesis and glycolysis”. In: *Frontiers in Immunology* 10 (2019), p. 115.

[48] V. Gerriets, R. Kishton, M. Johnson, et al. “Foxp3 and Toll-like receptor signaling balance Treg cell anabolic metabolism for suppression”. In: *Nature Immunology* 17.12 (2016), p. 1459.

[49] L. Berod, C. Friedrich, A. Nandan, et al. “De novo fatty acid synthesis controls the fate between regulatory T and T helper 17 cells”. In: *Nature Medicine* 20.11 (2014), p. 1327.

[50] H. Zeng, K. Yang, C. Cloer, et al. “mTORC1 couples immune signals and metabolic programming to establish Treg-cell function”. In: *Nature* 499.7459 (2013), p. 485.

[51] B. Raud, D. Roy, A. Divakaruni, et al. “Etomoxir actions on regulatory and memory T cells are independent of Cpt1a-mediated fatty acid oxidation”. In: *Cell Metabolism* 28.3 (2018), pp. 504–515.

[52] N. Pavlova and C. Thompson. “The emerging hallmarks of cancer metabolism”. In: *Cell Metabolism* 23.1 (2016), pp. 27–47.

[53] T. Murakami, T. Nishiyama, T. Shirotani, et al. “Identification of two enhancer elements in the gene encoding the type 1 glucose transporter from the mouse which are responsive to serum, growth factor, and oncogenes.” In: *Journal of Biological Chemistry* 267.13 (1992), pp. 9300–9306.

[54] P. Gao, I. Tchernyshyov, T. Chang, et al. “c-Myc suppression of miR-23a/b enhances mitochondrial glutaminase expression and glutamine metabolism”. In: *Nature* 458.7239 (2009), pp. 762–765.

[55] M. Conrad and H. Sato. “The oxidative stress-inducible cystine/glutamate antiporter, system x (c) (-): cystine supplier and beyond”. In: *Amino Acids* 42.1 (2012), pp. 231–246.

[56] D. Wu et al. “Hydrogen sulfide in cancer: friend or foe?” In: *Nitric Oxide* 50 (2015), pp. 38–45.

[57] K. Rajagopalan and R. DeBerardinis. “Role of glutamine in cancer: therapeutic and imaging implications”. In: *Journal of Nuclear Medicine* 52.7 (2011), pp. 1005–1008.

[58] J. Kamphorst, J. Cross, J. Fan, et al. “Hypoxic and Ras-transformed cells support growth by scavenging unsaturated fatty acids from lysophospholipids”. In: *Proceedings of the National Academy of Sciences* 110.22 (2013), pp. 8882–8887.

[59] H. Ying, A. Kimmelman, C. Lyssiotis, et al. “Oncogenic Kras maintains pancreatic tumors through regulation of anabolic glucose metabolism”. In: *Cell* 149.3 (2012), pp. 656–670.

[60] R. Nilsson, M. Jain, N. Madhusudhan, et al. “Metabolic enzyme expression highlights a key role for MTHFD2 and the mitochondrial folate pathway in cancer”. In: *Nature Communications* 5 (2014), p. 3128.

[61] H. Christofk, M. Vander Heiden, M. Harris, et al. “The M2 splice isoform of pyruvate kinase is important for cancer metabolism and tumour growth”. In: *Nature* 452.7184 (2008), p. 230.

[62] B. Chaneton, P. Hillmann, L. Zheng, et al. “Serine is a natural ligand and allosteric activator of pyruvate kinase M2”. In: *Nature* 491.7424 (2012), p. 458.

[63] K. Sircar, H. Huang, L. Hu, et al. “Integrative molecular profiling reveals asparagine synthetase is a target in castration-resistant prostate cancer”. In: *The American Journal of Pathology* 180.3 (2012), pp. 895–903.

[64] J. Zhang, J. Fan, S. Venneti, et al. “Asparagine plays a critical role in regulating cellular adaptation to glutamine depletion”. In: *Molecular Cell* 56.2 (2014), pp. 205–218.

[65] R. DeBerardinis and N. Chandel. “Fundamentals of cancer metabolism”. In: *Science Advances* 2.5 (2016), e1600200.

[66] T. Bowles, R. Kim, J. Galante, et al. “Pancreatic cancer cell lines deficient in argininosuccinate synthetase are sensitive to arginine deprivation by arginine deiminase”. In: *International Journal of Cancer* 123.8 (2008), pp. 1950–1955.

[67] C. Yoon, Y. Shim, E. Kim, et al. “Renal cell carcinoma does not express argininosuccinate synthetase and is highly sensitive to arginine deprivation via arginine deiminase”. In: *International Journal of Cancer* 120.4 (2007), pp. 897–905.

[68] J. Cunningham, M. Moreno, A. Lodi, et al. “Protein and nucleotide biosynthesis are coupled by a single rate-limiting enzyme, PRPS2, to drive cancer”. In: *Cell* 157.5 (2014), pp. 1088–1103.

[69] S. Eberhardy and P. Farnham. “c-Myc mediates activation of the cad promoter via a post-RNA polymerase II recruitment mechanism”. In: *Journal of Biological Chemistry* 276.51 (2001), pp. 48562–48571.

[70] D. Hanahan and L. Coussens. “Accessories to the crime: functions of cells recruited to the tumor microenvironment”. In: *Cancer Cell* 21.3 (2012), pp. 309–322.

[71] T. Whiteside. “The tumor microenvironment and its role in promoting tumor growth”. In: *Oncogene* 27.45 (2008), p. 5904.

[72] I. Kareva and P. Hahnfeldt. “The emerging ”hallmarks”of metabolic reprogramming and immune evasion: distinct or linked?” In: *Cancer Research* 73.9 (2013), pp. 2737–2742.

[73] M. Binnewies, E. Roberts, K. Kersten, et al. “Understanding the tumor immune microenvironment (TIME) for effective therapy”. In: *Nature Medicine* 24.5 (2018), p. 541.

[74] T. Gajewski, H. Schreiber, and Y. Fu. “Innate and adaptive immune cells in the tumor microenvironment”. In: *Nature Immunology* 14.10 (2013), p. 1014.

[75] U. Martinez-Outschoorn, R. Balliet, D. Rivadeneira, et al. “Oxidative stress in cancer associated fibroblasts drives tumor-stroma co-evolution: A new paradigm for understanding tumor metabolism, the field effect and genomic instability in cancer cells”. In: *Cell Cycle* 9.16 (2010), pp. 3276–3296.

[76] Y. Rattigan, B. Patel, E. Ackerstaff, et al. “Lactate is a mediator of metabolic cooperation between stromal carcinoma associated fibroblasts and glycolytic tumor cells in the tumor microenvironment”. In: *Experimental Cell Research* 318.4 (2012), pp. 326–335.

[77] C. Uyttenhove, L. Pilotte, I. Théate, et al. “Evidence for a tumoral immune resistance mechanism based on tryptophan degradation by indoleamine 2, 3-dioxygenase”. In: *Nature Medicine* 9.10 (2003), p. 1269.

[78] F. Chen, X. Zhuang, L. Lin, et al. “New horizons in tumor microenvironment biology: challenges and opportunities”. In: *BMC Medicine* 13.1 (2015), p. 45.

[79] S. Chirasani, P. Leukel, E. Gottfried, et al. “Diclofenac inhibits lactate formation and efficiently counteracts local immune suppression in a murine glioma model”. In: *International Journal of Cancer* 132.4 (2013), pp. 843–853.

[80] S. Kouidhi, F. Ben Ayed, and A. Benammar Elgaaied. “Targeting tumor metabolism: a new challenge to improve immunotherapy”. In: *Frontiers in Immunology* 9 (2018), p. 353.

[81] K. Patra, Q. Wang, P. Bhaskar, et al. “Hexokinase 2 is required for tumor initiation and maintenance and its systemic deletion is therapeutic in mouse models of cancer”. In: *Cancer Cell* 24.2 (2013), pp. 213–228.

[82] M. Sukumar, J. Liu, Y. Ji, et al. “Inhibiting glycolytic metabolism enhances CD8+ T cell memory and antitumor function”. In: *The Journal of Clinical Investigation* 123.10 (2013), pp. 4479–4488.

[83] Y. Xiang, Z. Stine, J. Xia, et al. “Targeted inhibition of tumor-specific glutaminase diminishes cell-autonomous tumorigenesis”. In: *The Journal of Clinical Investigation* 125.6 (2015), pp. 2293–2306.

[84] T. Eleftheriadis, G. Pissas, A. Karioti, et al. “Dichloroacetate at therapeutic concentration alters glucose metabolism and induces regulatory T-cell differentiation in alloreactive human lymphocytes”. In: *Journal of Basic and Clinical Physiology and Pharmacology* 24.4 (2013), pp. 271–276.

[85] V. Balachandran, M. Cavnar, S. Zeng, et al. “Imatinib potentiates antitumor T cell responses in gastrointestinal stromal tumor through the inhibition of IDO”. In: *Nature Medicine* 17.9 (2011), p. 1094.

[86] U. Martinez-Outschoorn, M. Peiris-Pages, R. Pestell, et al. “Cancer metabolism: a therapeutic perspective”. In: *Nature Reviews Clinical Oncology* 14.1 (2017), p. 11.

[87] E. Klipp, R. Herwig, A. Kowald, et al. *Systems biology in practice: concepts, implementation and application*. John Wiley & Sons, 2008.

[88] E. Stalidzans, A. Seiman, K. Peebo, et al. "Model-based metabolism design: constraints for kinetic and stoichiometric models". In: *Biochemical Society Transactions* (2018), BST20170263.

[89] F. Llaneras and J. Picó. "Stoichiometric modelling of cell metabolism". In: *Journal of Bioscience and Bioengineering* 105.1 (2008), pp. 1–11.

[90] S. Klamt and E. Gilles. "Minimal cut sets in biochemical reaction networks". In: *Bioinformatics* 20.2 (2004), pp. 226–234.

[91] J. Papin, N. Price, S. Wiback, et al. "Metabolic pathways in the post-genome era". In: *Trends in Biochemical Sciences* 28.5 (2003), pp. 250–258.

[92] C. Wagner and R. Urbanczik. "The geometry of the flux cone of a metabolic network". In: *Biophysical Journal* 89.6 (2005), pp. 3837–3845.

[93] F. Llaneras and J. Picó. "A procedure for the estimation over time of metabolic fluxes in scenarios where measurements are uncertain and/or insufficient". In: *BMC Bioinformatics* 8.1 (2007), p. 421.

[94] J. Orth, I. Thiele, and B. Palsson. "What is flux balance analysis?" In: *Nature Biotechnology* 28.3 (2010), p. 245.

[95] N. Lewis, K. Hixson, T. Conrad, et al. "Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models". In: *Molecular Systems Biology* 6.1 (2010), p. 390.

[96] R. Mahadevan and C. Schilling. "The effects of alternate optimal solutions in constraint-based genome-scale metabolic models". In: *Metabolic Engineering* 5.4 (2003), pp. 264–276.

[97] Q. Beg, A. Vazquez, J. Ernst, et al. "Intracellular crowding defines the mode and sequence of substrate uptake by Escherichia coli and constrains its metabolic activity". In: *Proceedings of the National Academy of Sciences* 104.31 (2007), pp. 12663–12668.

[98] J. Park, T. Kim, and S. Lee. "Prediction of metabolic fluxes by incorporating genomic context and flux-converging pattern analyses". In: *Proceedings of the National Academy of Sciences* 107.33 (2010), pp. 14931–14936.

[99] J. Edwards, R. Ramakrishna, and B. Palsson. "Characterizing the metabolic phenotype: a phenotype phase plane analysis". In: *Biotechnology and Bioengineering* 77.1 (2002), pp. 27–36.

[100] D. Segrè, D. Vitkup, and G. Church. "Analysis of optimality in natural and perturbed metabolic networks". In: *Proceedings of the National Academy of Sciences* 99.23 (2002), pp. 15112–15117.

[101] T. Shlomi, O. Berkman, and E. Ruppin. "Regulatory on/off minimization of metabolic flux changes after genetic perturbations". In: *Proceedings of the National Academy of Sciences* 102.21 (2005), pp. 7695–7700.

[102] N. Duarte, S. Becker, N. Jamshidi, et al. “Global reconstruction of the human metabolic network based on genomic and bibliomic data”. In: *Proceedings of the National Academy of Sciences* 104.6 (2007), pp. 1777–1782.

[103] H. Ma, A. Sorokin, A. Mazein, et al. “The Edinburgh human metabolic network reconstruction and its functional analysis”. In: *Molecular Systems Biology* 3.1 (2007), p. 135.

[104] T. Hao et al. “Compartmentalization of the Edinburgh human metabolic network”. In: *BMC Bioinformatics* 11.1 (2010), p. 393.

[105] I. Thiele, N. Swainston, R. Fleming, et al. “A community-driven global reconstruction of human metabolism”. In: *Nature Biotechnology* 31.5 (2013), p. 419.

[106] A. Mardinoglu, R. Agren, C. Kampf, et al. “Integration of clinical data with a genome-scale metabolic model of the human adipocyte”. In: *Molecular Systems Biology* 9.1 (2013), p. 649.

[107] R. Agren, S. Bordel, A. Mardinoglu, et al. “Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT”. In: *PLoS Computational Biology* 8.5 (2012), e1002518.

[108] P. Romero, J. Wagg, M. Green, et al. “Computational prediction of human metabolic pathways from the complete human genome”. In: *Genome Biology* 6.1 (2005), R2.

[109] M. Kanehisa, M. Furumichi, M. Tanabe, et al. “KEGG: new perspectives on genomes, pathways, diseases and drugs”. In: *Nucleic Acids Research* 45.D1 (2016), pp. D353–D361.

[110] A. Mardinoglu, R. Agren, C. Kampf, et al. “Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease”. In: *Nature Communications* 5 (2014), p. 3083.

[111] N. Swainston, K. Smallbone, H. Hefzi, et al. “Recon 2.2: from reconstruction to model of human metabolism”. In: *Metabolomics* 12.7 (2016), p. 109.

[112] E. Brunk, S. Sahoo, D. Zielinski, et al. “Recon3D enables a three-dimensional view of gene variation in human metabolism”. In: *Nature Biotechnology* 36.3 (2018), p. 272.

[113] I. Thiele, S. Sahoo, A. Heinken, et al. “When metabolism meets physiology: Harvey and Harvetta”. In: *Manuscript submitted for publication* ().

[114] J. L. Robinson et al. “An atlas of human metabolism”. In: *Science Signaling* 13.624 (2020).

[115] M. Uhlén, L. Fagerberg, B. Hallström, et al. “Tissue-based map of the human proteome”. In: *Science* 347.6220 (2015), p. 1260419.

[116] S. Becker and B. Palsson. “Context-specific metabolic networks are consistent with experiments”. In: *PLoS Computational Biology* 4.5 (2008), e1000082.

[117] T. Shlomi, M. Cabili, M. Herrgård, et al. “Network-based prediction of human tissue-specific metabolism”. In: *Nature Biotechnology* 26.9 (2008), p. 1003.

[118] H. Zur, E. Ruppin, and T. Shlomi. “iMAT: an integrative metabolic analysis tool”. In: *Bioinformatics* 26.24 (2010), pp. 3140–3142.

[119] R. Agren, A. Mardinoglu, A. Asplund, et al. “Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling”. In: *Molecular Systems Biology* 10.3 (2014), p. 721.

[120] K. Yizhak et al. “Phenotype-based cell-specific metabolic modeling reveals metabolic liabilities of cancer”. In: *Elife* 3 (2014), e03641.

[121] A. Schultz and A. Qutub. “Reconstruction of tissue-specific metabolic networks using CORDA”. In: *PLoS Computational Biology* 12.3 (2016), e1004808.

[122] L. Jerby, T. Shlomi, and E. Ruppin. “Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism”. In: *Molecular Systems Biology* 6.1 (2010), p. 401.

[123] Y. Wang, J. Eddy, and N. Price. “Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE”. In: *BMC Systems Biology* 6.1 (2012), p. 153.

[124] N. Vlassis, M. Pacheco, and T. Sauter. “Fast reconstruction of compact context-specific metabolic network models”. In: *PLoS Computational Biology* 10.1 (2014), e1003424.

[125] E. Özcan and T. Çakır. “Reconstructed metabolic network models predict flux-level metabolic reprogramming in glioblastoma”. In: *Frontiers in Neuroscience* 10 (2016), p. 156.

[126] E. Motamedian, G. Ghavami, and S. Sardari. “Investigation on metabolism of cisplatin resistant ovarian cancer using a genome scale metabolic model and microarray data”. In: *Iranian Journal of Basic Medical Sciences* 18.3 (2015), p. 267.

[127] R. Metri et al. “Modelling metabolic rewiring during melanoma progression using flux balance analysis”. In: *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2017, pp. 134–137.

[128] F. Shen et al. “Systematic investigation of metabolic reprogramming in different cancers based on tissue-specific metabolic models”. In: *Journal of Bioinformatics and Computational Biology* 14.05 (2016), p. 1644001.

[129] F.-S. Wang et al. “Genome-Scale Metabolic Modeling with Protein Expressions of Normal and Cancerous Colorectal Tissues for Oncogene Inference”. In: *Metabolites* 10.1 (2020), p. 16.

[130] B. ter Braak et al. “Insulin-like growth factor 1 receptor activation promotes mammary gland tumor development by increasing glycolysis and promoting biomass production”. In: *Breast Cancer Research* 19.1 (2017), p. 14.

[131] P. Ghaffari et al. “Identifying anti-growth factors for human cancer cell lines through genome-scale metabolic modeling”. In: *Scientific Reports* 5.1 (2015), pp. 1–10.

[132] K. Yizhak et al. “A computational study of the Warburg effect identifies metabolic targets inhibiting cancer migration”. In: *Molecular Systems Biology* 10.8 (2014), p. 744.

[133] S. Sahoo et al. “Metabolite systems profiling identifies exploitable weaknesses in retinoblastoma”. In: *FEBS Letters* 593.1 (2019), pp. 23–41.

[134] R. Agren et al. “Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling”. In: *Molecular Systems Biology* 10.3 (2014), p. 721.

[135] F. Gatto, R. Ferreira, and J. Nielsen. “Pan-cancer analysis of the metabolic reaction network”. In: *Metabolic Engineering* 57 (2020), pp. 51–62.

[136] M. Uhlen et al. “A pathology atlas of the human cancer transcriptome”. In: *Science* 357.6352 (2017).

[137] F. Han et al. “Genome-wide metabolic model to improve understanding of CD4+ T cell metabolism, immunometabolism and application in drug design”. In: *Molecular BioSystems* 12.2 (2016), pp. 431–443.

[138] B. L. Puniya et al. “Integrative computational approach identifies drug targets in CD4+ T-cell-mediated immune disorders”. In: *NPJ Systems Biology and Applications* 7.1 (2021), pp. 1–18.

[139] A. M. Abdel-Haleem et al. “Model-Driven Analysis of Gene Expression Data: Application to Metabolic Re-programming during T-cell Activation”. In: *Proceedings of the International Conference on Bioinformatics & Computational Biology (BIOCOMP)*. The Steering Committee of The World Congress in Computer Science. 2015, p. 91.

[140] A. Bordbar et al. “Insight into human alveolar macrophage and M. tuberculosis interactions via metabolic reconstructions”. In: *Molecular Systems Biology* 6.1 (2010).

[141] A. Bordbar et al. “Model-driven multi-omic data analysis elucidates metabolic immunomodulators of macrophage activation”. In: *Molecular Systems Biology* 8.1 (2012).

[142] K. Yu and M. Snyder. “Omics profiling in precision oncology”. In: *Molecular & Cellular Proteomics* 15.8 (2016), pp. 2525–2536.

[143] C. Manzoni, D. Kia, J. Vandrovcova, et al. “Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences”. In: *Briefings in Bioinformatics* 19.2 (2016), pp. 286–302.

[144] R. Bumgarner. “Overview of DNA microarrays: types, applications, and their future”. In: *Current Protocols in Molecular Biology* 101.1 (2013), pp. 22–1.

[145] J. Shendure. “The beginning of the end for microarrays?” In: *Nature Methods* 5.7 (2008), pp. 585–587.

[146] S. Liu and C. Trapnell. “Single-cell transcriptome sequencing: recent advances and remaining challenges”. In: *F1000Research* 5 (2016).

[147] L. Jerby and E. Ruppin. “Predicting drug targets and biomarkers of cancer via genome-scale metabolic modeling”. In: *Clinical Cancer Research* 18.20 (2012), pp. 5572–5584.

[148] H. Nam et al. “A systems approach to predict oncometabolites via context-specific genome-scale metabolic networks”. In: *PLoS Computational Biology* 10.9 (2014), e1003837.

[149] J. Gagneur and S. Klamt. “Computation of elementary modes: a unifying framework and the new binary approach”. In: *BMC Bioinformatics* 5.1 (2004), p. 175.

[150] L. Jerby and E. Ruppin. “Predicting drug targets and biomarkers of cancer via genome-scale metabolic modeling”. In: *Clinical Cancer Research* 18.20 (2012), pp. 5572–5584.

[151] O. Folger et al. “Predicting selective drug targets in cancer through metabolic networks”. In: *Molecular Systems Biology* 7.1 (2011), p. 501.

[152] C. Mattiuzzi, F. Sanchis-Gomar, and G. Lippi. “Concise update on colorectal cancer epidemiology”. In: *Annals of Translational Medicine* 7.21 (2019).

[153] P. Rawla, T. Sunkara, and A. Barsouk. “Epidemiology of colorectal cancer: incidence, mortality, survival, and risk factors”. In: *Przeglad Gastroenterologiczny* 14.2 (2019), p. 89.

[154] I. A. for Research on Cancer. *Cancer Today - Cancer Fact Sheets*. 2020. url: `http://gco.iarc.fr/today/fact-sheets-cancers`.

[155] Y. Hao et al. “Integrated analysis of multimodal single-cell data”. In: *Cell* 184.13 (2021), pp. 3573–3587.

[156] J. Guinney et al. “The consensus molecular subtypes of colorectal cancer”. In: *Nature Medicine* 21.11 (2015), pp. 1350–1356.

[157] Y. Hao et al. “Integrated analysis of multimodal single-cell data”. In: *Cell* 184.13 (2021), pp. 3573–3587.

[158] P. Hoffman. *SeuratDisk: Interfaces for HDF5-Based Single Cell File Formats*. 2021. url: `https://github.com/mojaveazure/seurat-disk`.

[159] R. Satija et al. *SeuratObject: Data Structures for Single Cell Data*. R package version 4.0.2. 2021. url: `https://CRAN.R-project.org/package=SeuratObject`.

[160] L. Zappia and A. Oshlack. “Clustering trees: a visualization for evaluating clusterings at multiple resolutions”. In: *GigaScience* 7.7 (2018). doi: `10.1093/gigascience/giy083`. url: `https://doi.org/10.1093/gigascience/giy083`.

[161] M. Mircea. *SIGMA: A clusterability measure for scRNA-seq data*. R package version 0.0.0.1. 2021.

[162] T. Tickle et al. *inferCNV of the Trinity CTAT Project*. Klarman Cell Observatory, Broad Institute of MIT and Harvard. Cambridge, MA, USA, 2019. url: `https://github.com/broadinstitute/inferCNV`.

[163] P. W. Eide et al. “CMScaller: an R package for consensus molecular subtyping of colorectal cancer pre-clinical models”. In: *Scientific Reports* 7 (2017), p. 16618. doi: 10.1038/s41598-017-16747-x.

[164] J. Qian et al. “A pan-cancer blueprint of the heterogeneous tumor microenvironment revealed by single-cell profiling”. In: *Cell research* 30.9 (2020), pp. 745–762. doi: 10.1038/s41422-020-0355-0.

[165] H. Lee et al. “Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer”. In: *Nature Genetics* 52.6 (2020), pp. 594–603. doi: 10.1038/s41588-020-0636-z.

[166] C. Smillie et al. “Intra-and inter-cellular rewiring of the human colon during ulcerative colitis”. In: *Cell* 178.3 (2019), pp. 714–730. doi: 10.1016/j.cell.2019.06.029.

[167] A. Loregger et al. “The E3 ligase RNF43 inhibits Wnt signaling downstream of mutated β-catenin by sequestering TCF4 to the nuclear membrane”. In: *Science Signaling* 8.393 (2015), ra90–ra90.

[168] S Salahshor and J. Woodgett. “The links between axin and carcinogenesis”. In: *Journal of Clinical Pathology* 58.3 (2005), pp. 225–236.

[169] S. Razak et al. “Screening and computational analysis of colorectal associated non-synonymous polymorphism in CTNNB1 gene in Pakistani population”. In: *BMC Medical Genetics* 20.1 (2019), pp. 1–12.

[170] C. F. Rochlitz, R. Herrmann, and E. de Kant. “Overexpression and amplification of c-myc during progression of human colorectal cancer”. In: *Oncology* 53.6 (1996), pp. 448–454.

[171] Z. Wang et al. “The prognostic and clinical value of CD44 in colorectal cancer: a meta-analysis”. In: *Frontiers in Oncology* 9 (2019), p. 309.

[172] S.-M. Wang et al. “Clinical significance of MLH1/MSH2 for stage II/III sporadic colorectal cancer”. In: *World Journal of Gastrointestinal Oncology* 11.11 (2019), p. 1065.

[173] J. L. Robinson et al. “An atlas of human metabolism”. In: *Science Signaling* 13.624 (2020), eaaz1482.

[174] R. Agren et al. “Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling”. In: *Molecular Systems Biology* 10.3 (2014), p. 721.

[175] A. Ebrahim et al. “COBRApy: constraints-based reconstruction and analysis for python”. In: *BMC Systems Biology* 7.1 (2013), pp. 1–6.

[176] J. Ferreira et al. “Troppo-A Python framework for the reconstruction of context-specific metabolic models”. In: *International Conference on Practical Applications of Computational Biology & Bioinformatics*. Springer. 2019, pp. 146–153.

[177] N. Vlassis, M. P. Pacheco, and T. Sauter. “Fast reconstruction of compact context-specific metabolic network models”. In: *PLoS Computational Biology* 10.1 (2014), e1003424.

[178] V. Vieira, J. Ferreira, and M. Rocha. “A pipeline for the reconstruction and evaluation of context-specific human metabolic models at a large-scale”. In: *PLOS Computational Biology* 18.6 (2022), e1009294.

[179] A. Richelle, C. Joshi, and N. E. Lewis. “Assessing key decisions for transcriptomic data integration in biochemical networks”. In: *PLoS Computational Biology* 15.7 (2019), e1007185.

[180] D. Wishart, Y. Feunang, A. Marcu, et al. “HMDB 4.0: the human metabolome database for 2018”. In: *Nucleic Acids Research* 46.D1 (2017), pp. D608–D617.

[181] M. K. Aurich et al. “Prediction of intracellular metabolic states from extracellular metabolomic data”. In: *Metabolomics* 11.3 (2015), pp. 603–619.

[182] M. Mir et al. “Optical measurement of cycle-dependent cell growth”. In: *Proceedings of the National Academy of Sciences* 108.32 (2011), pp. 13124–13129.

[183] M. Beck et al. “The quantitative proteome of a human cell line”. In: *Molecular Systems Biology* 7.1 (2011), p. 549.

[184] E. H. Chapman, A. S. Kurec, and F. Davey. “Cell volumes of normal and malignant mononuclear cells.” In: *Journal of Clinical Pathology* 34.10 (1981), pp. 1083–1090.

[185] *Optimization of Human T Cell Expansion Protocol: Effects of Early Cell Dilution*. Technical Bulletin. STEMCELL Technologies, 2020.

[186] J. Gu et al. “Metabolomics analysis in serum from patients with colorectal polyp and colorectal cancer by 1H-NMR spectrometry”. In: *Disease Markers* 2019 (2019).

[187] C. Zhang et al. “Metabolomic profiling identified serum metabolite biomarkers and related metabolic pathways of colorectal cancer”. In: *Disease Markers* 2021 (2021).

[188] Y. Qiu et al. “Serum metabolite profiling of human colorectal cancer using GC- TOFMS and UPLC-QTOFMS”. In: *Journal of Proteome Research* 8.10 (2009), pp. 4844–4850.

[189] S. Nishiumi et al. “A novel serum metabolomics-based diagnostic approach for colorectal cancer”. In: *PloS One* 7.7 (2012), e40459.

[190] B. Tan et al. “Metabonomics identifies serum metabolite markers of colorectal cancer”. In: *Journal of Proteome Research* 12.6 (2013), pp. 3000–3009.

[191] J. Zhu et al. “Colorectal cancer detection using targeted serum metabolic profiling”. In: *Journal of Proteome Research* 13.9 (2014), pp. 4120–4130.

[192] M. Kuhn. *caret: Classification and Regression Training*. R package version 6.0-88. 2021. url: `https://CRAN.R-project.org/package=caret`.

[193] S. E. Knowles et al. “Production and utilization of acetate in mammals”. In: *Biochemical Journal* 142.2 (1974), pp. 401–411.

[194] D. B. Njoku, H. V. Chitilian, and K. Kronish. “Hepatic physiology, pathophysiology, and anesthetic considerations”. In: *Miller's Anesthesia* (2020), pp. 420–443.

[195] E. Molnár et al. “Cholesterol and sphingomyelin drive ligand-independent T-cell antigen receptor nanoclustering”. In: *Journal of Biological Chemistry* 287.51 (2012), pp. 42664–42674.

[196] M. Aguilar-Ballester et al. “Impact of cholesterol metabolism in immune cell function and atherosclerosis”. In: *Nutrients* 12.7 (2020), p. 2021.

[197] A. Bietz et al. “Cholesterol metabolism in T cells”. In: *Frontiers in Immunology* 8 (2017), p. 1664.

[198] S. J. Bensinger et al. “LXR signaling couples sterol metabolism to proliferation in the acquired immune response”. In: *Cell* 134.1 (2008), pp. 97–111.

[199] Y. Kidani et al. “Sterol regulatory element–binding proteins are essential for the metabolic programming of effector T cells and adaptive immunity”. In: *Nature immunology* 14.5 (2013), pp. 489–499.

[200] X. Hu et al. “Sterol metabolism controls TH17 differentiation by generating endogenous RORγ agonists”. In: *Nature Chemical Biology* 11.2 (2015), pp. 141–147.

[201] J. Surls et al. “Increased membrane cholesterol in lymphocytes diverts T-cells toward an inflammatory response”. In: *PloS One* 7.6 (2012), e38733.

[202] B. L. Farrugia et al. “The role of heparan sulfate in inflammation, and the development of biomimetics as anti-inflammatory strategies”. In: *Journal of Histochemistry & Cytochemistry* 66.4 (2018), pp. 321–336.

[203] Y. Naparstek et al. “Activated T lymphocytes produce a matrix-degrading heparan sulphate endoglycosidase”. In: *Nature* 310.5974 (1984), pp. 241–244.

[204] T. R. Theodoro et al. “Heparanase expression in circulating lymphocytes of breast cancer patients depends on the presence of the primary tumor and/or systemic metastasis”. In: *Neoplasia* 9.6 (2007), pp. 504–510.

[205] A. J. Mayfosh, N. Baschuk, and M. D. Hulett. “Leukocyte heparanase: a double-edged sword in tumor progression”. In: *Frontiers in Oncology* 9 (2019), p. 331.

[206] A. M. Lone and K. Taskén. “Proinflammatory and immunoregulatory roles of eicosanoids in T cells”. In: *Frontiers in Immunology* 4 (2013), p. 130.

[207] S. L. Tilley, T. M. Coffman, B. H. Koller, et al. “Mixed messages: modulation of inflammation and immune responses by prostaglandins and thromboxanes”. In: *The Journal of Clinical Investigation* 108.1 (2001), pp. 15–23.

[208] J. M. Cook-Moreau et al. “Expression of 5-lipoxygenase (5-LOX) in T lymphocytes”. In: *Immunology* 122.2 (2007), pp. 157–166.

[209] M. G. Cifone et al. “Diacylglycerol lipase activation and 5-lipoxygenase activation and translocation following TCR/CD3 triggering in T cells”. In: *European Journal of Immunology* 25.4 (1995), pp. 1080–1086.

[210] M. Los et al. “IL-2 gene expression and NF-kappa B activation through CD28 requires reactive oxygen production by 5-lipoxygenase.” In: *The EMBO Journal* 14.15 (1995), pp. 3731–3740.

[211] G. R. Warnes, B. Bolker, and T. Lumley. *gtools: Various R Programming Tools*. R package version 3.9.2. 2021. url: `https://CRAN.R-project.org/package=gtools`.

[212] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2021. url: `https://www.R-project.org/`.

[213] Y. Benjamini and Y. Hochberg. “Controlling the false discovery rate: a practical and powerful approach to multiple testing”. In: *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995), pp. 289–300.

[214] I. Pacella et al. “Fatty acid metabolism complements glycolysis in the selective regulatory T cell expansion during tumor growth”. In: *Proceedings of the National Academy of Sciences* 115.28 (2018), E6546–E6555.

[215] A. Elahi et al. “Biotin deficiency induces Th1-and Th17-mediated proinflammatory responses in human CD4+ T lymphocytes via activation of the mTOR signaling pathway”. In: *The Journal of Immunology* 200.8 (2018), pp. 2563–2570.

[216] T. Yamaguchi et al. “Control of immune responses by antigen-specific regulatory T cells expressing the folate receptor”. In: *Immunity* 27.1 (2007), pp. 145–159.

[217] G. Pozzi et al. “Buffering adaptive immunity by hydrogen sulfide”. In: *Cells* 11.3 (2022), p. 325.

[218] R. Yang et al. “Hydrogen sulfide promotes Tet1-and Tet2-mediated Foxp3 demethylation to drive regulatory T cell differentiation and maintain immune homeostasis”. In: *Immunity* 43.2 (2015), pp. 251–263.

[219] C. Courtemanche et al. “Folate deficiency inhibits the proliferation of primary human CD8+ T lymphocytes in vitro”. In: *The Journal of Immunology* 173.5 (2004), pp. 3186–3192.

[220] I. Orzołek, J. Sobieraj, and J. Domagała-Kulawik. “Estrogens, Cancer and Immunity”. In: *Cancers* 14.9 (2022), p. 2265.

[221] F. C. Navarro and S. K. Watkins. “Estrogen stimulation differentially impacts human male and female antigen-specific T cell anti-tumor function and polyfunctionality”. In: *Gender and the Genome* 1.4 (2017), pp. 1–13.

[222] S. J. Cronin et al. “The metabolite BH4 controls T cell proliferation in autoimmunity and cancer”. In: *Nature* 563.7732 (2018), pp. 564–568.

[223] C.-H. Yao et al. “Identifying off-target effects of etomoxir reveals that carnitine palmitoyltransferase I is essential for cancer cell proliferation independent of β-oxidation”. In: *PLoS Biology* 16.3 (2018), e2003782.

[224] J. Ye et al. “L-carnitine attenuates H2O2-induced neuron apoptosis via inhibition of endoplasmic reticulum stress”. In: *Neurochemistry International* 78 (2014), pp. 86–95.

[225] E. De Bousser et al. “Human T cell glycosylation and implications on immune therapy for cancer”. In: *Human Vaccines & Immunotherapeutics* 16.10 (2020), pp. 2374–2388.

[226] T. Le Bourgeois et al. “Targeting T cell metabolism for improvement of cancer immunotherapy”. In: *Frontiers in Oncology* 8 (2018), p. 237.

[227] C. Dumitru, A. M. Kabat, and K. J. Maloy. “Metabolic adaptations of CD4+ T cells in inflammatory disease”. In: *Frontiers in Immunology* 9 (2018), p. 540.

[228] D. Loeffler, P. Juneau, and S Masserant. “Influence of tumour physico-chemical conditions on interleukin-2-stimulated lymphocyte proliferation”. In: *British Journal of Cancer* 66.4 (1992), pp. 619–622.

[229] R. Tripmacher et al. “Human CD4+ T cells maintain specific functions even under conditions of extremely restricted ATP production”. In: *European Journal of Immunology* 38.6 (2008), pp. 1631–1642.

[230] K. Carswell, J. Weiss, and E. Papoutsakis. “Low oxygen tension enhances the stimulation and proliferation of human T lymphocytes in the presence of IL-2”. In: *Cytotherapy* 2.1 (2000), pp. 25–37.

[231] H. Haddad et al. “Molecular understanding of oxygen-tension and patient-variability effects on ex vivo expanded T cells”. In: *Biotechnology and Bioengineering* 87.4 (2004), pp. 437–450.

[232] K. R. Atkuri, L. A. Herzenberg, and L. A. Herzenberg. “Culturing at atmospheric oxygen levels impacts lymphocyte function”. In: *Proceedings of the National Academy of Sciences* 102.10 (2005), pp. 3756–3759.

[233] A. Naldini et al. “Hypoxia affects cytokine production and proliferative responses by human peripheral mononuclear cells”. In: *Journal of Cellular Physiology* 173.3 (1997), pp. 335–342.

[234] S. S. Kolan et al. “Cellular metabolism dictates T cell effector function in health and disease”. In: *Scandinavian Journal of Immunology* 92.5 (2020), e12956.

[235] E. H. Ma et al. “Metabolic profiling using stable isotope tracing reveals distinct patterns of glucose utilization by physiologically activated CD8+ T cells”. In: *Immunity* 51.5 (2019), pp. 856–870.

[236] P. Y. Ting et al. “Guide Swap enables genome-scale pooled CRISPR–Cas9 screening in human primary cells”. In: *Nature Methods* 15.11 (2018), pp. 941–946.

[237] E. Shifrut et al. “Genome-wide CRISPR screens in primary human T cells reveal key regulators of immune function”. In: *Cell* 175.7 (2018), pp. 1958–1971.

[238] T. Zhang et al. “The role of glycosphingolipids in immune cell functions”. In: *Frontiers in Immunology* 10 (2019), p. 90.

[239] G. McDonald et al. “Normalizing glycosphingolipids restores function in CD4+ T cells from lupus patients”. In: *The Journal of clinical investigation* 124.2 (2014), pp. 712–724.

[240] Y. He et al. “Gut microbial metabolites facilitate anticancer therapy efficacy by modulating cytotoxic CD8+ T cell immunity”. In: *Cell Metabolism* 33.5 (2021), pp. 988–1000.

[241] J. van den Bulk et al. “Neoantigen-specific immunity in low mutation burden colorectal cancers of the consensus molecular subtype 4”. In: *Genome Medicine* 11.1 (2019), pp. 1–15.

[242] H. Aliee and F. J. Theis. “AutoGeneS: Automatic gene selection using multi-objective optimization for RNA-seq deconvolution”. In: *Cell Systems* 12.7 (2021), pp. 706–715.

[243] B. Jew et al. “Accurate estimation of cell composition in bulk expression through robust integration of single-cell information”. In: *Nature Communications* 11.1 (2020), pp. 1–11.

[244] M. Baron et al. “A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure”. In: *Cell Systems* 3.4 (2016), pp. 346–360.

[245] C. B. Steen et al. “Profiling cell type abundance and expression in bulk tissues with CIBERSORTx”. In: *Stem Cell Transcriptional Networks*. Springer, 2020, pp. 135–157.

[246] C. Torroja and F. Sanchez-Cabo. “Digitaldlsorter: deep-learning on scRNA-Seq to deconvolute gene expression data”. In: *Frontiers in Genetics* 10 (2019), p. 978.

[247] D. Tsoucas et al. “Accurate estimation of cell-type composition from gene expression data”. In: *Nature Communications* 10.1 (2019), pp. 1–9.

[248] X. Sun, S. Sun, and S. Yang. “An efficient and flexible method for deconvoluting bulk RNA-seq data with single-cell RNA-seq data”. In: *Cells* 8.10 (2019), p. 1161.

[249] X. Wang et al. “Bulk tissue cell type deconvolution with multi-subject single-cell expression reference”. In: *Nature Communications* 10.1 (2019), pp. 1–9.

[250] K. Menden et al. “Deep learning–based cell composition analysis from tissue expression profiles”. In: *Science Advances* 6.30 (2020), eaba2619.

[251] M. Dong et al. “SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references”. In: *Briefings in Bioinformatics* 22.1 (2021), pp. 416–427.

[252] K. Zaitsev et al. “Complete deconvolution of cellular mixtures based on linearity of transcriptional signatures”. In: *Nature Communications* 10.1 (2019), pp. 1–16.

[253] O. A. Sosina et al. “Strategies for cellular deconvolution in human brain RNA sequencing data”. In: *F1000Research* 10.750 (2021), p. 750.

[254] K. Ushey, J. Allaire, and Y. Tang. *reticulate: Interface to 'Python'*. R package version 1.20. 2021. url: `https://CRAN.R-project.org/package=reticulate`.

[255] F. Avila Cobos et al. “Benchmarking of cell type deconvolution pipelines for transcriptomics data”. In: *Nature Communications* 11.1 (2020), pp. 1–14.

[256] B. S. White et al. “Community assessment of methods to deconvolve cellular composition from bulk gene expression”. In: *bioRxiv* (2022).

A

Supplementary Figures

This appendix contains all supplementary figures, separated by chapters.

A.1 Chapter 3

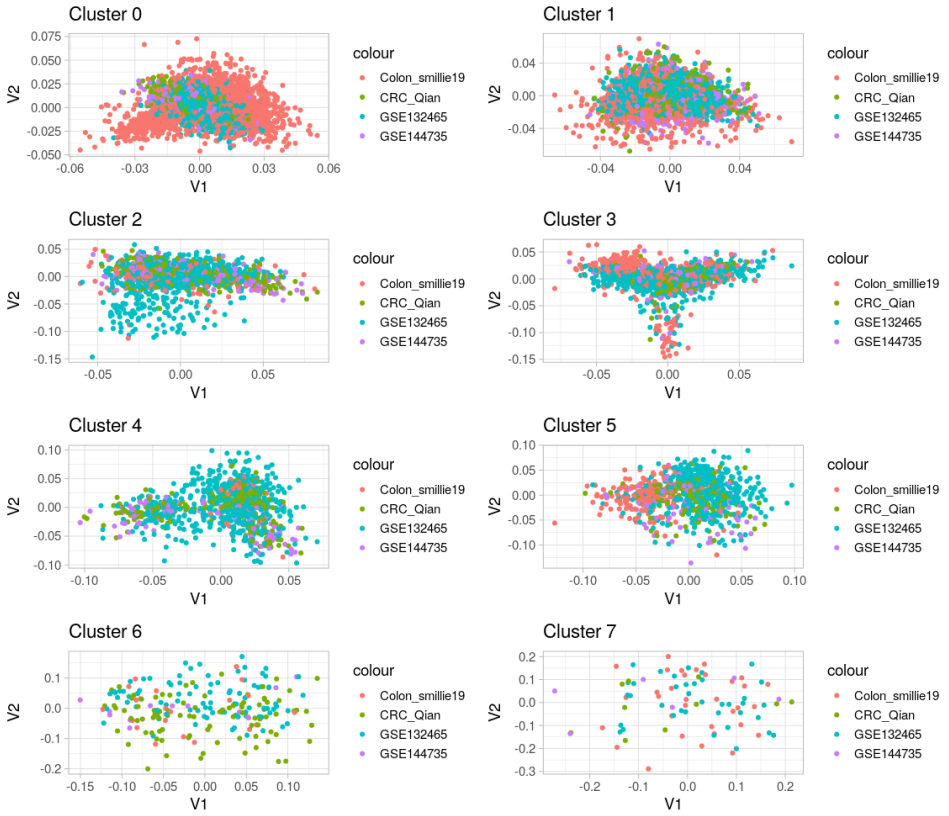


Figure 51: Clusterability results from SIGMA. Each dot corresponds to a cell, which are coloured by dataset of origin. The clusterability of the clusters was not dictated by the dataset of origin.

Figure 52: Heatmap of CNV predictions for the tumour cells of patient KUL21.

Figure 53: Heatmap of CNV predictions for the tumour cells of patient SMC04.

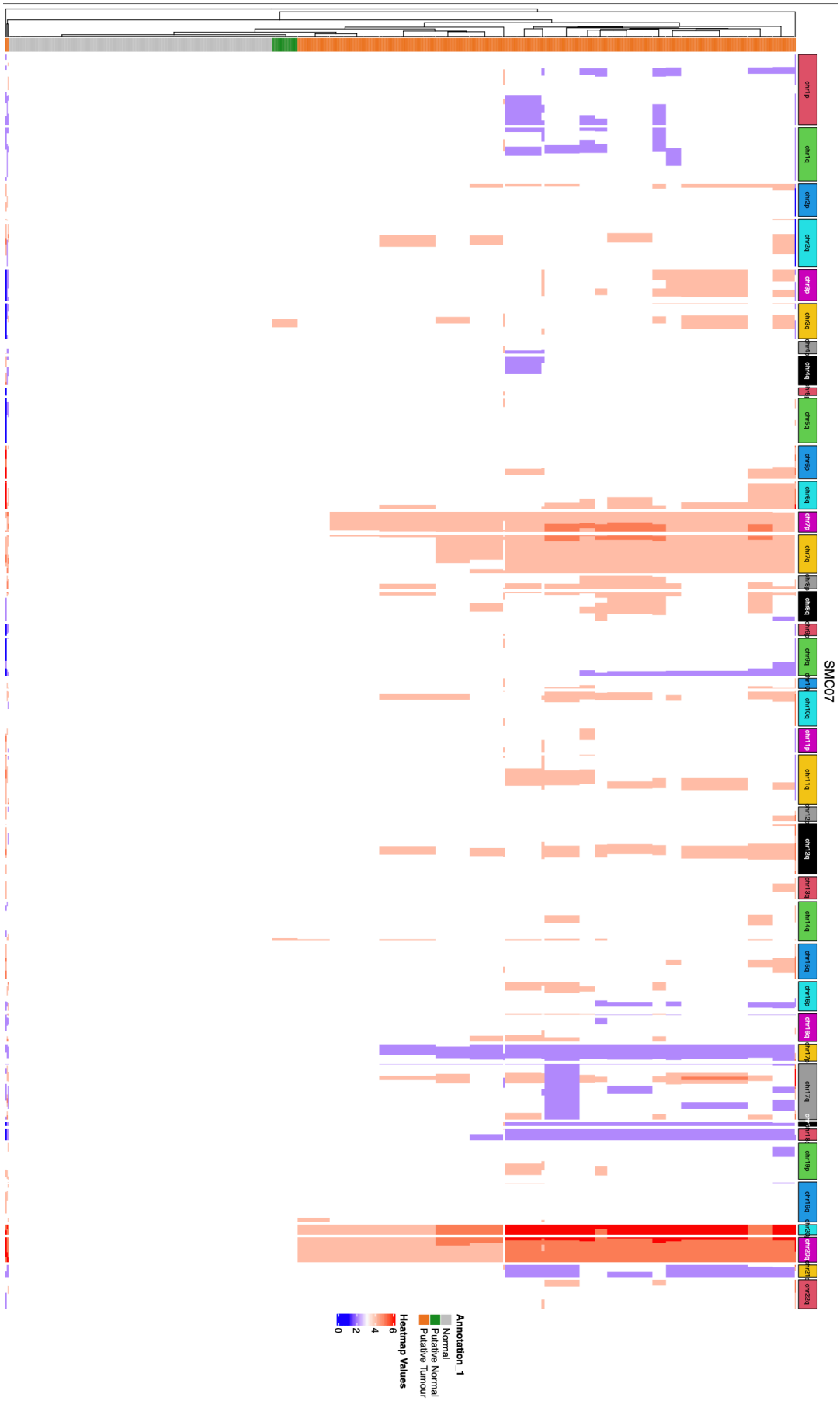


Figure 54: Heatmap of CNV predictions for the tumour cells of patient SMC07.

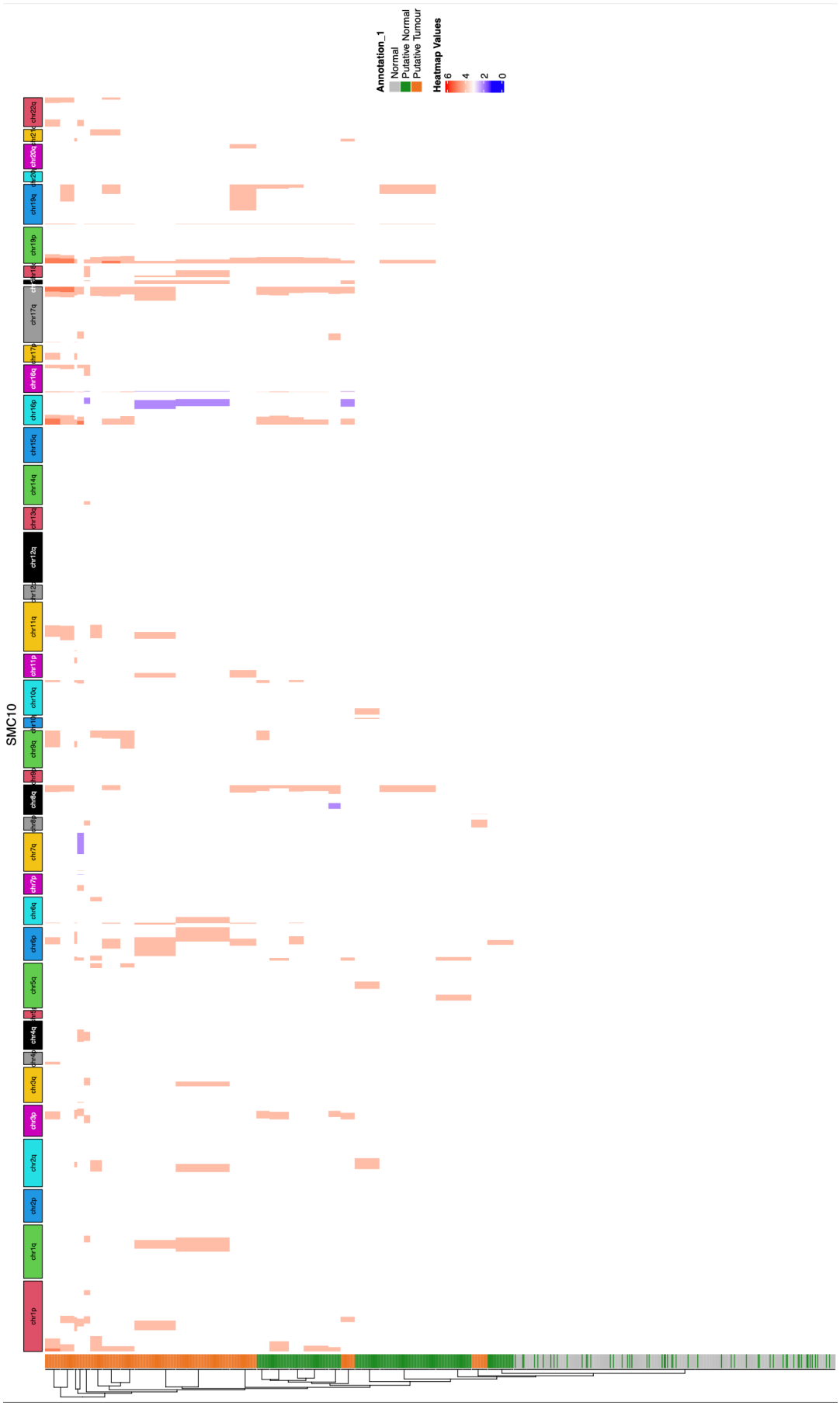


Figure 55: Heatmap of CNV predictions for the tumour cells of patient SMC10.

A.2 Chapter 4

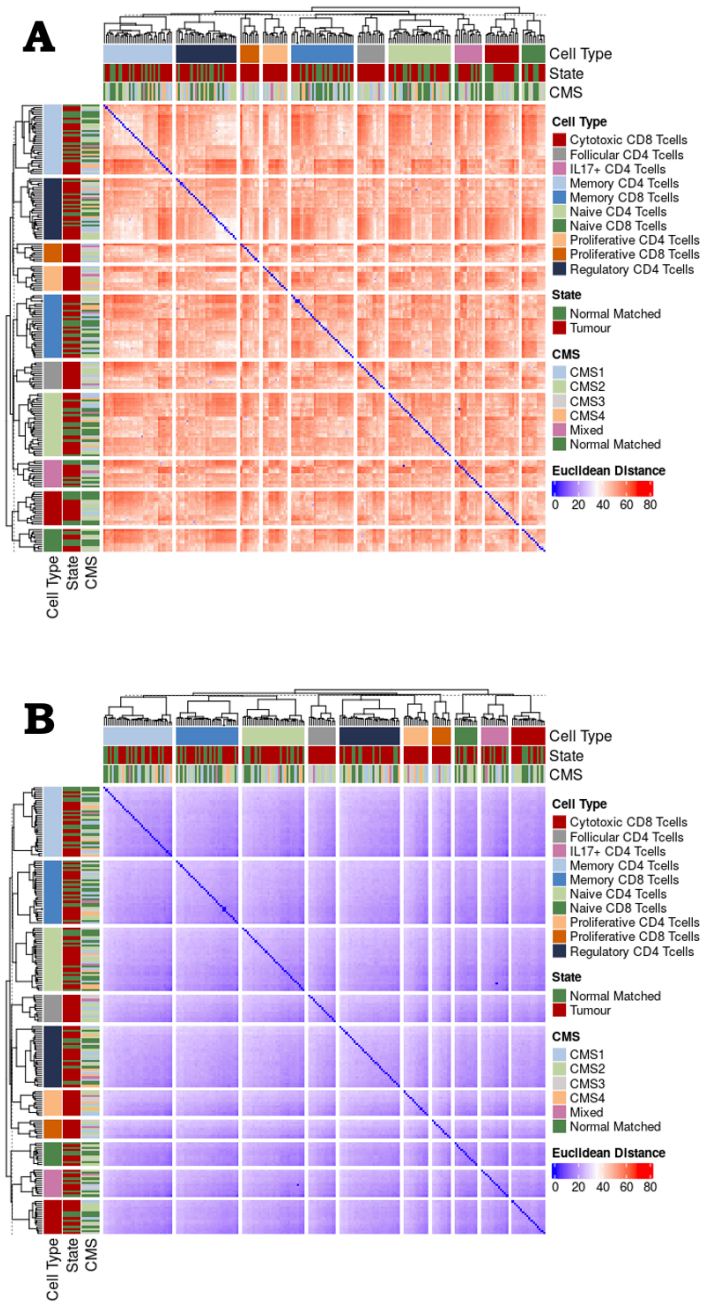


Figure 56: Similarity between (A) the structure of the models (i.e., reaction presence/absence) and (B) predicted fluxes under normal human blood medium. The smaller Euclidean distance is, the smaller the similarity is.

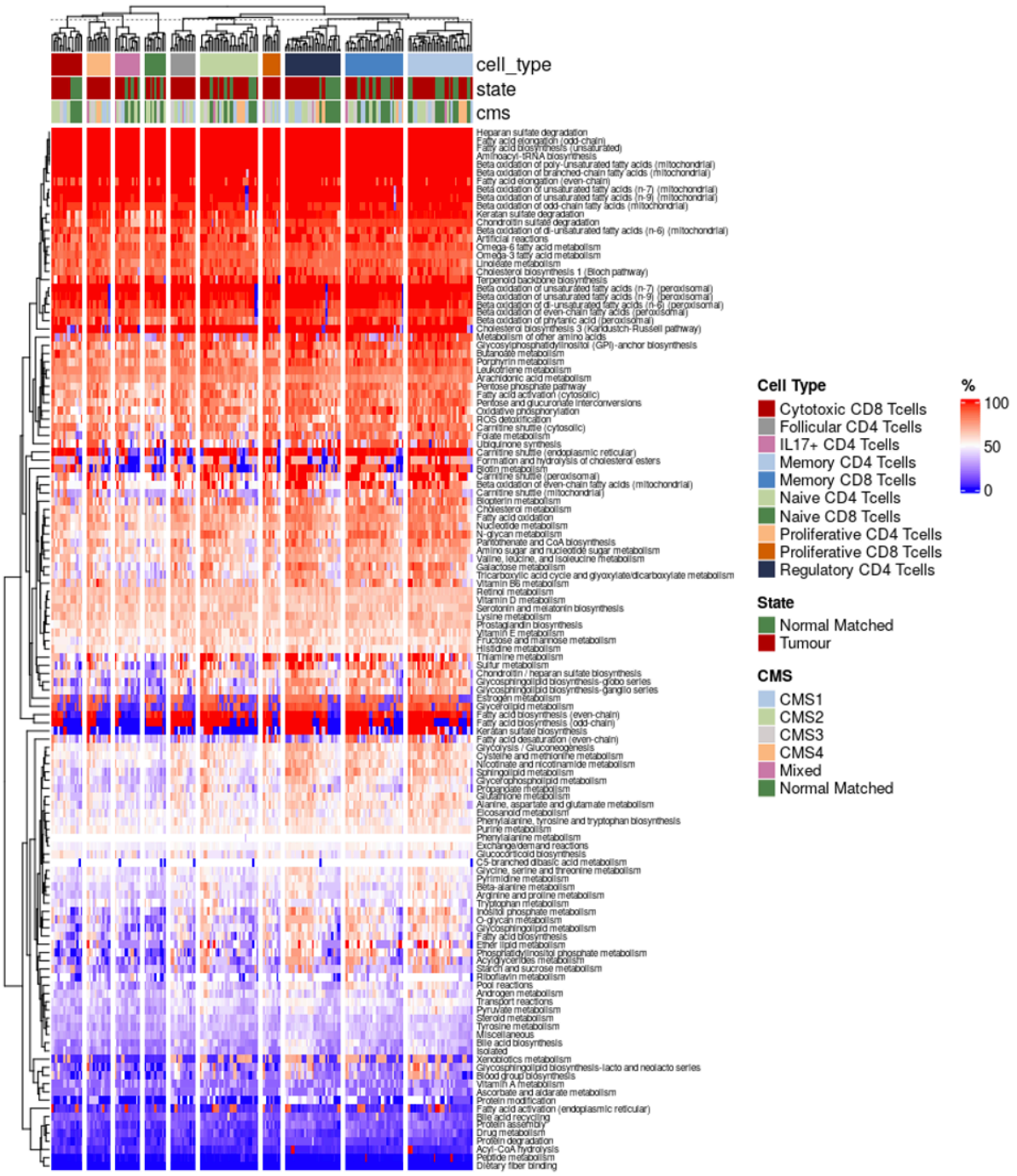


Figure 57: Pathway coverage (%).

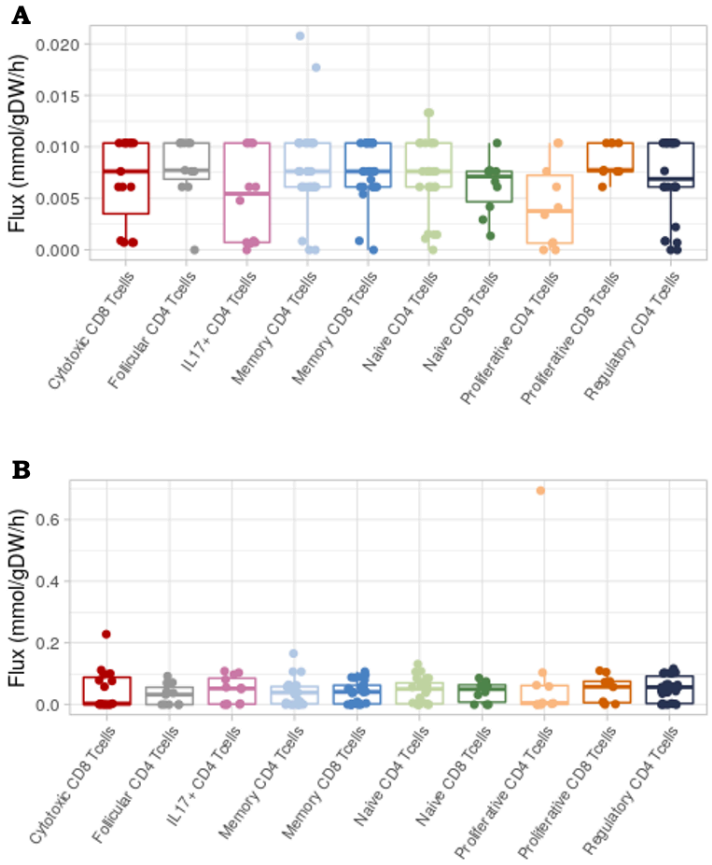


Figure 58: Biomass (A) and ATP (B) production when biomass was set as the only objective for all models.

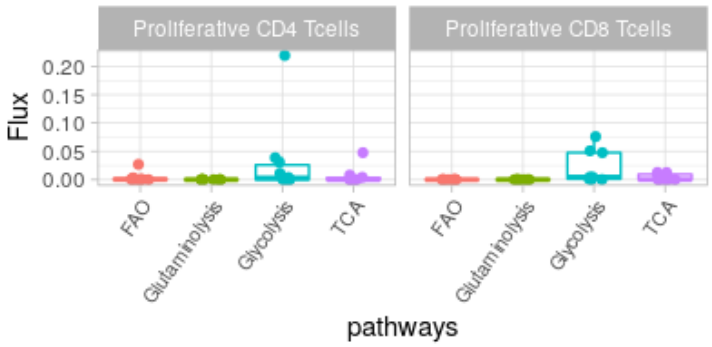


Figure 59: Cumulative fluxes (mmol/gDW/h) of the reactions that produce NADH, from all source pathways, for proliferative CD4 and CD8 T-cell models

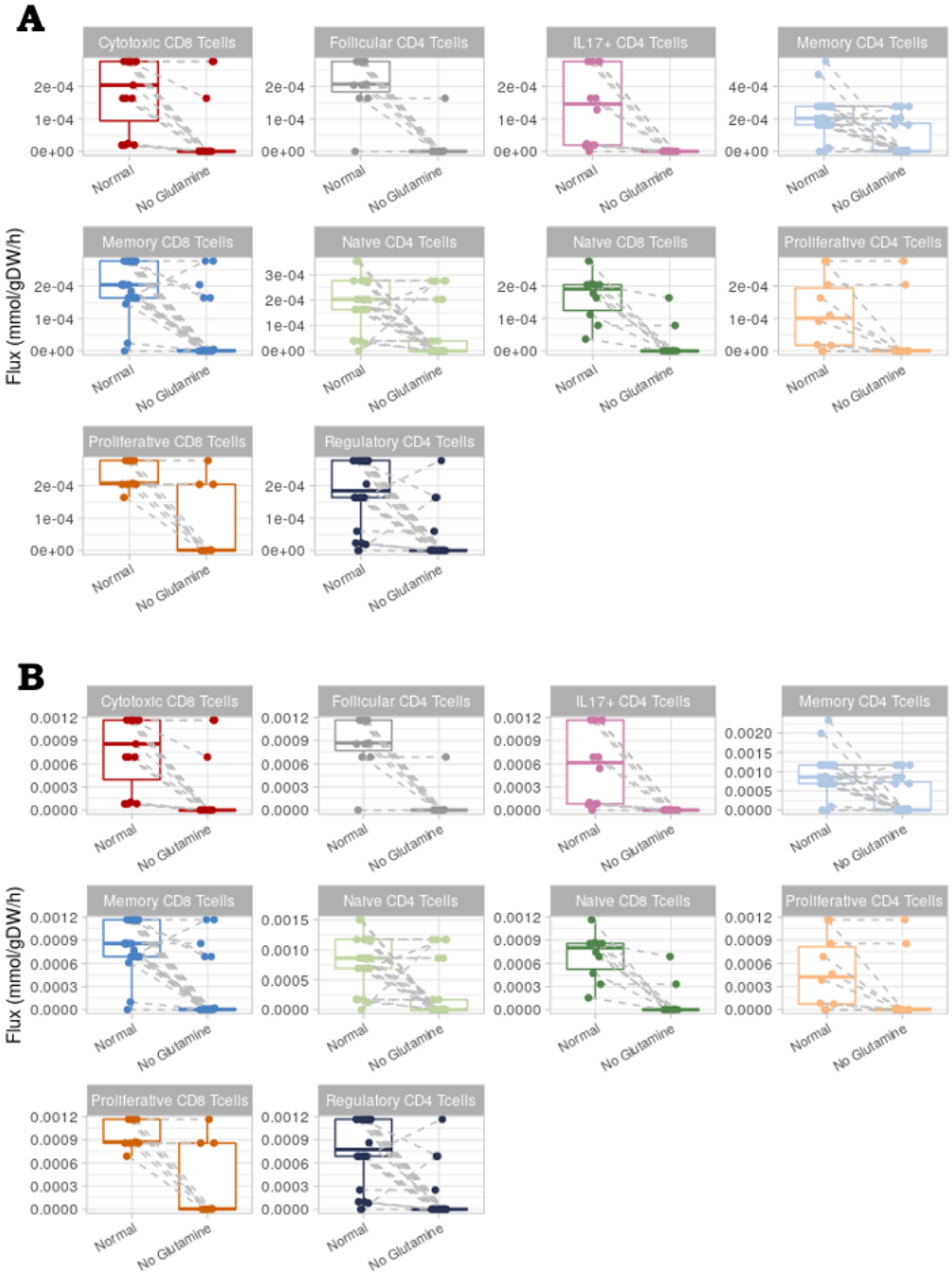


Figure 60: Distribution of (A) DNA and (B) RNA production of the T-cell types, with and without glutamine in the medium.

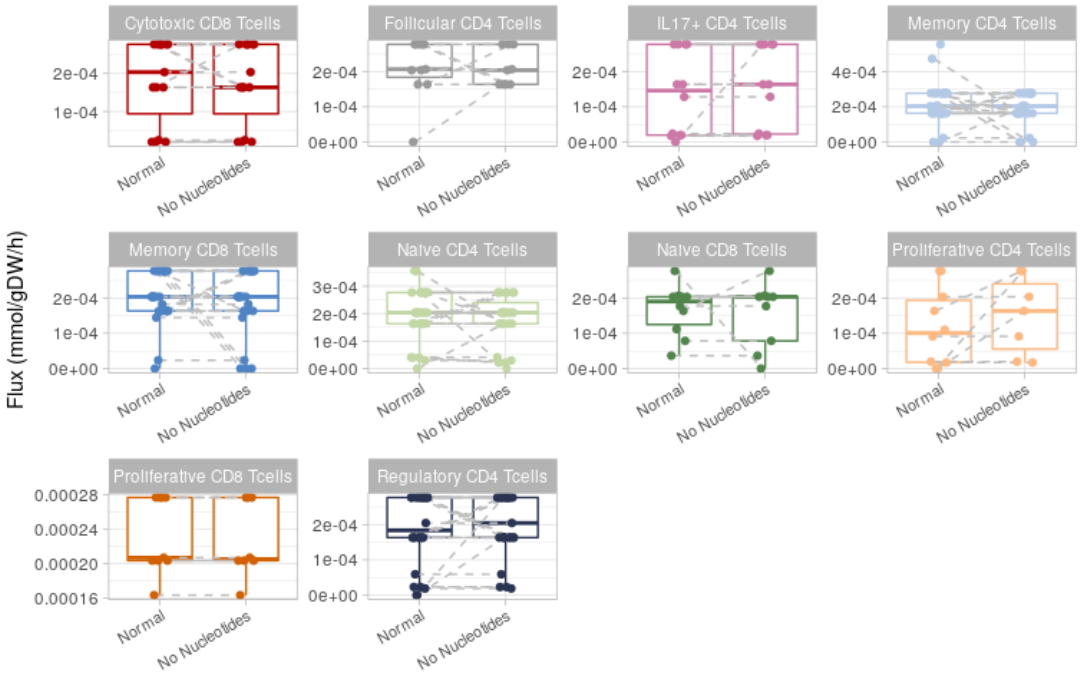


Figure 61: Distribution of DNA production of the T-cell types, with and without nucleotides in the medium.

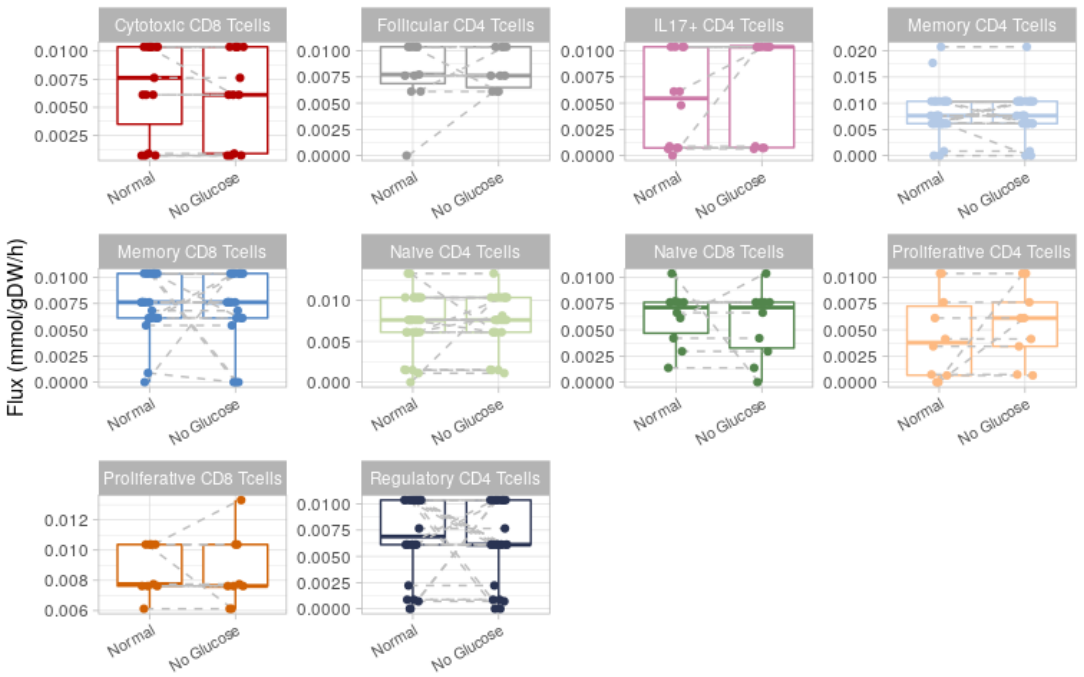


Figure 62: Distribution of the biomass flux of the T-cell types, with and without glucose in the medium.

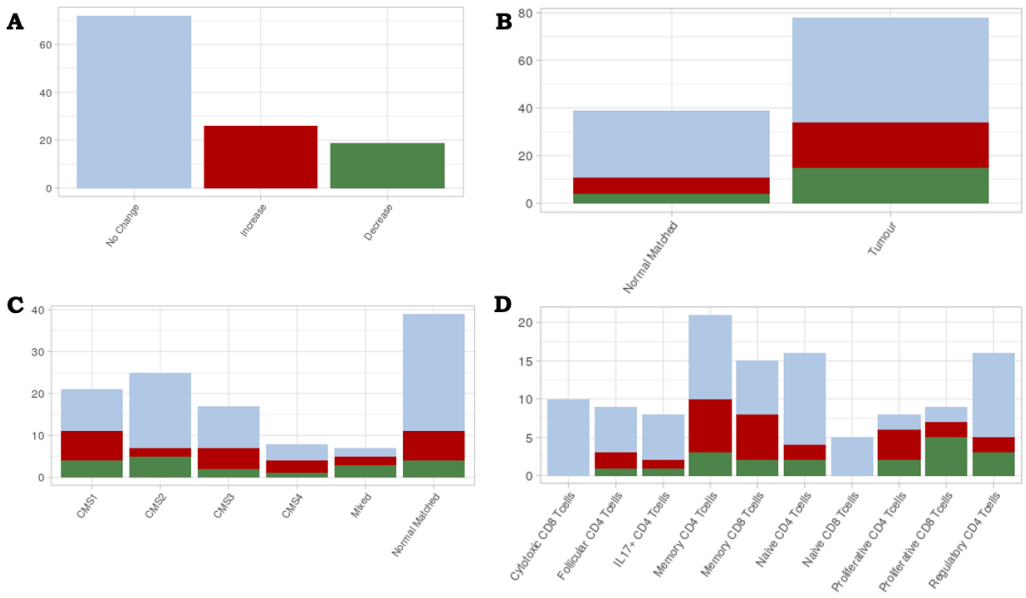


Figure 63: (A) Number of models where biomass flux increases, decreases or suffers no change. This information is further showed by (B) tissue of origin, (C) CMS type, and (D) cell-type.

A.3 Chapter 5

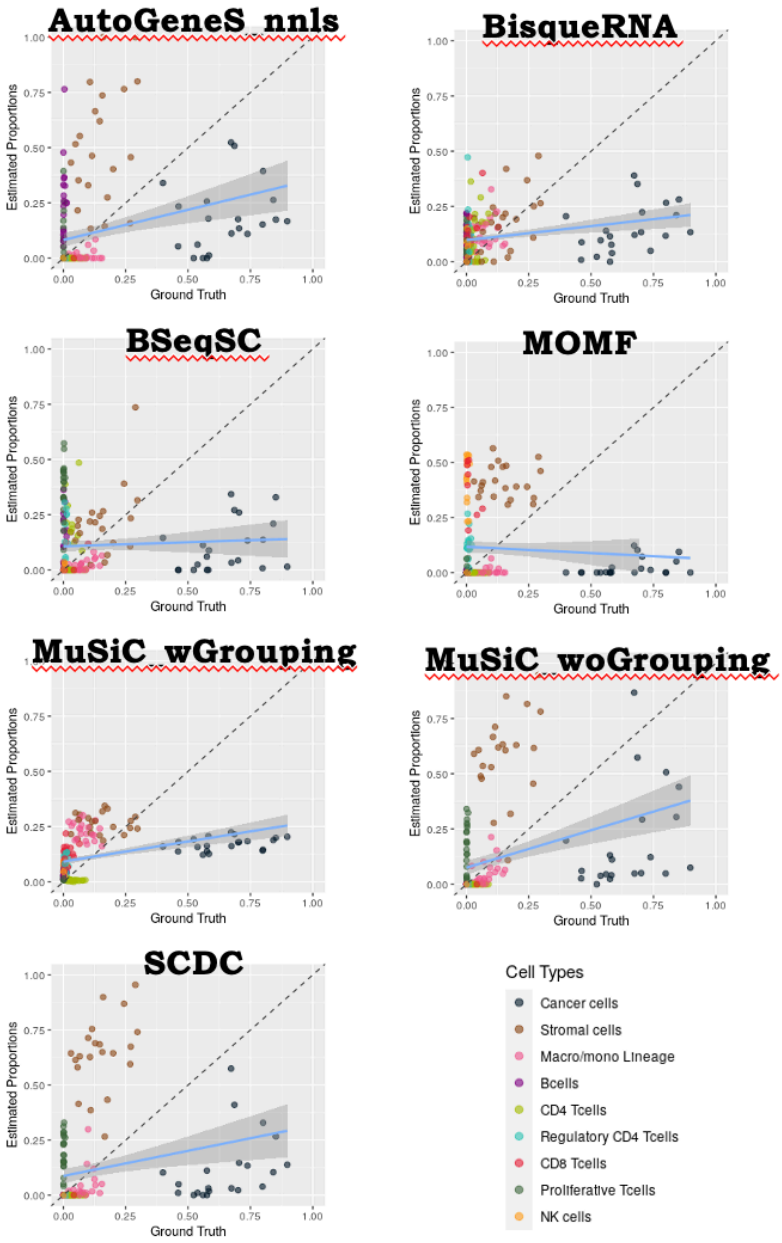


Figure 64: Scatter plots of estimated vs ground-truth proportions for the remaining methods that are not present in figure 42.

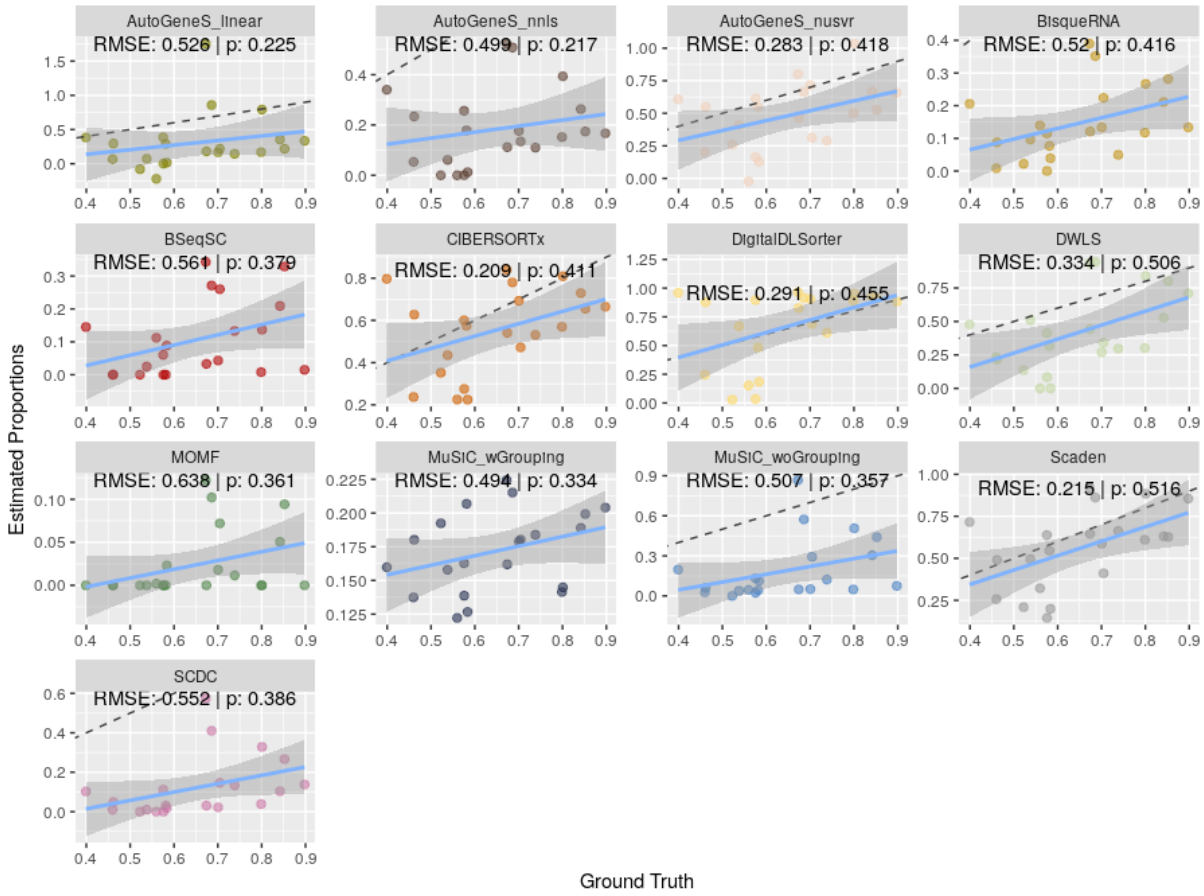


Figure 65: Scatter plots of estimated vs ground-truth proportions of all methods for cancer cells.

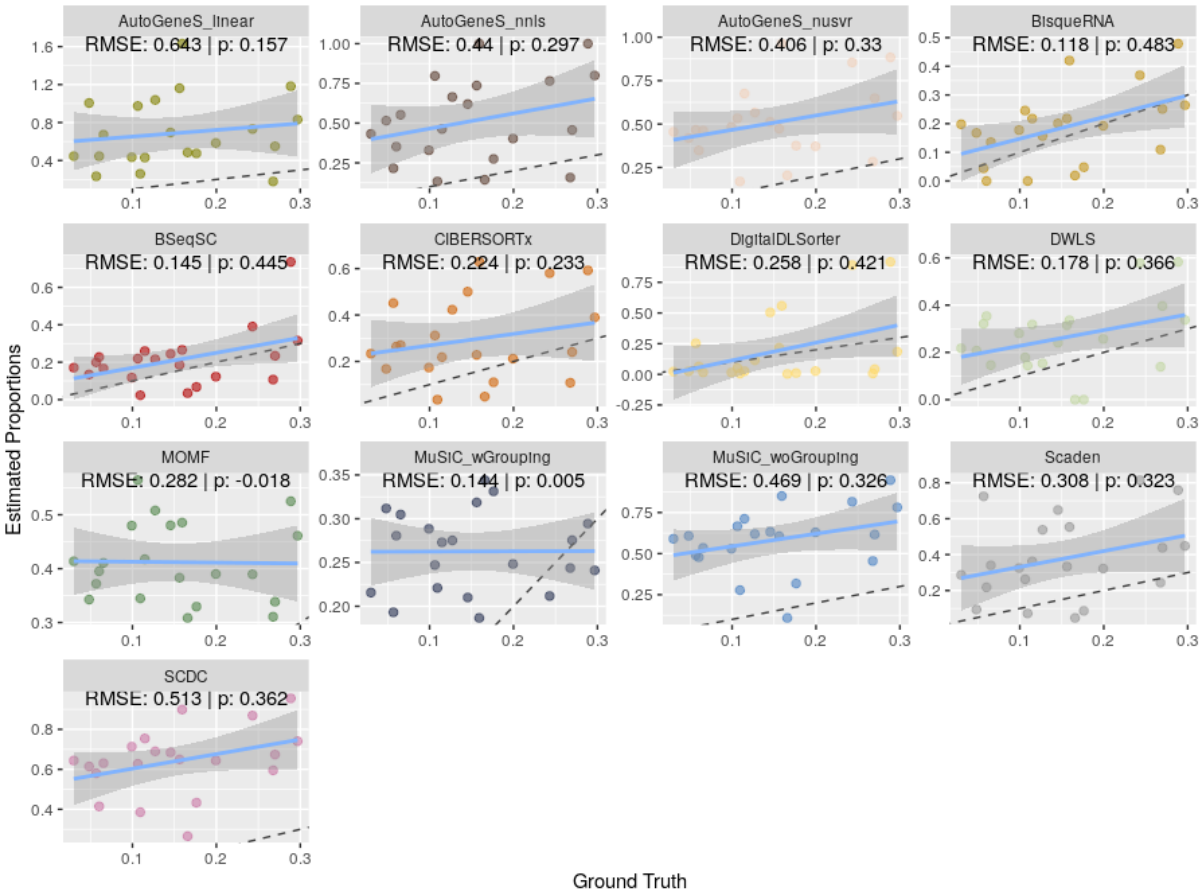


Figure 66: Scatter plots of estimated vs ground-truth proportions of all methods for stromal cells.

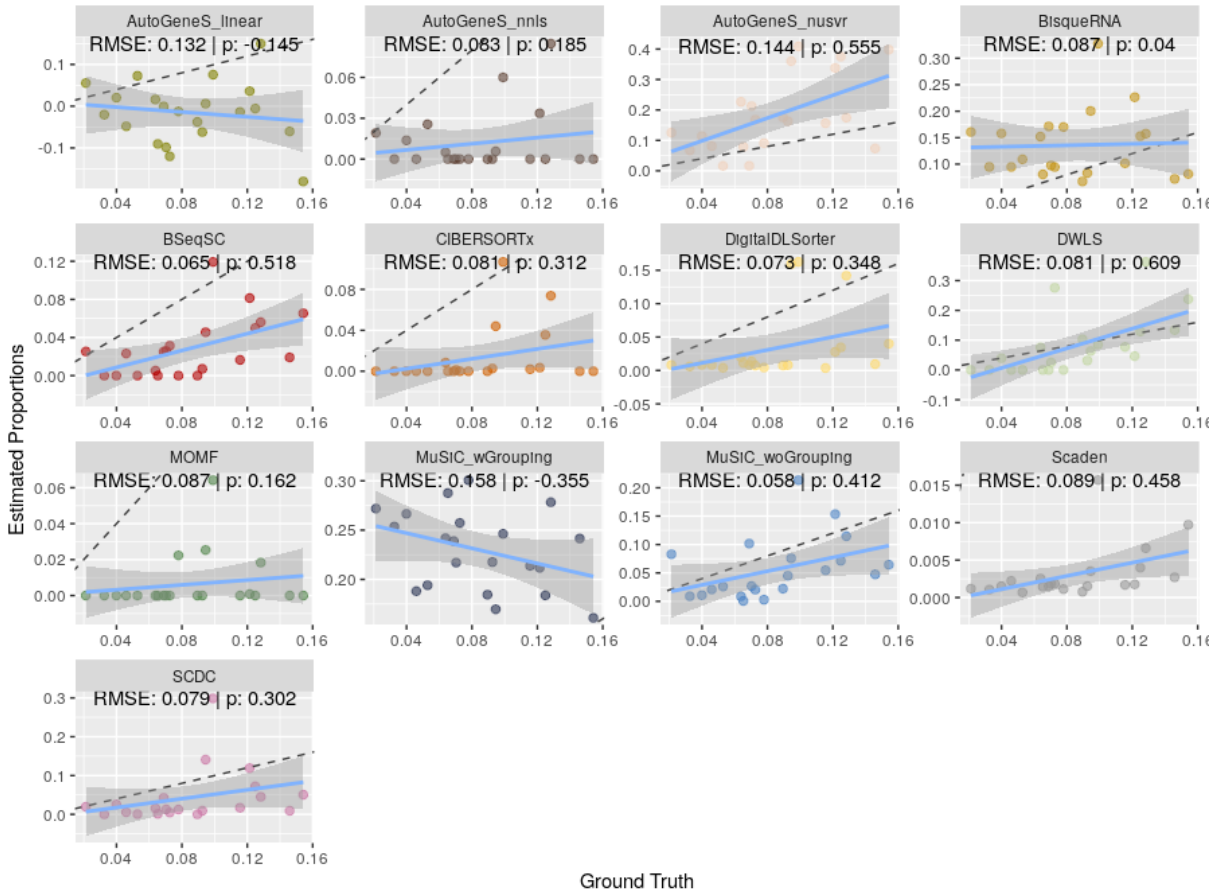


Figure 67: Scatter plots of estimated vs ground-truth proportions of all methods for macro/mono lineage cells.

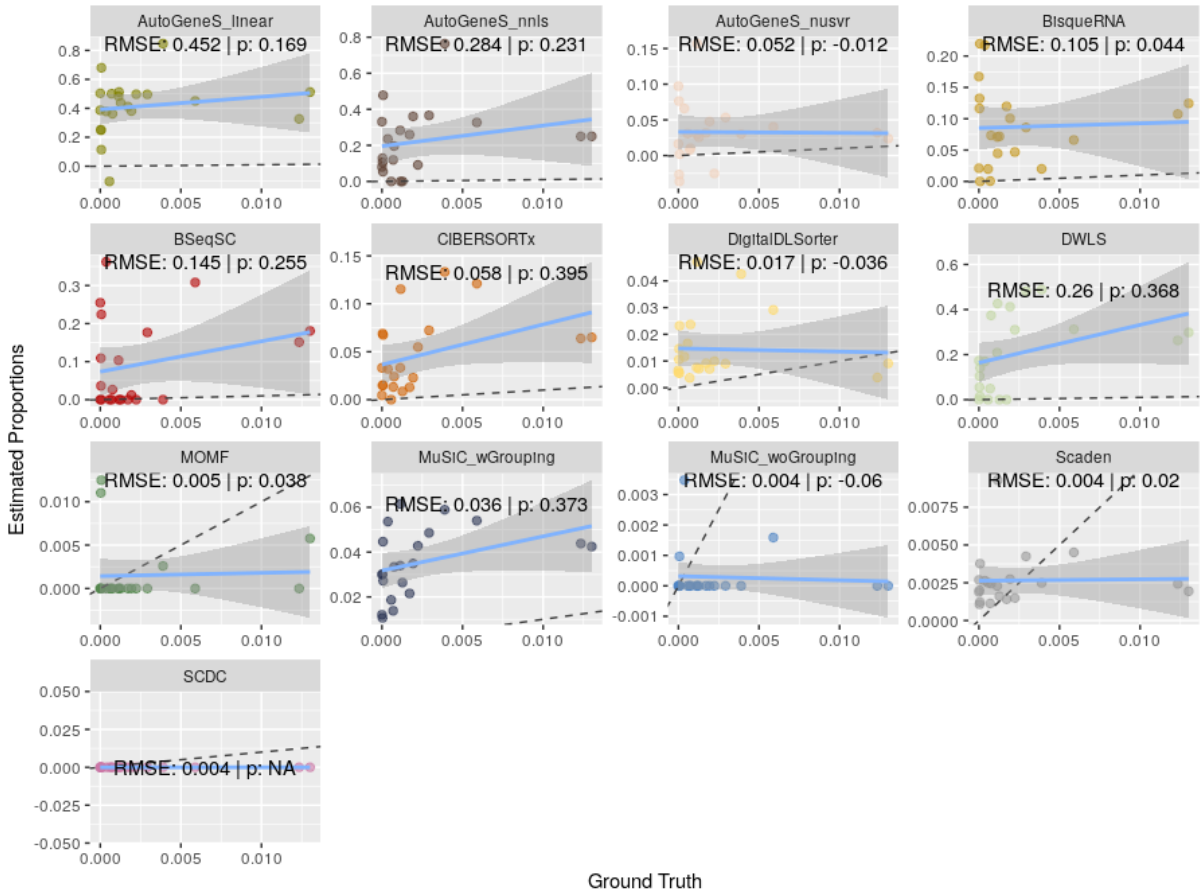


Figure 68: Scatter plots of estimated vs ground-truth proportions of all methods for B-cells.

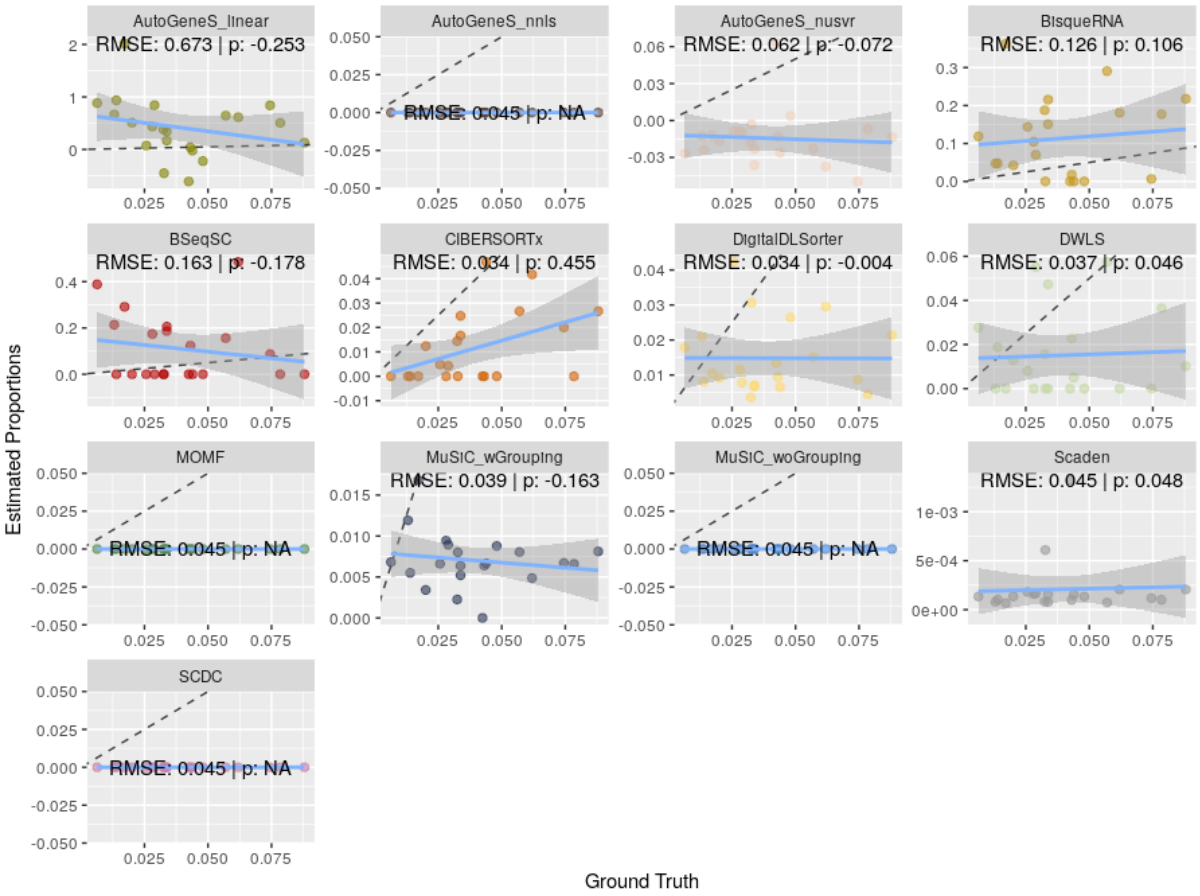


Figure 69: Scatter plots of estimated vs ground-truth proportions of all methods for CD4 T-cells.

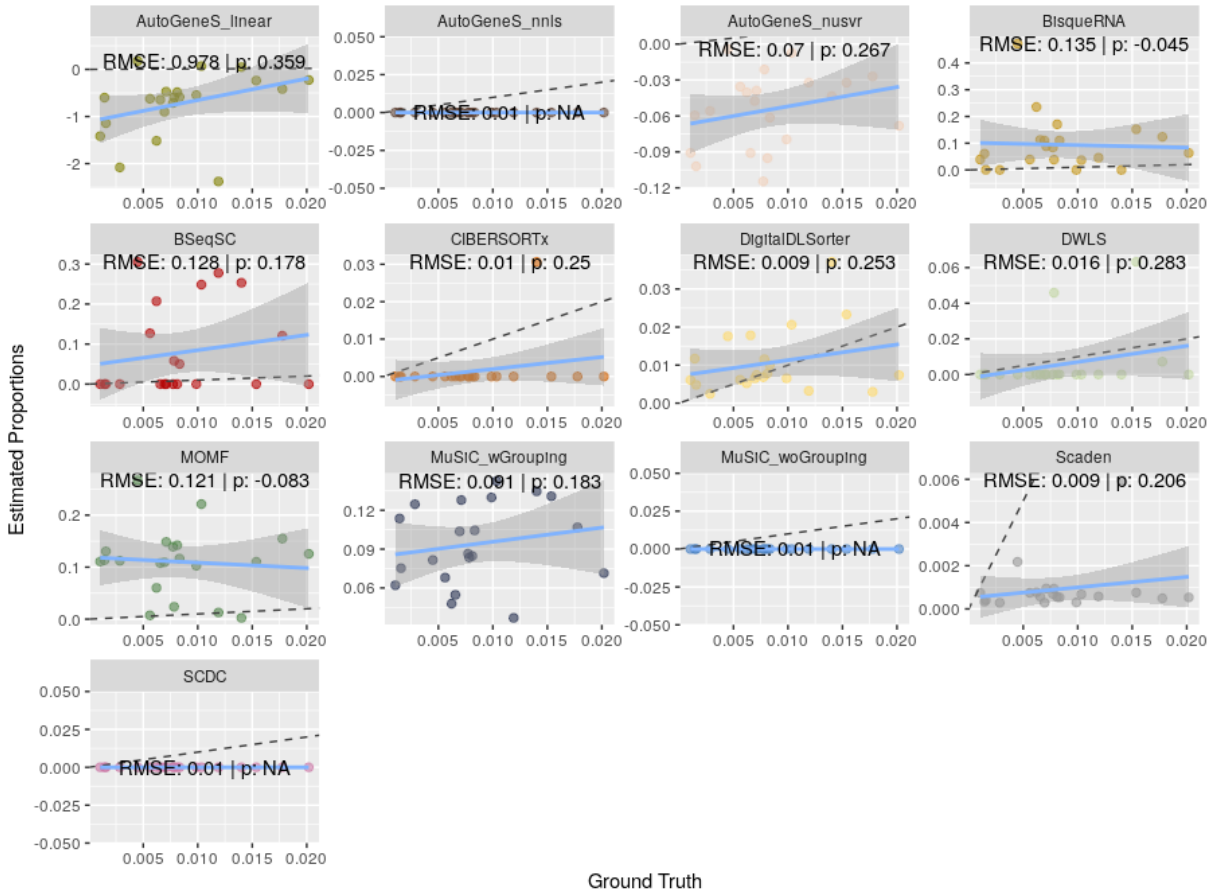


Figure 70: Scatter plots of estimated vs ground-truth proportions of all methods for regulatory CD4 T-cells.

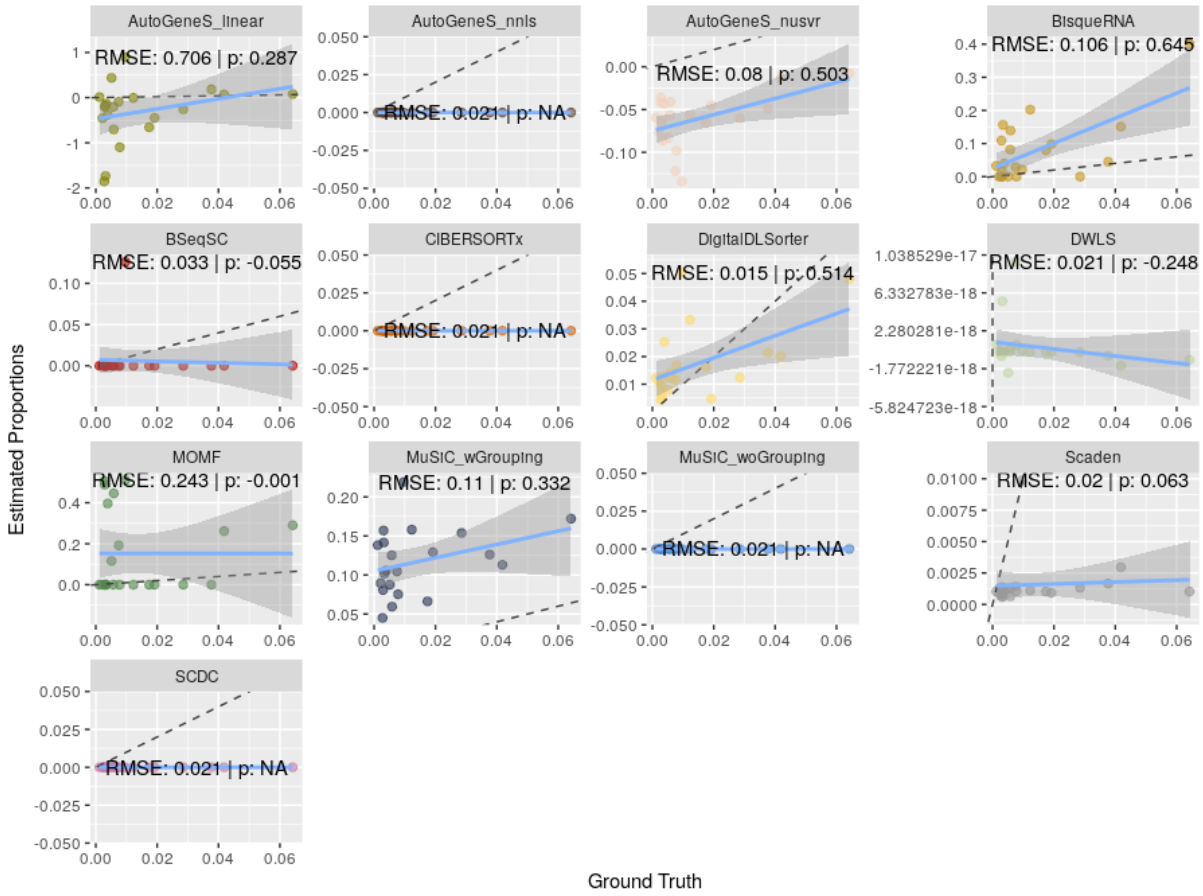


Figure 71: Scatter plots of estimated vs ground-truth proportions of all methods for CD8 T-cells.

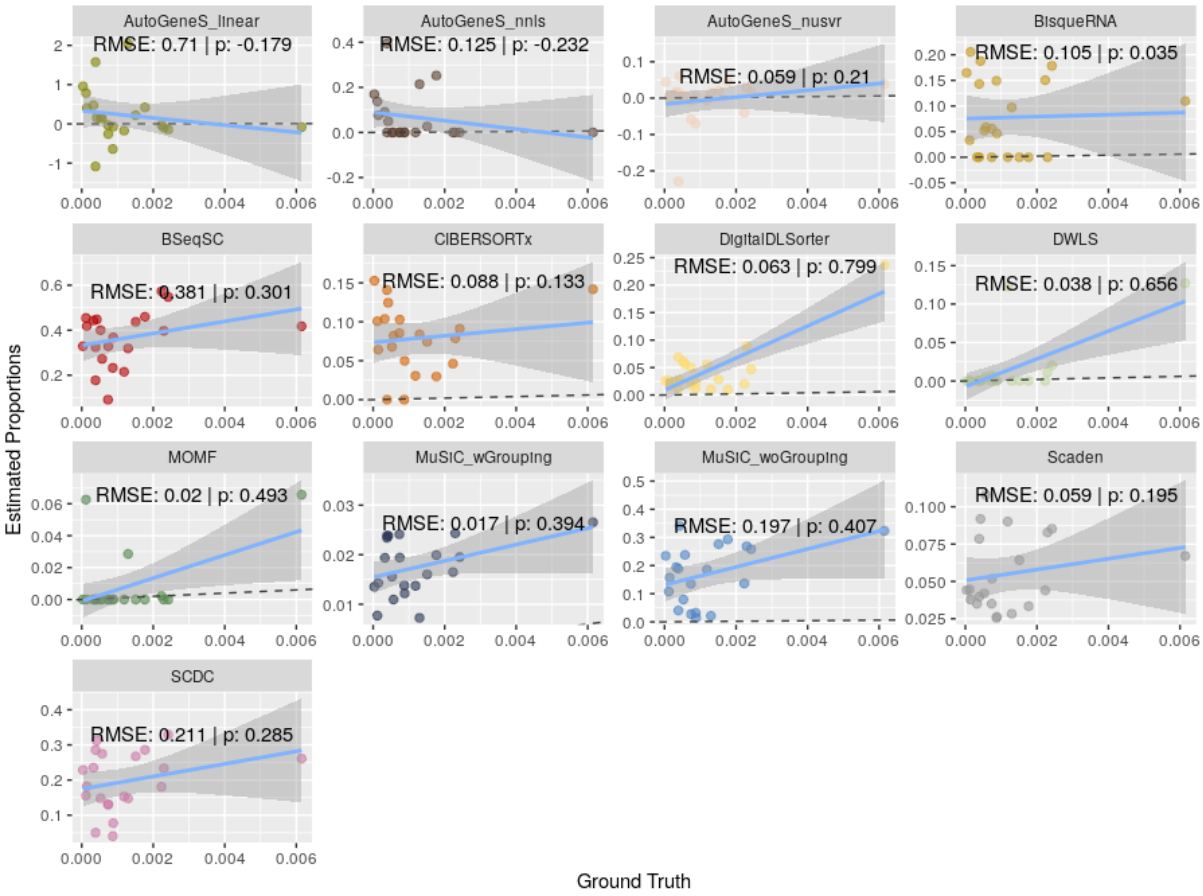


Figure 72: Scatter plots of estimated vs ground-truth proportions of all methods for proliferative T-cells.

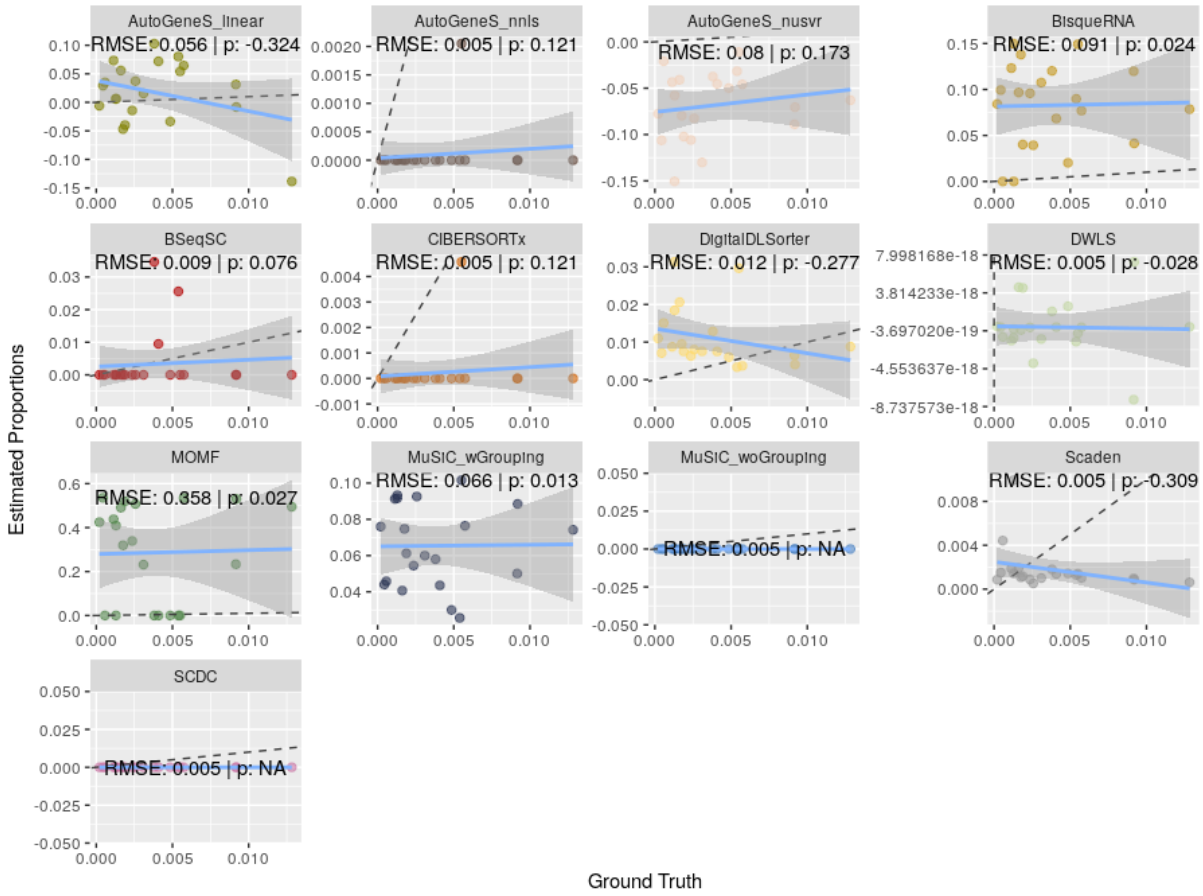


Figure 73: Scatter plots of estimated vs ground-truth proportions of all methods for NK cells.

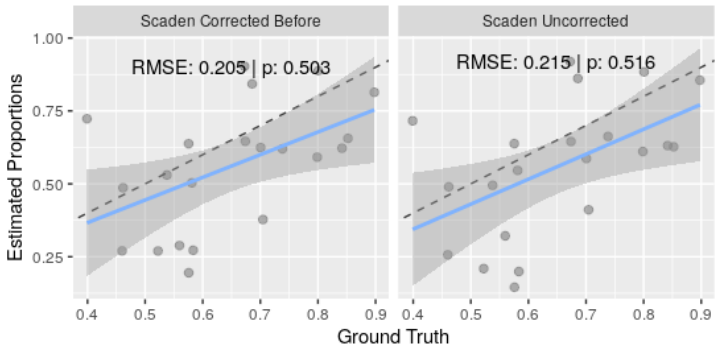


Figure 74: For cancer cells, scatter plots of estimated vs ground-truth proportions of best uncorrected method and the corrected methods whose RMSE improved relative to the best uncorrected method.

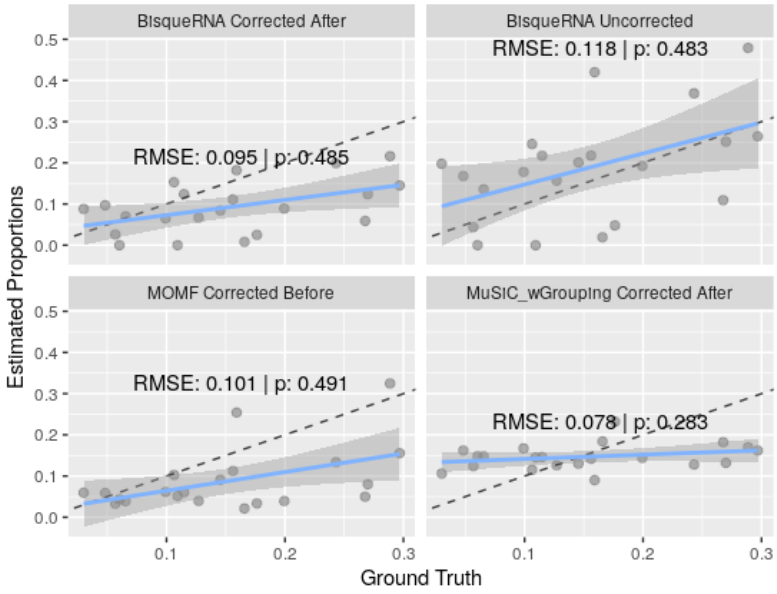


Figure 75: For stromal cells, scatter plots of estimated vs ground-truth proportions of best uncorrected method and the corrected methods whose RMSE improved relative to the best uncorrected method.

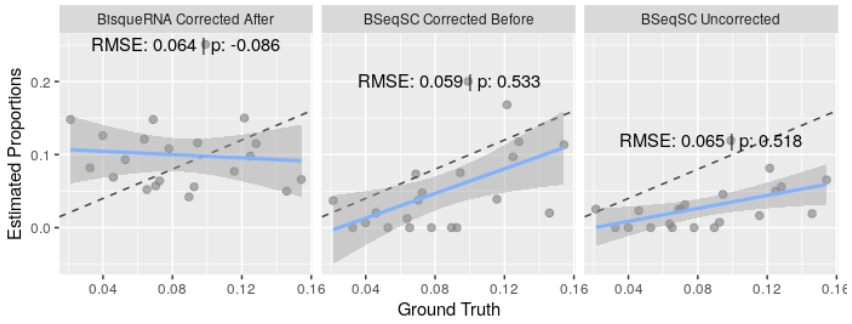


Figure 76: For macro/mono lineage cells, scatter plots of estimated vs ground-truth proportions of best uncorrected method and the corrected methods whose RMSE improved relative to the best uncorrected method.

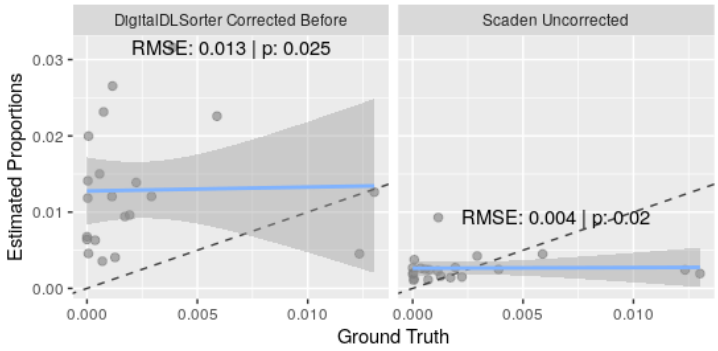


Figure 77: For B-cells, scatter plots of estimated vs ground-truth proportions of best uncorrected method and the corrected methods whose RMSE improved relative to the best uncorrected method.

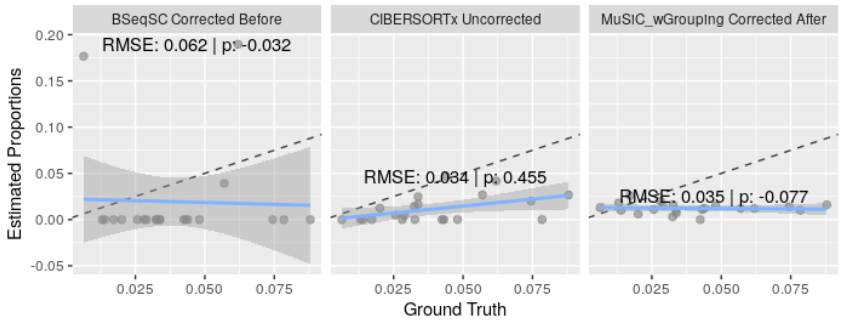


Figure 78: For CD4 T-cells, scatter plots of estimated vs ground-truth proportions of best uncorrected method and the corrected methods whose RMSE improved relative to the best uncorrected method.

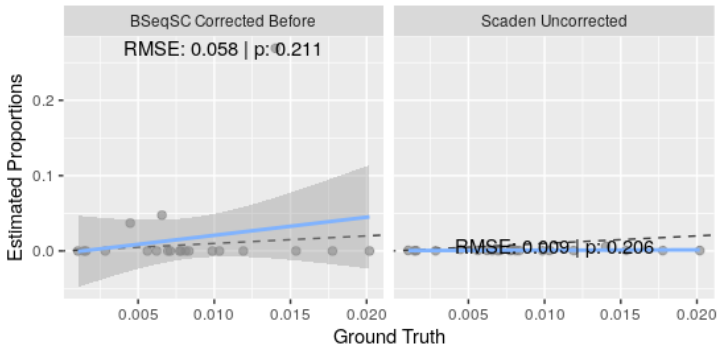


Figure 79: For regulatory CD4 T-cells, scatter plots of estimated vs ground-truth proportions of best uncorrected method and the corrected methods whose RMSE improved relative to the best uncorrected method.

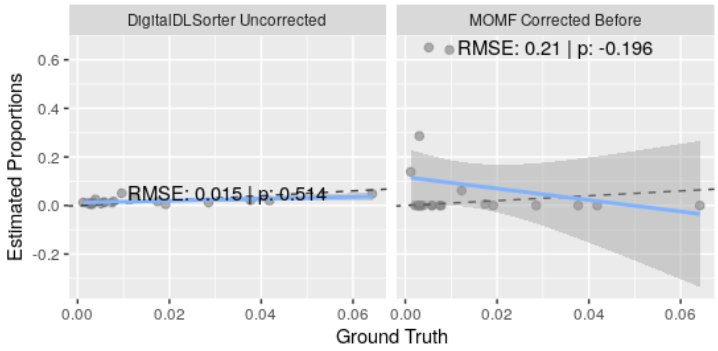


Figure 80: For CD8 T-cells, scatter plots of estimated vs ground-truth proportions of best uncorrected method and the corrected methods whose RMSE improved relative to the best uncorrected method.

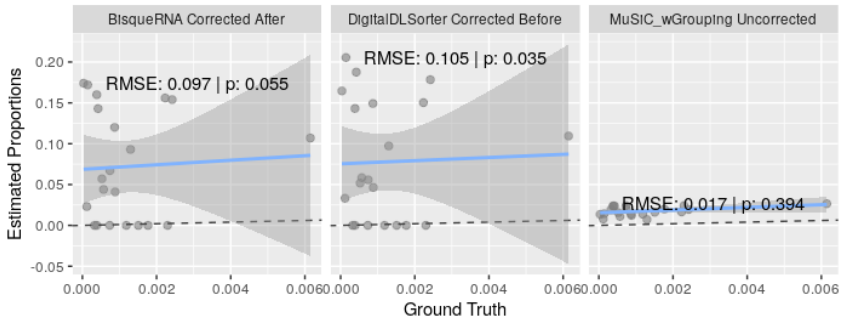


Figure 81: For proliferative T-cells, scatter plots of estimated vs ground-truth proportions of best uncorrected method and the corrected methods whose RMSE improved relative to the best uncorrected method.

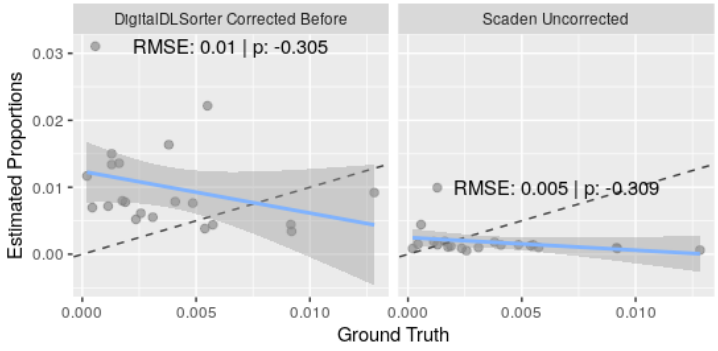


Figure 82: For NK cells, scatter plots of estimated vs ground-truth proportions of best uncorrected method and the corrected methods whose RMSE improved relative to the best uncorrected method.

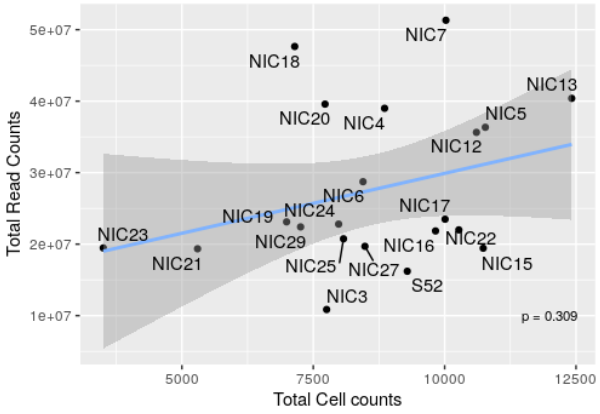


Figure 83: Scatter plot of samples' total read counts vs total cell counts.

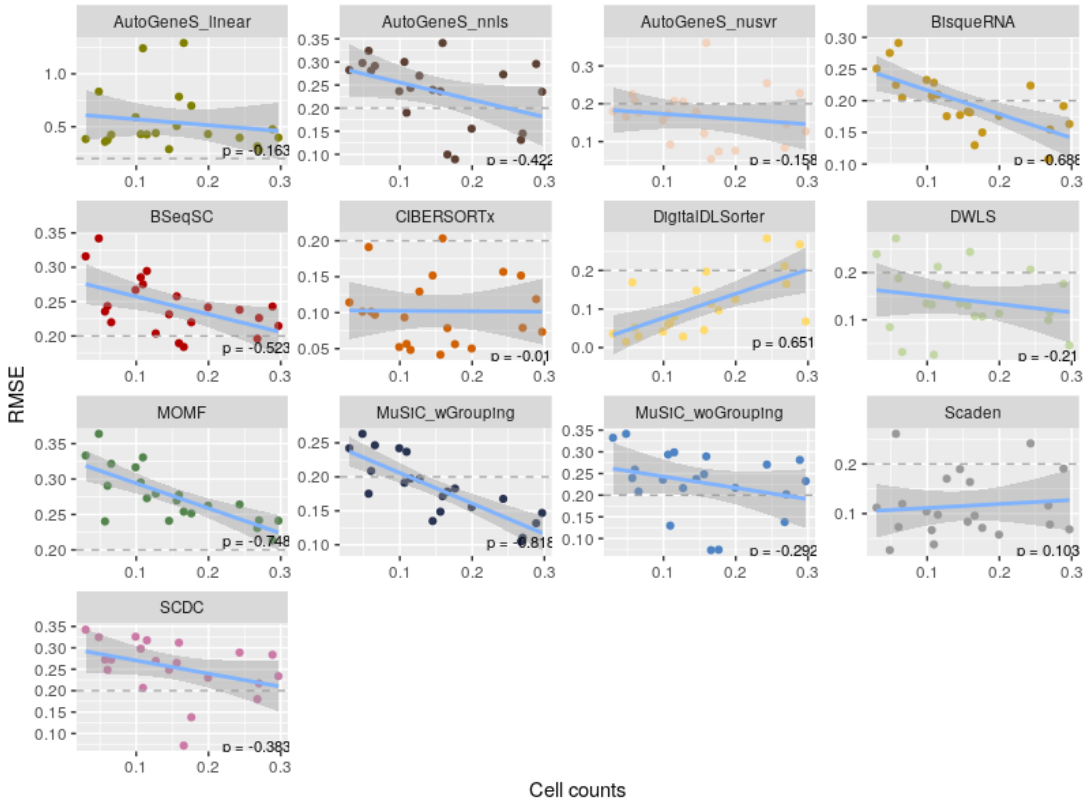


Figure 84: For each method, RMSE of the samples is mapped against their corresponding proportion of stromal cells.

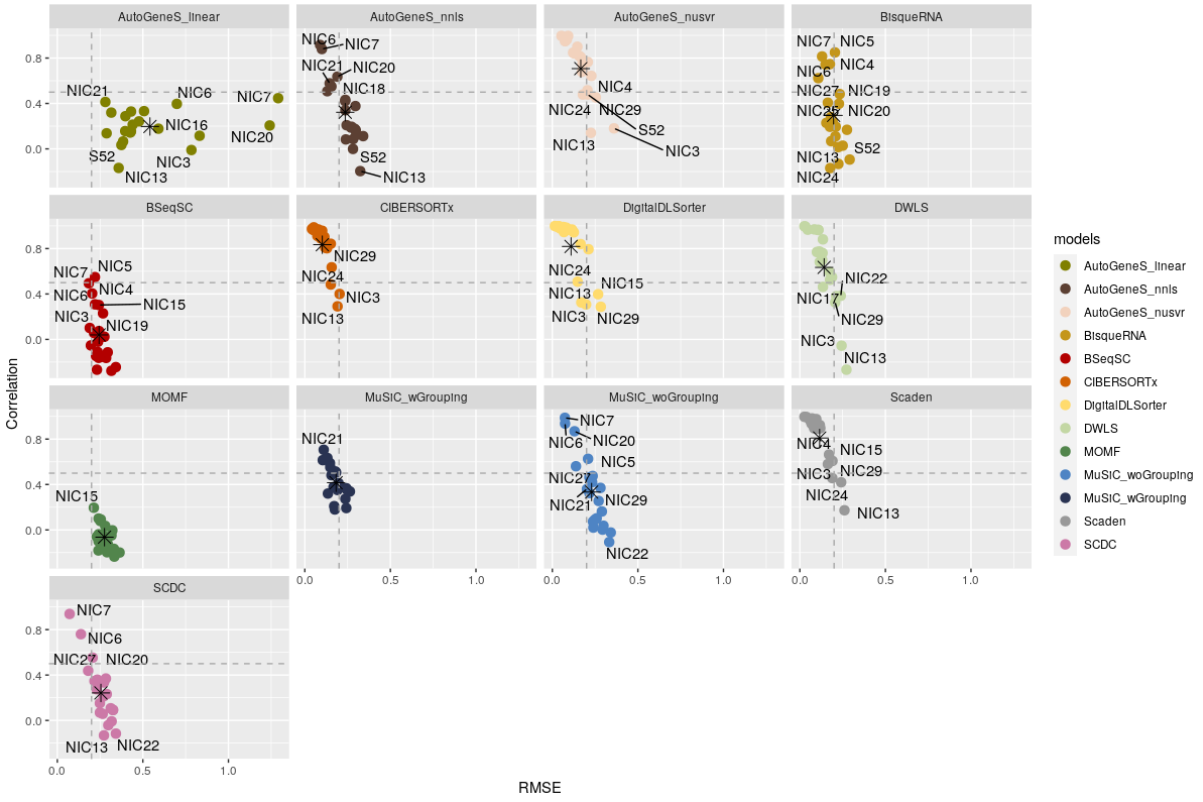


Figure 85: For each uncorrected method, scatter plot of pearson correlation vs RMSE of the samples.

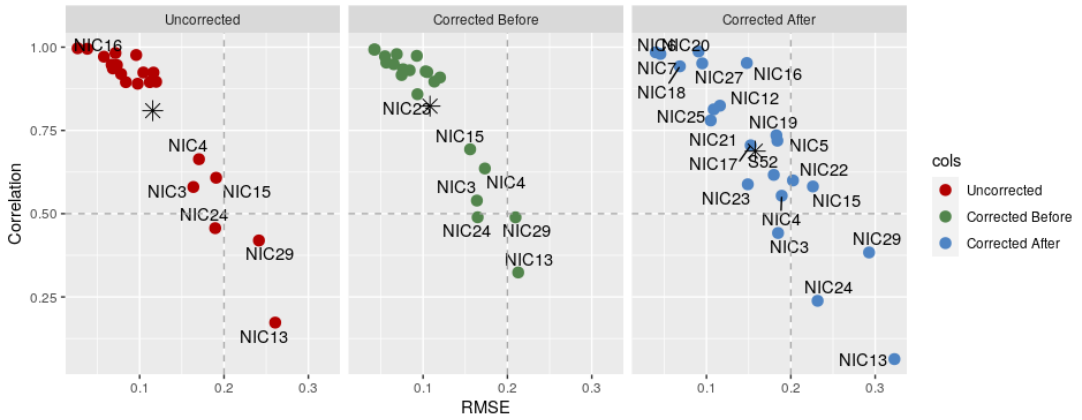


Figure 86: For *Scaden*, scatter plots of pearson correlation vs RMSE of the samples for each correction type.

B

Supplementary Tables

This appendix contains all supplementary tables, separated by chapters.

B.1 Chapter 4

Table 11: Metabolites used in the human blood medium, with the corresponding exchange reaction id from the model. Each metabolite has information on the average concentration (mM) in normal human blood, gathered from SMDB database, and the fluxes (mmol/gDW/h) used in normal and tumour human blood media.

Metabolite Name	Exchange Reaction	Concentrations Normal Blood	Fluxes Normal Blood	Fluxes Blood Tumour
(11Z,14Z,17Z)-eicosatrienoic acid	MAR13042	3,410E-04	1,076E-03	1,076E-03
(13Z)-eicosenoic acid	MAR13043	1,172E-02	3,700E-02	3,700E-02
(18R)-HEPE	MAR11907	1,380E-07	4,356E-07	4,356E-07
(R)-3-hydroxybutanoate	MAR09134	7,025E-01	2,217E+00	6,861E+00
(R)-mevalonate	MAR10262	3,230E-05	1,020E-04	1,020E-04
(S)-2-aminobutanoate	MAR10206	2,280E-02	7,197E-02	1,159E-01
(S)-Glycerate	MAR00603	2,000E-03	6,313E-03	8,207E-03
1-methylnicotinamide	MAR09104	4,300E-04	1,357E-03	1,357E-03
1,2-diacylglycerol-LD-TAG pool	MAR00574	5,567E-01	1,757E+00	1,757E+00
1,3-Diaminopropane	MAR11933	4,000E-05	1,263E-04	1,263E-04
10Z-Heptadecenoic acid	MAR13040	1,030E-03	3,251E-03	3,251E-03
11-Dehydro-thromboxane B2	MAR11911	2,750E-06	8,681E-06	8,681E-06
11-deoxycorticosterone	MAR11866	3,680E-05	1,162E-04	1,162E-04
11-deoxycortisol	MAR11865	3,900E-06	1,231E-05	1,231E-05
11,12-EET	MAR10215	5,420E-07	1,711E-06	1,711E-06
12-hydroxy-arachidonate	MAR11836	8,340E-05	2,633E-04	2,633E-04
12,13-hydroxyoctadec-9(z)-enoate	MAR11969	8,780E-06	2,771E-05	2,771E-05
12(13)-EpOME	MAR10218	6,450E-06	2,036E-05	2,036E-05
12(S)-HHT	MAR12128	2,030E-06	6,408E-06	6,408E-06
12(S)-HPETE	MAR10179	1,450E-06	4,577E-06	4,577E-06
13-cis-Retinoate	MAR11895	3,000E-06	9,470E-06	9,470E-06
13-cis-retinoyl-glucuronide	MAR09210	6,600E-06	2,083E-05	2,083E-05
13,16,19-docosatrienoic acid	MAR13051	4,000E-06	1,263E-05	1,263E-05
13(S)-HPODE	MAR10208	6,010E-06	1,897E-05	1,897E-05

Table 11 (cont.)

Metabolite Name	Exchange Reaction	Concentrations Normal Blood	Fluxes Normal Blood	Fluxes Blood Tumour
14,15-DiHETE	MAR10234	1,130E-06	3,567E-06	3,567E-06
14,15-EET	MAR10216	9,420E-07	2,973E-06	2,973E-06
15-deoxy-PGD2	MAR10229	1,940E-06	6,124E-06	6,124E-06
15-Keto-prostaglandin F2a	MAR10181	1,990E-07	6,282E-07	6,282E-07
15(R)-HEPE	MAR11830	5,990E-07	1,891E-06	1,891E-06
15(S)-HEPE	MAR10232	5,990E-07	1,891E-06	1,891E-06
15(S)-HETrE	MAR10227	4,610E-07	1,455E-06	1,455E-06
15(S)-HPETE	MAR10180	1,060E-06	3,346E-06	3,346E-06
16α-hydroxyestrone	MAR10491	7,800E-07	2,462E-06	2,462E-06
17α-hydroxypregnenolone	MAR11870	6,820E-06	2,153E-05	2,153E-05
17α-hydroxypregnenolone sulfate	MAR11876	2,000E-03	6,313E-03	6,313E-03
17α-hydroxyprogesterone	MAR11869	5,070E-06	1,600E-05	1,600E-05
18-hydroxy-arachidonate	MAR11839	2,730E-07	8,617E-07	8,617E-07
2-arachidonoylglycerol	MAR13031	7,800E-03	2,462E-02	2,462E-02
2-Hydroxy-Isovalerate	MAR11481	8,633E-03	2,725E-02	2,725E-02
2-Hydroxybutyrate	MAR09216	4,223E-02	1,333E-01	2,181E-01
2-Hydroxyestrone	MAR11860	1,700E-07	5,366E-07	5,366E-07
2-hydroxyphenylacetate	MAR11440	5,030E-04	1,588E-03	1,588E-03
2-methoxyestradiol-17β	MAR11863	1,000E-05	3,157E-05	3,157E-05
2-Methylcitrate	MAR09217	7,950E-05	2,509E-04	2,509E-04
2-oxo-3-methylvalerate	MAR09012	2,035E-02	6,424E-02	6,424E-02
2-oxobutyrate	MAR11391	7,110E-03	2,244E-02	2,244E-02
2-phospho-D-glycerate	MAR09842	1,600E-03	5,051E-03	5,051E-03
2,5-dihydroxybenzoate	MAR11901	8,250E-04	2,604E-03	2,604E-03
20α-hydroxy-4-pregnen-3-one	MAR09268	2,910E-05	9,186E-05	9,186E-05
21-hydroxyallopregnanolone	MAR11864	5,200E-06	1,641E-05	1,641E-05
24-Hydroxycholesterol	MAR11879	6,200E-05	1,957E-04	1,957E-04
25-Hydroxycholesterol	MAR11884	1,050E-05	3,314E-05	3,314E-05
25-Hydroxyvitamin D2	MAR09214	9,001E+00	2,841E+01	2,841E+01
26-Hydroxycholesterol	MAR11881	2,840E-04	8,965E-04	8,965E-04
3-(3-Hydroxy-Phenyl)Propionate	MAR10426	1,440E-04	4,545E-04	4,545E-04
3-Hydroxy butyryl carnitine	MAR04815	8,200E-05	2,588E-04	2,588E-04
3-Hydroxy Trans7,10-Hexadecadienoyl Carnitine	MAR04830	1,500E-05	4,735E-05	4,735E-05
3-Hydroxy-3-Methyl-Glutarate	MAR11494	4,600E-02	1,452E-01	1,452E-01
3-Hydroxy-glutarate	MAR11519	1,500E-04	4,735E-04	4,735E-04
3-Hydroxy-isovaleryl carnitine	MAR04824	2,660E-04	8,396E-04	8,396E-04
3-hydroxy-L-kynurenine	MAR09865	5,000E-05	1,578E-04	1,578E-04
3-Hydroxyanthranilate	MAR09863	7,900E-05	2,494E-04	2,494E-04
3-Hydroxyhexadecanoylcarnitine	MAR04823	1,250E-05	3,946E-05	3,946E-05
3-Hydroxyhexadecenoylcarnitine	MAR04822	1,000E-05	3,157E-05	3,157E-05
3-Hydroxyisobutyrate	MAR10185	2,050E-02	6,471E-02	6,471E-02
3-iodo-L-tyrosine	MAR11891	6,900E-07	2,178E-06	2,178E-06
3-Methoxytyramine	MAR10427	2,500E-06	7,891E-06	7,891E-06
3-methyl-2-oxobutyrate	MAR09011	1,250E-02	3,946E-02	3,946E-02
3-Methylhistidine	MAR10188	2,850E-03	8,996E-03	8,996E-03
3-O-methyldopa	MAR10225	8,900E-05	2,809E-04	2,809E-04
3-Phospho-D-glycerate	MAR09862	5,270E-02	1,664E-01	1,664E-01
3-phosphoserine	MAR10441	1,700E-02	5,366E-02	5,366E-02
3,4-Dihydroxymandelate	MAR11902	1,100E-05	3,472E-05	3,472E-05
3,4-dihydroxyphenylacetate	MAR11432	1,010E-05	3,188E-05	3,188E-05
3,4-Dihydroxyphenylethanol	MAR12114	1,600E-04	5,051E-04	5,051E-04

Table 11 (cont.)

Metabolite Name	Exchange Reaction	Concentrations Normal Blood	Fluxes Normal Blood	Fluxes Blood Tumour
3,4-dihydroxyphenylethyleneglycol	MAR09218	7,670E-06	2,421E-05	2,421E-05
3,5-Diiodo-L-tyrosine	MAR11893	3,600E-06	1,136E-05	1,136E-05
3α,12alpha-dihydroxy-5β-cholanate	MAR08646	4,500E-04	1,420E-03	1,420E-03
4-Acetamidobutanoate	MAR10190	5,000E-04	1,578E-03	1,578E-03
4-aminobutyrate	MAR09091	2,100E-04	6,629E-04	6,254E-04
4-androstene-3,17-dione	MAR11848	3,690E-06	1,165E-05	1,165E-05
4-coumarate	MAR11823	2,010E-04	6,345E-04	6,345E-04
4-hydroxy-2-nonenal	MAR11912	1,100E-04	3,472E-04	3,472E-04
4-hydroxy-2-quinolinecarboxylic acid	MAR09855	2,650E-05	8,365E-05	8,365E-05
4-Hydroxy-butyrate	MAR11568	2,460E-02	7,765E-02	7,765E-02
4-Hydroxy-debrisoquine	MAR09223	2,170E-04	6,850E-04	6,850E-04
4-Hydroxybenzoate	MAR11431	4,809E-03	1,518E-02	2,429E-02
4-hydroxyphenylacetate	MAR09224	9,881E-03	3,119E-02	3,119E-02
4-hydroxyphenyllactate	MAR10184	6,400E-04	2,020E-03	2,020E-03
4-Hydroxyphenylpyruvate	MAR09010	3,700E-04	1,168E-03	1,168E-03
4-methyl-2-oxopentanoate	MAR09013	2,650E-02	8,365E-02	8,365E-02
4-Pyridoxate	MAR09228	2,100E-05	6,629E-05	6,629E-05
5-α-dihydrotestosterone	MAR09229	1,940E-06	6,124E-06	6,124E-06
5-Aminolevulinate	MAR10428	3,500E-04	1,105E-03	1,105E-03
5-formyl-THF	MAR09100	2,500E-06	7,891E-06	7,891E-06
5-guanidino-2-oxopentanoate	MAR11847	1,230E-04	3,883E-04	3,883E-04
5-Hydroxy-L-tryptophan	MAR09094	1,800E-05	5,682E-05	5,682E-05
5-Hydroxyindoleacetate	MAR09843	5,160E-05	1,629E-04	1,629E-04
5-Hydroxytryptophol	MAR11952	9,000E-07	2,841E-06	2,841E-06
5-Methoxytryptophol	MAR11918	1,248E-07	3,938E-07	3,938E-07
5-methyl-THF	MAR09234	3,000E-04	9,470E-04	9,470E-04
5-oxoproline	MAR09025	1,950E-02	6,155E-02	6,155E-02
5,10-Methylene-THF	MAR11953	1,000E-05	3,157E-05	3,157E-05
5,15-DiHETE	MAR11909	3,400E-07	1,073E-06	1,073E-06
5,6-dihydrouracil	MAR10193	3,130E-04	9,880E-04	9,880E-04
5,6-EET	MAR10214	3,560E-06	1,124E-05	1,124E-05
5(S)-HEPE	MAR11908	7,280E-07	2,298E-06	2,298E-06
5(S)-HETE	MAR10209	5,010E-04	1,581E-03	1,581E-03
5α-androstane-3,17-dione	MAR11855	3,600E-07	1,136E-06	1,136E-06
5α-androstane-3α,17beta-diol	MAR11856	4,750E-07	1,499E-06	1,499E-06
5α-pregnane-3,20-dione	MAR11871	3,605E-05	1,138E-04	1,138E-04
6-Hydroxymelatonin	MAR11922	2,400E-07	7,576E-07	7,576E-07
6-oxo-prostaglandin F1alpha	MAR10219	2,690E-07	8,491E-07	8,491E-07
6-trans-LTB4	MAR10226	1,850E-07	5,840E-07	5,840E-07
8,9-EET	MAR10145	6,270E-07	1,979E-06	1,979E-06
9-cis-Retinoate	MAR11897	4,900E-06	1,547E-05	1,547E-05
9-Eicosenoic acid	MAR13054	1,172E-02	3,700E-02	3,700E-02
9,10-hydroxyoctadec-12(Z)-enoate	MAR11971	5,020E-05	1,585E-04	1,585E-04
9(10)-EpOME	MAR10217	3,050E-06	9,628E-06	9,628E-06
Acetaldehyde	MAR09242	1,000E-03	3,157E-03	3,157E-03
Acetate	MAR09086	4,674E-02	1,475E-01	1,741E-01
Acetoacetate	MAR09132	1,521E-01	4,800E-01	4,800E-01
Acetone	MAR09243	6,503E-02	2,053E-01	2,053E-01
Acetyl-glycine	MAR10196	8,957E-02	2,827E-01	2,126E-01
Adenine	MAR09253	4,700E-04	1,484E-03	1,736E-03
Adenosine	MAR09254	1,344E-03	4,242E-03	4,242E-03

Table 11 (cont.)

Metabolite Name	Exchange Reaction	Concentrations Normal Blood	Fluxes Normal Blood	Fluxes Blood Tumour
Adipic acid	MAR10200	9,000E-05	2,841E-04	2,841E-04
ADP	MAR09255	1,600E-01	5,051E-01	5,051E-01
adrenaline	MAR09095	7,380E-07	2,330E-06	2,330E-06
Adrenic acid	MAR00566	4,073E-03	1,286E-02	1,286E-02
AKG	MAR09259	1,050E-02	3,314E-02	3,314E-02
Alanine	MAR09061	3,586E-01	1,132E+00	1,132E+00
Aldosterone	MAR09261	6,780E-07	2,140E-06	2,140E-06
Allantoin	MAR10431	2,100E-03	6,629E-03	8,220E-03
Allopregnanolone	MAR11868	3,730E-06	1,177E-05	1,177E-05
α-D-Glucose 1,6-bisphosphate	MAR09461	9,800E-02	3,093E-01	3,093E-01
α-Tocopherol	MAR09151	3,243E-02	1,024E-01	1,024E-01
α-Tocotrienol	MAR09152	4,230E-03	1,335E-02	1,335E-02
Aminoacetone	MAR11956	6,000E-02	1,894E-01	1,894E-01
AMP	MAR09262	1,908E-02	6,021E-02	6,021E-02
Anandamide	MAR10213	2,039E-03	6,437E-03	6,437E-03
Androsterone	MAR09263	6,030E-05	1,903E-04	1,903E-04
Androsterone sulfate	MAR10230	1,110E-02	3,504E-02	3,504E-02
Androsterone-glucuronide	MAR09264	4,370E-04	1,379E-03	1,379E-03
anthranilate	MAR09028	1,550E-05	4,893E-05	4,893E-05
aquacob(III)alamin	MAR09269	2,720E-07	8,586E-07	8,586E-07
Arachidonate	MAR00568	1,711E-01	5,400E-01	5,400E-01
Arginine	MAR09066	8,581E-02	2,708E-01	2,708E-01
Argininosuccinate	MAR09919	2,100E-03	6,629E-03	6,629E-03
Ascorbate	MAR09158	4,361E-02	1,377E-01	1,377E-01
Asparagine	MAR09062	5,314E-02	1,677E-01	1,962E-01
aspartate	MAR09070	2,172E-02	6,857E-02	1,046E-01
ATP	MAR00569	1,793E+00	5,659E+00	5,659E+00
Azelaic acid	MAR11351	2,700E-02	8,523E-02	8,523E-02
Behenic acid	MAR04929	1,095E-02	3,457E-02	3,457E-02
Benzoate	MAR10475	2,079E-02	6,561E-02	6,561E-02
β-Alanine	MAR09260	2,635E-03	8,318E-03	8,318E-03
β-Carotene	MAR09276	3,526E-02	1,113E-01	1,113E-01
β-hydroxy-β-methylbutyrate	MAR10224	4,000E-03	1,263E-02	1,263E-02
Betaine	MAR09341	7,255E-02	2,290E-01	2,290E-01
Bilirubin	MAR09273	4,481E-02	1,415E-01	1,415E-01
Biotin	MAR09109	2,570E-05	8,112E-05	8,112E-05
Butyrate	MAR09809	1,000E-03	3,157E-03	3,157E-03
Calcidiol	MAR09215	6,840E-05	2,159E-04	2,159E-04
Calcitriol	MAR11965	8,000E-08	2,525E-07	2,525E-07
cAMP	MAR09275	8,500E-06	2,683E-05	2,683E-05
Carnosine	MAR08644	3,230E-03	1,020E-02	1,020E-02
CDP	MAR04096	3,600E-02	1,136E-01	1,136E-01
Ceramide pool	MAR04384	8,880E-03	2,803E-02	2,803E-02
cerotic acid	MAR00618	7,230E-04	2,282E-03	2,282E-03
cGMP	MAR09220	5,500E-06	1,736E-05	1,736E-05
Chenodeoxycholate	MAR04144	1,083E-03	3,417E-03	3,417E-03
Chenodiol	MAR10026	1,083E-03	3,417E-03	3,417E-03
Chloride	MAR09150	1,030E+02	3,253E+02	3,253E+02
Cholate	MAR09280	8,400E-04	2,652E-03	2,652E-03
Cholesterol	MAR09285	3,124E+00	9,861E+00	9,861E+00
Cholesterol-sulfate	MAR11889	5,400E-03	1,705E-02	1,705E-02

Table 11 (cont.)

Metabolite Name	Exchange Reaction	Concentrations Normal Blood	Fluxes Normal Blood	Fluxes Blood Tumour
Choline	MAR09083	1,328E-02	4,190E-02	5,028E-02
cis-cetoleic acid	MAR13056	2,950E-05	9,312E-05	9,312E-05
cis-erucic acid	MAR04927	7,164E-03	2,261E-02	2,261E-02
cis-gondoic acid	MAR10432	4,628E-03	1,461E-02	1,461E-02
cis-Vaccenic acid	MAR00704	9,860E-02	3,112E-01	3,112E-01
Cis,Cis-11,14-Eicosadienoic Acid	MAR10228	9,036E-03	2,852E-02	2,852E-02
Citrate	MAR09286	1,257E-01	3,968E-01	4,802E-01
Citrulline	MAR09201	4,189E-02	1,322E-01	1,322E-01
CO	MAR09288	7,200E-02	2,273E-01	2,273E-01
CO2	MAR09058	2,133E+01	6,733E+01	6,733E+01
Coproporphyrin I	MAR09701	8,100E-06	2,557E-05	2,557E-05
Coproporphyrin III	MAR09702	6,000E-06	1,894E-05	1,894E-05
Coproporphyrinogen I	MAR01965	1,500E-05	4,735E-05	4,735E-05
Corticosterone	MAR09294	2,700E-05	8,523E-05	8,523E-05
Cortisol	MAR09293	3,460E-04	1,092E-03	1,092E-03
Cortisone	MAR10235	4,720E-05	1,490E-04	1,490E-04
Creatine	MAR09290	5,045E-02	1,593E-01	1,593E-01
Creatinine	MAR09460	6,058E-02	1,912E-01	1,723E-01
Cyanate	MAR04334	4,500E-05	1,420E-04	1,420E-04
cys-gly	MAR09279	5,188E-02	1,638E-01	1,638E-01
Cysteine	MAR09065	1,407E-01	4,442E-01	4,442E-01
Cystine	MAR09363	8,146E-02	2,571E-01	2,880E-01
Cytidine	MAR09295	1,750E-04	5,524E-04	5,524E-04
Cytosine	MAR09291	6,400E-03	2,020E-02	2,020E-02
D-3-amino-isobutanoate	MAR09222	1,643E-03	5,186E-03	5,186E-03
D-Alanine	MAR09098	4,468E-01	1,410E+00	1,410E+00
D-Arabitol	MAR10429	1,500E-03	4,735E-03	6,392E-03
D-Aspartate	MAR09097	1,179E-02	3,720E-02	5,674E-02
D-glucitol	MAR09685	7,045E-03	2,224E-02	2,224E-02
D-gluconic acid	MAR11393	3,295E-03	1,040E-02	1,040E-02
D-Lactate	MAR09136	9,130E-03	2,882E-02	4,016E-02
D-Ornithine	MAR09454	8,900E-02	2,809E-01	2,809E-01
D-Xylose	MAR09203	2,443E+00	7,712E+00	1,026E+01
D-Xylulose	MAR11942	2,500E-03	7,891E-03	7,891E-03
debrisoquin	MAR09299	1,490E-04	4,703E-04	4,703E-04
decanoic acid	MAR09815	1,100E-02	3,472E-02	3,472E-02
Decanoyl carnitine	MAR04859	2,370E-04	7,481E-04	7,481E-04
Dehydroascorbic acid	MAR09301	5,772E-03	1,822E-02	1,822E-02
dehydroepiandrosterone	MAR11850	1,330E-05	4,198E-05	4,198E-05
dehydroepiandrosterone sulfate	MAR09302	7,547E-03	2,382E-02	2,382E-02
Deoxycytidine	MAR09296	2,000E-04	6,313E-04	6,313E-04
Deoxyuridine	MAR09310	2,840E-04	8,965E-04	1,363E-03
Desmosterol	MAR11887	1,834E-03	5,789E-03	5,789E-03
DHA	MAR00573	1,056E-01	3,334E-01	3,334E-01
DHAP	MAR01922	1,385E-02	4,372E-02	4,372E-02
dihomo-γ-linolenate	MAR00576	4,666E-02	1,473E-01	1,473E-01
dihydrobiopterin	MAR10495	4,180E-03	1,319E-02	1,319E-02
Dihydrofolate	MAR09303	5,000E-06	1,578E-05	1,578E-05
Dimethylglycine	MAR09848	2,500E-03	7,891E-03	6,468E-03
Dodecanedioic acid	MAR10240	1,099E-01	3,469E-01	3,469E-01
Dodecanedioyl carnitine	MAR04869	2,270E-04	7,165E-04	7,165E-04

Table 11 (cont.)

Metabolite Name	Exchange Reaction	Concentrations Normal Blood	Fluxes Normal Blood	Fluxes Blood Tumour
Dopamine	MAR09092	6,500E-08	2,052E-07	2,052E-07
Dopamine 3-O-sulfate	MAR09308	2,650E-05	8,365E-05	8,365E-05
Dopamine 4-O-sulfate	MAR12007	2,700E-06	8,523E-06	8,523E-06
DPA	MAR00571	3,073E-02	9,699E-02	9,699E-02
Ecgonine-methyl ester	MAR11833	2,370E-04	7,481E-04	7,481E-04
eicosanoate	MAR00567	7,278E-03	2,297E-02	2,297E-02
Elaidate	MAR00583	1,000E-01	3,157E-01	4,040E-01
EPA	MAR00701	1,434E-01	4,528E-01	4,528E-01
Estradiol-17β	MAR09314	2,720E-07	8,586E-07	8,586E-07
estradiol-17β 3-glucuronide	MAR09129	3,870E-08	1,222E-07	1,222E-07
Estriol	MAR09452	4,780E-06	1,509E-05	1,509E-05
Estrone	MAR11858	1,630E-07	5,145E-07	5,145E-07
Estrone 3-sulfate	MAR09317	2,330E-06	7,355E-06	7,355E-06
Ethanol	MAR09099	2,383E-02	7,523E-02	7,523E-02
Ethanolamine	MAR09084	2,695E-02	8,507E-02	8,507E-02
ethanolamine-phosphate	MAR09849	1,150E-02	3,630E-02	3,457E-02
Etiocholanolone	MAR09453	1,250E-06	3,946E-06	3,946E-06
FAD	MAR01939	7,133E-05	2,252E-04	2,252E-04
Fe2+	MAR09076	8,922E+00	2,816E+01	2,816E+01
Fe3+	MAR09096	1,778E-02	5,612E-02	5,612E-02
FMN	MAR08962	1,060E-05	3,346E-05	3,346E-05
Folate	MAR09146	2,370E-05	7,481E-05	7,481E-05
Formaldehyde	MAR01946	1,630E-02	5,145E-02	5,145E-02
Formate	MAR09318	1,032E-01	3,257E-01	3,257E-01
formyl-N-acetyl-5-methoxykynurenamine	MAR11904	6,500E-08	2,052E-07	2,052E-07
Fructose	MAR09139	3,950E-02	1,247E-01	1,247E-01
Fructose-1,6-bisphosphate	MAR11974	2,500E-03	7,891E-03	7,891E-03
Fumarate	MAR11400	1,223E-03	3,860E-03	4,208E-03
Galactitol	MAR11422	5,900E-04	1,862E-03	1,862E-03
Galactose	MAR09140	5,930E-02	1,872E-01	1,872E-01
γ-butyrobetaine	MAR10191	1,000E-02	3,157E-02	3,157E-02
γ-carboxyethyl-hydroxychroman	MAR04380	1,600E-04	5,051E-04	5,051E-04
γ-Glutamyl-cysteine	MAR11916	9,983E-03	3,151E-02	3,151E-02
γ-Linolenate	MAR00626	1,213E-02	3,829E-02	3,829E-02
γ-Tocopherol	MAR09153	1,156E-02	3,650E-02	3,650E-02
GDP	MAR09340	1,650E-02	5,208E-02	5,208E-02
globoside	MAR09336	2,150E-03	6,787E-03	6,787E-03
Glucosamine	MAR09168	2,900E-04	9,154E-04	1,172E-03
Glucose	MAR09034	4,688E+00	1,480E+01	3,152E+01
Glucosylceramide pool	MAR12043	9,600E-03	3,030E-02	3,030E-02
Glucuronate	MAR11424	1,650E-01	5,208E-01	6,979E-01
glutamate	MAR09071	6,542E-02	2,065E-01	2,722E-01
Glutamine	MAR09063	5,877E-01	1,855E+00	1,635E+00
Glutaryl carnitine	MAR04897	2,750E-05	8,681E-05	8,681E-05
Glyceraldehyde	MAR09851	1,476E+00	4,659E+00	6,243E+00
Glycerate	MAR09342	1,000E-02	3,157E-02	4,104E-02
Glycerol	MAR09085	1,407E-01	4,442E-01	1,159E+00
Glycine	MAR09067	4,567E-01	1,442E+00	1,888E+00
glycochenodeoxycholate	MAR09283	6,800E-04	2,146E-03	3,048E-03
Glycocholate	MAR09281	5,130E-04	1,619E-03	2,899E-03
Glycogen	MAR09729	4,120E-02	1,301E-01	1,301E-01

Table 11 (cont.)

Metabolite Name	Exchange Reaction	Concentrations Normal Blood	Fluxes Normal Blood	Fluxes Blood Tumour
Glycolate	MAR10436	3,405E-02	1,075E-01	1,215E-01
glycolithocholate	MAR04148	9,000E-06	2,841E-05	2,841E-05
Glycoursodeoxycholate	MAR04151	1,450E-04	4,577E-04	4,577E-04
Glyoxalate	MAR11448	3,050E-03	9,628E-03	9,628E-03
GM3	MAR12044	5,067E-03	1,599E-02	1,599E-02
GMP	MAR09343	9,700E-06	3,062E-05	3,062E-05
GSH	MAR09351	8,984E-02	2,836E-01	2,836E-01
GSSG	MAR09350	1,083E-02	3,417E-02	3,417E-02
GTP	MAR09352	5,600E-02	1,768E-01	1,768E-01
guanidinoacetate	MAR09852	4,890E-03	1,544E-02	1,544E-02
Guanosine	MAR09348	8,000E-04	2,525E-03	2,525E-03
H2O2	MAR09354	1,050E-02	3,314E-02	3,314E-02
H2S	MAR09103	5,165E-02	1,630E-01	1,630E-01
HCO3-	MAR09078	2,303E+01	7,269E+01	7,269E+01
Hepoxilin A3	MAR04409	1,140E-07	3,598E-07	3,598E-07
heptaglutamyl-folate(DHF)	MAR09239	7,600E-06	2,399E-05	2,399E-05
Hexadecanedioic acid	MAR10246	1,100E-04	3,472E-04	6,007E-04
Hexadecenoylcarnitine(9)	MAR11949	3,170E-05	1,001E-04	1,001E-04
hexanoic acid	MAR09811	1,700E-02	5,366E-02	5,366E-02
Hexanoy lcarnitine	MAR04899	6,030E-05	1,903E-04	1,903E-04
Hippurate	MAR10474	1,141E-02	3,603E-02	3,603E-02
Histamine	MAR00619	7,530E-05	2,377E-04	2,377E-04
Histidine	MAR09038	1,207E-01	3,809E-01	3,097E-01
Homocitrulline	MAR10249	5,000E-03	1,578E-02	1,578E-02
Homocysteine	MAR09853	8,808E-03	2,780E-02	2,780E-02
Homocysteine-thiolactone	MAR11914	2,820E-06	8,902E-06	8,902E-06
Homogentisate	MAR10247	4,300E-05	1,357E-04	1,245E-04
Homoserine	MAR09161	1,200E-02	3,788E-02	3,788E-02
Homovanillate	MAR09694	5,960E-05	1,881E-04	1,881E-04
Hyaluronate	MAR09122	5,300E-05	1,673E-04	1,673E-04
hydracrylate	MAR10186	4,100E-03	1,294E-02	1,294E-02
hydrogen-cyanide	MAR09160	4,850E-03	1,531E-02	1,531E-02
Hypoxanthine	MAR09358	6,322E-03	1,996E-02	1,996E-02
imidazole-4-acetate	MAR11951	1,000E-04	3,157E-04	3,157E-04
IMP	MAR09360	6,300E-02	1,989E-01	1,989E-01
Indoleacetate	MAR11419	9,830E-04	3,103E-03	3,103E-03
Inosine	MAR09362	1,388E-03	4,381E-03	4,381E-03
Inositol	MAR09361	2,353E-02	7,426E-02	8,837E-02
Isocitrate	MAR09854	6,000E-03	1,894E-02	1,894E-02
Isoleucine	MAR09039	6,371E-02	2,011E-01	2,645E-01
Isovaleryl carnitine	MAR04943	2,160E-04	6,818E-04	6,818E-04
Isovalerylglycine	MAR11457	1,700E-04	5,366E-04	5,366E-04
Keratan sulfate I	MAR09113	1,400E-04	4,419E-04	4,419E-04
Kynurenine	MAR09857	2,681E-03	8,463E-03	1,549E-02
L-2-aminoadipate	MAR09856	1,368E-03	4,318E-03	3,482E-03
L-3-amino-isobutanoate	MAR09221	1,643E-03	5,186E-03	5,186E-03
L-Arabinose	MAR09270	2,500E-03	7,891E-03	1,199E-02
L-Arabitol	MAR09241	2,000E-03	6,313E-03	8,523E-03
L-Carnitine	MAR09292	4,094E-02	1,292E-01	1,635E-01
L-Cystathionine	MAR09846	3,640E-04	1,149E-03	1,149E-03
L-Dopa	MAR09219	7,230E-06	2,282E-05	2,282E-05

Table 11 (cont.)

Metabolite Name	Exchange Reaction	Concentrations Normal Blood	Fluxes Normal Blood	Fluxes Blood Tumour
L-Homocystine	MAR11301	8,520E-03	2,689E-02	2,689E-02
L-hydroxylysine	MAR10301	1,500E-03	4,735E-03	4,735E-03
L-Lactate	MAR09135	1,613E+00	5,090E+00	7,092E+00
L-Metanephrine	MAR09373	1,600E-06	5,051E-06	5,051E-06
L-Octanoylcarnitine	MAR04918	2,100E-04	6,629E-04	6,629E-04
L-Palmitoylcarnitine	MAR09925	1,170E-04	3,693E-04	3,693E-04
L-Pipecolate	MAR10255	6,230E-03	1,967E-02	1,967E-02
L-xylulose	MAR11940	9,000E-03	2,841E-02	2,841E-02
LacCer pool	MAR10027	1,000E-02	3,157E-02	3,157E-02
Lanosterol	MAR11441	4,800E-04	1,515E-03	1,515E-03
Lathosterol	MAR10256	5,903E-03	1,863E-02	1,863E-02
lauric acid	MAR10434	5,822E-03	1,838E-02	2,297E-02
Leucine	MAR09040	1,333E-01	4,207E-01	3,082E-01
Leukotriene B4	MAR09365	7,830E-06	2,472E-05	2,472E-05
Leukotriene B5	MAR10233	7,900E-08	2,494E-07	2,494E-07
Leukotriene C4	MAR09366	4,190E-06	1,323E-05	1,323E-05
Leukotriene E4	MAR09368	7,530E-06	2,377E-05	2,377E-05
Leukotriene F4	MAR09163	3,050E-07	9,628E-07	9,628E-07
lignocerate	MAR00621	5,958E-03	1,881E-02	1,881E-02
Limonene	MAR09369	1,900E-04	5,997E-04	5,997E-04
Linoleate	MAR09035	6,374E-01	2,012E+00	2,012E+00
Linolenate	MAR09036	2,865E-02	9,044E-02	7,066E-02
lipoic acid	MAR09167	7,700E-05	2,431E-04	2,042E-04
Lipoxin A4	MAR10211	7,850E-08	2,478E-07	2,478E-07
Lithocholate	MAR04164	2,050E-04	6,471E-04	6,471E-04
LTD4	MAR09367	1,290E-05	4,072E-05	4,072E-05
Lysine	MAR09041	1,877E-01	5,926E-01	4,666E-01
malate	MAR11404	7,600E-03	2,399E-02	3,039E-02
malonic-dialdehyde	MAR11945	2,335E-03	7,371E-03	7,371E-03
Mannose	MAR09137	5,150E-02	1,626E-01	2,471E-01
margaric acid	MAR00620	2,666E-01	8,416E-01	8,416E-01
mead acid	MAR13058	3,458E-03	1,092E-02	1,092E-02
Melatonin	MAR11920	5,870E-07	1,853E-06	1,853E-06
Methanol	MAR09372	1,676E-01	5,290E-01	5,290E-01
Methionine	MAR09042	2,150E-01	6,788E-01	5,954E-01
Methylamine	MAR11934	1,000E-03	3,157E-03	3,157E-03
methylglyoxal	MAR09375	5,530E-02	1,745E-01	1,745E-01
Methylimidazoleacetic acid	MAR09090	8,460E-05	2,670E-04	2,670E-04
methylmalonate	MAR10243	1,740E-04	5,492E-04	5,492E-04
Myristic acid	MAR00702	2,026E-02	6,395E-02	4,919E-02
N-acetyl-L-cysteine	MAR11821	4,000E-03	1,263E-02	1,263E-02
N-acetylneuraminate	MAR11348	1,285E-03	4,056E-03	4,056E-03
N-Acetylornithine	MAR10198	1,093E-03	3,451E-03	4,486E-03
N-methylhistamine	MAR11930	3,400E-07	1,073E-06	1,073E-06
N1-Acetylspermidine	MAR11962	7,000E-06	2,210E-05	2,210E-05
N8-Acetylspermidine	MAR11917	5,000E-05	1,578E-04	1,578E-04
NAD+	MAR09376	2,430E-02	7,670E-02	7,670E-02
NADH	MAR12141	2,200E-02	6,944E-02	6,944E-02
NADP+	MAR09377	1,960E-02	6,187E-02	6,187E-02
Nervonic acid	MAR00637	2,600E-02	8,207E-02	6,313E-02
NH3	MAR09073	1,102E-01	3,478E-01	3,478E-01

Table 11 (cont.)

Metabolite Name	Exchange Reaction	Concentrations Normal Blood	Fluxes Normal Blood	Fluxes Blood Tumour
NH4+	MAR11420	3,850E-02	1,215E-01	1,215E-01
nicotinamide	MAR09378	2,350E-04	7,418E-04	6,182E-04
Nicotinate	MAR09142	4,308E-02	1,360E-01	1,360E-01
Nitrite	MAR09149	1,974E-02	6,230E-02	6,230E-02
NO	MAR09381	1,200E-08	3,788E-08	3,788E-08
noradrenaline	MAR09093	2,810E-06	8,870E-06	8,870E-06
Norepinephrine sulfate	MAR09382	8,000E-06	2,525E-05	2,525E-05
Normetanephrine	MAR10490	3,400E-07	1,073E-06	1,073E-06
O-Acetylcarnitine	MAR09920	6,473E-03	2,043E-02	2,656E-02
O-Butyrylcarnitine	MAR04889	2,380E-04	7,513E-04	7,513E-04
O-propanoylcarnitine	MAR09921	4,140E-04	1,307E-03	1,307E-03
octadecenoylcarnitine(5)	MAR09923	2,950E-04	9,312E-04	9,312E-04
octanoic acid	MAR09813	5,250E-03	1,657E-02	1,657E-02
Oleate	MAR00650	3,037E-01	9,585E-01	1,270E+00
omega-3-arachidonic acid	MAR00577	7,000E-06	2,210E-05	2,210E-05
Omeprazole	MAR09387	3,400E-04	1,073E-03	1,073E-03
Ornithine	MAR09087	7,827E-02	2,471E-01	2,471E-01
Orotate	MAR09690	2,945E-03	9,296E-03	9,296E-03
Oxalate	MAR09165	1,029E-02	3,248E-02	4,580E-02
Oxypurinol	MAR12693	5,000E-02	1,578E-01	1,578E-01
Palmitate	MAR00611	4,871E-01	1,537E+00	1,537E+00
Palmitolate	MAR00617	5,034E-02	1,589E-01	1,589E-01
Pantothenate	MAR09145	2,744E-03	8,662E-03	8,662E-03
PC-LD pool	MAR00655	3,711E+01	1,171E+02	1,171E+02
pentadecylic acid	MAR00662	1,008E-01	3,181E-01	3,181E-01
PEP	MAR09858	1,250E-02	3,946E-02	3,523E-02
peroxynitrite	MAR11906	3,868E-02	1,221E-01	1,221E-01
PG-CL pool	MAR00658	3,577E-02	1,129E-01	1,129E-01
Phenylacetate	MAR10439	5,500E-02	1,736E-01	1,736E-01
Phenylacetylglutamine	MAR09391	3,340E-03	1,054E-02	1,054E-02
Phenylalanine	MAR09043	6,945E-02	2,192E-01	2,192E-01
Phenylpyruvate	MAR11438	2,750E-03	8,681E-03	8,681E-03
phosphocholine	MAR09845	2,200E-03	6,944E-03	6,944E-03
physeteric acid	MAR13060	1,500E-03	4,735E-03	4,735E-03
Phytanate	MAR00659	4,697E-03	1,483E-02	1,483E-02
Phytanic acid	MAR09037	4,697E-03	1,483E-02	1,483E-02
Pi	MAR09072	9,028E-01	2,850E+00	3,306E+00
PI pool	MAR12127	3,524E-01	1,112E+00	1,112E+00
Picolinic acid	MAR11925	2,990E-04	9,438E-04	9,438E-04
Porphobilinogen	MAR11932	6,000E-05	1,894E-04	1,894E-04
PPi	MAR11405	1,800E-03	5,682E-03	5,682E-03
Pregnenolone	MAR11867	1,430E-05	4,514E-05	4,514E-05
Pregnenolone sulfate	MAR11875	1,300E-04	4,104E-04	4,104E-04
Pristanic acid	MAR11967	1,675E-03	5,287E-03	5,287E-03
Progesterone	MAR09393	2,368E-03	7,475E-03	7,475E-03
Proline	MAR09068	1,989E-01	6,277E-01	6,277E-01
propane-1,2-diol	MAR11936	8,933E-03	2,820E-02	2,820E-02
Propanoate	MAR09808	1,250E-03	3,946E-03	3,946E-03
Prostaglandin A1	MAR04203	7,400E-08	2,336E-07	2,336E-07
Prostaglandin A2	MAR04213	9,550E-07	3,015E-06	3,015E-06
Prostaglandin B2	MAR04217	4,460E-07	1,408E-06	1,774E-06

Table 11 (cont.)

Metabolite Name	Exchange Reaction	Concentrations Normal Blood	Fluxes Normal Blood	Fluxes Blood Tumour
Prostaglandin D2	MAR09395	1,930E-07	6,092E-07	6,092E-07
Prostaglandin E1	MAR09396	3,700E-09	1,168E-08	1,168E-08
Prostaglandin E2	MAR09397	7,310E-07	2,307E-06	2,307E-06
Prostaglandin F1alpha	MAR04234	3,760E-07	1,187E-06	1,187E-06
Prostaglandin F2alpha	MAR09398	2,510E-07	7,923E-07	7,923E-07
Prostaglandin J2	MAR10210	4,950E-08	1,563E-07	1,563E-07
protoporphyrin	MAR11954	7,600E-04	2,399E-03	2,399E-03
provitamin D3	MAR10195	5,000E-03	1,578E-02	1,578E-02
PRPP	MAR01984	5,310E-03	1,676E-02	1,676E-02
PS-LD pool	MAR00661	1,748E-02	5,518E-02	5,518E-02
Putrescine	MAR11426	1,560E-04	4,924E-04	4,282E-04
Pyridoxal	MAR09400	2,510E-04	7,923E-04	7,923E-04
Pyridoxal-phosphate	MAR09691	3,440E-05	1,086E-04	1,086E-04
Pyridoxamine	MAR09399	1,640E-04	5,177E-04	5,177E-04
Pyridoxine	MAR09144	2,500E-05	7,891E-05	7,891E-05
Pyruvate	MAR09133	7,269E-02	2,295E-01	3,748E-01
Quinolinate	MAR09859	4,700E-04	1,484E-03	1,484E-03
quinonoid dihydrobiopterin	MAR11957	6,000E-06	1,894E-05	1,894E-05
Retinal	MAR10492	1,550E-04	4,893E-04	4,893E-04
retinoate	MAR09404	1,070E-04	3,378E-04	3,378E-04
retinol	MAR09147	8,488E-02	2,679E-01	2,679E-01
Retinoyl-glucuronide	MAR09405	6,600E-06	2,083E-05	2,083E-05
Retinyl palmitate	MAR13067	5,610E-05	1,771E-04	1,771E-04
Retinyl-ester	MAR00666	1,030E-04	3,251E-04	3,251E-04
Ribitol	MAR09401	1,167E-03	3,684E-03	5,231E-03
Riboflavin	MAR09143	3,500E-04	1,105E-03	1,105E-03
Ribose	MAR09406	2,300E-03	7,260E-03	5,149E-03
SAH	MAR09026	2,440E-04	7,702E-04	7,702E-04
Salsolinol	MAR11927	1,310E-06	4,135E-06	4,135E-06
SAM	MAR10202	8,550E-05	2,699E-04	2,699E-04
Sarcosine	MAR09131	7,000E-04	2,210E-03	3,403E-03
Sebacicacid	MAR10321	9,100E-05	2,872E-04	2,872E-04
Selenomethionine	MAR11961	6,900E-04	2,178E-03	2,178E-03
Serine	MAR09069	1,397E-01	4,410E-01	5,600E-01
Serotonin	MAR09412	7,540E-04	2,380E-03	2,380E-03
SM pool	MAR11281	1,084E+00	3,423E+00	3,423E+00
sn-glycerol-3-PC	MAR09850	3,400E-02	1,073E-01	1,073E-01
sn-glycerol-3-phosphate	MAR09868	3,000E-02	9,470E-02	9,470E-02
Spermidine	MAR11427	4,776E-03	1,508E-02	1,508E-02
Spermine	MAR09715	2,604E-03	8,220E-03	8,220E-03
Sphinganine	MAR11959	1,100E-05	3,472E-05	1,417E-04
Sphinganine-1-phosphate	MAR09410	5,500E-05	1,736E-04	1,736E-04
Sphingosine	MAR11947	5,000E-05	1,578E-04	1,578E-04
Sphingosine-1-phosphate	MAR09411	2,780E-04	8,775E-04	8,775E-04
Squalene	MAR11842	1,900E-03	5,997E-03	5,997E-03
Stearate	MAR00639	2,863E-01	9,037E-01	9,037E-01
Stearidonic acid	MAR00695	2,120E-04	6,692E-04	6,692E-04
Stearoylcarnitine	MAR09924	5,000E-05	1,578E-04	1,578E-04
Suberic acid	MAR10339	1,870E-03	5,903E-03	5,903E-03
Succinate	MAR09415	1,428E-02	4,506E-02	5,542E-02
Succinylacetone	MAR11466	8,150E-05	2,573E-04	2,573E-04

Table 11 (cont.)

Metabolite Name	Exchange Reaction	Concentrations Normal Blood	Fluxes Normal Blood	Fluxes Blood Tumour
Sucrose	MAR09416	1,800E-03	5,682E-03	5,682E-03
Sulfate	MAR09074	3,528E-01	1,113E+00	1,113E+00
Sulfite	MAR09088	1,230E-03	3,883E-03	3,883E-03
sulfochenodeoxycholate	MAR04153	1,000E-04	3,157E-04	3,157E-04
sulfoglycolithocholate	MAR04178	6,000E-05	1,894E-04	1,894E-04
TAG-extraction	MAR09023	7,136E+00	2,253E+01	2,253E+01
Taurine	MAR09418	1,036E-01	3,269E-01	1,121E+00
Taurochenodesoxycholate	MAR09284	3,000E-04	9,470E-04	9,470E-04
taurocholate	MAR09282	2,400E-04	7,576E-04	7,576E-04
Taurodeoxycholate	MAR08959	6,200E-05	1,957E-04	1,957E-04
taurolithocholate	MAR04146	1,212E-03	3,826E-03	3,826E-03
Tauroursodeoxycholate	MAR04150	2,000E-03	6,313E-03	6,313E-03
Testosterone	MAR09429	5,930E-06	1,872E-05	1,872E-05
tetradecenoylcarnitine(5)	MAR10345	7,700E-05	2,431E-04	2,431E-04
Tetrahydrobiopterin	MAR10496	1,100E-05	3,472E-05	3,472E-05
THF	MAR09421	2,500E-06	7,891E-06	7,891E-06
Thiamin	MAR09159	1,600E-04	5,051E-04	5,051E-04
Thiocyanate	MAR09420	3,183E-02	1,005E-01	1,005E-01
Thiosulfate	MAR09432	5,115E-02	1,615E-01	1,615E-01
Threonine	MAR09044	1,478E-01	4,665E-01	3,332E-01
Thromboxane A2	MAR09433	2,400E-07	7,576E-07	7,576E-07
thromboxane B2	MAR10346	1,470E-04	4,640E-04	4,640E-04
Thymidine	MAR09423	2,050E-04	6,471E-04	6,471E-04
Thyroxine	MAR09424	6,518E-03	2,057E-02	2,057E-02
Tiglyl carnitine	MAR04895	5,080E-05	1,604E-04	1,604E-04
trans-4-hydroxy-L-proline	MAR08386	2,157E-02	6,807E-02	6,807E-02
Tricosanoic acid	MAR13061	3,300E-05	1,042E-04	1,042E-04
triiodothyronine	MAR09427	9,860E-07	3,112E-06	3,112E-06
tryptophan	MAR09045	6,108E-02	1,928E-01	1,377E-01
Tyramine	MAR11439	5,000E-06	1,578E-05	1,578E-05
Tyrosine	MAR09064	7,779E-02	2,455E-01	1,846E-01
Ubiquinol	MAR08967	4,000E-04	1,263E-03	1,263E-03
ubiquinone	MAR08966	1,179E-03	3,722E-03	3,722E-03
UDP	MAR09435	4,100E-02	1,294E-01	1,294E-01
UDP-glucose	MAR01986	1,550E-01	4,893E-01	4,893E-01
UMP	MAR09436	1,840E-01	5,808E-01	5,808E-01
Uracil	MAR09437	1,135E-03	3,583E-03	3,583E-03
urate	MAR09075	2,830E-01	8,934E-01	8,272E-01
Urea	MAR09438	3,749E+00	1,183E+01	8,453E+00
Uridine	MAR09439	9,107E-03	2,875E-02	1,916E-02
Urocanate	MAR10347	4,300E-04	1,357E-03	1,357E-03
Uroporphyrin I	MAR11929	6,300E-06	1,989E-05	1,989E-05
Ursodeoxycholate	MAR04173	1,600E-04	5,051E-04	5,051E-04
Ursodeoxycholic acid 3-sulfate	MAR12233	1,910E-02	6,027E-02	6,027E-02
Valeric acid	MAR09810	6,000E-04	1,894E-03	1,894E-03
Valine	MAR09046	2,166E-01	6,836E-01	5,121E-01
Vanillylmandelate	MAR11446	3,500E-05	1,105E-04	1,105E-04
Vitamin D2	MAR09441	2,750E-06	8,681E-06	8,681E-06
Vitamin D3	MAR09442	4,590E-05	1,449E-04	1,449E-04
Xanthine	MAR11428	1,973E-02	6,228E-02	6,228E-02
Xanthosine	MAR09861	5,080E-03	1,604E-02	4,349E-01

Table 11 (cont.)

Metabolite Name	Exchange Reaction	Concentrations Normal Blood	Fluxes Normal Blood	Fluxes Blood Tumour
Xanthurenate	MAR09693	1,100E-05	3,472E-05	3,472E-05
xylitol	MAR09138	1,589E-03	5,016E-03	6,821E-03
H2O	MAR09047	-	1,000E+03	1,000E+03
O2	MAR09048	-	1,000E+03	1,000E+03
H+	MAR09079	-	1,000E+03	1,000E+03

Table 12: Fold changes between tumour and normal blood reported by the studies and calculated average fold change. Study sinalised with a † reported results for both GC-TOFMS and GCMS-QP201.

Metabolite Name	Exchange Reaction	Fold Changes from Studies [186]	[187]	[188]†	[188]†	[189]	[190]	[191]	Average Fold Change
L-2-aminoadipate	MAR09856	0	0	0	0	0	0	0,806	0,806
(S)-2-aminobutanoate	MAR10206	0	0	0	0	0	1,61	0	1,61
Deoxyuridine	MAR09310	0	0	0	0	1,52	0	0	1,52
2-Hydroxybutyrate	MAR09216	0	0	1,4	0	1,42	2,09	0	1,637
(R)-3-hydroxybutanoate	MAR09134	1,59	2,01	1,4	0	1,88	8,59	0	3,094
4-aminobutyrate	MAR09091	0	0	0	0	0	0	0,944	0,944
Acetate	MAR09086	1,18	0	0	0	0	0	0	1,18
O-Acetylcarnitine	MAR09920	0	0	0	0	0	1,3	0	1,3
Adenine	MAR09253	0	0	0	0	0	1,17	0	1,17
Allantoin	MAR10431	0	0	0	0	0	0	1,24	1,24
L-Arabinose	MAR09270	0	0	0	0	1,52	0	0	1,52
L-Arabitol	MAR09241	0	0	0	0	1,35	0	0	1,35
D-Arabitol	MAR10429	0	0	0	0	1,35	0	0	1,35
Asparagine	MAR09062	0	0	0	0	1,17	0	0	1,17
aspartate	MAR09070	0	0	0	0	1,68	0	1,37	1,525
D-aspartate	MAR09097	0	0	0	0	1,68	0	1,37	1,525
L-Carnitine	MAR09292	0	0	0	1,3	0	1,23	0	1,265
Choline	MAR09083	1,2	0	0	0	0	0	0	1,2
Citrate	MAR09286	1,21	0	0	0	0	0	0	1,21
Creatinine	MAR09460	0	0	0	0	0,901	0	0,901	0,901
Cystine	MAR09363	0	0	0	0	0	1,12	0	1,12
Dimethylglycine	MAR09848	0	0	0	0	0	0	0,820	0,820
Elaidate	MAR00583	0	0	0	0	1,14	1,42	0	1,28
Fumarate	MAR11400	0	0	0	0	0	0	1,09	1,09
Glucosamine	MAR09168	0	0	0	0	1,28	0	0	1,28
Glucose	MAR09034	1,53	0	0	0	0	0	2,73	2,13
Glucuronate	MAR11424	0	0	0	0	1,34	0	0	1,34
glutamate	MAR09071	1,2	0	0	0	1,54	0	1,22	1,318
Glutamine	MAR09063	0,847	0	0	0	0	0	0,917	0,881
Glyceraldehyde	MAR09851	0	0	0	0	0	0	1,34	1,34
Glycerate	MAR09342	0	0	0	0	1,3	0	0	1,3
(S)-Glycerate	MAR00603	0	0	0	0	1,3	0	0	1,3
Glycerol	MAR09085	0	0	0	0	0	2,61	0	2,61
Glycine	MAR09067	1,46	0	0	0	1,16	0	0	1,31
glycochenodeoxycholate	MAR09283	0	0	0	0	0	0	1,42	1,42
Glycocholate	MAR09281	0	0	0	0	0	0	1,79	1,79
Glycolate	MAR10436	0	0	0	0	1,13	0	0	1,13
Hexadecanedioic acid	MAR10246	0	1,73	0	0	0	0	0	1,73
Histidine	MAR09038	0	0	0	0	0	0	0,813	0,813
Homogentisate	MAR10247	0	0	0	0	0	0	0,917	0,917

Table 12 (cont.)

Metabolite Name	Exchange Reaction	Fold Changes from Studies							Average Fold Change
		[186]	[187]	[188]†	[188]†	[189]	[190]	[191]	
Inositol	MAR09361	0	0	0	0	1,19	0	0	1,19
Isoleucine	MAR09039	1,28	0	0	0	1,35	0	0	1,315
Kynurenine	MAR09857	0	0	0	0	1,83	0	0	1,83
L-Lactate	MAR09135	1,48	0	1,3	1,4	0	0	0	1,393
D-Lactate	MAR09136	1,48	0	1,3	1,4	0	0	0	1,393
lauric acid	MAR10434	0	0	0	0	1,25	0	0	1,25
Leucine	MAR09040	0,813	0	0,667	0	0	0	0	0,733
Linolenate	MAR09036	0	0	0	0	0	0	0,781	0,781
Lipoic acid	MAR09167	0,840	0	0	0	0	0	0	0,840
Lysine	MAR09041	0	0	0,714	0	0	0	0,877	0,787
malate	MAR11404	0	0	1,3	0	1,37	0	1,13	1,267
Mannose	MAR09137	0	0	0	0	1,52	0	0	1,52
Methionine	MAR09042	0	0	0	0	0	0	0,877	0,877
Myristic acid	MAR00702	0	0	0	0,769	0	0	0	0,769
Acetyl-glycine	MAR10196	0	0	0	0	0	0	0,752	0,752
N-Acetylornithine	MAR10198	0	0	0	0	1,3	0	0	1,3
Nervonic acid	MAR00637	0	0	0	0,769	0	0	0	0,769
nicotinamide	MAR09378	0	0	0	0,833	0	0	0	0,833
Oleate	MAR00650	0	0	1,1	0	0	1,55	0	1,325
ethanolamine-phosphate	MAR09849	0	0	0	0	0,952	0	0	0,952
Oxalate	MAR09165	0	0	0	0	1,7	0	1,12	1,41
Pi	MAR09072	0	0	0	0	1,16	0	0	1,16
PEP	MAR09858	0	0	0	0	0	0	0,893	0,893
4-Hydroxybenzoate	MAR11431	0	0	0	0	1,6	0	0	1,6
Prostaglandin B2	MAR04217	0	1,26	0	0	0	0	0	1,26
Putrescine	MAR11426	0	0	0	0	0,870	0	0	0,870
Pyruvate	MAR09133	0	0	2,1	2	0	1,3	0	1,633
Ribitol	MAR09401	0	0	0	0	1,42	0	0	1,42
Ribose	MAR09406	0	0	0	0	0,709	0	0	0,709
Sarcosine	MAR09131	0	0	0	0	1,54	0	0	1,54
Serine	MAR09069	1,18	0	0	0	1,36	0	0	1,27
Sphinganine	MAR11959	0	0	0	0	0	4,08	0	4,08
Succinate	MAR09415	1,23	0	0	0	0	0	0	1,23
Taurine	MAR09418	0	0	0	0	3,43	0	0	3,43
Threonine	MAR09044	0	0	0,714	0	0	0	0	0,714
tryptophan	MAR09045	0	0	0,625	0,833	0	0	0	0,714
Tyrosine	MAR09064	0,840	0	0,667	0,769	0	0	0	0,752
urate	MAR09075	0	0	0	0	0	0	0,926	0,926
Urea	MAR09438	0	0	0,714	0	0	0	0	0,714
Uridine	MAR09439	0	0	0,588	0,769	0	0	0	0,667
Valine	MAR09046	0,855	0	0,667	0	0	0	0	0,749
Xanthosine	MAR09861	0	0	0	0	0	27,12	0	27,12
xylitol	MAR09138	0	0	0	0	1,36	0	0	1,36
D-Xylose	MAR09203	0	0	0	0	1,33	0	0	1,33

Table 13: Number of cells present in each sample for each T-cell subtype considered. Those subtypes with 5 or less cells in a sample were not considered for model reconstruction in that sample. Those with a † did not pass the gap-fill and were not analysed. The last line is the total number of models reconstructed for each T-cell subtype.

Individual	Sample	State	Cytotoxic CD8	Follicular CD4	IL17+ CD4	Memory CD4	Memory CD8	Naive CD4	Naive CD8	Proliferative CD4	Proliferative CD8	Regulatory CD4
31	scrEXT001	Tumour (core)	4	5	10	81	20	17	0	7	1	138
	scrEXT002	Tumour (border)	8†	22	4	69	51	30	1	6	8	157
	scrEXT003	Normal Matched	5	0	20	78	66	10	0	0	3	3
32	scrEXT009	Tumour (core)	21	81	8	223	171	245	11	19	31	328
	scrEXT010	Tumour (border)	28	101	4	224	165	373	14	19	15	239
	scrEXT011	Normal Matched	10	0	0	49	36	10†	1	2	1	23
33	scrEXT012	Tumour (core)	4	4	1	62	19†	39	0	9	1	61
	scrEXT013	Tumour (border)	4	2	2	63	26	25	1	10	3	72
	scrEXT014	Normal Matched	2	0	8	181	91	92	4	1	0	15
35	scrEXT018	Tumour (core)	1	30	53	118	171	84	3	67	60	209
	scrEXT019	Tumour (border)	0	18	22	78	133	26†	0	24	25	54
	scrEXT020	Normal Matched	1	0	1	65	45	18	1	2	3	13
KUL01	KUL01-T	Tumour (core)	4	4	10	76	20	17	0	1	0	117
	KUL01-B	Tumour (border)	8†	17†	4	61	48	29	2	0	0	128
	KUL01-N	Normal Matched	4	0	21	78	64	10	0	0	0	3
KUL19	KUL19-T	Tumour (core)	30	92	2	227	166	350	17	1	3	214
	KUL19-B	Tumour (border)	21	73	3	216	166	235	12	2	1	304
	KUL19-N	Normal Matched	10	0	0	49	35	9†	1	2	2	24
KUL21	KUL21-T	Tumour (core)	4	2	2	63	26†	19	1	4	0	68
	KUL21-B	Tumour (border)	4	4	1	61	19	42	0	2	1	59
	KUL21-N	Normal Matched	2	0	8	187	90	85	4	0	0	15
SMC01	SMC01-T	Tumour	60	95	32	175	104	88	1	15	24	280
	SMC01-N	Normal Matched	12	0	1	423	169	86	14	0	0	28†
SMC04	SMC04-T	Tumour	23	7	3	90	31	69	6	2	1	153
	SMC04-N	Normal Matched	6	1	6	204	124	57	16	2	0	8†
SMC06	SMC06-T	Tumour	35	39	2	16	62	24	1	2	10	405
	SMC06-N	Normal Matched	2	1	1	310	169	177	31	0	0	34
SMC07	SMC07-T	Tumour	53	45	9	568	188	396	22	17	20	302
	SMC07-N	Normal Matched	4	0	1	355	146	159	11	0	0	35
SMC08	SMC08-T	Tumour	83	41	31	256	103	57	2	7	1	169
	SMC08-N	Normal Matched	12	3	1	187	206†	50	3	0	0	33
SMC10	SMC10-T	Tumour	76	16	2	70	51	79	4	19	13	272
	SMC10-N	Normal Matched	10	0	0	192	124	60	8	1	0	15
	Final Number of Models		16	13	13	33	30	30	11	12	9	29

Table 14: Essential genes that catalase uptake of metabolites. Cell-types codes: *Cyto*: cytotoxic CD8; *Fol*: follicular CD4; *IL17*: IL17+ CD4; *Mem4*: memory CD4; *Mem8*: memory CD8; *N4*: naive CD4; *N8*; *Prol4*: proliferative CD4; *Prol8*: proliferative CD8; *Regs*: regulatory CD4. The essentiality reported by the two CRISPR-Cas9 studies is provided: -: gene not tested in study; *Essential*: gene tested and reported as essential; *Not essential*: gene tested and reported as not essential.

Gene	Cell-types	CD4 Study [236]	CD8 Study [237]	Reactions Affected
CUBN	IL17; Fol; Prol8	-	-	MAR06884 multiple vitamine D metabolism reactions
FASN	Cyto; IL17; Mem4; Mem8; N4; N8; Prol4; Regs	Not essential	-	MAR04844
SLC5A8	Prol8	-	Not essential	MAR08515 MAR09874 MAR09896 MAR09900 MAR09902
SLC6A5	Cyto; Fol; IL17; Mem8; N4; N8; Prol4; Regs	Not essential	-	MAR07739 MAR11456 MAR11453 MAR11450 MAR05457
SLC6A6	Fol; Prol8	-	Essential	MAR02388 MAR04986 MAR05451 MAR05452 MAR05454
SLC7A5	Cyto; Fol; IL17; Mem4; Mem8; N4; N8; Prol4; Prol8; Regs	Not essential	-	Multiple (97)
SLC7A7	IL17	-	Not essential	MAR11804
SLC7A8	Cyto; Mem4; N8; Regs	-	-	MAR09402
SLC7A11	Cyto; Fol; IL17; Mem4; Mem8; Prol8; Regs	Not essential	-	MAR04931
SLC10A6	Fol; N8	-	-	MAR09621
SLC12A3	Cyto; Fol; IL17; Mem8; N4; N8; Prol4; Prol8; Regs	Not essential	-	MAR06524
SLC12A7	Cyto; Regs	-	-	MAR11785
SLC22A3	Cyto; Fol; IL17; Mem8; N4; N8; Prol4; Regs	Not essential	-	MAR11782
SLCO2A1	Fol; Prol8	Not essential	-	Multiple (28)

Table 15: Differentially covered pathways between normal- and tumour- derived regulatory CD4 T-cell models.

Pathway	Adjusted p-value	Fold Change
Acylglycerides metabolism	2,209E-04	3,211
Bile acid biosynthesis	2,209E-04	2,035
Drug metabolism	2,209E-04	3,086
Galactose metabolism	2,209E-04	1,848
Glycerophospholipid metabolism	2,209E-04	1,619
Glycosphingolipid biosynthesis-globo series	2,209E-04	1,585
Glycosphingolipid biosynthesis-lacto and neolacto series	2,209E-04	4,029
Isolated	2,209E-04	2,014
Tricarboxylic acid cycle and glyoxylate/dicarboxylate metabolism	2,209E-04	1,730
Glycosphingolipid metabolism	2,371E-04	1,876
Glycerolipid metabolism	3,308E-04	3,508
Protein assembly	3,578E-04	2,488
Ether lipid metabolism	4,348E-04	2,093
Pool reactions	7,253E-04	1,724
Fatty acid biosynthesis (even-chain)	9,362E-04	3,822
Blood group biosynthesis	1,204E-03	2,239
Protein degradation	1,204E-03	1,780
O-glycan metabolism	1,597E-03	1,588
Biotin metabolism	1,964E-03	1,598
Phosphatidylinositol phosphate metabolism	2,111E-03	2,541
Protein modification	2,111E-03	2,823
Sulfur metabolism	2,111E-03	1,786
Folate metabolism	2,180E-03	1,610
Keratan sulfate biosynthesis	2,733E-03	3,627
Fatty acid biosynthesis	3,264E-03	1,814
Carnitine shuttle (mitochondrial)	3,746E-03	1,885
Fatty acid biosynthesis (odd-chain)	4,754E-03	3,600
Riboflavin metabolism	1,322E-02	1,740
Thiamine metabolism	1,439E-02	1,614
Acyl-CoA hydrolysis	2,317E-02	2,505
Estrogen metabolism	3,400E-02	2,430
Peptide metabolism	4,275E-02	-4,074

Table 16: Differentially covered pathways between normal- and tumour- derived cytotoxic CD8 T-cell models.

pathway	pval_adjust	fold_change
Glycosphingolipid biosynthesis-ganglio series	1,036E-02	1,965
Glycosphingolipid metabolism	1,036E-02	2,833
Pantothenate and CoA biosynthesis	1,036E-02	1,693
Riboflavin metabolism	1,036E-02	5,600
Folate metabolism	1,648E-02	2,161
Inositol phosphate metabolism	1,648E-02	3,489
Oxidative phosphorylation	1,648E-02	1,540
Sphingolipid metabolism	1,648E-02	1,645
Bile acid biosynthesis	3,864E-02	1,673
Biopterin metabolism	3,864E-02	1,611
Estrogen metabolism	4,036E-02	2,977
Carnitine shuttle (peroxisomal)	4,578E-02	1,674

Table 17: Differentially covered pathways between naive and proliferative CD8 T-cell models.

pathway	pval_adjust	fold_change
Acylglycerides metabolism	5,338E-02	-2,049
Blood group biosynthesis	1,620E-02	-2,287
Carnitine shuttle (endoplasmic reticular)	2,279E-02	-2,004
Carnitine shuttle (mitochondrial)	1,222E-03	-2,161
Drug metabolism	1,805E-02	-2,424
Glycosphingolipid biosynthesis-globo series	9,207E-03	-2,241
Glycosphingolipid biosynthesis-lacto and neolacto series	4,716E-01	-2,048
Glycosphingolipid metabolism	2,646E-02	-2,160
Keratan sulfate biosynthesis	5,338E-02	-4,578
O-glycan metabolism	1,222E-03	-2,750
Phosphatidylinositol phosphate metabolism	1,222E-03	-4,617
Protein modification	1,840E-01	-6,111
Sulfur metabolism	1,222E-03	-3,463
Thiamine metabolism	2,265E-03	-2,556
Xenobiotics metabolism	1,703E-01	-5,248

Table 18: Differentially covered pathways between IL17+ and regulatory CD4 T-cell models.

pathway	pval_adjust	fold_change
Biotin metabolism	5,730E-04	-4,692
Fatty acid biosynthesis (even-chain)	1,529E-03	-2,970
Fatty acid biosynthesis (odd-chain)	7,421E-03	-4,034
Glycerolipid metabolism	1,442E-03	-3,684
Keratan sulfate biosynthesis	5,730E-04	-21,782
Peptide metabolism	3,257E-02	4,068
Protein assembly	5,730E-04	-2,619
Protein degradation	5,730E-04	-2,221
Protein modification	5,730E-04	-3,985
Xenobiotics metabolism	1,117E-01	-2,354

B.2 Chapter 5

Table 19: Cell-types used for tumour deconvolution, and respective overlap with CRC atlas and ground-truth phenotypes.

CRC atlas cell-types	Deconvolution cell-types	Ground-truth phenotypes
Tumour Epithelial cells	Cancer cells	Apoptotic_tum
		Bcat+_tum
		CD15+_Bcat+_tum
		CD15+_HLA-DR+_Bcat+_tum
		CD15+_tum
		HLA-DR+_Bcat+_tum
		IDO+_Bcat+_HLA-DR+_tum
		P16ink4a+_tum
		Prol_apoptotic_tum
		Prol_Bcat+_tum
		Prol_CD15+_Bcat+_tum
		Prol_HLA-DR+_Bcat+_tum
		Prol_HLA-DR+_TGFb+_Bcat+_tum
		Prol_HLA-DR+_tum
		Prol_IDO+_Bcat+_HLA-DR+_tum
		Prol_TGFb+_Bcat+_tum
		Prol_TGFb+_tum
		Prol_tum
		TGFb+_Bcat+_CD15+_tum
		TGFb+_Bcat+_tum
		TGFb+_tum
Stromal cells	Stromal cells	CD39+_vessels
		D2-40+_vessels
		Prol_vessels
		TGFb+_vessels
		Vessels
		D2-40+_fibroblasts
		CD39+_D2-40+_fibroblasts
		CD56+_CD39+_D2-40+_fibroblasts

Table 19 (cont.)

CRC atlas cell-types	Deconvolution cell-types	Ground-truth phenotypes
		CD57+_D2-40+_fibroblasts
		TGFb+_fibroblasts
		TGFb+_D2-40+_fibroblasts
		Firboblasts
		TGFb+_monocytes
		HLA-DR+_CD163+_macrophages
		HLA-DR+_monocytes
		HLA-DR+_macrophages
Anti-Inflammatory macro/mono	Macro/Mono Lineage	VISTA+_monocytes
Pro-Inflammatory macro/mono		Macrophages_undefined
		Monocytes
		CD11c+_macrophages
		CD45RO_undefined
Naïve B-cells		
Memory B-cells	B-cells	Bcells
Proliferative B-cells		
Naïve CD4+ T-cells		CD39+_CD8-_Tcells
IL22+ CD4+ T-cells		CD4+_Tcells
IL17+ CD4+ T-cells	CD4+ T-cells	CD57+_CD8-_Tcells
Memory CD4+ T-cells		Intra_CD8-_Tcells
Follicular CD4+ T-cells		CD8-_Tcells
Regulatory CD4+ T-cells	Regulatory T-cells	Regulatory_Tcells
		ICOS+_Regulatory_Tcells
Naïve CD8+ T-cells		CD8+_Tcell
CXCL13+ CD8+ T-cells		CD57+_CD8+_Tcells
Memory CD8+ T-cells	CD8+ T-cells	Intra_CD39+_CD8+_Tcells
Cytotoxic CD8+ T-cells		Intra_CD8+_Tcells
		Intra_GZMB+_CD8+_Tcell
Proliferative CD4+ T-cells	Proliferative T-cells	Prol_Tcells
Proliferative CD8+ T-cells		
		NK_cells
NK cells	NK cells	CD7+_CD3-_cells
		CD56+_D2-40+_cells
Mast cells		Apoptotic_cells
cDCs		VISTA+_CD31+_CD38+_cells
pDCs		TGFb+_CD31+_CD38+_cells
unknown myeloid cells		CD31+_CD38+_cells
IgG+ Plasma cells	Other cells	Granulocytes
IgA+ Plasma cells		TGFb+_granulocytes
Unconventional T-cells		
LTi cells		
Double-Negative T-cells		