THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Utilization of single-cell RNA-Seq and genome-scale modeling for investigating cancer metabolism

JOHAN GUSTAFSSON

CHALMERS
UNIVERSITY OF TECHNOLOGY

Systems and Synthetic Biology
Department of Biology and Biological Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2022

Utilization of single-cell RNA-Seq and genome-scale modeling for investigating cancer metabolism
JOHAN GUSTAFSSON
ISBN: 978-91-7905-651-3
Löpnummer: 5117

Doktorsavhandlingar vid Chalmers tekniska högskola
Ny serie (ISSN0346-718X)

Division of Systems and Synthetic Biology
Department of Biology and Biological Engineering
Chalmers University of Technology
SE-412 96 Gothenburg
Sweden
Telephone + 46 (0)31-772 1000

Cover illustration: Conceptual description of the storyline in this thesis - single-cell RNA-Seq is used as input to genome-scale metabolic modeling, which in turn is used to study cancer.

Printed by Chalmers digitaltryck
Gothenburg, Sweden 2022

Utilization of single-cell RNA-Seq and genome-scale modeling for investigating cancer metabolism

Johan Gustafsson
Department of Biology and Biological Engineering
Chalmers University of Technology

Abstract

Cancer remains a leading cause of death worldwide, and its dysregulated metabolism is a promising target for therapy. However, metabolism is complex to study – the metabolism of a cell involves the interplay of thousands of chemical reactions that are combined in different ways across tissues and cell types. Genome-scale metabolic models (GEMs), where the reaction networks of cells are described using a mathematical formulation, have been developed to help in such studies.

In this thesis, methods were developed for determining the active metabolic network (the context-specific model) in individual cell types, followed by studies of cancer metabolism. To enable identification of the active metabolic network per cell type, single-cell RNA sequencing (scRNA-Seq) was employed to detect the presence of individual genes. However, the technical and biological variation in scRNA-Seq data poses a major challenge to the identification of the active reaction network in a cell type. The variability of gene expression due to technical and biological factors was therefore examined, concluding that data from thousands of cells is often required to provide enough stability for robust model generation. An improved quantification method for scRNA-Seq data, called BUTTERFLY, was also developed and implemented as part of the kallisto-bustools scRNA-Seq workflow. A new optimized version of tINIT, which enables generation of context-specific models, was also developed. It allowed for generation of models based on bootstrapped cell populations, which were used to acquire the statistical uncertainty of models generated from scRNA-Seq data. Finally, the method was applied to a lung cancer dataset, identifying both known and unknown features of cancer metabolism.

To further explore cancer metabolism, a study was conducted to investigate the most optimal metabolic behavior under different degrees of hypoxia. To this end, a diffusion-based model for estimating nutrient availability was developed, as well as a light-weight version of the tool GECKO that enables constraining the total enzyme usage in the model. The model could explain the glutamine addiction phenomenon in cancers and was used to show that metabolic collaboration between cell types in tumors is likely not important for growth.

Keywords: single-cell RNA-Seq, genome-scale metabolic modeling, metabolism, cancer.

List of Publications

This thesis is based on the following publications and manuscripts:

Paper I: Sources of variation in cell-type RNA-Seq profiles.
Gustafsson J, Held F, Robinson J, Björnson E, Jörnsten R & Nielsen J. PLOS ONE 15, e0239495 (2020).

Paper II: DSAVE: Detection of misclassified cells in single-cell RNA-Seq data
Gustafsson J, Robinson J, Inda-Díaz J, Björnson E, Jörnsten R & Nielsen J. PLOS ONE 15, e0243360 (2020).

Paper III: Addressing the pooled amplification paradox with unique molecular identifiers in single-cell RNA-seq.
Gustafsson J, Robinson J, Nielsen J & Pachter L. Genome Biology 22, 174 (2021).

Paper IV: Generation and analysis of context-specific genome-scale metabolic models derived from single-cell RNA-Seq data.
Gustafsson J, Robinson JL, Roshanzamir F, Jörnsten R, Kerkhoven E & Nielsen J. bioRxiv 2022.04.25.489379 (2022)

Paper V: Cellular limitation of enzymatic capacity explains glutamine addiction in cancers.
Gustafsson J, Roshanzamir F, Hagnestål A, Robinson JL & Nielsen, J. bioRxiv 2022.02.08.479584 (2022)

Additional papers and manuscripts not included in this thesis:

Paper VI: An atlas of human metabolism.
Robinson JL†, Kocabaş P†, Wang H†, Cholley PE†, Cook D, Nilsson A, Anton M, Ferreira R, Domenzain I, Billa V, Limeta A, Hedin A, **Gustafsson J**, Kerkhoven E, Svensson T, Palsson BO, Mardinoglu A, Hansson L, Uhlén M & Nielsen J. Science Signaling 13, (2020).

Paper VII: Genome-scale metabolic network reconstruction of model animals as a platform for translational research
Wang H, Robinson JL, Kocabas P, **Gustafsson J**, Anton M, Cholley PE, Huang S, Gobom J, Svensson T, Uhlen M, Zetterberg H & Nielsen J. Proceedings of the National Academy of Sciences 118, (2021).

Paper VIII: Expansion of the Yeast Modular Cloning Toolkit for CRISPR-Based Applications, Genomic Integrations and Combinatorial Libraries
Otto M, Skrekas C, Gossing M, **Gustafsson J**, Siewers V, and David F. ACS Synthetic Biology 10, 3461–3474 (2021)

† co-first authorship

Contribution summary

Paper I. I co-designed the study, co-developed the methods, collected and analyzed the data, and wrote the manuscript.

Paper II. I co-designed the study, developed the methods, collected and analyzed the data, co-implemented the software package, and wrote the manuscript.

Paper III. I co-designed the study, developed the methods, collected and analyzed the data, and co-wrote the manuscript.

Paper IV. I designed the study, developed the methods, collected and analyzed the data, and wrote the manuscript.

Paper V. I designed the study, co-developed the diffusion model, developed the methods, collected and analyzed the data, and wrote the manuscript.

Preface

This dissertation serves as partial fulfillment of the requirements to obtain the degree of Doctor of Philosophy at the Department of Biology and Biological Engineering at Chalmers University of Technology. The PhD studies were carried out between September 2017 and June 2022 at the division of Systems and Synthetic Biology (SysBio) under the supervision of Jens Nielsen. The project was co-supervised by Jonathan Robinson and Thomas Svensson and examined by Ivan Mijakovic. The project was funded by the Knut and Alice Wallenberg Foundation.

Johan Gustafsson
April 2022

Table of Contents

Abbreviations

BTM	Biological, technical, and misclassification
cDNA	Complementary DNA
CPM	Counts per million
CU	Copies per UMI
DSAVE	Downsampling-based variation estimation
ETC	Electron transport chain
FSCM	Fraction of single-copy molecules
GECKO	GEM with Enzymatic Constraints using Kinetic and Omics data
GEM	Genome-scale metabolic model
GLM	Generalized linear model
GPR	Gene-protein-reaction, gene rule per reaction
MCC	Matthews correlation coefficient
NADH	Reduced nicotinamide adenine dinucleotide
NGAM	Non-growth associated maintenance
NK cells	Natural killer cells
OXPHOS	Oxidative phosphorylation
PCA	Principal component analysis
REDOX	Reduction and oxidation
RNA	Ribonucleic acid
RNA-Seq	RNA sequencing
RT	Reverse transcription
scRNA-Seq	Single-cell RNA sequencing
SNO	Sampling noise only
t-SNE	t-Distributed Stochastic Neighbor Embedding
TCA cycle	Tricarboxylic acid cycle
tINIT	Task-driven integrative network inference for tissues
ftINIT	Fast task-driven integrative network inference for tissues
TME	Tumor Microenvironment
TMM	Trimmed mean of M values
TPM	Transcripts per million
UMAP	Uniform manifold approximation and projection
UMI	Unique molecular identifier
UMICF	UMI copy fraction

Acknowledgements

The journey towards a doctoral degree is in many ways a personal endeavor, where a person's own thoughts, theories, and ideas should be a central part of the final thesis. However, at SysBio at Chalmers, I have never felt alone, not even through the isolation vested upon us during the pandemic. There are many persons that have helped me through both enthusiasm and despair, stress and confusion, who deserve special mention in this thesis.

First of all, I would like to thank my supervisor Jens Nielsen for his solid support and for the opportunity to work with him at SysBio. You have allowed me to follow my own path and to freely explore even my craziest ideas, and always supplied constructive criticism in a positive way. Likewise, you keep high moral standards in the research, which I hope I will be able to take with me for the rest of my life.

I would also like to thank my two co-supervisors Jonathan Robinson and Thomas Svensson for their support over the last 5 years, I have deeply appreciated your help in my efforts towards this thesis. A special thanks goes to Jonathan. You have spent a tremendous effort in making the best out of my projects and provided me with solid support all the way through my PhD, despite the fact that you left SysBio halfway through this work.

I have also had the honor to be involved in a deep collaboration with Lior Pachter, and I think we have complemented each other in a really good way in our projects. I have truly enjoyed working with you, and I would like to thank you for all the help and enthusiasm you have provided me with; you have almost been like a second supervisor at times. You have also opened up my eyes to the single-cell world, which has likely set the direction of my future research career.

I would also like to thank Rebecka Jörnsten for your solid support through my thesis. You have helped me with numerous mathematical questions over these years, and deeply analyzed my issues and concerns, and I am very grateful for the help you have provided.

The research community at SysBio has been a great environment to work in. I would like to thank all coauthors of my papers. Fariba, Elias, Felix, Juan, Anders, and Eduard, I deeply appreciate the help you have provided me with, and I have much enjoyed your company in our projects. I would also like to thank many people at SysBio for helping me with my projects and my thesis, for being allowed to be part in your projects, and not the least for being good colleagues and friends. Rasool, Mihail, Hao, Angelo, Sinisa, Avlant, Daniel, Raphael, Feiran, Yu, Pinar, Dimitra, Christos, Max, and Pierre, thank you, and I hope to keep contact with many of you in the future.

Finally, I would like to thank my wife Matilda and my children Siri, Alice and Iris, for supporting me through these sometimes fantastic and sometimes stressful years, and for giving me something else to think about than work. In the end, you are the most important thing in my life, and having you constantly backing me up helps more than you may realize. I hope that we will continue to thrive together in the new adventures that await us all!

1. Background

1.1. Metabolism in cells

Metabolism is defined as the set of chemical reactions occurring in living cells to sustain life [1]. These reactions serve a multitude of purposes, mainly extraction of energy from metabolites, generation of the different building blocks needed in the cell, and disposal of waste products. To feed the reactions with substrates, the cell takes up metabolites from the surroundings, such as sugars, amino acids, lipids, and oxygen. Much of the metabolism is conserved across species, even between animals, bacteria, fungi, and plants [1]. This thesis is primarily focused on energy metabolism and generation of building blocks for cellular growth in human cancers.

1.1.1. Energy metabolism

The primary goal of energy metabolism is to extract chemical energy in a format that can be used in other reactions in the cell. While there are many molecules that can be used to store energy in the cell, a central energy metabolite is adenosine triphosphate (ATP), which serves as an energy currency that can be used by most energy-demanding reactions. Energy is extracted from ATP by the removal of one phosphate group, turning ATP into adenosine diphosphate (ADP). Likewise, ADP can be turned to ATP when coupled with the degradation of energy-rich substrates [1].

An important aspect of energy metabolism is the maintenance of the reduction-oxidation (REDOX) balance in the cell. The cell needs both reducing and oxidizing agents for various tasks and the levels of these compounds need to be regulated for the cell to function properly. The most common such agents come in pairs: an oxidizing agent that when reduced is turned into a reducing agent and vice versa. The most important such redox pairs for energy metabolism are nicotinamide adenine dinucleotide (NAD^+) and its corresponding reducing agent NADH, nicotinamide adenine dinucleotide phosphate ($NADP^+$/NADPH), and flavin adenine dinucleotide (FAD/$FADH_2$).

The main processes for generation of ATP in human cells are glycolysis, the tricarboxylic acid (TCA) cycle, and oxidative phosphorylation (OXPHOS) (Fig. 1). Glycolysis is the process in which glucose is converted into pyruvate at the net conversion of 2 ADP to ATP and 2 NAD^+ to NADH. Glycolysis can be run independently of other processes by the conversion of the produced pyruvate to lactate, in which the NADH produced is converted back to NAD^+, followed by lactate export. To yield more ATP from the substrates, the pyruvate and NADH can instead be supplied to the TCA cycle and OXPHOS [1].

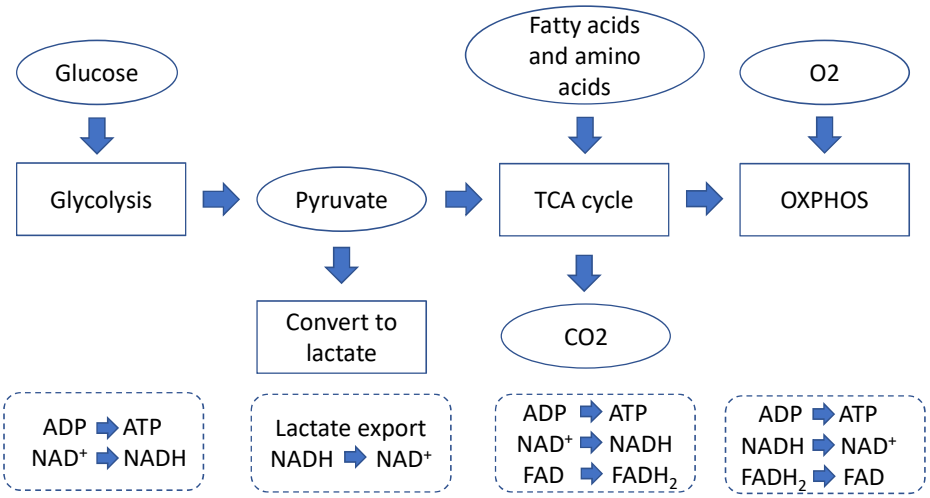

Fig. 1: Overview of energy metabolism in human cells. *Glucose is passed through glycolysis to generate ATP, NADH and pyruvate. Pyruvate can either be exported as lactate or enter the TCA cycle together with amino acids and fatty acids (via beta oxidation). The TCA cycle generates ATP (or GTP), $FADH_2$ and NADH. The $FADH_2$ and NADH are oxidized via oxidative phosphorylation, generating the main amount of ATP.*

At full oxidation of glucose, the pyruvate generated by glycolysis enters the TCA cycle in the mitochondria, in which the carbon in the pyruvate is fully oxidized to form carbon dioxide (CO_2). In the process, the cycle generates one ATP (henceforth meaning it is converted from ADP), or alternatively one guanine triphosphate, GTP. In addition, three NADH and one $FADH_2$ are generated [2]. All NADH and $FADH_2$ is then under normal aerobic conditions oxidized via OXPHOS, where the complexes of the electron transport chain (ETC) oxidize the NADH and FADH2 and use the released energy to pump protons out of the mitochondrial matrix. The proton gradient generated is then used by complex V of the ETC to generate ATP [3]. The total ATP generated from one glucose molecule cannot easily be calculated due to various factors such as the leakiness of mitochondrial membrane with regard to protons and may vary between conditions. Extracellular measurements have quantified the ATP production rate to around 33 ATP per molecule of glucose, where OXPHOS generates 29 ATP, emphasizing that OXPHOS is the main process for producing ATP in the cells [4].

The TCA cycle can be fueled by various substrates, not just pyruvate from glucose. Other examples are fatty acids, which can enter the TCA cycle as acetyl-CoA after undergoing a process known as beta oxidation, and amino acids [2].

As mentioned above, glycolysis can be run independently of the TCA cycle and OXPHOS if combined with lactate export. Such an approach is useful in hypoxia, where oxygen is limited [5]. In addition, an important difference between glycolysis and OXPHOS is the difference in the mass of enzymes required to catalyze the reactions. While OXPHOS produces far more ATP per glucose molecule, OXPHOS requires a larger allocation of enzyme mass per ATP molecule produced to catalyze the chemical reactions. Thus, glycolysis can also be used in a stand-alone fashion to maximize the total ATP production in a cell under conditions where oxygen is not in shortage, often termed “aerobic glycolysis” [6], [7].

1.1.2. Cellular growth

Cells multiply (proliferate) through cell division, in which they undergo a series of transformations known as the cell cycle. The process is tightly regulated and cells pass through several stages in the process [8]. The cell requires a large variety of building blocks to be able to grow, which in some cases can be acquired from the surroundings but often need to be synthesized by the cell. In many cases, these building blocks are synthesized from intermediate metabolites in processes mainly associated with energy metabolism, namely glycolysis and the TCA cycle [2], [9].

The substances needed for growth are a source of carbon, such as glucose, and a source of nitrogen, such as amino acids, and a collection of additional essential metabolites, cofactors, and vitamins. Examples of essential metabolites are the essential amino acids, which human cells cannot synthesize *de novo*. In addition, growing cells in general need to produce substantial amounts of ATP, both for cell maintenance and for biosynthesis of new biomass during growth [10].

1.2. Cancer metabolism

Despite the enormous scientific effort invested in cancer research, cancer still remains a leading cause of death worldwide. The cancer research field is multifaceted, addressing different goals such as prevention, early detection, and treatment of cancer, all with the common goal of reducing the total death and suffering caused by cancer [11]–[13]. Within cancer treatment, scientists address different aspects of cancer biology to target cancer, such as immune system evasion, rapid proliferation, and cancer metabolism [14], [15]. This thesis is focused on finding aspects of cancer metabolism that are targetable for therapy.

Dysregulated metabolism has been proposed as an emerging hallmark of cancer [14], [15], and the differences in metabolism are driven by a combination of lack of metabolites, adaptations to conditions, and optimizations to increase growth [15]. Solid tumors commonly suffer from a leaky and irregular vasculature, which leads to a high internal tissue pressure. The increase in pressure incapacitates the microcirculation and blocks the lymph vessels, severely reducing the flux of fluid and nutrients through a large portion of the tumor [16]–[18]. The main remaining mechanism for transport of metabolites from the blood into the tumor is therefore diffusion [19]. The uneven distribution of blood vessels in solid tumors leads to large differences in nutrient availability, where some regions are severely hypoxic or even necrotic with a low influx of nutrients, while nutrient and oxygen availability is plentiful in other regions. Hypoxic and necrotic regions tend to be more common in the center (core) of tumors, while the edges of a tumor in general have a better availability of metabolites [20].

Cancer cells are exposed to a complex selection pressure, involving a plethora of factors [15], [21]. One such factor is cellular growth; cells that harbor a trait that increases proliferation while not causing any negative effects will eventually dominate a population, given enough time. Cancer cells have often therefore developed traits to increase the growth rate, and of particular interest for this thesis, the ATP production rate. A commonly observed behavior of cancer cells is to rely more on glycolysis and less on the TCA cycle and OXPHOS for ATP production, combined with secretion of lactate [7]. While such a behavior is directly understandable in hypoxia due to the lack of oxygen for driving

OXPHOS, the behavior is also observed in well oxygenated regions of the tumors. The latter is called the Warburg effect (also aerobic glycolysis), and is commonly observed in both tumors and cell lines [7], [22]. A potential explanation for this behavior is the difference in enzyme usage between OXPHOS and glycolysis, where OXPHOS allocates more enzyme mass per ATP production, enabling a higher total ATP production in a cell when relying on glycolysis [6].

The amino acid metabolism in cancer is different from that of normal cells [23]. Cancer cells tend to use glutamine as substrate for the TCA cycle instead of pyruvate, a phenomenon commonly called "glutamine addiction" [24], which increases the lactate export of cancer cells even further. As part of this process, cancer cells also tend to export proline [25]. In addition, some cancers are known to secrete glutamate, potentially to support nucleotide synthesis [25], [26]. The reason for these changes in metabolism is in general poorly understood, which calls for additional studies in the field.

A tumor is in many aspects similar to an organ, with multiple cell types fulfilling different roles, forming the tumor microenvironment (TME) [27]. Over the last decade, the interplay between cell types in the TME has received increasing attention in cancer research, where this thesis focuses on metabolic interactions. For example, cancer-associated fibroblasts (CAFs) are thought to secrete metabolites such as lactate and ketone bodies to supply cancer cells with these resources for increasing growth [28]–[30]. Likewise, macrophages in the tumor microenvironment have the ability to clean up dead cells and debris in the TME [31] and eventually produce metabolites useful for the cancer cells.

1.3. Genome-scale metabolic models

1.3.1. Genome-scale metabolic models and flux balance analysis

Fluxes through individual chemical reactions in living cells are generally difficult to measure, and genome scale metabolic models (GEMs) have therefore been developed to enable prediction of such fluxes using a mathematical approach. Supported by advances in genome sequencing, the first GEM was published in 1999, modeling in total 488 metabolic reactions of *Haemophilus influenzae* Rd [32]. Since then, a plethora of GEMs have been developed for different organisms, for example Yeast-GEM [33] for *Saccharomyces cerevisiae* and Human1 [34] for human metabolism. GEMs have been proven useful in a number of areas, for example metabolic engineering [35], evolutionary systems biology [36], uncovering of metabolic behaviors in cells [6], and understanding of human disease [25].

A GEM is a mathematical representation of the available chemical reactions in a cell, stored as a stoichiometric coefficient matrix (Fig. 2 A-B). GEMs are often used together with flux balance analysis (FBA) [37]. FBA operates under a pseudo steady-state assumption, where the derivatives of metabolite concentrations with respect to time are assumed to be zero, meaning that there is no accumulation or depletion of metabolite concentrations. The goal of FBA is to solve the equation $S*v = 0$, where S is the matrix of stoichiometric coefficients describing the metabolic reactions and v is the unknown vector representing the fluxes through those reactions.

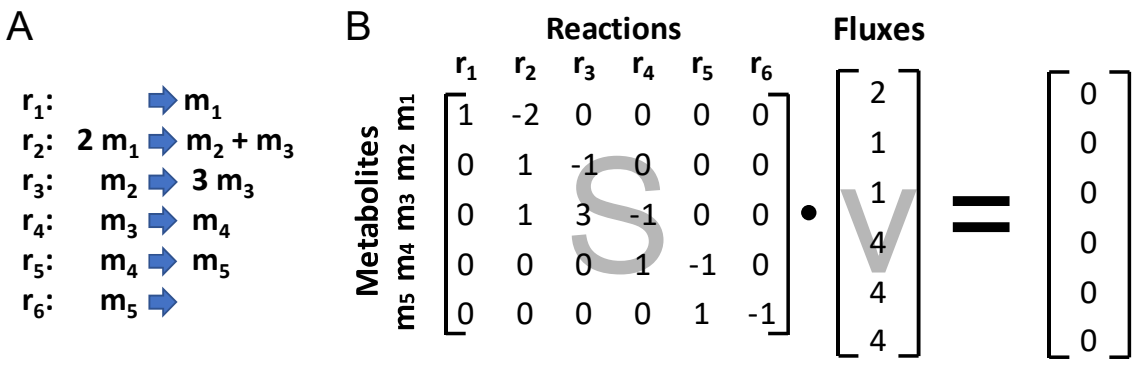

Fig. 2: The genome-scale metabolic model as a mathematical representation of the reaction network in a cell. *A. Example of a metabolic network with 6 reactions (r_1-r_6) and 5 metabolites (m_1-m_5). B. The reaction network in A represented in matrix form, with reactions as columns and metabolites as rows. In flux balance analysis, the goal is to identify the fluxes through the reactions that yield a net change of zero in metabolite concentrations over time.*

The flux balance equation system is in most practical cases underdetermined, which means that there are many solutions to the problem. The reason for this is twofold: 1) there are usually fewer equations than unknowns, and 2) the right side of the equation is zero, meaning that the *v* vector can be multiplied by an arbitrary scalar and still provide a solution to the problem. Linear constraints are therefore imposed on the problem, which restricts the possible solutions to a volume often referred to as the solution space (Fig. 3). For example, linear constraints are commonly based on experimental measurements of the uptake rates of metabolites into the cell. However, linear constraints are in most cases not enough to determine a single solution. To address this issue the cell is assumed to in an optimal way pursue a certain objective, defined as the objective function. A common objective is to maximize cellular growth (biomass production), which is a reasonable assumption for some cells such as cancer cells and cell lines, for which there is a selection pressure for proliferation. The objective is defined as a linear combination of reaction fluxes and can be either maximized or minimized, where an optimal objective can always be found in a corner of the solution space. The problem is solved computationally using linear programming, in which the problem is solved using software such as Gurobi [38].

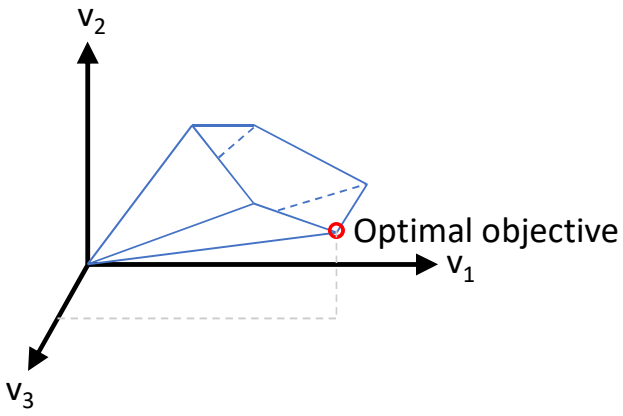

Fig. 3: The solution space in flux balance analysis. *The solution space is bounded by linear constraints, but still allows for an infinite number of solutions within that volume. Optimizing towards an objective defined by the objective function can be used to select a single solution of biological relevance.*

1.3.2. Enzyme usage constraints

Most reactions in the cell are facilitated by enzymes, and the presence of such enzymes in adequate concentrations is therefore needed for the reactions to be able to carry a certain amount of flux. The required enzyme concentration to uphold a certain flux can under certain assumptions be estimated via the Michaelis-Menten equation:

$$v = V_{max} \frac{[S]}{K_m + [S]} \quad (1)$$

where v is the flux through the reaction, V_{max} is the maximal flux through the reaction, $[S]$ is the substrate concentration, and K_m is the Michaelis constant, which is specific per enzyme and substrate. It directly follows that the flux cannot exceed V_{max}, which can then be used as an upper flux constraint [39]. V_{max} can be expressed as

$$V_{max} = k_{cat}[E]_0 \quad (2)$$

where k_{cat} is the turnover rate of the enzyme and $[E]_0$ is the enzyme concentration. As proposed in the GECKO method [39], it is possible to impose an enzyme cost C_r to each reaction r, such that

$$C_r = v_r \frac{M_w}{k_{cat}} \quad (3)$$

where v_r is the flux through reaction r and M_w is the molecular weight of the enzyme. C_r can then be constrained either per enzyme (using quantification of enzyme concentrations) or collectively by imposing a total enzyme usage constraint. While the cost C_r is the minimum enzyme cost to uphold the flux v_r, it is a reasonable approximation of the actual cost when K_m is much smaller than the substrate concentration. Measured k_{cat} values for enzymes can be downloaded from the BRENDA database [40]. In this thesis, a total enzyme usage constraint is used.

The Michaelis-Menten kinetics is is based on certain assumptions, such as operations in quasi-steady-state (where substrate and enzyme concentrations are constant over time), that the enzyme concentration is much smaller than the substrate concentration, that the enzyme and substrate(s) are in rapid equilibrium with the complex they form during the reaction, and that the reactions are irreversible (which many reactions in cells in practice are, since the formed product is rapidly consumed by another reaction) [41]. These assumptions are normally acceptable for enzyme kinetics in cells, although it remains important to keep in mind that Michaelis-Menten is an approximation of the actual kinetics.

1.3.3. Genome-scale models for human metabolism

To facilitate genome-scale metabolic modeling we have developed the genome-scale model Human1 [34], which contains in total more than 13,000 metabolic reactions. In addition, the model has been used as template for generating GEMs for model animals such as mouse and rat [42], and is compatible with the GECKO toolbox [39], [43] for applying enzyme usage constraints.

1.4. Generation of context-specific genome-scale metabolic models

In complex multicellular organisms such as humans the active reaction networks vary across cell types. While all enzymes are encoded in the genome for all cells, not all enzymes are expressed in all cell types. Determining the active reaction network is a means to reduce the solution space in for example FBA and thereby yield more accurate predictions of metabolism in cell types. In addition, it provides a method for comparison of metabolism across cell types based on omics data.

Today many methods are available for prediction of the active metabolic networks of cell types and organs [44]–[46]. The resulting models are often referred to as context-specific models and in this thesis I have used the method tINIT [46] for generating such models (Fig. 4). The reactions are associated with genes via gene-protein-reaction (GPR) rules, and the tINIT method scores each reaction based on evidence of presence of the associated enzymes from either proteomics or gene expression data. The reactions are given a negative score if there is little or no evidence that the associated enzymes are present, and a positive score otherwise. The algorithm then strives to identify the reactions to include by maximizing the sum of the scores of the included reactions, under the constraint that all included reactions must be able to carry flux. In practice, this constraint means that some reactions with negative score must be added to enable inclusion of other reactions with positive scores. The problem is solved by mixed integer linear programming (MILP) using a solver such as Gurobi [38].

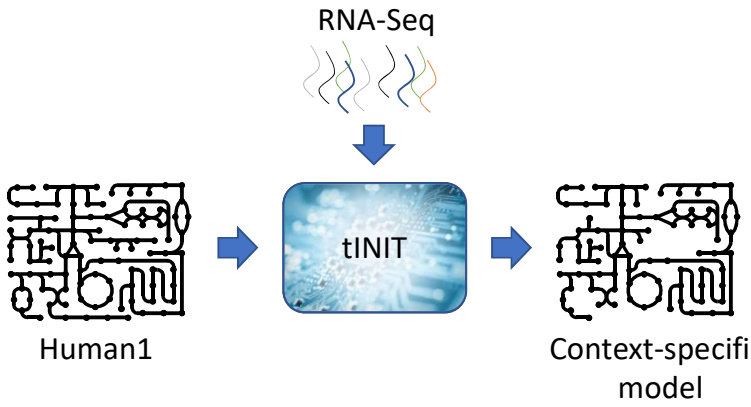

Fig. 4: Generation of context-specific models with tINIT. *tINIT takes as input a full model such as Human1 and data containing evidence of the presence of enzymes, here RNA sequencing data, and generates a context-specific model with a reduced set of reactions.*

1.5. Means to identify enzyme presence

Although some metabolic reactions in cells are spontaneous, most reactions present in the metabolic models have a negligible spontaneous reaction rate and rely on enzymes. In genome-scale metabolic modeling the genes of primary interest are those directly coding for metabolic enzymes or parts of metabolic enzyme complexes, in total around 3,000 in Human1. Generation of context-specific models requires evidence of the presence of enzymes, which typically involves measurements of the proteome (proteomics) or gene expression data (transcriptomics). The goal is to measure the absolute proteome of a cell since the enzyme concentration is approximately proportional to the maximal flux capacity of a reaction.

Already in 1975, 2-dimensional gel electrophoresis was developed to separate more than a thousand different species of proteins. In the eighties, techniques were developed to measure proteomics with protein mass spectrometry [47], [48], and antibody-based techniques have also been developed for estimating the proteome [49]. Proteomics is today well-established for bulk data (a sample containing of a large number of cells), but while significant advances have also been made in the single-cell field, single-cell proteomics still struggles with many challenges, for example lack of methods for amplification of the small amount of starting material from a single cell [50]. An alternative to proteomics is to quantify the gene expression profile, known as the *transcriptome*. Studies of the transcriptome began in the nineties, and techniques such as quantitative PCR (qPCR), microarrays, and later RNA sequencing (RNA-Seq) have emerged as well established methods [51], [52]. During the last 15 years, the development of RNA-Seq methods have advanced rapidly. The first studies in 2008 covered the transcriptome of a few samples [53], [54], but over the last decade, efforts such as the cancer genome atlas (TCGA) has sequenced over 10,000 patient samples. The advances in the field of single-cell RNA sequencing (scRNA-Seq) are even more impressive – in ten years, it has gone from its infancy to enabling the profiling of hundreds of thousands of cells in a single dataset [55].

While the correlation between transcriptome and proteome is modest [56], RNA-Seq can still be used for detecting the presence of enzymes and thereby the presence of chemical reactions. There are several advantages of RNA-Seq compared to proteomics: Mass spectrometry suffers from large sources of noise and covers fewer genes than RNA-Seq, RNA-Seq cost less, single-cell RNA-Seq is a widely used technology while single-cell proteomics is less developed, and the number of public datasets is vast [57], [58]. Proteomics is however more advantageous when absolute protein measurements are required, for example for constraining individual reactions, or if additional information such as protein phosphorylation is desired.

Bulk RNA-Seq can be useful for investigating the average gene expression in a cell population but fails to describe the heterogeneity within that population. FACS-sorting [59] can be used to sort cells into cell types defined by cell surface proteins, which partly helps in investigating the transcriptome of individual cell types with bulk RNA-Seq. However, FACS-sorting is limited to sorting on cell surface protein abundancies and is difficult to use for separating cells into more subtle categories than cell types and subtypes. In addition, publicly available datasets are often not FACS-sorted in categories suitable for other experiments. Single-cell RNA Sequencing can overcome these limitations by measuring the transcriptome of individual cells and thereby offers possibilities to investigate cell heterogeneity in detail.

1.6. Bulk RNA Sequencing

Bulk RNA-Seq is usually performed on transcripts originating from a large pool of cells (typically $> 10^5$ in a biopsy sample [60], although it depends on the size of the sample), yielding a fair amount of RNA as starting material. Fig. 5 describes a typical workflow for bulk RNA-Seq [61], although such protocols can vary much across experiments. The workflow starts with extraction of RNA from tissue. The extraction strategy varies substantially depending on tissue - it is for example easier to extract RNA from blood than from solid tissues, and commercial kits are available for such purposes.

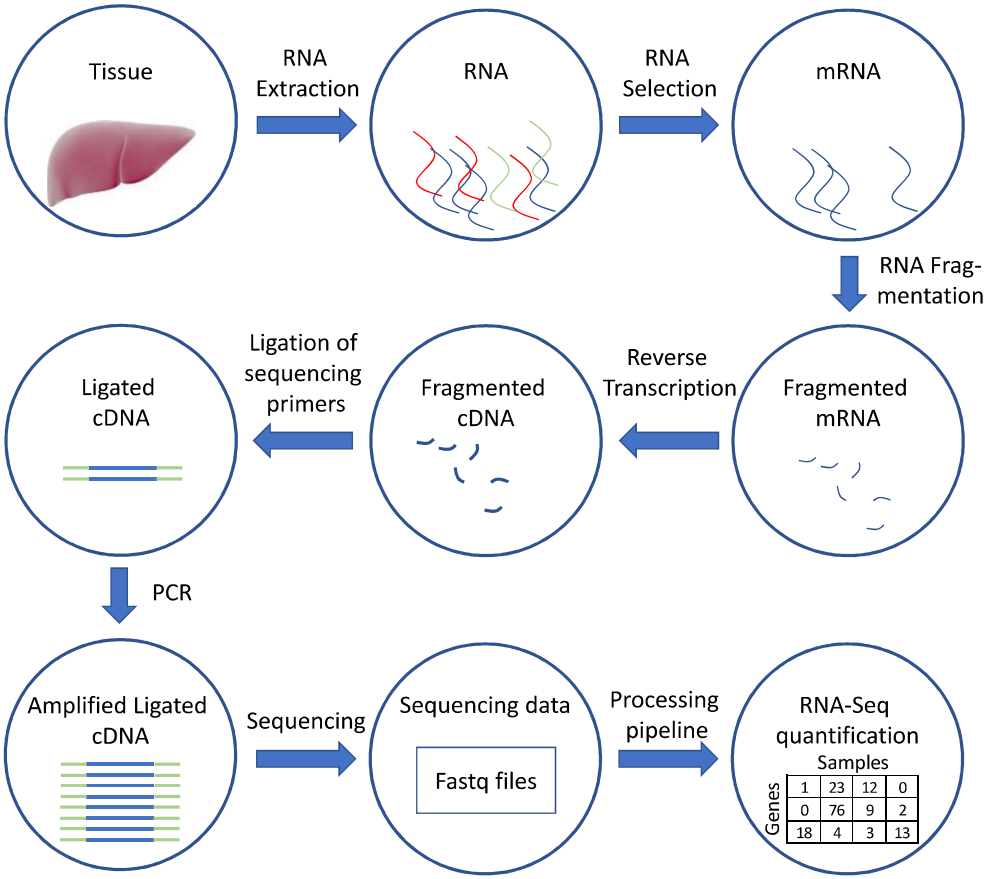

Fig. 5: Typical bulk RNA-Seq workflow. *The figure describes one possible workflow; there are however many different variants of workflows. For example, fragmentation can be done after reverse transcription.*

In the extracted RNA, the mRNA molecules are usually the ones of interest, and ribosomal RNA is abundant and needs to be removed to avoid sequencing reads of little interest. Common methods are poly-A selection, which enriches for RNA molecules with a poly-A tail and thereby removes rRNA, and rRNA depletion, which selectively removes rRNA molecules. The subsequent fragmentation step splits the mRNA into shorter sequences to simplify sequencing. Reverse transcription then creates complementary DNA (cDNA) from the RNA using reverse transcriptase [62], an enzyme naturally present in retroviruses. cDNA is much more stable than RNA and can also be used in polymerase chain reaction (PCR) [63], which is almost always required to yield enough material for sequencing. To enable PCR amplification, primer sequences are ligated at the ends of the cDNA molecules, which also serve as barcodes enabling pooling of different samples when sequencing. Commonly, about 10 cycles of PCR are run before sequencing is commenced, but the number of required cycles may vary depending on the amount of starting material. The sequence of a molecule affects the efficiency of the PCR, resulting in an uneven amplification across genes and samples, leading to technical batch effects [64]–[67]. During sequencing, the sequencer can perform single-end or paired-end sequencing. In single-end sequencing, the sequencer reads one end of the molecule, while in paired-end sequencing, both ends are read, providing more data for each molecule.

1.7. Single-cell RNA-Sequencing

Single-cell RNA-Seq seeks to capture the transcriptomic profile of individual cells. The number of mRNAs within a human cell varies across cell types and cell states and has been estimated as somewhere between 50,000 to 300,000 molecules [68], which is very small compared to the total number of molecules in for example a bulk biopsy sample. Therefore, scRNA-Seq typically requires more PCR amplification cycles than its bulk counterpart. Measurements of individual cells alone are in general noisy – the small number of mRNA molecules per cell in combination with PCR biases and other technical limitations yields a sparse transcriptional profile that diverges substantially from the average expression over many cells of the same cell type. In addition, the transcriptome of a cell varies stochastically with time due to a process called transcriptional bursting [69]. For many analyses, including genome-scale modeling, it is therefore desirable to identify similar cells and perform analyses on groups of cells.

There exist a large variety of single-cell RNA-Seq technologies. The technologies can be divided into high-throughput methods, having the advantage of lower cost per analyzed cell, and methods that focus on high data capture per cell. High-throughput methods are often based on capturing cells in oil droplets (droplet-based methods), while methods focusing on high data capture often rely on FACS sorting of individual cells into plates (plate-based methods).

1.7.1. Barcoding

Individual mRNA molecules can be barcoded by attaching additional nucleotide sequences, commonly at the poly-A tail of the transcripts. Barcodes can be used for several purposes; the two most important are unique molecular identifiers (UMIs) and cell barcodes. A typical strategy is to combine barcoding with paired-end reads, where one read contains the barcodes and the other contains the biological read.

UMIs are random sequences that are used to identify original molecules [70]. Since PCR amplifies the molecule fragments, several reads could originate from the same molecule. To reduce PCR biases and sampling noise, a common method is to count detected molecules rather than sequenced read counts, where all reads with the same UMI sequence (and sometimes also the same cell barcode and gene) are collapsed to a single molecule in a process called *UMI collapsing*.

Cell barcodes are used to identify which cell an mRNA molecule originates from, and are commonly used in droplet-based methods. Cell barcodes enable pooling of mRNAs from multiple cells (on the order of thousands in droplet-based methods). The molecules from all pooled barcoded cells are then processed together and demultiplexed in the computational pipeline to form gene expression profiles for single cells, reducing both costs and unwanted technical variation across cells.

1.7.2. Extraction of cells from complex tissue

Single-cell RNA-Seq requires isolation of single cells for further processing. Depending on the tissue to investigate, this procedure may be challenging. In some cases, such as peripheral blood mononuclear cells (PBMCs), extraction is easy since the cells are already separated and freely available in a fluid. However, for complex tissues (such as brain, lung, etc.) many of the cells are tightly attached to the extracellular matrix and need to be freed

before processing. Commonly, the tissue is first dissected into smaller pieces, mechanically minced, and then treated with enzymes (such as dispase, collagenase, and trypsin) to break down the extracellular matrix, often followed by further mechanical treatment [71]. The extraction step introduces technical biases, since the cells may start to change their gene expression during the procedure as an adaptation to the new environment [71]. In addition, some cell types may be extracted more efficiently than others, which is one of the reasons why the fraction of cell types in the single-cell data cannot be assumed to reflect the fractions of the different cell types in the tissue [71].

1.7.3. Droplet-based methods

Droplet-based methods allow for processing of thousands of cells in a single run at low cost. An example of a droplet technology is *10x Chromium NextGEM*, where each cell is partly processed in its own oil droplet (Fig. 6). As described in the manual [72], cells are extracted from the tissue and input into a flow-cell together with beads containing barcoded reverse transcription (RT) primers. The cells attach to the beads, and both assemble in an oil drop together with a reverse transcription solution. The cells are lysed and the beads dissolve, producing oil drops containing all necessary ingredients to perform reverse transcription with barcoding. RT is then performed, generating barcoded full-length cDNA from RNA species with a polyA tail, where each molecule contains both a cell barcode and a UMI. Subsequently, the oil drops are dissolved, and the oil is removed. The barcoded full-length cDNA is then amplified in a preamplification PCR step. The rest of the processing resembles that of bulk RNA-Seq; the amplified full-length molecules are fragmented, followed by PCR amplification. Since the barcodes are attached to the polyA tail, only the fragments containing this tail will have barcodes, and only such fragments are selected for in the amplification. The cDNA library is then sequenced and processed, where the reads are demultiplexed by cell barcode and multiple copies of the same molecule, as identified by UMI, are discarded [72].

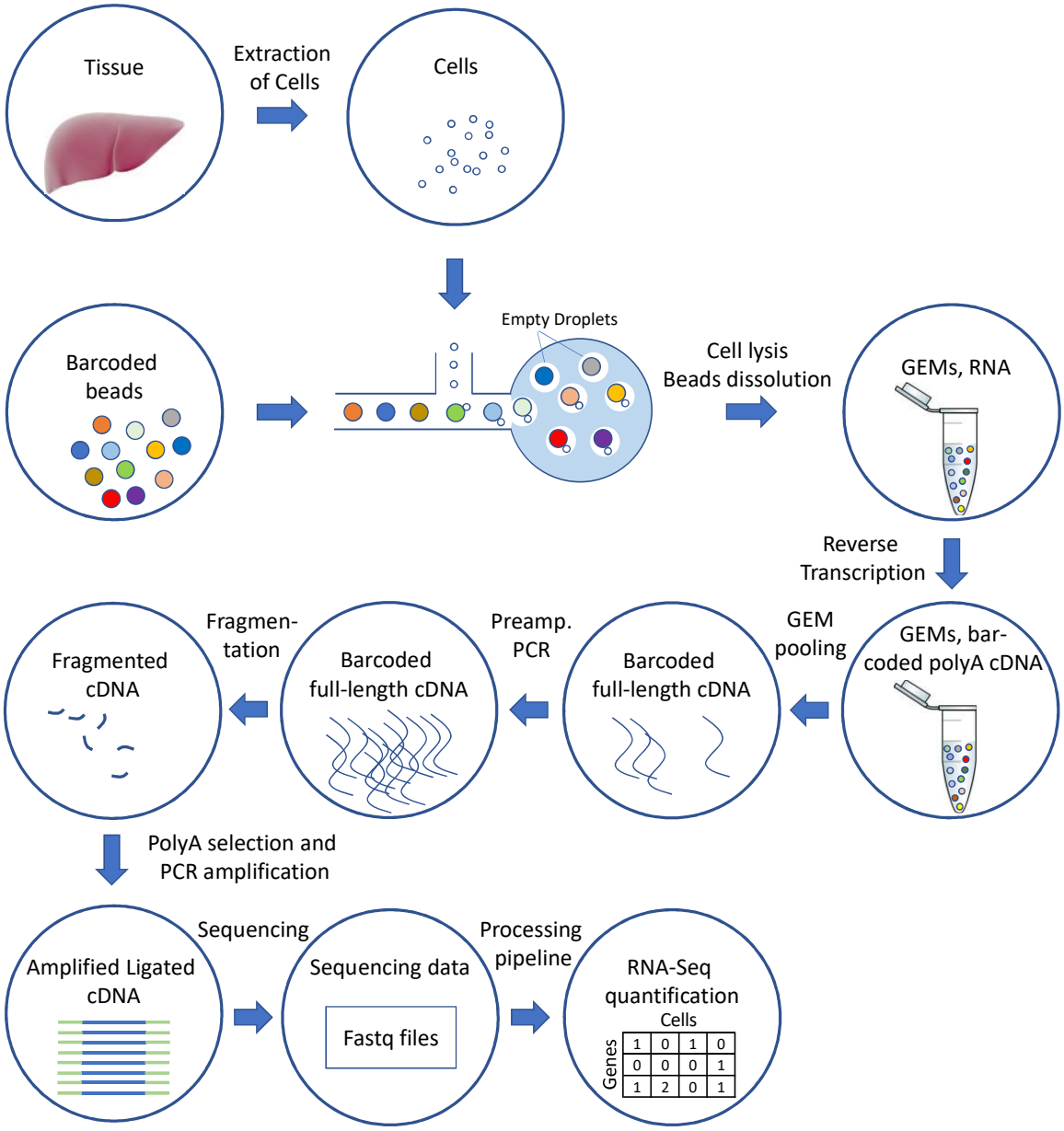

Fig. 6: Overview of the 10x Chromium workflow. *Cells are attached to barcoded beads and separated into droplets in which processing up to reverse transcription is performed, producing barcoded full-length transcripts. The transcripts from all cells are then pooled, preamplified, and then processed in a similar way as bulk RNA-Seq. In this context, GEM corresponds to gel-beads in emulsion.*

In droplet-based methods, reads are concentrated to a limited part of the gene, since most amplified fragments are short and close to the polyA tail. It is therefore difficult to identify different splice variants of a gene from data of droplet-based methods.

The most commonly used droplet-based technologies to date are 10X Chromium [73] and DropSeq [74].

1.7.4. Plate-based methods

In plate-based methods, single cells are typically sorted into plate wells using FACS. The most common plate-based method to date is Smart-seq2 [75]. The protocol does not support unique molecular identifiers; the count matrix produced thus does not represent original molecules, but fragment counts after PCR. The two large advantages of Smart-seq2 compared to droplet-based techniques is that it 1) captures much more original mRNA molecules and 2) captures fragments from the whole transcript, like in bulk, which enables identification of different splice variants of genes. Although such molecules cannot be uniquely identified, it is evident that more molecules are captured since the count matrix produced by this protocol is much less sparse than matrices from droplet-based technologies. Smart-seq2, being a plate-based protocol where each cell is treated in similar ways as bulk RNA-Seq samples, is much more expensive to perform per cell. The overall workflow is similar to that of bulk, although with more amplification.

A new version of Smart-Seq, Smart-Seq3 [76], has recently been developed. Smart-Seq3 supports unique molecular identifiers and has been shown to be able to capture up to 150k molecules per cell [76], which is at least an order of magnitude more than most droplet-based datasets. This new version is a promising candidate for capturing the transcriptome of individual cells at high resolution.

1.8. Analysis of single-cell RNA sequencing data

1.8.1. Processing of sequence files

The final output from sequencing is usually a list of text files in the FASTQ format [77], containing sequences from individual transcript fragments. The goal of *alignment* is to map these sequences to a reference genome and thereby enable the generation of a count matrix with samples as columns and genes or transcripts as rows. Common aligners are STAR [78] and HISAT2 [79], which adopt different algorithms for alignment. Additional options for alignment include pseudo-aligners such as kallisto [80] and Salmon [81], which rely on efficient mapping of kmers to the genome. Pseudoalignment and full alignment give similar results; however, a recent evaluation between STAR (full alignment) and kallisto showed that full alignment using STAR gives more mapped reads but is much slower and consumes much more memory [82].

Droplet-based data require additional processing in the form of cell demultiplexing and UMI collapsing, and often include correction methods for sequence errors in both UMIs and cell barcodes. 10X Chromium data can be processed using CellRanger, which uses the STAR aligner and is provided by 10X Genomics. Alternatives are the kallisto-bustools [83] and Salmon-Alevin [84] workflows, which are based on pseudoalignment, as well as STAR Solo [85], which is tightly integrated with the STAR aligner.

1.8.2. Statistical properties of single-cell RNA-Seq data

Bulk RNA-Seq data is often modeled using a negative binomial distribution for the counts for each gene, for example in DESeq2 [86]. Practically, this means that many mathematical methods that assume normally distributed data, such as least squares regression, T tests etc., should be used with caution on count data. A common method to make the data approximately normally distributed is to log transform the data before further processing. Although single-cell data resembles bulk data in many ways, there are several additional aspects to take into consideration when working with such data. 1) UMI-based scRNA-

Seq behaves differently than technologies relying on direct count data (such as Smart-seq2), in that several copies of the same molecule are encountered in the direct counts, creating grouping in the data and dependence across reads. While UMI-counts for a cell can be described with a simple sampling model [87], grouping in the data (here referring to the presence of multiple reads from the same original molecule amplified with PCR) will lead to excess zeros in the total counts matrix as compared to what would be expected from sampling [88]. This phenomenon, which is termed zero-inflation, only applies to direct counts and makes the data deviate from the expected sampling (multinomial) distribution. The same effect should in theory appear in bulk data; however, the number of copies per molecule encountered in a typical bulk sample is much smaller, making this effect less important. 2) Due to the limited number of molecules in a cell, many lowly and moderately expressed genes will by chance become zero in many cells, although they are most likely expressed. This property of single-cell data is termed "dropouts" [89] and causes problems for many computational methods.

Generation of single-cell RNA-Seq data involves many factors that can affect the final gene expression values obtained, for example sample preparation method, different personnel, different equipment, and even technology platform [67]. Together, all these effects create systematic differences across datasets, commonly termed "batch effects". For integration of single-cell datasets, batch correction, which attempts to remove such effects, is an important step of the analysis [67].

1.8.3. Processing of gene count data

The processing of single-cell gene count matrices is today fairly standardized in software packages such as Seurat [90]. The standard steps for the analysis are outlined in Table 1. In addition, more analysis options, such as differential expression analysis, are supported. Normalization, data transformation and scaling can be replaced with a method called scTransform, which is based on using Pearson residuals from a generalized linear model (GLM), and it can reduce biases across cells induced by library size [91]. To present an overview of the dataset, single-cell data is often finally visualized using uniform manifold approximation and projection (UMAP) [92], which when used with single-cell data often (as in Seurat) uses the output from PCA as input (Fig. 7). Additional steps can include removal of contaminating transcripts from broken cells [93] and removal of doublets (droplets/wells containing more than one cell) [94].

Table 1: The standard Seurat Workflow. *The table describes a typical step-by-step performed with the software package Seurat.*

Step	Description
Filtering of low-quality cells	Filtering of empty droplets/wells and dead cells.
Normalization	Library size normalization.
Data transformation	Each normalized value is transformed to $ln(c + 1)$, where ln is the natural logarithm and c is the normalized count value in the matrix.
Finding variable genes	It is advantageous to filter the genes before processing, only including the most variable genes. Genes expressed similarly across cells mostly adds noise to the analysis and worsens the results.
Data scaling	Involves centering of the data and scaling to yield the same standard deviation for all genes.
PCA	Principal component analysis, used to reduce the number of dimensions of the data.
Clustering of cells	Clusters the cells, usually into cell types and subtypes.
UMAP	Uniform manifold approximation and projection, used for visualizing the cells in two dimensions.
Visualization	Various figures to visualize different aspects of the data.

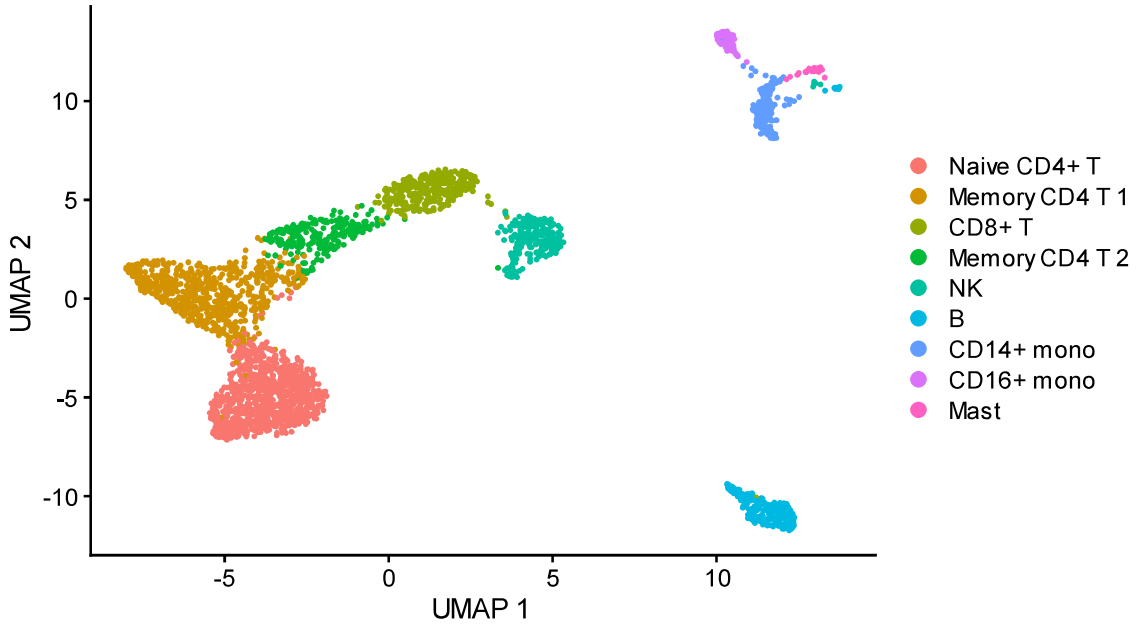

Fig. 7: Visualization of single-cell data using the UMAP projection. *The dataset presented is 2,900 frozen peripheral blood mononuclear (PBMC) cells, sample A [73].*

1.8.4. Filtering of low-quality cells

Single-cell RNA-Seq produces low quality data for some cells that need to be filtered out before data analysis. There are three common ways to identify such cells: 1) Cells with too few or too many total counts/UMI counts, where too few UMI counts often represent empty droplets in droplet data; 2) Cells with too few detected genes (often called features); and 3) Cells with too high fraction of mitochondrial gene content, which may indicate dead cells (Fig 8) [95], [96]. 1 and 2 above are often highly correlated (Fig 8B), and 2 is often

excluded from the filtering process. The threshold values for these metrics vary across datasets and must be estimated from the data. For droplet-based methods, many "cells" represent empty droplets, containing few mRNA molecules. The mRNA molecules found here originate from broken cells and are spread throughout the cell solution [93] and the aim is to find the threshold value where the vast majority of the empty droplets are below the threshold. Typical threshold values here may vary from 200 to 1000 UMI counts, whereas typical values for the mitochondrial content threshold may range between 4-10%. Estimating these values is supported by software packages such as Seurat [90].

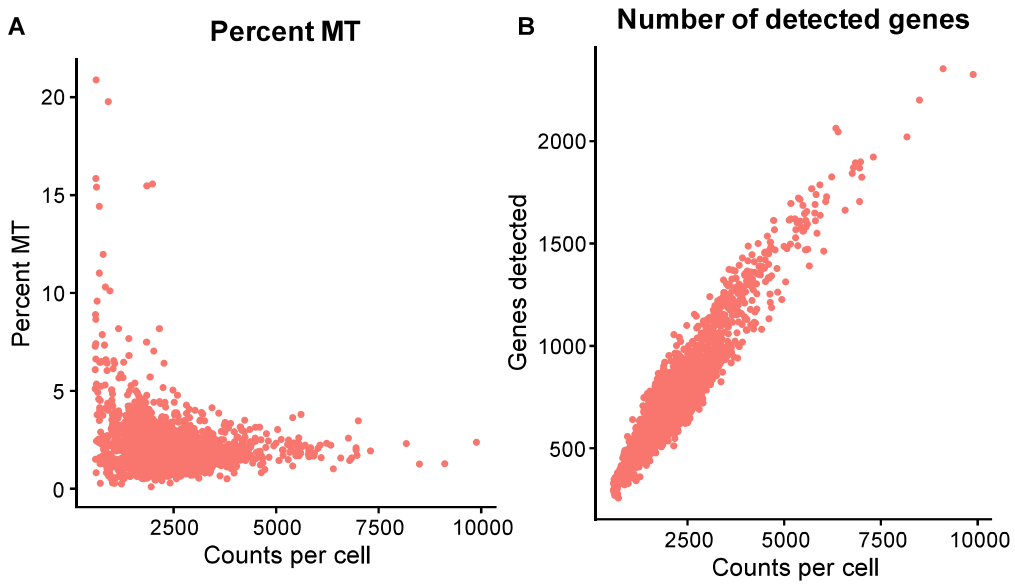

Fig. 8: Filter criteria for single cells. *A. Mitochondrial (MT) content per cell (fraction of total counts that belongs to mitochondrial genes). A suitable threshold for this dataset could be 4.5%. B. Number of detected genes per cell. Suitable filtering for this dataset could be to keep cells with between 200 and 2,000 detected genes. The dataset presented contains 2,900 frozen peripheral blood mononuclear (PBMC) cells, sample A [73].*

1.8.5. Normalization

Bulk RNA-Seq and other full-length protocols such as Smart-seq2 suffer from gene length bias, since the number of cDNA fragments generated per molecule is approximately proportional to the gene length. To remedy the bias, the counts can be divided by gene length, and are often scaled to a total sum of 10^6, yielding transcripts per million (TPM). An alternative method is FPKM/RPKM [97], where library size scaling is performed before compensation for gene length. A comparable normalization for droplet-based data, for which there is no obvious source of gene length bias, is to simply scale the total counts of a cell to a sum of 10^6, forming counts per million (CPM). These methods are called library size normalization methods and are useful for normalization across genes (TPM and FPKM/RPKM).

Library size methods have been shown to be less effective for normalization across samples in bulk RNA-Seq samples [98]. Therefore, more advanced methods, such as the trimmed mean of M values (TMM) [98] and the median of ratios normalization performed by DESeq2 [86], have been developed to control for this effect. These methods assume that most genes are not differentially expressed across samples and scale the samples to fit this assumption. A downside with these approaches is that it is difficult to compensate for gene length, since they operate on direct counts, which makes them less suitable for comparisons across genes. For datasets with technical biases that change the distribution

of counts per gene, quantile normalization [99] can be applied, which makes the counts distributions equal across samples.

Normalization of the transcriptomes of single cells in scRNA-Seq is more challenging due to the high number of zeros in the data. Methods commonly applied to bulk data, such as TMM and the normalization performed by DESeq2, fails due to the high zero content. Library size normalization is therefore still commonly applied to single-cell data, for example in Seurat [90], despite its shortcomings. An advanced method for normalizing single-cell data is based on repeatedly normalizing the total transcriptome of groups of cells, and then applying a deconvolution strategy to single out the scale factor for individual cells [100]. With this method, TMM or the normalization performed by DESeq2 can be applied. The method is implemented in the Bioconductor scran package [101].

1.8.6. Clustering of cells

The purpose of clustering is to divide the cells into groups, called clusters, where the cells in a cluster share common properties such as cell type/subtype and/or cell state. There are numerous clustering algorithms available for this purpose and there is no clear recipe for which algorithm to choose for a certain dataset. Most algorithms adopt an unsupervised approach, meaning that they do not use any previous knowledge about the expected transcriptional profiles of clusters. The data is often preprocessed, by normalization, log transformation and dimensionality reduction methods such as principal component analysis (PCA). The algorithms can be grouped into categories, for example k-means clustering, hierarchical clustering, and community-detection-based algorithms [102]. K-means identifies a specified number of cluster centers in multidimensional space and assigns each cell to the closest cluster. Hierarchical clustering creates a tree structure where the nodes are clusters, where each cell is assigned to a node at each tree level. Close to the root, there are few, large clusters, while the leaves represent a larger collection of clusters with fewer cells. A downside with hierarchical clustering is the required computational resources and memory, which both scale at least quadratically with the number of cells [102]. Community-detection-based algorithms, such as the Louvain algorithm, which for example is implemented in Seurat [90], work on a k-nearest-neighbors graph and are generally fast. All clustering methods mentioned here require the user to specify parameters that either directly or indirectly specify the number of clusters desired.

1.9. Use of single-cell RNA-Seq with genome-scale metabolic modeling

While single-cell RNA-Seq has not been extensively used together with genome-scale metabolic modeling, there are a few examples. Some methods focus on generating context-specific models from single cells, but due to lack of data use small simplified models containing highly expressed enzymes [103]. Other methods employ different strategies to integrate data across neighboring cells, which gives a more stable gene expression, but still only investigate highly expressed pathways [104]. Others have generated context-specific GEMs from pooled pseudo-bulk RNA-Seq profiles derived from single-cell data, but have not fully investigated the effects of sparsity on the model quality [105], [106]. While these methods are useful, none of them fully address the problem of generating full (non-simplified) context-specific models, where the uncertainty of the sparsity from scRNA-Seq is considered. These models can then be used in advanced simulations including constraints on enzyme usage and metabolite uptake rates.

1.10. Aims and significance

Genome-scale metabolic modeling holds promise to unravel the metabolism of human cells in health and disease [107]. However, the method is limited by difficulties in determining the presence of individual enzymes in the different cell types in the human body. Previously, such analyses from public datasets have mainly been limited to whole tissues [46], or potentially to cell types for which data is available in FACS-sorted datasets. The metabolic interplay between different cell types in organs is therefore largely unknown and dysregulation of such interactions may play an important role in human disease. With the arrival of single-cell RNA sequencing, a new opportunity has arisen to generate context-specific GEMs for individual cell types.

Despite the vast scientific effort spent on cancer research, cancer remains a leading cause of death worldwide, and dysregulated metabolism has been identified as an emerging hallmark of cancer [15]. The metabolism in the tumor microenvironment of solid tumors is to date not fully understood. Previous research has implicated metabolic collaboration between stromal cell types and cancer cells in the TME [28]–[30], but no such effects have been quantified and it is therefore unknown if they exist. Genome-scale metabolic modeling is ideal for quantifying such a collaboration, which if proven to exist could motivate further research in the area. Likewise, proving the collaboration to be nonexistent could help by avoiding spending more research effort on the subject.

In this thesis, the two primary aims have been to develop a method that utilizes single-cell RNA-Seq data for generation of context-specific genome-scale models and to investigate cancer metabolism in the TME using genome-scale models. Secondary aims have been to develop methods for improved quantification of single-cell RNA-Seq data and to improve methods for applying enzyme usage constraints to GEMs.

2. Addressing variation in RNA Sequencing data

Part of this thesis is centered around generation of context-specific models from RNA-Seq data, particularly by utilizing RNA-Seq data from single cells. As a first step on the road towards generation of context-specific models from such data, methods for normalization and batch correction were examined (**Paper I**). As a second step the sources of variation in the data were examined (**Paper I**). The third part concerned investigation of the variation across cells in single-cell data (**Paper II**). The fourth part focused on the possibility to address sparsity in single-cell data by pooling cells (**Paper II**). The fifth part of this work explored the possibility to utilize bulk RNA-Seq data together with mathematical deconvolution [108] to estimate the gene expression of individual cell types in cancers (**Paper I**).

2.1. Evaluation of normalization and batch correction methods

Normalization of individual cell profiles in single-cell RNA-Seq data is nowadays part of the standard workflow in tools such as Seurat [90], and is well understood. However, normalization of profiles from pooled single-cell populations is less investigated, especially how they compare to bulk RNA-Seq. To evaluate normalization methods for bulk RNA-Seq profiles and RNA-Seq profiles generated by pooling data from populations of cells in scRNA-Seq, three normalization methods were applied on a collection of 105 profiles from 10 publicly available datasets (Fig. 9). The samples were either B or T cells, so a certain deviation between the samples was expected, but the effect of different normalization methods was still apparent. Library size methods, such as TPM (or CPM), failed to properly normalize the samples. While it is tempting to use more advanced normalization methods, there is a catch – such methods operate on count data. Count data is not comparable between droplet-based single-cell methods and full-length protocols such as bulk RNA-Seq and Smart-seq2, since the counts for the former category should not be compensated for gene length, while the counts for the latter should. To circumvent this issue, the TPM and count values were used together for bulk/Smart-seq2 data, where the TPM was scaled to the same library size as the total number of counts for each sample, generating pseudo-counts that were used instead of counts. TMM improved the normalization to an acceptable level, and although quantile normalization improved the results further, quantile normalization introduced much unwanted changes in the data and should be avoided if possible.

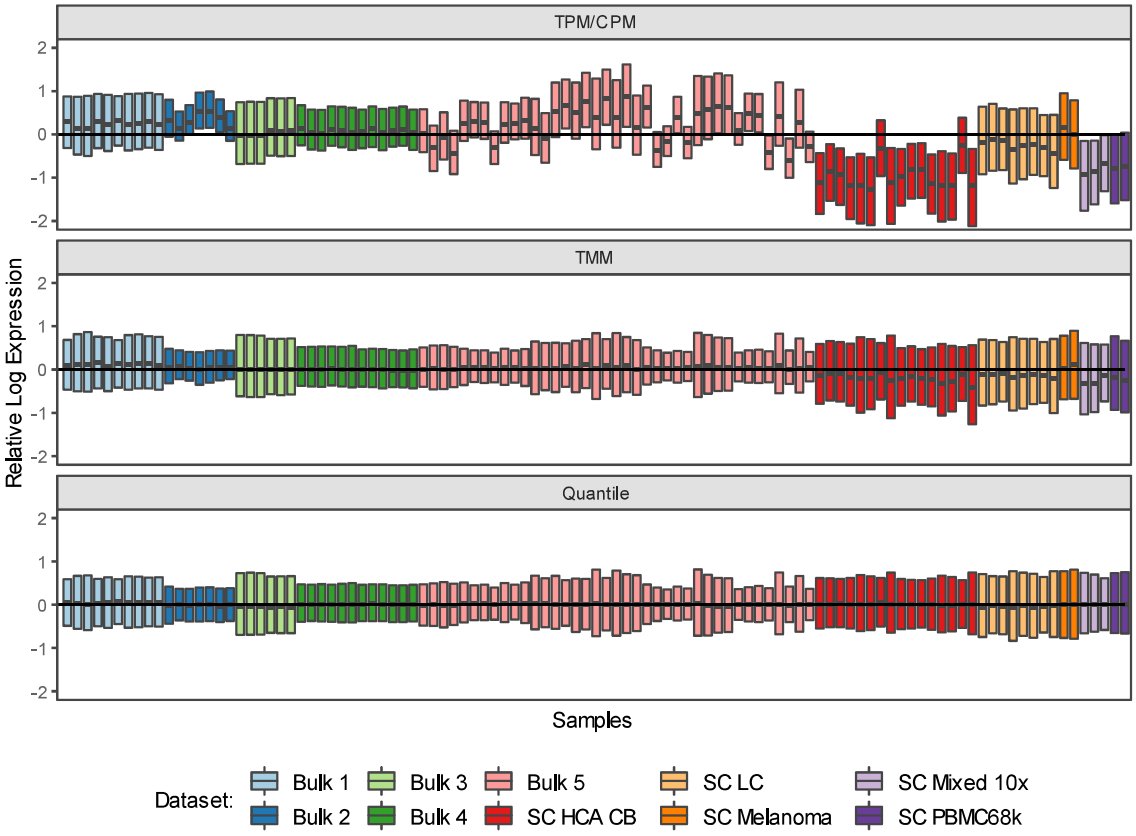

Fig. 9: The effect of different normalization methods. *Relative log expression is calculated as the $\log_2$ fold change (with a pseudo-count of 1 in the $\log_2$ transformation) between the expression of each gene and the median expression of that gene across all samples, presented as one boxplot of all genes per sample. The samples consist of a mix of B and T cells from 10 datasets, of which 5 are single-cell datasets. For single-cell datasets, the samples consist of RNA-Seq profiles generated by pooling cells of the same cell type. SC Melanoma is Smart-Seq2 data, the rest of the single-cell data was generated using 10X Chromium.*

To investigate the effect of the failure to normalize samples properly, the samples were compared using PCA, where the two first components were deemed to be technical (PC1) and cell type (PC2) (Fig 10). TMM clearly reduced the technical component compared to TPM (or CPM), while quantile normalization only yielded a small improvement compared to TMM. To test the effect of batch correction, ComBat [65] was applied, instructed to preserve differences in cell type. As expected, ComBat removes most technical effects across datasets.

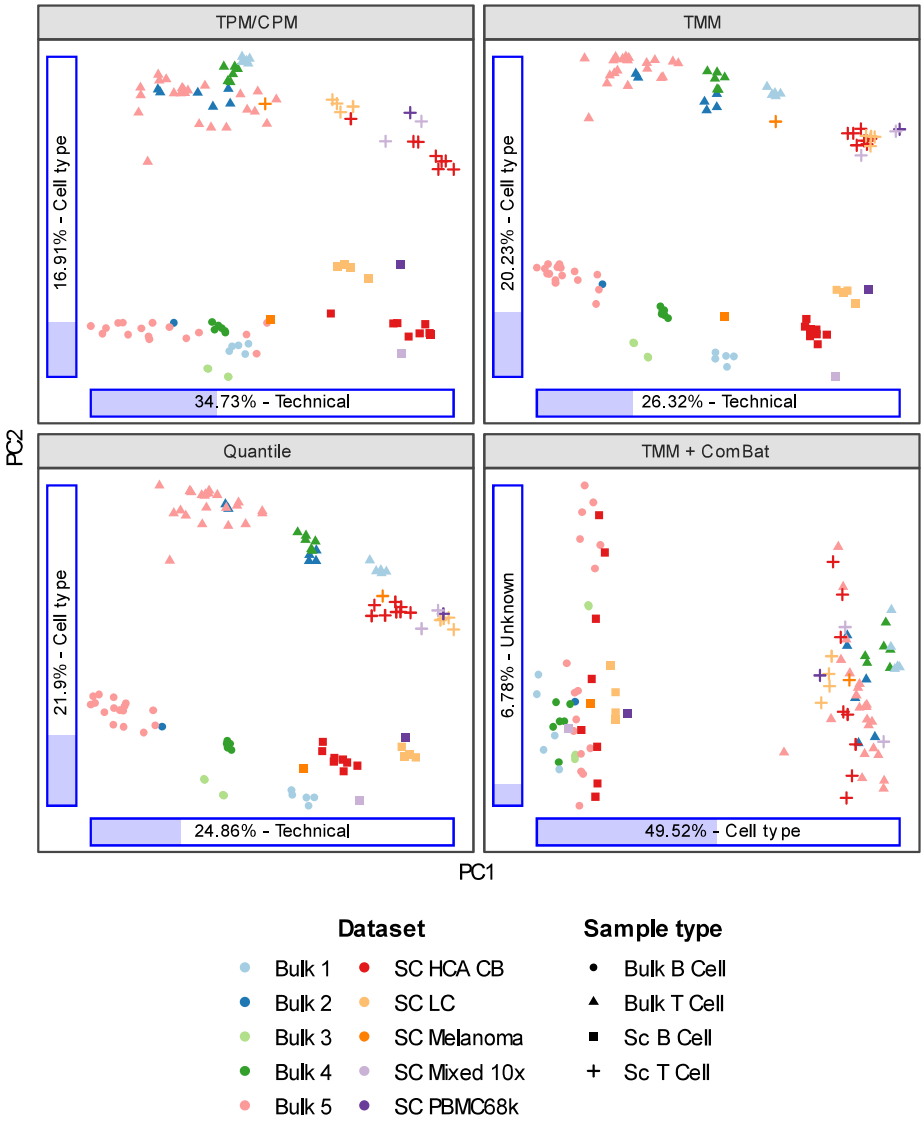

Fig. 10: The effect of normalization on PCA. *The bars indicate the variance explained by the principal components.*

2.2. Sources of variation in RNA-Seq profiles

To investigate what explains the differences in gene expression between samples, we quantified the effect of several factors on TMM-normalized gene expression (the same data as used for evaluation of normalization above). Specifically, the R package variancePartition [109] was used, where all factors were modeled as random effects in a mixed linear model. The factors investigated were the laboratory that produced the data, cell type (B or T cell), and tissue of origin (i.e., the tissue the biological sample was collected from).

Across all genes in bulk data, the laboratory was the most important factor (Fig. 11A), likely partly since many genes don't vary substantially across different types of lymphocytes. For housekeeping genes, laboratory as expected became even more important since the biological differences are expected to be small (Fig. 11B). Interestingly, tissue of origin explained more variance than cell type in these cases, which may suggest that local adaptations to tissue occurs for many genes, while fewer genes are actually different across B and T cells. When focusing on genes that are different between B and T cells (denoted LM22S, as identified in the LM22 matrix in CIBERSORTx [108], filtered on having an absolute log fold change > 1 between the cell types), cell type became the most important factor (Fig. 11C).

For pooled single-cell profiles, lab was still the most important factor across all genes, although the residuals are large (Fig. 11D). The high residuals, which represent either random noise or factors not accounted for, could potentially be explained by sampling noise arising from data sparsity, since pooled single-cell profiles in many cases have much fewer counts than a bulk sample. For the LM22S gene subset, cell type was the most important factor, although the residuals were still high (Fig. 11E). For a mix of bulk and pooled single-cell profiles, cell type was still the most important factor for the LM22S gene subset (Fig. 11F).

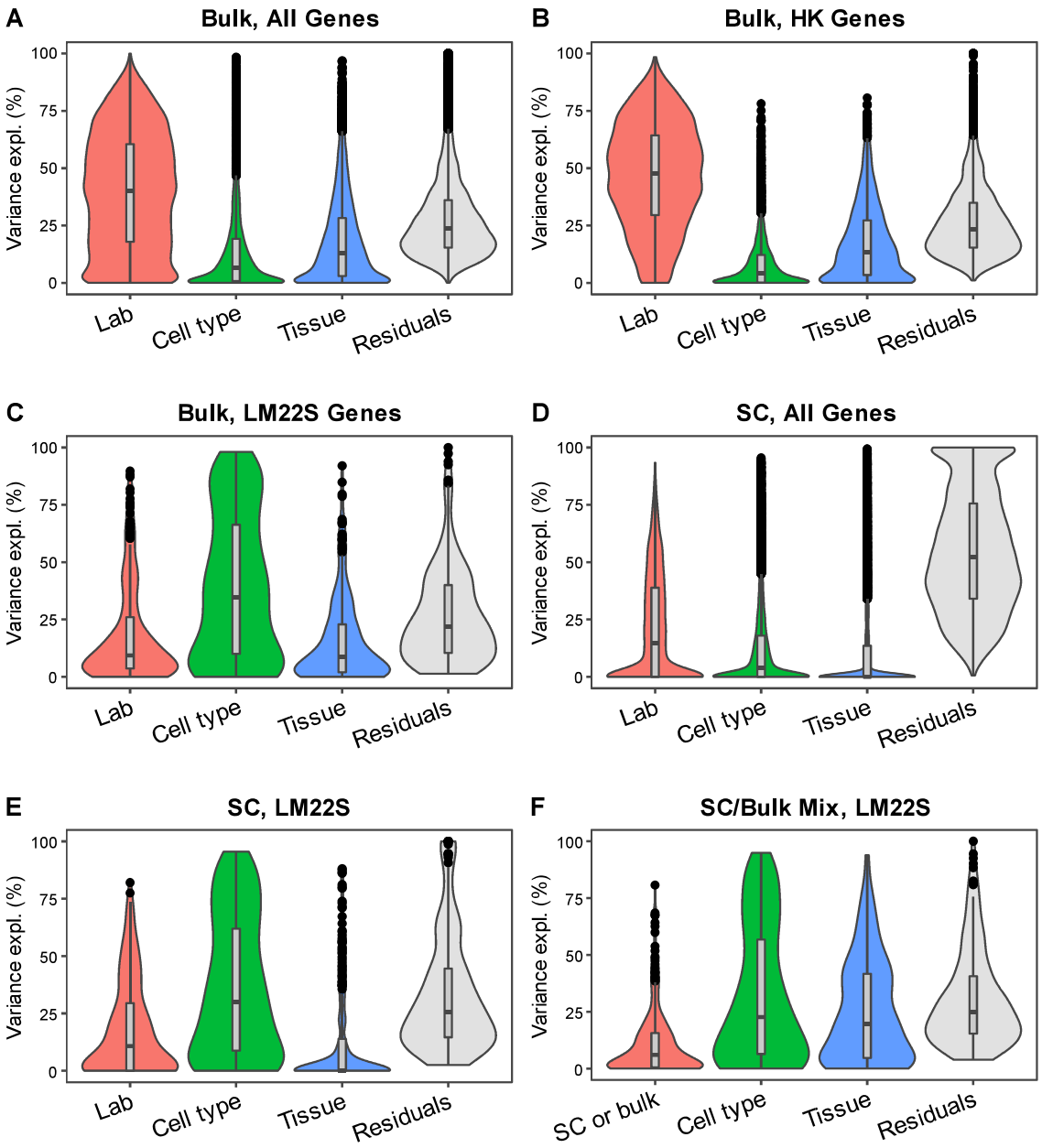

Fig. 11: Factors explaining the variation in RNA-Seq profiles. *A collection of B and T cell RNA-Seq profiles from both bulk and pooled single-cell was used (see* ***Paper I****). A. Explained variances across all genes (12072 genes) in bulk samples. B. Explained variances for the housekeeping genes (3393 genes) in bulk samples. C. Explained variances for the LM22S genes (274 genes) in bulk samples. D. Explained variances for all genes in pooled single-cell RNA-Seq profiles. E. Explained variances for the LM22S genes in pooled single-cell RNA-Seq profiles. F. Explained variances for the LM22S genes in a mix of bulk and pooled single-cell RNA-Seq profiles. "Residuals" represent the fraction of the variance not explained by the other factors in the figure.*

2.3. Cell-to-cell variation in single-cell data

Cell-to-cell variation in single-cell data is dominated by sampling noise for most genes. To investigate the part of the variation that does not originate from sampling noise we developed the DSAVE (DownSAmpling-based Variation Estimation) method, which is available as an R package (**Paper II**). The method is based on partitioning the variation into two components; sampling noise and BTM (Biological, Technical and cell Misclassifications) variation, where the BTM variation corresponds to all variation not related to sampling. The method is designed to be used with UMI-based data – for non-UMI-based data such as Smart-seq2, the assumptions regarding the sampling process are incorrect, which will lead to an overestimation of the BTM variation.

The sampling noise represents the variation that originates from data sparsity and approaches zero as the number of counts approaches infinity. For most genes (except highly expressed genes) in most datasets, the sampling noise is the dominating component. A problem with the sampling noise is that it varies across datasets and sometimes between different cell populations within a dataset, since it is dependent on the number of counts per cell (and its distribution across cells). To estimate the sampling noise per gene expression in each cell population, we generated *in silico* datasets by sampling cells with the same number of counts from the mean gene expression of the cell population. These generated datasets are called sampling noise only (SNO) datasets and always have less or equal variation compared to the real datasets (Fig. 12A). Since the sampling noise is different for the cell populations, it is difficult to compare the BTM variation between populations.

To make the BTM variation comparable across datasets we developed a down-sampling based method called *cell population alignment*, in which a template distribution of counts per cell is generated from a template dataset. To compare the BTM variation of cell populations, all such populations are first aligned to the template, which gives them virtually the same sampling noise (Fig. 12B). Any difference in variation between aligned populations must therefore be explained by differences in BTM variation. To generate a variation metric that is reasonably stable across gene expression ranges, the BTM variation is expressed as the total variation subtracted by the sampling noise, which shows a large variation across cell populations (Fig. 12C). To convert the BTM variation into a single number, we defined the BTM score as the average BTM variation across the gene expression range.

To single out the factors causing the BTM variation, we calculated the BTM variation score for in total 68 cell populations from 5 datasets, 5 cell types and 9 tissues. The results were used as input in a relative importance analysis where the contribution to the total variance from each factor was estimated (Fig. 12D). The estimated difference in variation between cell types was small. The four variables that explained most of the variation were all associated with the BC dataset. Although the variation could in theory be associated with tissue, the most likely explanation is that the BC dataset has a much BTM higher variation than the other datasets, since all samples from the tissues with high variation come from that dataset. When BC dataset samples were removed from the analysis the PBMC68k dataset became the most important factor, which strengthens this theory (**Paper II**). It is likely that technical effects such as strong batch effects are present in the BC dataset, which could cause such an increase in BTM variation.

Details of the method can be found in **Paper II**.

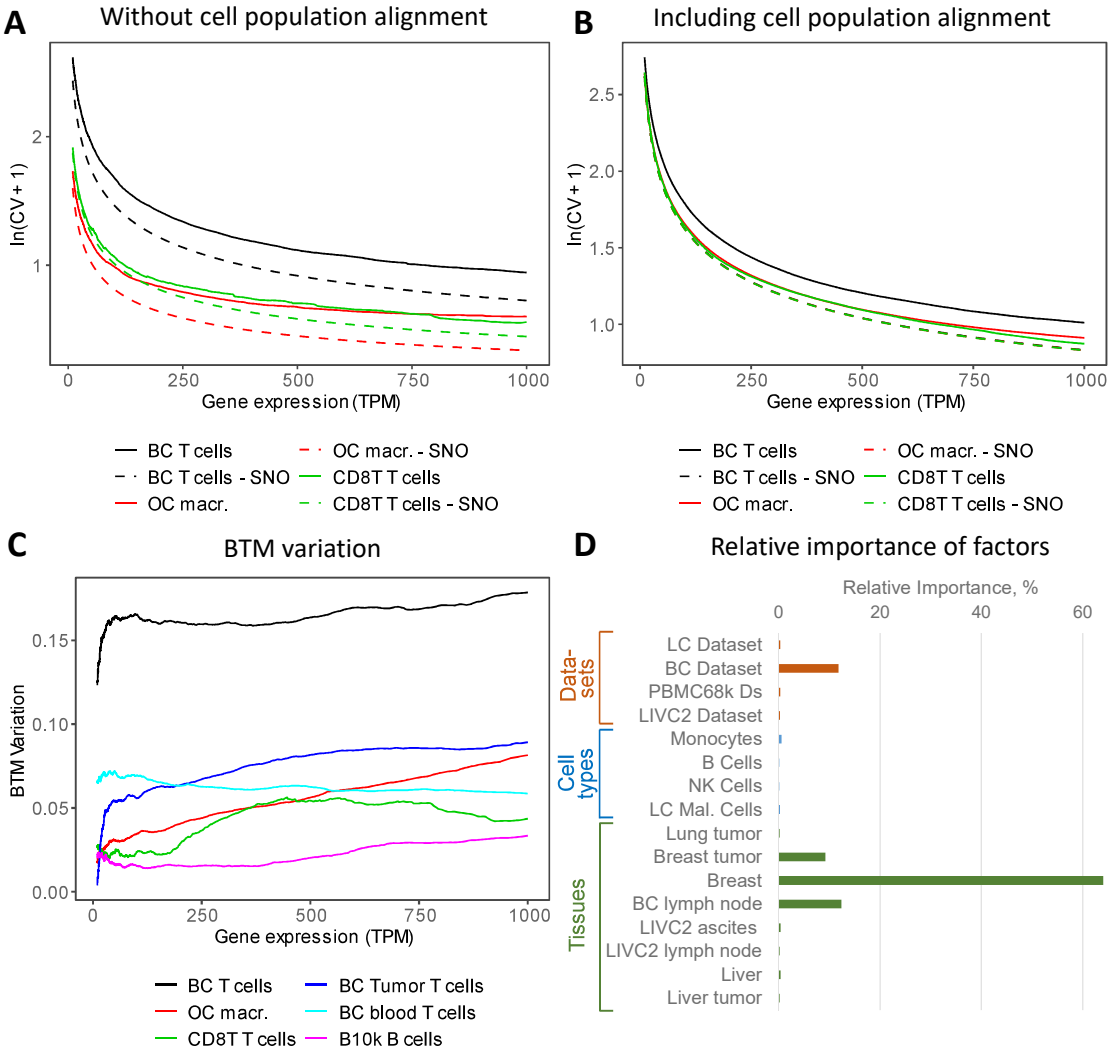

Fig. 12: Cell-to-cell variation in single-cell data. *A. Comparison of the cell-to-cell variation (within genes) between different cell populations and their SNO counterparts. No cell population alignment was performed. B. Identical to A, but with cell population alignment. All SNO curves are now virtually identical. C. BTM variation per gene expression. D. Relative important analysis of how different factors affect the BTM score. One variable type per factor (dataset/cell type/tissue) is used as intercept in the calculation and is therefore not included in the figure.*

2.4. The relationship between pool size and variation

To generate context-specific GEMs from single-cell data, multiple cells must be pooled due to the sparsity of the data. To initially examine the order of magnitude of number of cells required, we developed the total variation metric (R_{mean}) as part of the DSAVE method. The purpose of the method is to estimate the number of cells needed in a cell pool to obtain the same expected variation in gene expression as between typical bulk samples. In this method, pairs of non-overlapping subpopulations of cells are selected from a population to investigate, and pooled into two RNA-Seq profiles, followed by a comparison of the two profiles. The process is run for different pool sizes and repeated many times for each size to reduce the influence of randomness. Similarly, the metric is calculated between pairs of reference bulk samples (T cells). The difference is large across datasets, and the overall result suggests that the pool size needed to obtain the same variation as in bulk is on the order of thousands of cells (Fig. 13A). For the dataset generated by Smart-seq2 (LIVC), which can be expected to be less sparse, many cells are still needed, although fewer than for most droplet-based datasets. For highly expressed genes, much fewer cells are needed, typically on the order of 100 cells (Fig. 13B).

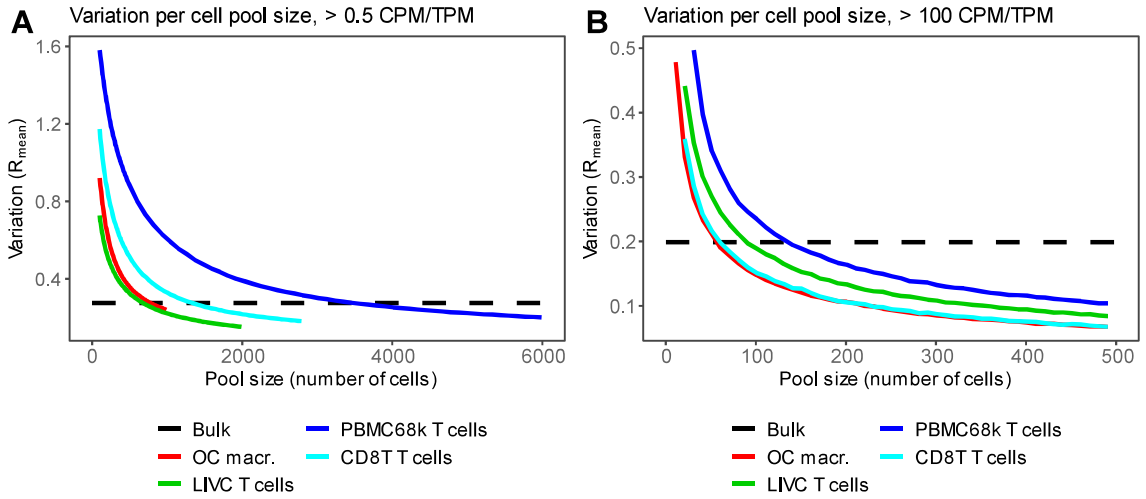

Fig. 13: Variation per pool size in RNA-Seq profiles assembled from pools of single-cells. *A. Variation for genes > 0.5 CPM/TPM, which covers most expressed genes. B. Variation for highly expressed genes, > 100 CPM/TPM. The LIVC dataset was generated using Smart-Seq2, while the other datasets were generated using droplet-based technologies. The bulk variation presented is the average across all pairs from 8 samples.*

2.5. Mathematical deconvolution for estimating cell type proportions in bulk data

The number of publicly available single-cell RNA-Seq datasets is today growing at an unprecedented rate, but the number of patients involved is still small compared to that available for bulk RNA-Seq. For example, the cancer genome atlas (TCGA) contains RNA-Seq profiles from more than 10,000 patients. It would therefore be of interest to utilize such bulk data for generating context-specific models of individual cell types if possible. CIBERSORTx [108] supports a method called digital cytometry, in which the gene expression profiles of individual cell types can be extracted from bulk data. Part of the method uses cell-type-specific RNA-Seq profiles derived from other datasets, either single-cell RNA-Seq or FACS-sorted bulk data, to determine the abundances of different cell types in bulk data. To investigate the potential for this technology for generating context-specific GEMs, the performance of this part of the algorithm, often termed *mathematical deconvolution* (although the solution in CIBERSORTx is technically based on support vector machines), was evaluated.

To evaluate the performance of CIBERSORTx on the B and T cell data previously used for evaluation of normalization methods, the data was divided into different sets of cell type profiles and in-silico mixtures of B and T cells, 50% each. The errors in estimation of cell type abundances were generally substantial, but much worse when single-cell and bulk data were combined (Fig. 14). Although batch correction helps, the performance is still at best modest. CIBERSORT (the predecessor of CIBERSORTx) performed well compared to other methods in a benchmark study [110], suggesting that the problem as such is difficult. In the benchmark study the performance was better. However, the study authors generated profiles and mixtures from the same dataset, thereby avoiding much of the unwanted variation. In a real use case, the profiles are in most cases generated from other datasets, and in most evaluations this complication is simply ignored. The performance observed in this experiment, which is a simplification in that there are only two cell types involved, was not deemed sufficient for extracting cell-type profiles for generation of cell-type-specific GEMs from bulk data.

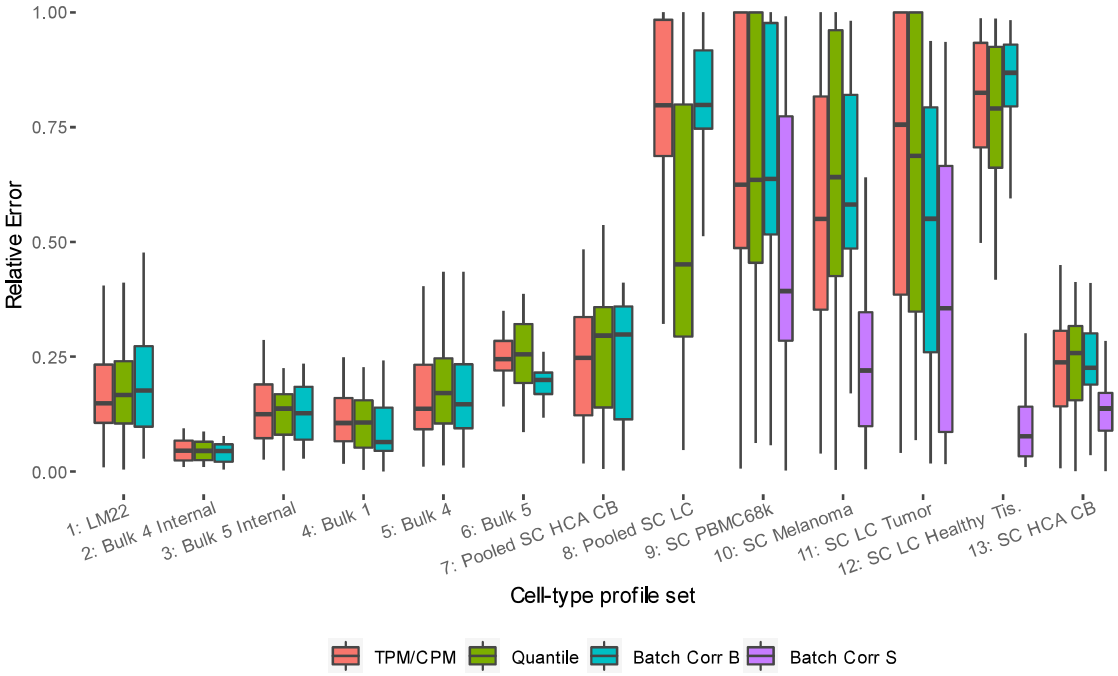

Fig. 14: Deconvolution performance. *Deconvolution of cell type abundances of 40 in-silico mixtures containing 50% B cells and 50% T cells generated from bulk RNA-Seq profiles. The relative error presented is the deviation from ground truth, i.e., 50% B cells (and the other half T cells) in the mixture. 13 different sets of cell type profiles were used: 1: Cell type profiles from LM22 from CIBERSORT were used (the fractions of all B and T cell subtypes were summed), mixtures were generated from bulk samples. 2-3: Bulk samples from the same lab were split in two groups used for cell type profiles and mixtures, respectively. 4-6: Cell type profiles were generated from bulk data from one lab and mixtures were created from bulk data from other labs. 7-13: Cell type profiles were generated from single-cell datasets and mixtures were generated from bulk data. CIBERSORTx supports two different normalization methods, TPM (or CPM) and quantile normalization, and two different batch correction methods, B and S, both used with TPM (or CPM) data.*

2.6. Summary

In this chapter, I have shown that the technical variation in bulk RNA-Seq is high, where much of the variation can be attributed to differences between the labs in which the datasets were produced. Reproducibility has previously been reported to be high, for example in the Geuvadis project [111], while others have reported the need for batch effect correction [64], [112]. A difference is that in the Geuvadis project, the labs were given strict instructions about the laboratory procedures to maximize the similarity, which is not the typical case when using data from different sources. For gene expression profiles generated from single cells, we have shown that the technical variation is even greater, including severe effects from data sparsity. Typically, more than a thousand cells are required to obtain a stable profile, and my initial ambition to generate GEMs for single cells was simply not possible with the data at hand. While I have not explicitly examined batch effects in single-cell data, such effects are also reported to be substantial, and need to be taken into consideration [67]. While at this point it is not clear which level of expected variation is acceptable in an RNA-Seq profile used for generating context-specific models, I have identified a dilemma: there are usually not enough cells in a single batch (here referring to 10X Chromium) to supply thousands of cells for many cell types. In general, batch effects will therefore in most cases be present in the data when we seek to generate context-specific GEMs from single-cell data.

The high technical variation in RNA-Seq data, potentially in combination with a high biological variation, makes mathematical methods such as digital cytometry [108] very challenging. I cannot at this point see that they are useful for generating context-specific models by extracting gene expression profiles for cell types from bulk mixes, the uncertainties are simply too high. Neither have I found any examples in the literature where digital cytometry is used for this purpose.

3. Detection of misclassified cells in single-cell RNA-Seq data

The previous chapter concerned the variation in RNA-Seq profiles from both bulk and single-cell data. For single-cell data, the assembly of single cells into clusters based on similarity is an important part of the analysis and has a profound effect on the RNA-Seq profiles. In this chapter, I have investigated the presence of misclassified cells in clusters and provide a method for detecting such cells.

3.1. Method and method evaluation

Clustering is a challenging task due to data sparsity [102] and it is especially challenging to separate cell subtypes, for example subtypes of B and T cells. To investigate to what extent misclassified cells are present in clusters, we as part of DSAVE developed a method to detect misclassified cells (**Paper II**). For each cell in a population, the method calculates the probability to acquire the observed gene expression counts by sampling from the mean gene expression of the cell population. The probability is transformed into a cell divergence metric, which increases the more a cell diverges from the population. In addition, the method identifies which genes that exhibit the most diverging gene expression for each cell, which is useful information when trying to understand the reason why cells diverge from the mean gene expression of the population.

To simplify the examination of single cells, we as part of DSAVE developed an interactive tool that supports investigation of individual cells (Fig. 15A). The tool can show the most divergent genes for each cell, often giving a hint about whether the cell is of a different cell type or potentially a doublet. For example, the dendritic cell population in Fig. 15A, as identified by the authors of the publication where the dataset was published [73], contains both natural killer (NK) cells and megakaryocytes.

To further investigate the presence of misclassified cells in clusters, we applied DSAVE to a lung cancer dataset with associated cell type classifications [113]. First, we investigated the cells classified as T cells (Fig. 15B). We identified three classes of cells that were divergent in some aspect. Some cells had high expression of hemoglobin-related genes, suggesting red blood cell precursors (nucleated red blood cells, NRBC). Other cells had high expression of immunoglobins, suggesting plasma cells. T cells with a high expression of lactate dehydrogenase A are an example of cells with the right cell type classification, but with some diverging genes. Such T cells have likely entered a program where more ATP is needed, such as T cell activation [114]. We then investigated clustering into cell subtypes, where the clusters are more similar. Specifically, we investigated a population of follicular B cells, and found misclassified cells or doublets showing the phenotypes of cytotoxic T cells, NRBCs, and plasma cells (Fig. 15C). As expected, we found more misclassified cells for this subpopulation compared to the T cells.

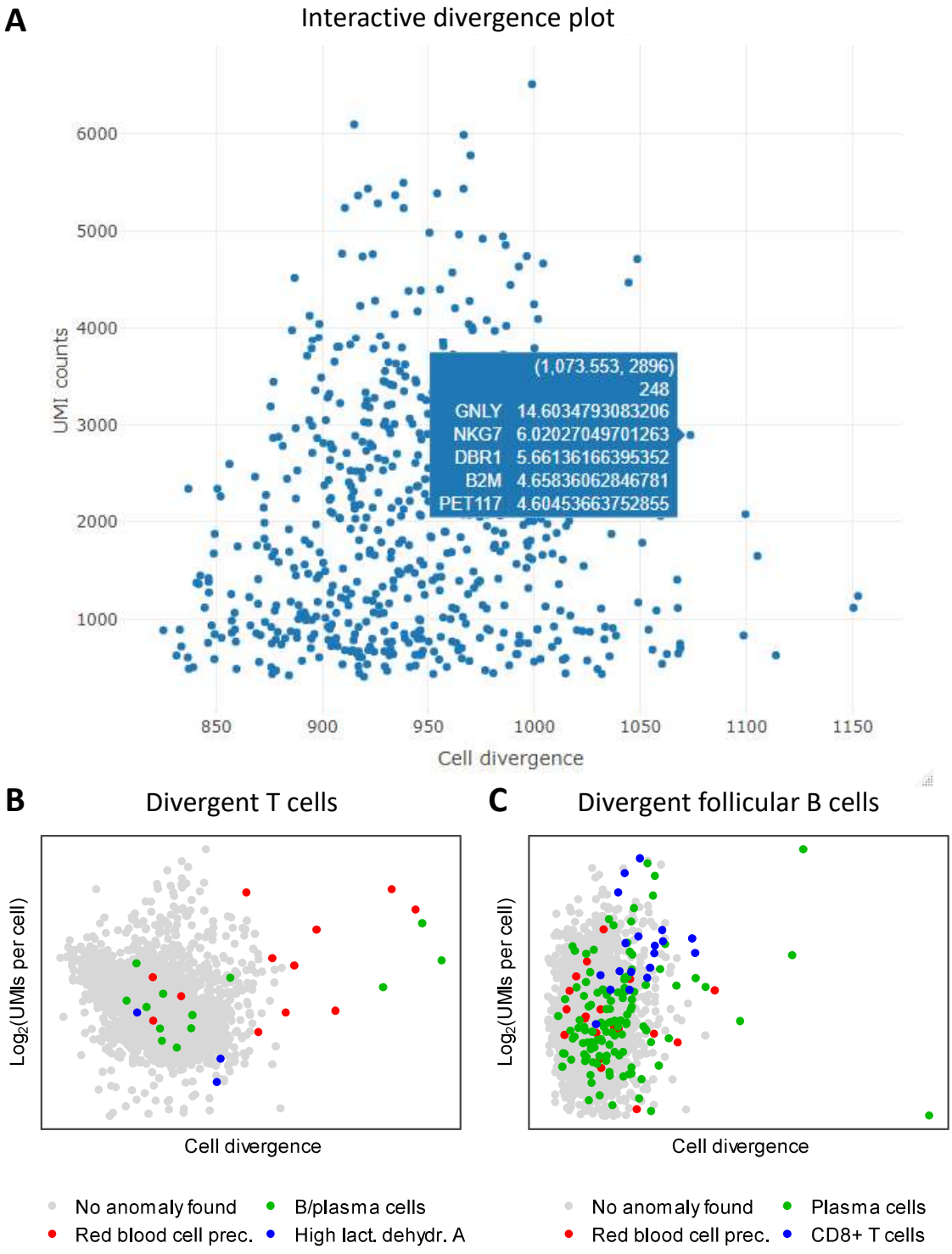

Fig. 15: Detection of misclassified cells. *A. Screen shot from the interactive divergence plot available in DSAVE, here investigating a population of dendritic cells from PBMC (The PBMC68k dataset, see paper II). Divergence details are shown when hovering over the cells with the mouse, where the 5 most divergent genes are shown. Here, a potential NK cell is found in the population. B. Divergent cells in a T cell population from lung cancer. C. Divergent cells in a population of follicular B cells from lung cancer.*

3.2. Summary

It is evident that clustering results in misclassified cells. DSAVE provides a semi-automatic method for finding such cells, and DSAVE finds misclassified cells not found by other tools (**Paper II**). An alternative approach to handle part of this issue is to computationally identify and remove doublets [94]. Given misclassified cells, an intriguing question that remains is to what extent such cells affect the generation of context-specific GEMs for cell types.

4. Improved quantification of UMI-based RNA-Seq data

In the previous two chapters, I have addressed the problem of forming reliable RNA-Seq profiles from clusters of single-cell data, based on a count matrix from single-cell data. However, such work is entirely dependent on the quality of the count matrix, and I have taken a particular interest in biases introduced by PCR. While much attention has been given to reducing such bias across cells in single-cell RNA-Seq where UMI collapsing is adequate, bias across genes, which is more important for generation of pooled RNA-Seq profiles, have been given less attention. In **paper I**, we discovered that measuring the number of counts per UMI affects the quantification and can be used to regress out technical differences between bulk and single-cell data. We therefore in **paper III** set out to investigate this problem in more detail and developed a correction method called BUTTERFLY.

4.1. Discovery of the problem

In **paper I**, we investigated a dataset processed by a customized single-cell pipeline (based on STAR) that produced both UMI-collapsed count matrices and raw counts matrices [115]. Specifically, we investigated the differences explained by different covariates between 10X Chromium (v2) data from mouse cortex and matching bulk samples. Since we had access to both raw counts and UMI counts, we could calculate the average number of copies per UMI for each gene, which explained much of the difference between 10X Chromium and bulk data (Fig. 16A). Likewise, we investigated the effect of gene length (Fig. 16B) and GC content (Fig. 16C). GC content has previously been used to regress out technical PCR bias across genes in RNA-Seq [81], [116], but interestingly, the number of copies per UMI seemed to explain more differences between single-cell and bulk, although the combination of both is helpful (Fig. 16D). This discovery called for an in-depth investigation of the problem.

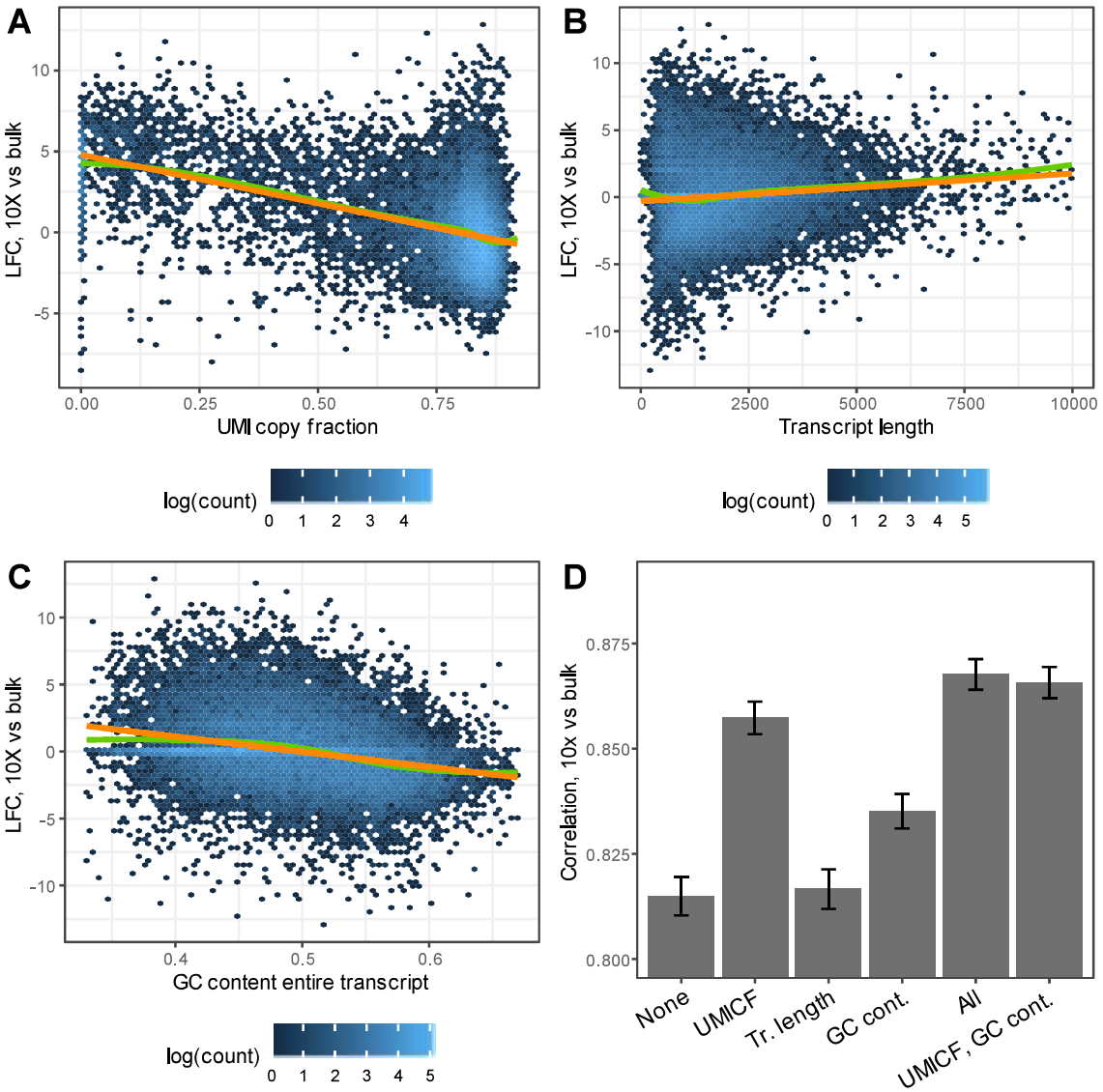

Fig. 16: Technical differences in RNA-Seq data between pooled cells from 10X chromium and bulk. *UMI copy fraction (UMICF) for a gene is defined as (total counts − UMI counts)/total counts. A-C. Log2 fold change between pooled cells from 10X Chromium data and bulk across genes plotted against different covariates, fitted to linear (orange) and LOESS (green) curves. D. Improvement in correlation between 10X Chromium and bulk after regressing out different combinations of the covariates from A-C. Specifically, the difference in LFC between single-cell and bulk was regressed out in the single-cell data. The labels on the x axis describe which covariates were regressed out.*

4.2. Problem definition

As mentioned in the background section, UMI-based single-cell data is commonly quantified using UMI collapsing, where all sequencing reads that belong to the same UMI, cell, and sometimes gene are treated as the same molecule. We discovered an effect that we have termed “the pooled amplification paradox” that can lead to amplification biases across genes. The effect introduces biases in cases when there is a systematic difference in amplification across genes and the library is incompletely sequenced, where the latter is practically always the case (Fig. 17). In **paper III**, we hypothesized that a method often used in ecology, called prediction of unseen species [117], [118], could be used to correct the bias. We implemented an unseen molecules correction method based on fitting a zero-truncated negative binomial distribution to the distribution of copies per UMI within each gene.

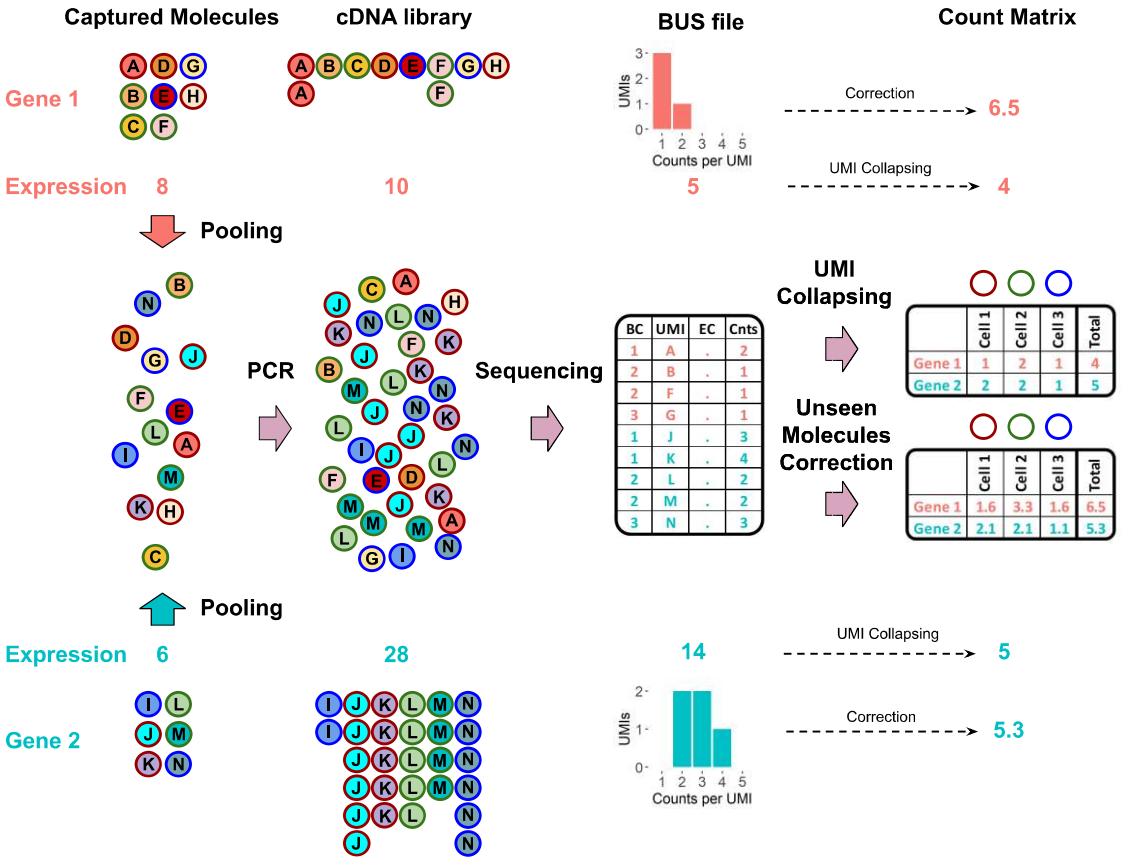

Fig. 17: The pooled amplification paradox. *Description of a seemingly paradoxical reversal in gene expression that arises from differences in amplification across genes. In this example, more molecules have been captured for gene 1 compared to gene 2, but the molecules in gene 2 are more highly amplified. Since the sequencing process only captures a subset of all molecules available after PCR, some original molecules (having distinct UMIs and cell barcodes) will not be sequenced at all. The number of unseen molecules differs across genes, where highly amplified genes have fewer unseen molecules, introducing a bias yielding a reversal in gene expression despite the bias reduction attained by UMI collapsing. The bias can be partly remedied using unseen molecules correction.*

To investigate the effect of the pooled amplification in sequencing data, we began by evaluating the differences in amplification per gene. For this purpose, we defined two metrics: the fraction of single-copy molecules for the gene across all cells in a dataset (FSCM), representing the fraction of the molecules found that only have a single read, and the average number of copies (reads) per UMI for the gene across all cells in the dataset (CU). FSCM varied substantially across genes within a dataset, and highly expressed genes in general tended to be more amplified (Fig. 18A). We measured the FSCM metric for a collection of datasets, and found that datasets generated by the same technology and from the same tissue had similar relative amplification across genes (Fig 18B). However, differing sequencing depth between datasets can introduce a skew in that relationship, since more reads per UMI on average in a dataset leads to lower FSCM values for all genes. Datasets generated from the same tissue but using different technologies exhibited substantially larger differences in amplification per gene (Fig. 18C).

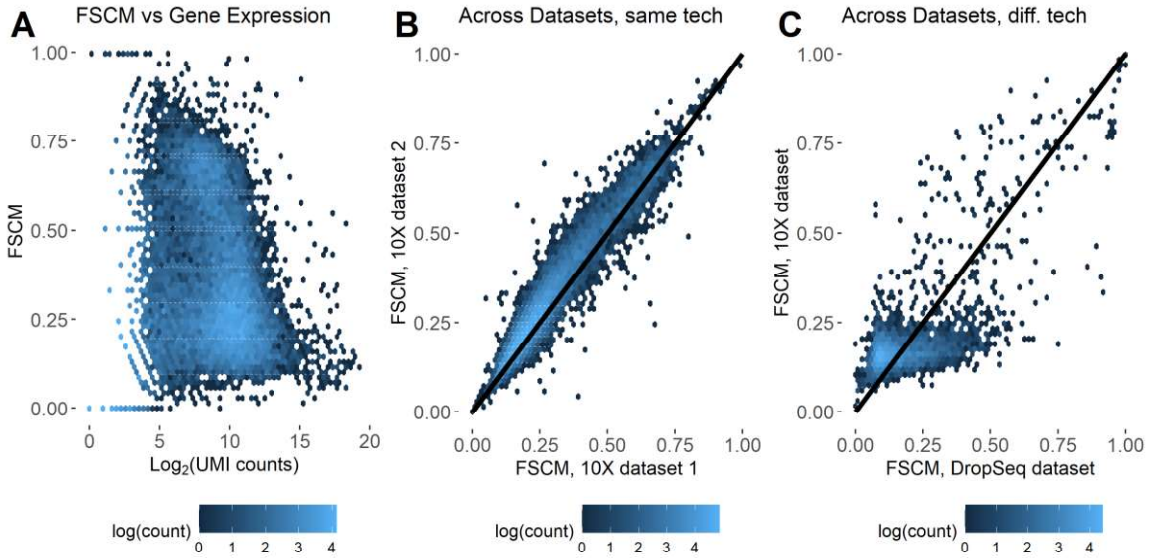

Fig. 18: Variation in amplification across genes and datasets. *A. FSCM vs gene expression for all genes in the PBMC_V3_3 dataset (see* ***paper III****). B. Differences in FSCM per gene across datasets generated using the same technology (10X Chromium v3, PBMC_V3_2 vs PBMC_V3_3). C. Differences in FSCM per gene across datasets generated with different technologies (10X Chromium v3 vs DropSeq, EVALPBMC vs EVALPBMC_DS).*

4.3. Description and evaluation of the correction algorithm

To predict the number of molecules that would be detected given deeper sequencing, we used the zero-truncated negative binomial (ZTNB) method previously implemented in the R package PreSeqR [119]–[121]. The method is based on fitting a ZTNB distribution to the CU histogram of each gene. It is not possible to measure the number of molecules with zero copies (which is ultimately what we seek), hence the zero truncation of the negative binomial distribution. While it may seem natural to predict the gene expression at an infinite number of reads (and thereby include all predicted molecules with zero copies), such a strategy turned out to introduce much technical noise. The CU histograms of some genes do not perfectly follow a negative binomial distribution and the sampling noise is high for lowly expressed genes, leading to errors in the parameter estimations. A conservative approach is therefore to predict the gene expression at a higher, but not infinite, sequencing depth. First, the ZTNB is fitted to the CU histogram for each gene, estimating the negative binomial parameters μ (mean) and *size* (reflecting the dispersion). To predict number of molecules with non-zero copies given x times as many reads, the mean of the distribution is simply multiplied with x, since the mean number of copies per UMI (including molecules with zero copies) is proportional to the total number of reads (Figs. 19 A, B, D, E). Both genes in the figure get more molecules in the prediction, but for gene 2 a substantially higher relative number of new molecules are detected. When downsampling the predicted histogram to the original number of reads, the negative binomial parameters are roughly preserved, motivating that the size parameter is reasonably independent of the sequencing depth and can be kept constant during prediction (Figs. 19 C, F).

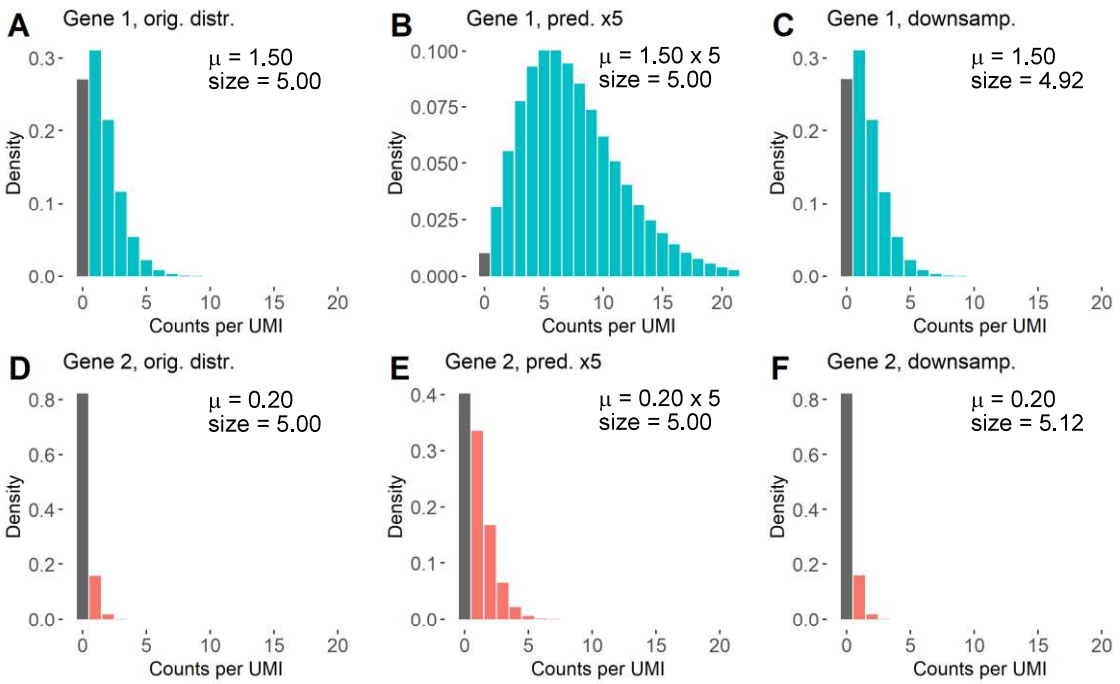

Fig. 19: Illustration of ZTNB prediction. *The ZTNB prediction of unseen molecules assumes that CU diagrams reasonably well follow a negative binomial distribution. The figure shows in-silico generated data sampled from a negative binomial distribution for two imaginary genes with different amplification (10^6 molecules each). A. The original CU histogram for gene 1. B. The CU histogram when predicting to 5 times the number of reads. C. The histogram in B downsampled to reflect the number of reads in A, including the negative binomial parameters fitted to the distribution. D-F: Same as A-C, but the data is simulated with a smaller mean parameter, representing a lower-expressed gene (gene 2).*

To investigate the effect of the pooled amplification paradox on individual genes we conducted a downsampling experiment on a mouse cortex dataset (EVAL, see **paper III**). We compared the gene expression of two genes with different amplification (Fig. 20A) and found that the gene expression gradually changed in favor of the highly amplified gene as the data was downsampled (Fig. 20B). Applying the unseen molecules correction, where the downsampled data is predicted to the same number of reads as the full dataset, largely removes the bias (Fig 20C).

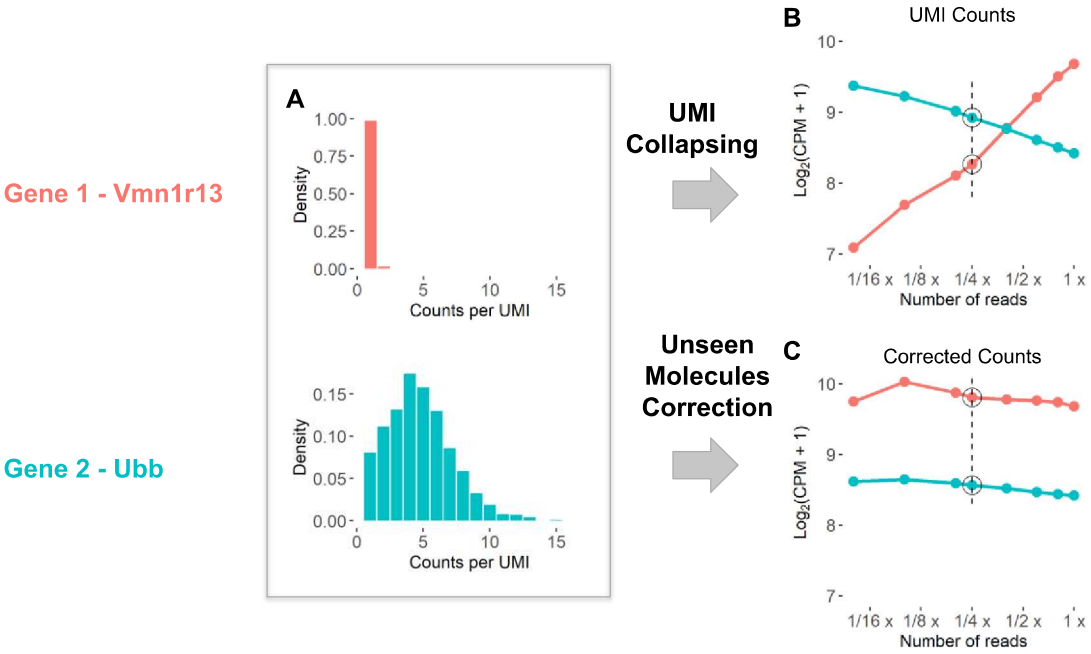

Fig. 20: Example of the pooled amplification paradox in real data. A. *CU histograms for two differently amplified genes in a mouse dataset generated using 10X Chromium, v2 chemistry (the EVAL dataset, see* ***Paper III****). The histograms were generated using data downsampled to a quarter of the original reads, as visualized by the dashed line in B and C. B. The normalized gene expression generated using UMI collapsing changes with sequencing depth (simulated using downsampling) for the two genes – Vmn1r13 has more unseen molecules, which worsens with lower sequencing depth. C. Application of the unseen molecules correction, where the number of molecules at each point is predicted up to the full reads without downsampling.*

To evaluate the performance of the prediction, we conducted downsampling experiments on real data, where we evaluated the performance of several prediction methods, including ZTNB and the PreSeq DS method [119]–[121]. In general, the methods performed similarly (**Paper III**), and we choose to proceed with the ZTNB method. For the PBMC_V3_3 dataset (see **paper III**), the method for example increased the concordance correlation between downsampled and original data (Fig. 21A). The improvement is largest for the highly expressed genes, likely since the CU histograms for those genes are based on more data and hence are more stable. Interestingly, when comparing datasets produced from the same biological sample but with different technologies, the correlation increases with correction (Fig. 21B). These results suggest that differences in amplification across genes play a part in the batch effects often experienced when combining technologies.

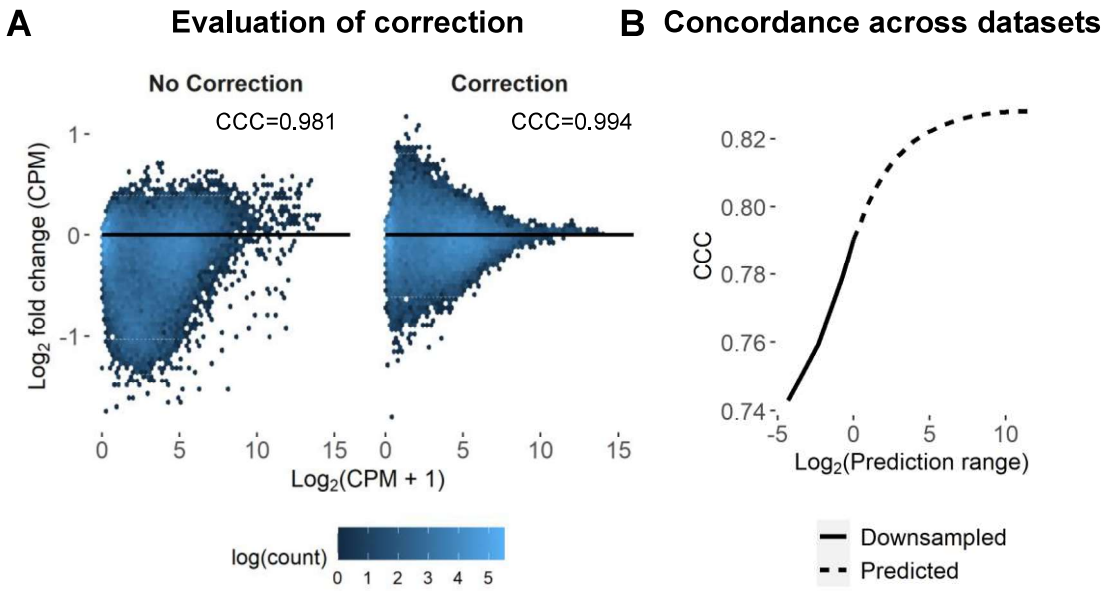

Fig. 21: Performance of the unseen molecules correction. *A. The difference (log fold change) in gene expression between original and downsampled data (to 10% of the UMIs) for the PBMC_V3_3 dataset (see* ***paper III****). Two cases are displayed: the downsampled data is either uncorrected or corrected. B. Correlation (Lin's concordance correlation) between a DropSeq and a 10X Chromium dataset generated from the same sample, as a function of prediction range. For negative prediction ranges, the data has been downsampled and not predicted.*

The correction method, called BUTTERFLY, has been implemented as part of the kallisto/bustools [80], [83] workflow. The method has not been tested together with other data processing workflows such as Cellranger or Salmon [81] since the copies per UMI information is not available as output from these pipelines.

4.4. Batch effects from different sequencing depth

Due to the pooled amplification paradox, there will be batch effects between datasets with different sequencing depth. To investigate this effect in real datasets, we mixed cells from two datasets generated from the same biological sample and processed the mixed data using Seurat. The uncorrected data exhibited clear batch effects (Fig. 22 A). To correct for the differences, we did not use prediction, but a method we termed *binomial downsampling*, which in short resembles downsampling but calculates the expected outcome after downsampling of reads. This choice is motivated by that downsampling is more stable than prediction, and that it is therefore better to downsample the more deeply sequenced dataset than to predict the lower sequenced dataset. The effort increased the correlation (CCC increased from 0.991 to 0.994) and gave a better visual overlap between cells (Fig 22 B).

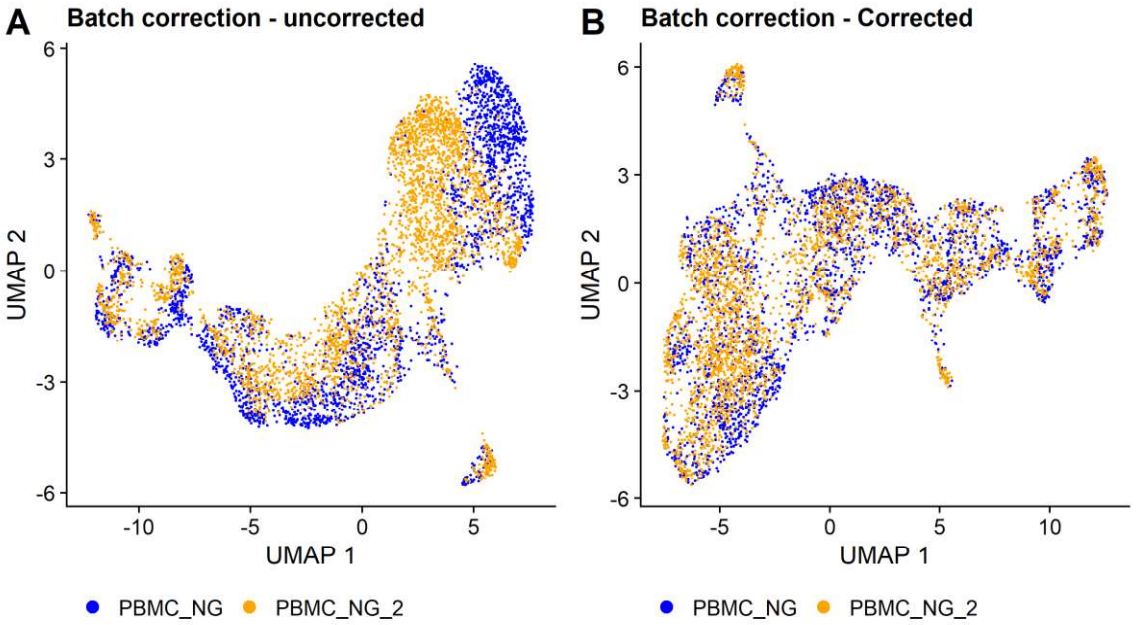

Fig. 22: Batch correction of datasets with different sequencing depth. *A. Uncorrected data. B Data batch-corrected using binomial downsampling. The datasets are generated from the same biological sample. The figures show a subselection of the clusters in the data.*

4.5. Differences in amplification across clusters

We also investigated the difference in amplification across clusters in the single-cell data. Interestingly, the amplification of some genes varied across cell types (Fig. 23 A-B), which we expect will introduce a bias when comparing gene expression across clusters. The CU histograms of genes clearly do not resemble a negative binomial distribution. We therefore applied the PreSeq DS method for prediction for these genes, showcasing that the unseen molecules correction can assist in reducing the bias between such genes (Fig. 23 C-F).

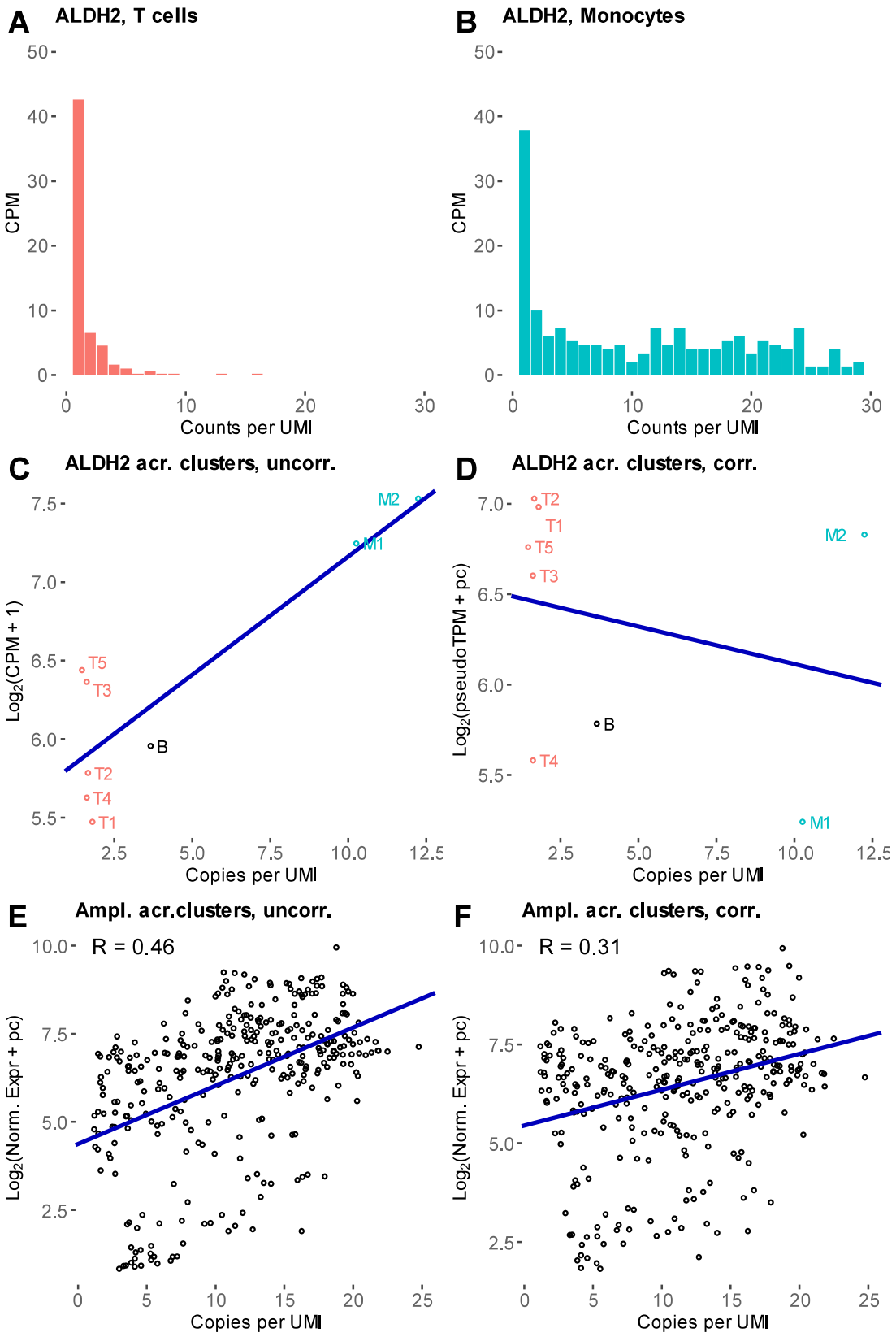

Fig. 23: Gene amplification differs across clusters. *A-B:CU histograms of the same gene for two different cell types. C-D: Gene expression of 8 different clusters from B cells, T cells and monocytes, before and after unseen molecules correction. E-F: Gene expression of the 50 genes with most variation in CU across clusters, plotted versus CU. Each dot represents a combination of cluster and gene. E shows uncorrected gene expression, F corrected.*

4.6. Summary

In this part of the thesis, we have shown that the gene expression in single-cell data is dependent on the distribution of copies per UMI and sequencing depth, described by the pooled amplification paradox. We also showed that the distribution of copies per UMI can be used to predict the gene expression at greater sequencing depth, and thereby be used for correction of gene abundance estimations. However, we have not provided clear evidence that the correction increases the similarity between the measured and true gene expression, although the result in Fig. 21 B points in that direction. We have recently discovered that many reads with low CU values are falsely aligned, and origin from non-exon parts of the genome. The problem arises because pseudo-alignment usually only maps reads to the transcriptome, and some sequences in exons are shared with introns from other genes or other parts of the genome. Such non-transcriptome reads are poorly but sometimes falsely aligned, and most copies are therefore discarded, yielding a low CU value for such molecules. This problem is known from literature, and is corrected in for example Salmon using decoys [122], and a similar correction is underway in kallisto. While BUTTERFLY correctly increases the abundances of such transcripts, this behavior is not desired. These molecules are false positives that should be filtered out, not scaled up, and it is very possible that this effect can explain the differences in amplification across clusters. However, the differences in copies per UMI across genes is not purely an artefact arising from false positives in pseudo-alignment, since the same effect was observed in **paper I**, where alignment was done by STAR. It is also important to realize that regardless of the source of the observed differences in CU across genes, they will induce batch effects depending on sequencing depth. To conclude, it is recommended to remove these false reads before running BUTTERFLY. With the removal of such reads, BUTTERFLY will lead to an improved gene expression matrix.

5. Generation of context-specific models from scRNA-Seq

In **paper IV**, the learnings from **paper I** and **II** were put into use for the generation of context-specific GEMs from single-cell RNA-Seq data. While the ultimate goal would be to generate one such model per cell, there is simply not enough data per cell to generate such models with reasonable quality. To understand why this is the case, let us first look at the threshold level used in tINIT, which is the method used in this thesis for generation of context-specific GEMs. In tINIT a reaction is considered to be "on" if the gene expression is above 1 TPM (or CPM), which corresponds to one molecule out of a million. The number of mRNAs in a cell has been estimated to somewhere between 50,000 to 300,000 [68], but in practice, the captured number of molecules by the scRNA-Seq technologies is much lower, for example typically 1,000-10,000 per cell in 10X Chromium data. It is simply not possible, even if all molecules of a cell are captured, to get a stable gene expression for genes close to the tINIT threshold. The expression value of these genes in a cell will largely be determined by randomness, since the gene is only expected to be detected in a small percentage of the cells in a population even if it is expressed in all, due to the sparsity in the data. We are therefore left with little choice but to combine the gene expression over multiple cells to generate context-specific models for a certain phenotype.

5.1. Method and method evaluation

To investigate the metabolism in individual cell types, we formed populations of single cells that we used as input for generation of context-specific GEMs (Fig. 24A). The cell populations in the single-cell data can be identified by using tools such as Seurat to process and cluster the data. Here, we used public datasets with associated cell type classifications, which simplified the analysis. To estimate the uncertainty in the generated models, we generated 100 bootstraps (subsamples of the data of the same size as the original population, sampled with replacement) per cell type, and the UMIs of the cells in each such bootstrap were pooled (summed) to form a gene expression profile for each bootstrap in each cell type. We then applied tINIT on the gene expression profiles from the bootstraps, using Human1 as template model. The models were then both compared directly (structural comparison) and used in network analyses where the ability of the cell to perform certain tasks was determined. A task was here defined as the ability to produce a specific set of products given a set of substrates. Statistics were then applied across the bootstrap models to determine for a cell type if a task could be performed, if it was absent, or if the results were uncertain.

Since the strategy using bootstraps generates thousands of models, the execution time of tINIT became a serious issue. tINIT typically takes between 20 minutes and 3 hours to run, which for thousands of models results in a substantial computational effort. To reduce the execution time, we developed a new version of tINIT, called ftINIT, that runs substantially faster, but also uses a slightly different algorithm. Substantial changes were applied to the tINIT algorithm, which is based on mixed-integer linear programming (MILP). In short, the number of integer variables included in the problem were substantially reduced, the optimization was split up into several substeps to reduce the complexity of the problem, and reactions without gene associations were treated differently. To ensure that the quality of the generated context-specific models was retained, we evaluated both algorithms in several ways, which showed similar performance. For example, we generated models from

RNA-Seq profiles of 15 cell lines from DepMap [123], [124] with both methods. We then compared the ability of the resulting models to predict essential genes, as compared to ground truth available as CRISPR screens (Fig. 24B). To evaluate the gain in computation time by ftINIT, we generated 10 models from the genotype-tissue expression project (GTEx) [125]. We measured the computation time required to generate each model on a standard laptop computer (Intel Core i7-6600U, 2.60 GHz, 2+2 cores), which showed a substantial reduction in computation time for ftINIT compared to the previous version of tINIT (Fig. 24C).

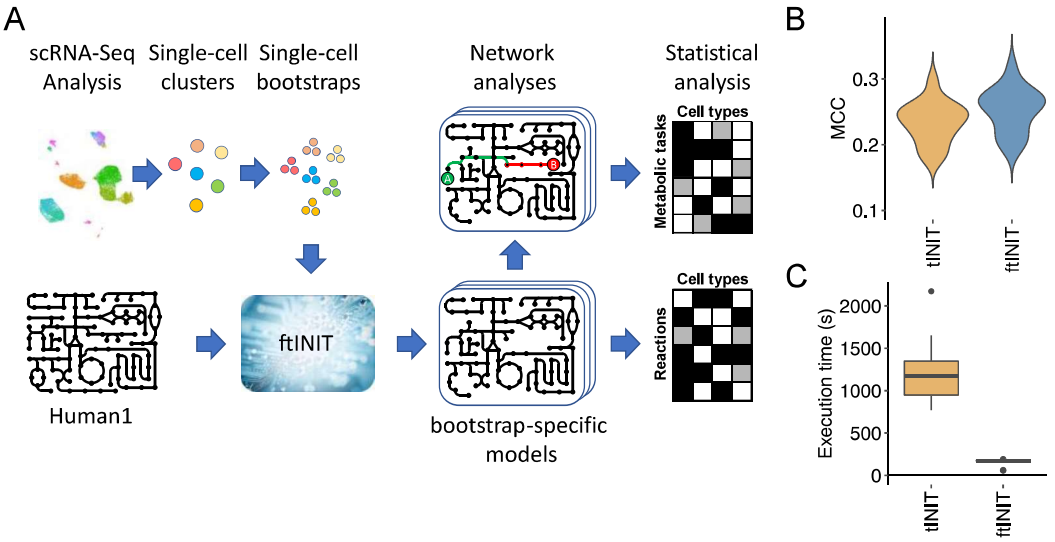

Fig. 24: Generation of context-specific GEMs from single-cell RNA-Seq data. *A. Overview of the method. Clusters of cells are formed from single-cell data and bootstraps are then generated from each cluster. The bootstrapped cell populations are then pooled to form RNA-Seq profiles, which are used together with a template model such as Human1 to form context-specific GEMs, one per bootstrap. The cell clusters can then either be compared structurally or be used further in for example network analyses that determines the ability to perform metabolic tasks. In both cases, the bootstraps are used to determine the uncertainty of the results. B. Gene essentiality analysis, where the new version of tINIT (ftINIT) is compared to the previous (tINIT). The ability to predict essential genes by models generated from RNA-Seq from 15 cell lines from DepMap is compared to ground truth values from CRISPR knockout screens from the same cell lines. The performance is measured using Matthews correlation coefficient (MCC). C. Evaluation of the execution time of the previous and new version of tINIT on 10 RNA-Seq profiles from GTEx.*

When trying to draw statistical inference from modeling results from bootstrap models, two challenges arise: 1) In for example differential expression analysis in DESeq2 [86], the uncertainty from data sparsity (i.e., total number of counts) can directly be estimated by modeling the gene expression using the negative binomial distribution. It is very challenging (probably not possible) to estimate the uncertainty in the output from ftINIT from the number of counts. Although bootstrapping helps, it does not fully address the problem. For example, let us consider a case where we pool 10 cells from 10X Chromium data. Most lowly expressed genes will by chance be zero in this data. Zeros will also be zero in all bootstraps, thereby giving a false certainty that the gene is not expressed. While differential expression analysis can handle such sparse samples by directly estimating the uncertainty from sparsity, our method cannot. It is therefore important to use cell pools large enough to reduce this problem to an acceptable level. 2) There is often substantial variation across batches and biological samples. The total variation in the data becomes underestimated if all cells are treated as they belong to the same cell population. It has recently been shown that single-cell differential expression analysis methods that do not take the sample origin of cells into account underestimates the uncertainty in the data and

thereby produce false positives [126]. However, in most datasets there is not enough data to be able to create large enough cell populations per cell type and sample. We are therefore in most cases not able to estimate this uncertainty.

Given the limitations mentioned above, our approach for handling the uncertainty in model generation is based on bootstrapping, where we ensure that the cell populations have sufficient size to avoid any large effects from challenge 1. In addition, we use low p value thresholds for significance to counter for the remaining effects for which we cannot properly estimate the uncertainty. We began by measuring the uncertainty in model generation by comparing pairs of models generated from random non-overlapping cell subpopulations of certain sizes (Fig 25A). Similar to our previous result from **paper II**, thousands of cells were needed to reach the same similarity as between bulk samples, and the number of cells needed varied across datasets. To estimate the required pool size this way is very computationally demanding, which calls for a more practical approach. Seemingly, the DSAVE total variation score developed in **paper II** yielded similar results (Fig. 25B). As a rule of thumb, we therefore recommend to pool at least the number of cells that yields the same DSAVE total variation score as between the bulk reference samples (see **paper II**).

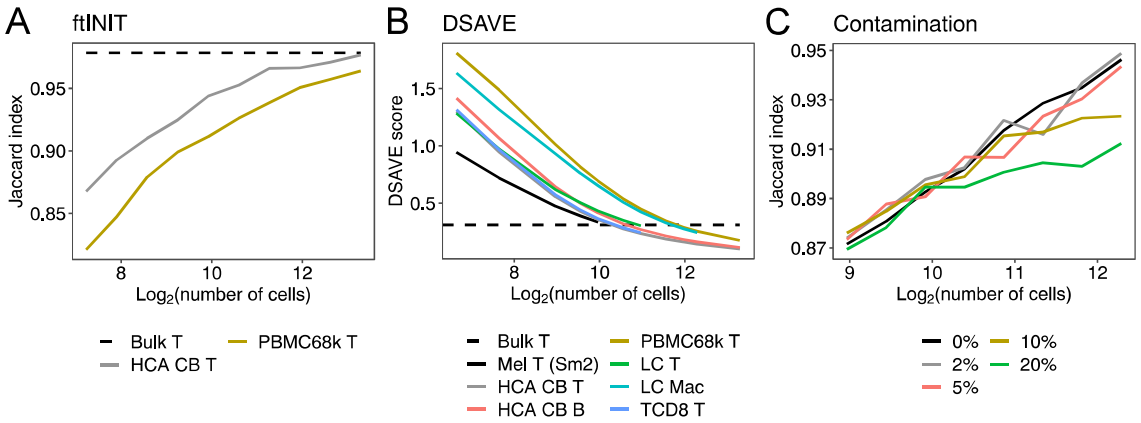

Fig. 25: Evaluation of required cell population size for generation of context-specific models from single-cell data. *A. Reproducibility of context-specific models generated using ftINIT, with cell populations of varying size. Pairs of non-overlapping cell populations of the same size were sampled from the same population of cells. Context-specific models were then generated from the pooled transcriptomic profile of each population in the pair, followed by a structural comparison. For each size, the procedure was repeated 30 times; the figure shows the average Jaccard index across the repetitions for each population size. B. DSAVE total variation score for a collection of datasets, including the datasets used in A. C. The effect of misclassified cells on model structure. T cells from the LC dataset (see **paper IV**) were contaminated with a varying fraction of cancer cells from the same dataset, followed by model generation. The contaminated models were then compared to pure models by comparing the reaction scores generated, where positive reaction scores are interpreted as reaction presence. The reaction scores calculations are part of tINIT (See **paper IV**).*

To investigate the impact of cell misclassification on model generation we made a similar comparison of variation between pairs of cell subpopulations of a certain size, but where one of the T cell populations used was contaminated with a certain fraction of cancer cells (Fig 25C). Surprisingly, the effect of contamination is rather small compared to other error sources, and it is first at levels of 10-20% of contamination that we can observe a decline in model similarity. Thus, while we showed in **paper II** that misclassified cells are common, and in that paper developed a method for detecting such cells, they only pose a concern if there is a large portion of misclassified cells. For clusters containing top level cell types (such as fibroblasts, T cells, etc.), this problem can therefore likely be ignored, while it may be of concern for clusters at the cell subtype level.

To compare models generated from bulk RNA-Seq and single-cell data, we generated 5 models for each of 53 tissues from GTEx bulk RNA-Seq data [125] and additional models from different single-cell datasets (Fig. 26A). The bulk models originating from the same tissue and technology clustered together, while the models generated from all single cells from the same tissue (L4 samples) showed less agreement. The models generated from individual cell types in single-cell data clustered by cell type similarity, and interestingly, immune cells clustered with GTEx blood, which has a high immune cell content. Likewise, the L4 spleen sample clustered with blood instead with spleen, which may suggest that a higher portion of the captured cells from spleen in single-cell data are immune cells. The different groups of cell types are well separated, suggesting that their metabolism is clearly different, which motivates our approach. Further structural comparisons confirm that technology introduces bias (Fig. 26B). Similar to our previous results from **paper I**, we conclude that TMM normalization reduces variation in bulk data (using pseudo-counts as described in the normalization section) (Fig. 26C). However, TMM normalization does not help between clusters, likely because the cells in these clusters have been processed together and thereby have similar biases. Although quantile normalization reduces the variation further, we concluded in **paper IV** that it also introduces new biases and that the models no longer group as well on tissue. We therefore do not recommend quantile normalization for generation of context-specific models.

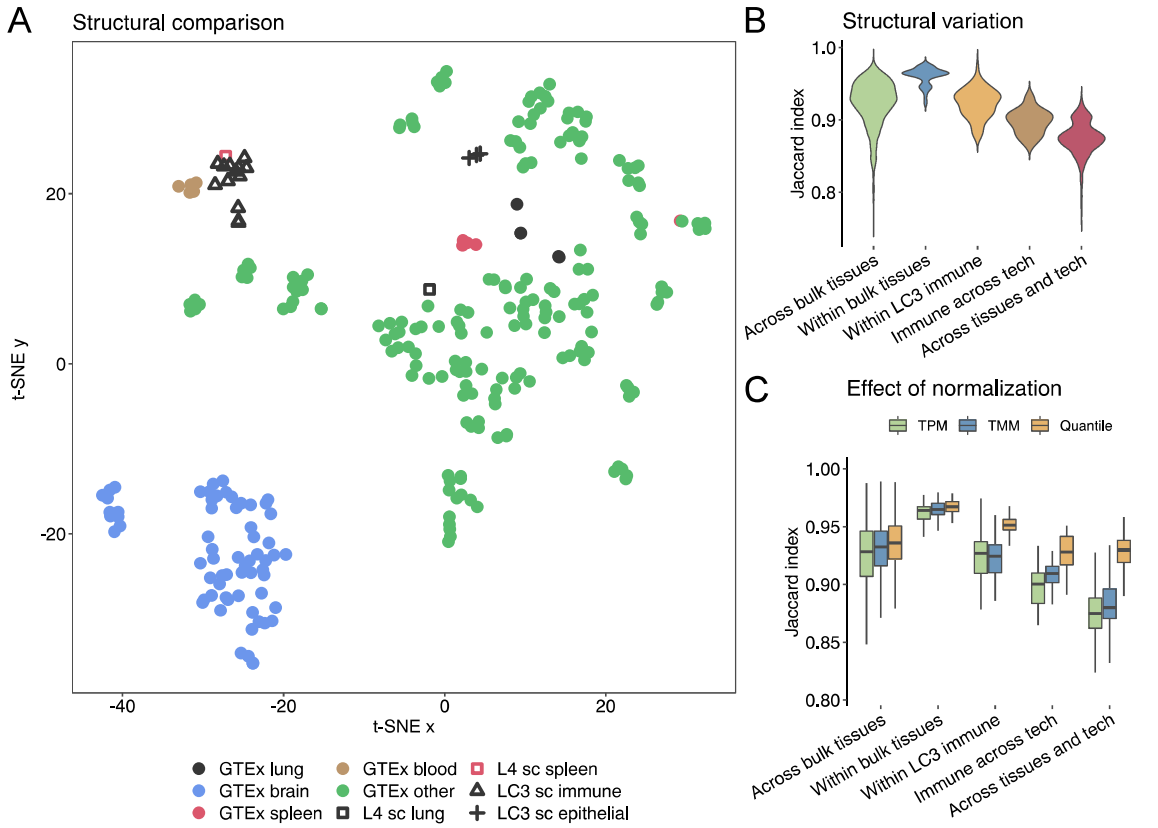

Fig 26: Structural comparison of models generated from bulk and single-cell RNA-Seq data. *A. t-SNE projection of models generated from GTEx bulk data and cell populations from different single-cell datasets (see **paper IV**) For the 2 models from the L4 dataset, all cells from each tissue (lung or spleen) were pooled together, while the 16 models from the LC3 dataset were generated from individual cell types from tumor (10) and healthy tissue (6). B. Structural variation within and across different model groups. C. The effect of normalization on structural variation.*

5.2. Application: Mouse primary motor cortex

To evaluate the utility of our method, we generated context-specific GEMs from a deeply sequenced scRNA-Seq dataset from the mouse primary motor cortex [127] and the model Mouse-GEM [42]. We used DSAVE to estimate the required pool size, resulting in the selection of 17 predefined neuron subpopulations (provided together with the dataset) with at least 450 cells, where we only used samples processed in one batch (labeled 4/26/2019). The UMAP projection was in good agreement with the predefined cell types (Fig. 27A). The cell types still separated well in the UMAP when only metabolic genes were used, suggesting that each neuron subtype has a unique metabolism (Fig. 27B). A structural comparison (PCA) of the generated models revealed that they grouped primarily by neuron type (IT: inferial temporal, NP: near-projecting, CT: corticothalamic, and Lamp5-expressing neurons), and not by cortex layer (Fig. 27C) (See **paper IV** for details).

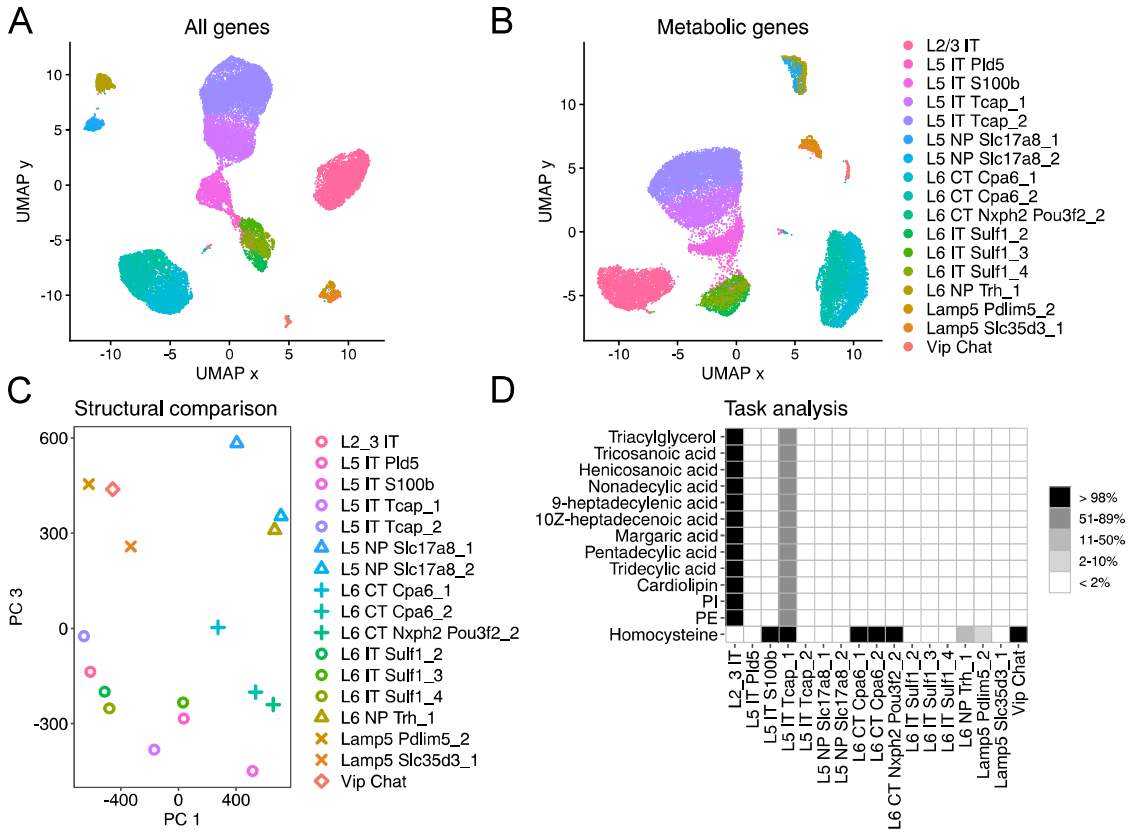

Fig. 27. Generation of context-specific models for different neuron types in the primary mouse motor cortex. *A. UMAP projection of the single cells from the 17 largest clusters in the MCOR3 dataset (see **paper IV**), colored by cell subtype from classifications published together with the data. The full gene set is used. B. Similar to A, but only metabolic genes present in the Mouse-GEM model are used. C. Structural comparison (PCA) of the GEMs generated from the neuron subtypes shown in A. The symbols group the models into neuron types (IT/NP/CT/Lamp5/Vip). PC 3 was used instead of PC 2, since PC 2 represented an unknown factor for which we could see no pattern. D. Task analysis of bootstraps (100 per cell subtype). The color represents for how many bootstraps of a cell subtype that the task could be completed. Only tasks that could be completed for at least 99 bootstraps in one cell subtype while only being completed in a maximum of 1 bootstrap for another cell subtype are shown.*

To evaluate differences in metabolic capabilities between the neuron subtypes, we generated 100 bootstraps per subtype, followed by model generation by ftINIT and task

analysis (Fig. 27D). We defined a task to be available (on) if it could be completed in at least 99 of 100 bootstraps and unavailable (off) if it could be completed in one or zero bootstrap models, and we have shown that the difference between on and off is statistically significant (see Note S1 in **paper IV**). We found 13 tasks that were considered on in at least one cell type while off in at least one other. Most such tasks were related to *de novo* synthesis of fatty acids, phospholipids (PE and PI) and cardiolipin. Interestingly, the importance of lipids as signaling molecules in the brain have recently been highlighted, where deficiencies in lipid metabolism is associated with neurogenerative diseases and cognitive problems [128]. Likewise, we detected significant differences in homocysteine synthesis capabilities, and high levels of homocysteine in the blood is associated with neurological disorders [129]. While homocysteine levels in blood are mainly regulated by the liver, it seems that the capability of homocysteine synthesis is available for some neuron subtypes, but not for others, which may be of interest to investigate further.

5.3. Application: Tumor microenvironment

To investigate metabolic differences across cell types in the tumor microenvironment, we generated context-specific models for 16 cell subtypes from a lung adenocarcinoma dataset [130]. 10 cell subtypes were extracted from the tumor and 6 from healthy lung tissue, and all had at least 1,600 cells (the limit determined by DSAVE). We used cell subtype classifications distributed with the dataset, and the classifications were in good agreement both for the cell subtypes found in the tumor (Fig. 28A) and healthy lung tissue (Fig. 28B). The cell subtypes were similar across patients except for the neoplastic cells (tS1 and tS2), which have unique mutations for each patient (Fig. 28C). A structural comparison (PCA) of the models generated from the cell subtypes revealed that they grouped with cell subtypes of similar class (myeloid cells, epithelial cells, lymphocytes, and mast cells) rather than tissue (Fig. 28D). This is an interesting finding, since the tumor microenvironment in contrast to other tissues is often nutrient-deprived, hypoxic, and acidic, at least in some regions [131]. Apparently, these differences in conditions had less of an effect on the metabolism of cells than the cell subtype for these tumors.

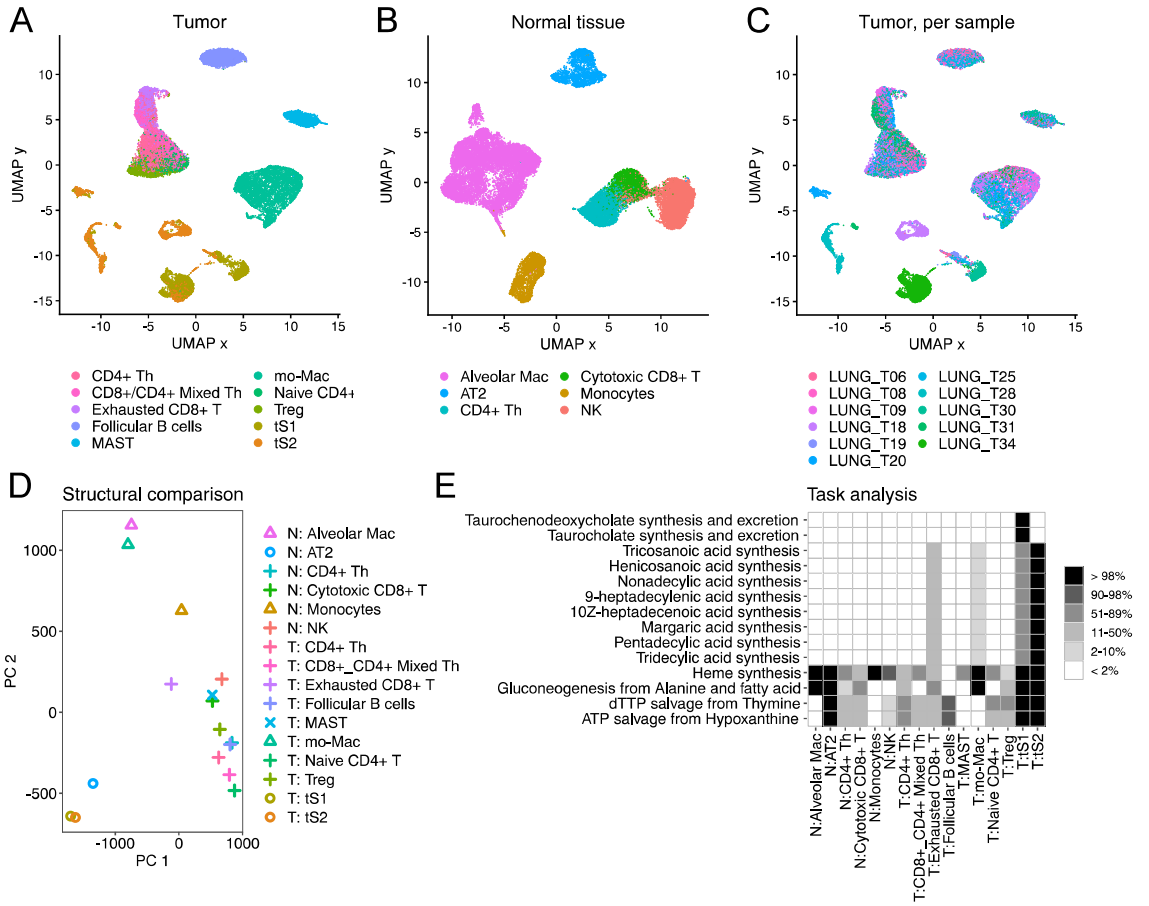

Fig. 28: Generation and analysis of context-specific models in the tumor microenvironment. *A. UMAP projection of the single cells from lung adenocarcinoma tumor samples (The LC3 dataset, see* ***paper IV****, contains cells from several patients). The cells are colored by cell subtype classification published together with the dataset. Only cell subtypes with at least 1,600 cells were included in the analysis. The full gene set is used. B. Similar to A but shows healthy tissue samples extracted from the cancer patients. C. Similar to A, but the cells are colored per sample instead of cell subtype. D. Structural comparison (PCA) of the GEMs generated from the cell subtypes shown in A and B. The symbols group the models into cell type classes (myeloid, epithelial, lymphocytes, mast cells). E. Task analysis of bootstraps (100 per cell subtype). The color represents for how many bootstraps of a cell subtype the task could be completed. Only tasks that could be completed for at least 99 bootstraps in one cell subtype while only being able to complete in maximum 1 bootstrap for another cell subtype are shown.*

We next investigated the ability of the different cell subtypes to perform metabolic tasks. We found 14 tasks with a significant difference in the ability to perform tasks (i.e., the task could be completed for at least 99 of the bootstrap models in one cell subtype and could be completed for no more than 1 bootstrap model in another cell subtype) (Fig. 28E). The two transcriptional states of neoplastic cells (tS1 and tS2) show differences in bile acid metabolism (taurochenodeoxcholate and taurocholate synthesis and excretion), despite containing cells from multiple patients with clearly different transcriptional programs (Fig. 28A, E). The role of bile acids in cancer has been a topic of interest for decades [132], but its role in lung cancer is unknown, and may be interesting to investigate further. Another observation of interest is the ability of the cancer cells to synthesize heme. Heme synthesis, coupled to degradation of the just-synthesized heme and export of bilirubin, provides means to degrade succinyl-CoA from the TCA cycle without involving fumarate hydratase. While we cannot see any benefit from using this pathway for this purpose in healthy cells, it has been proven vital for cancer cell lines with dysfunctional fumarate

hydratase, allowing for generation of mitochondrial NADH by running parts of the TCA cycle [133]. The role of this pathway in lung cancer is to my knowledge unknown.

5.4. Summary

It is not possible to generate reliable context-specific GEMs that cover the entire metabolism from the transcriptomic profile of a single cell due to data sparsity – it is often required to pool thousands of cells to yield reasonably reproducible results. Statistical comparison of model properties is generally challenging. To handle such comparisons, we in this work propose three things: 1) to use the DSAVE total variation score to estimate the required number of cells, which reduces the uncertainty from sparsity; 2) to use bootstrapping to estimate the uncertainty from sampling; and 3) to only accept very low p values as significant, which makes it possible to without comparison between models determine if a reaction or task is "on" with statistical significance.

6. A light-weight approach to enzyme usage constraints

As mentioned in the introduction, GECKO toolbox [39], [43] supports the addition of enzyme usage constraints to a GEM. However, during my PhD, I encountered three problems when using the GECKO Toolbox with the Human1 model. 1) Addition of enzyme constraints using the GECKO Toolbox made the model substantially larger, which led to long run times for flux balance analyses, especially on large, combined models where several cell types are interacting with each other. 2) The solver encountered numerical issues when using GECKO models, and in some cases, it could not solve the problem. The issue was related to the stoichiometry in the GECKO model, where some fluxes became very small. 3) The run time of the GECKO pipeline was substantial, which slowed down my research efforts. As part of **paper V**, I therefore investigated the possibility to remedy these issues.

6.1. The method

GECKO was designed to allow for constraining the total enzyme usage in a cell as well as constraining individual enzymes by proteomics data, and the latter makes the models generated by GECKO larger and more complex. In most projects on human metabolism, only the total enzyme usage constraint is used. I therefore set out to create a simplified version of GECKO, called GECKO Light (**Paper V**), that runs faster and generates smaller models without numerical issues. Consequently, GECKO Light does not support constraining the fluxes through individual enzymes, but only allows for constraining the total enzyme usage. The strategy of GECKO Light is similar to that described in sMOMENT [134].

A complication when applying enzyme usage constraints are isozymes, i.e., parallel enzymes that can catalyze the same reaction. The enzyme usage $c_{e,r}$ per flux unit for the enzyme e in reaction r is in Gecko modeled as

$$c_{e,r} = \frac{M_{w,e}}{k_{cat,e,r}} \quad (4)$$

where $M_{w,e}$ is the molecular weight of the enzyme e and $k_{cat,e,r}$ is the turnover rate for enzyme e in reaction r. These values typically vary between isozymes, and in the full GECKO model, each of those can be constrained separately. However, if support for constraining of individual enzymes is omitted, we can assume that the isozyme with the lowest cost will always be used, since that enzyme will always be used in an optimization where enzyme usage is limiting. While GECKO Toolbox builds up a complex network of reactions to reflect all possible paths through different isozymes, it is possible to simplify the network and just use one reaction, by defining the cost c_r for the reaction r as

$$c_r = \min\left(\frac{\sum_i M_{w,i,e}}{k_{cat,e,r}}\right), e \in E, i \in I_e \quad (5)$$

where E is the set of available isozymes for reaction r, $M_{w,i,e}$ is the molecular weight of subunit i in enzyme e, $k_{cat,e,r}$ is the turnover rate for enzyme e when catalyzing reaction r, and I_e is the set of subunits that enzyme e consists of. With this approach, which will give the same results as GECKO when only constraining the total protein usage, the total enzyme usage constraint can be implemented by adding one metabolite and one reaction.

We added the metabolite “prot_pool” with the stoichiometric coefficient $-c_r$ to each enzyme-catalyzed reaction and a reaction “prot_pool_exchange” that produces the metabolite prot_pool. The reaction “prot_pool_exchange” can then be constrained to limit the total enzyme usage of the model.

GECKO Light has two advantages compared to the original GECKO approach – the generated models are smaller, and we no longer experience numerical issues with the solver (Gurobi). The reason for the latter is that the stoichiometry of GECKO Light avoids the very low fluxes sometimes produced by the original GECKO models. To speed up the execution time of the model generation, the code was also optimized, which reduced the execution time by an order of magnitude – GECKO Light typically finishes within 2 minutes on a standard laptop computer (Intel Core i7-6600U, 2.60 GHz, 2+2 cores) for adding enzyme constraints to Human1. The same optimizations were as part of this work implemented in the original GECKO workflow as well, which yielded a similar improvement in execution speed.

6.2. Summary

The use of enzyme-constrained models can be divided into two use cases: 1) when constraining individual enzymes is of interest, and 2) when only a global enzyme usage constraint is applied. GECKO Light cannot be used for the first case, but is the better approach for the second, especially when generating large models containing several cell types.

7. Genome-scale metabolic modeling of the tumor microenvironment

The complex metabolism in the tumor microenvironment is difficult to study. While it is often possible to measure the uptake rates of different metabolites for experiments performed on cell lines *in vitro* [25], such measurements are very difficult to perform in living tissue. In **paper V**, we therefore set out to investigate this metabolism using a modeling approach, where the maximum uptake rates of metabolites were defined by a theoretical diffusion model based on diffusion coefficients and metabolite concentrations in blood. While we in **paper IV** worked with context-specific models, we in this work used the full Human1 model (curated), to investigate the theoretically most optimal behavior of cells in the TME. Specifically, we 1) set out to investigate the optimal metabolic behavior of cancer cells at different pseudo-distances from blood vessels, 2) investigated the optimal amino acid metabolism for the tumor conditions in detail, and 3) investigated if the common belief that stromal cells help tumor cells by providing them with energy-rich resources such as lactate is truly beneficial for cancer cell growth.

7.1. A diffusion model for constraining metabolite uptake rates

It is challenging to estimate the maximum uptake rates of metabolites in cells in the TME. The maximum influx of a metabolite to a cell depends on many factors such as the distance to blood vessels and blood vessel permeability, and to estimate the absolute maximum influx to a particular cell becomes a very complex task. We therefore instead set out to estimate the relative maximum uptake rates of the different metabolites in the tumor, based on metabolite concentrations in blood and their diffusion coefficients.

As mentioned in the Background section, the main mechanism for metabolite influx from a blood vessel into solid tumors is diffusion. The influx can be modeled as an axisymmetrical two-dimensional model, since we assume that there is no concentration gradient along the blood vessel axis (Fig. 29). We assume a radial diffusion flux from the capillary. Under certain assumptions, for example steady state conditions, the uptake bound U_i for a metabolite i can then be estimated as

$$U_i = aD_i c_{b,i}$$

where D_i is the diffusion coefficient for metabolite i, $c_{b,i}$ its concentration in blood, and a is a proportionality constant that is related to the distance from the capillary to the point for which the uptake bound is to be estimated.

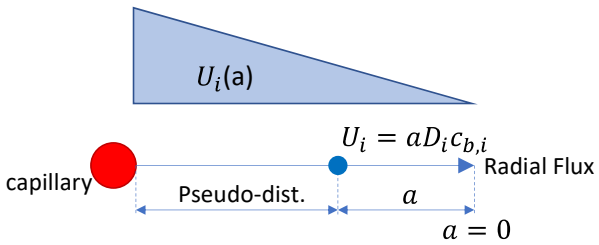

Fig. 29: The diffusion model used to estimate the maximum metabolite flux uptake for a metabolite. *The influx of the metabolite is proportional to the diffusion coefficient and the concentration of the metabolite in blood and assumed to be 0 at a distance far away. The proportionality constant a is the same for all metabolites, and is related to the distance to the capillary, although not linearly. We can define a pseudo-distance from the capillary as the maximum value of a minus its current value.*

The model should be viewed in the light that it is an approximation that provides uptake constraints on the right order of magnitude. The modeling results presented later in this thesis are not sensitive to small changes in the metabolite uptake constraints. The model is justified in detail in Note S2 in **paper V**.

7.2. The optimal metabolic behavior for cellular growth in the TME

As described in the background section, the metabolism of the tumor microenvironment is dysregulated, partly since the metabolite availability is different compared to other tissues. To investigate the optimal metabolic behavior under these conditions we set up a modeling scenario with the diffusion model and the Human1 GEM, curated and extended with an enzyme usage constraint (Fig. 30A). In addition, non-growth associated maintenance (NGAM) was added as an ATP cost of 1.833 mmol gDW^{-1} h^{-1}, which was derived from literature [135], [136]. Metabolite blood concentrations for in total 69 metabolites were collected from several sources [137]–[140]. We used the concentration of free oxygen, which is free to diffuse, excluding roughly 98% of the total oxygen concentration in blood, which is bound to hemoglobin and therefore cannot diffuse. Lipids were grouped into 2 groups: sterols (represented by cholesterol) and other lipids (fatty acyls, glycerolipids, glycerophospholipids, and sphingolipids, represented as a mix of fatty acids). Diffusion coefficients for 18 metabolites were downloaded from several sources [141]–[143]. For lipids, we used the diffusion coefficient of albumin, since they diffuse bound to either albumin or a lipoprotein, while the rest of the diffusion coefficients were predicted using a linear model based on molecular weight [144].

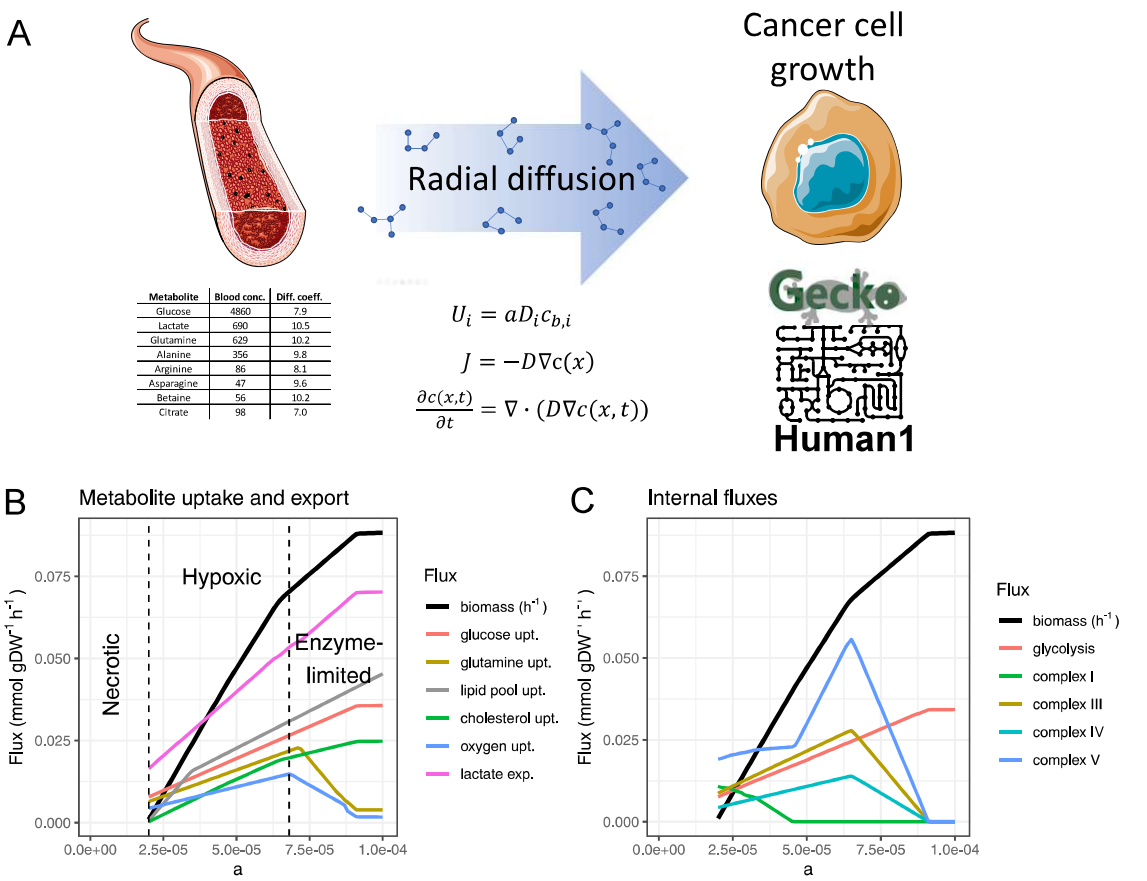

Fig. 30: Modeling of cancer cell growth in the TME. *A. Modeling setup. The model Human1 is prepared with enzyme usage constraints using GECKO Light. The metabolite uptake rates are limited by the diffusion model. B. Simulated specific growth rate and metabolite uptake at different* a *values, which is inversely proportional to the distance from blood vessels. C. Fluxes through glycolysis and the enzyme complexes in the electron transport chain.*

The metabolism of the different regions of the tumor, represented by different values of *a*, were simulated using FBA, optimizing for biomass (growth) (Fig. 30B). At large distances from blood vessels (low *a*) the model could not produce enough ATP for maintaining the cell (NGAM), resulting in necrosis. When moving closer to the capillaries (moderate *a* values), the metabolism was dominated by hypoxia and lack of nutrients, which limited the growth. At small distances (high *a* values), the enzyme usage constraint was the main factor limiting growth, and lack of oxygen and nutrients was less of a problem. Interestingly, the model predicted the Warburg effect [22], where lactate is exported despite that all oxygen is not used, which can be explained by the higher enzyme usage cost per produced molecule of ATP in the electron transport chain. The fluxes of the different complexes vary with the constant *a*, predicting an early bypass of complex I, which is consistent with a previous report modeling muscle tissue [6] (Fig. 30C). Likewise, the model predicted high use of glutamine, which we will examine in more detail later.
The modeling results here need to be taken for what they are, a simplified model. The behavior of the model is extreme – in real cells, OXPHOS and the TCA cycle are not reduced to zero at high glucose availability. For example, 80% of the ATP has been reported to be generated by oxidative phosphorylation in some highly proliferative cells [145]. However, the mechanisms are still of interest to study, and many of them are present in real cells [22], [24], although real cells only partly employ the strategies fully adopted by the model.

To estimate which metabolites were limiting for growth, we ran the simulation in a mode where the uptake rate of a single metabolite was reduced to 90% of its original value (Fig. 31A). Glucose was the most important metabolite for growth, followed by oxygen, and we also saw a small effect from glutamine. The effect from reducing lipids, cholesterol, lactate and albumin (not shown) was negligible. To understand which parts of the biomass reaction limit growth, we conducted simulations where different components of the biomass were removed (Fig. 31B). ATP production was the limiting process for growth; only removal of components requiring ATP increased growth. After removing the ATP costs, lipids became limiting. Interestingly, the direct use of amino acids for protein synthesis was small compared to the availability – the only reason that glutamine was limiting for growth is because it can be used to increase ATP production.

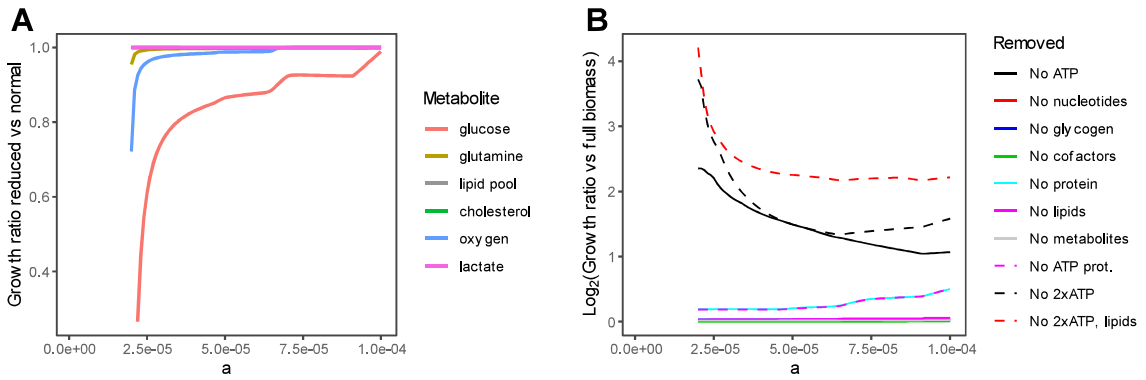

Fig 31: Growth limitation from metabolites. *A. Changes in specific growth rate when the uptake constraint of a specific metabolite is reduced to 90% of its original value (as defined by the diffusion model). B. Specific growth rate ratio between models with original and reduced biomass reaction (some components have been removed), where the model is optimized for biomass. "No ATP prot." means removal of the ATP cost from building proteins from amino acids, while "No 2xATP" refers to having both this protein generation ATP cost and the direct ATP cost removed from the biomass reaction. For "No 2xATP, lipids", the consumption of lipids have also been removed in addition to the ATP costs.*

7.3. Amino acid metabolism in the TME.

Amino acid metabolism in tumors is not fully understood, and to shed light on the subject we investigated the fluxes of amino acids predicted by the model. The model predicted large uptake rates of glutamine, glycine, serine, and threonine, and large export of aspartate and proline (Fig. 32A). In addition, arginine, asparagine, cysteine, glutamate, histidine, and valine showed irregular uptake patterns that vary with *a* (Fig. 32B), while the rest of the amino acids are taken up at rates proportional to the specific growth rate and were used primarily for protein synthesis (not shown). Glutamate secretion, which is observed in many cell lines [25], [146], was not observed in the model. Such a behavior has previously been linked to nucleotide synthesis [25], which our model does not predict, and also to signaling effects that increase growth in glioblastoma, where glutamate acts as a signaling molecule [147], [148].

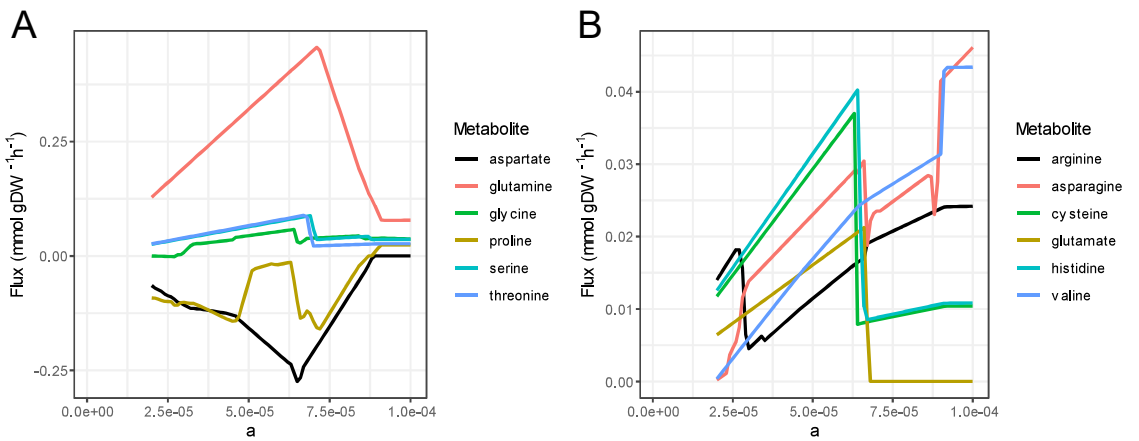

Fig. 32: Amino acid uptake and export. *A. Uptake and export of amino acids with large fluxes. Negative flux corresponds to export of the metabolite. B. Uptake (no export observed) of amino acids with small but irregular fluxes. The fluxes for the amino acids not presented in A or B are proportional to the specific growth rate.*

The model predicts two interesting behaviors: use of glutamate (mainly converted from glutamine) to feed the TCA cycle, and the export of proline. To understand these behaviors, we need to first understand the effects of hypoxia and limitations of enzyme usage in the TCA cycle and OXPHOS. The TCA cycle generates NADH and $FADH_2$, which is oxidized during OXPHOS to generate ATP. While many carbon-based metabolites via conversions can enter the TCA cycle for ATP production, the process requires oxygen, and most such metabolites are useless for ATP generation if oxygen cannot be used. When oxygen is limited, it is therefore crucial to maximize the ATP production per oxygen molecule spent, and not per carbon-based metabolite. In Human1, which reflects the latest understanding of the stoichiometry in OXPHOS, complex I pumps 4 protons out of the mitochondrial matrix per oxidized NADH. Complex V produces 1 ATP per 3 such protons, so in practice, complex I produces 1.33 ATP per NADH oxidized. More protons are then pumped out of the mitochondrial matrix using complex III and IV. Complex II (which oxidizes $FADH_2$) on the other hand does not pump any protons (but the same amount as for NADH is pumped by complex III and IV, i.e., in total 6 H^+). This means that if oxygen is used to oxidize $FADH_2$, 1.33 less ATP will be produced compared to if the same amount of oxygen was used to oxidize NADH. It is therefore of interest to maximize the ratio of NADH vs $FADH_2$. It could be hypothesized that it could be beneficial to find ways to get rid of NADH through other pathways, and thereby increase the flux through the TCA cycle if oxygen is limiting, since less oxygen would then be required per round of the cycle. However, as long as the $FADH_2$ cannot be oxidized, such a strategy will not work. Each round of the TCA cycle will produce 1 ATP and 1 $FADH_2$, and for each $FADH_2$ molecule that is oxidized instead of a NADH molecule, 1.33 ATP is lost, yielding a net ATP loss of 0.33 ATP per cycle round.

When enzyme usage is limiting, the situation is quite different. In such cases, it is beneficial to minimize the use of OXPHOS since the complexes in the ETC are large and slow and thereby constitute much of the total enzyme pool. The most optimal behavior in the model to maximize ATP production is to run glycolysis alone and export lactate. However, in the middle range of *a*, where glucose is still in shortage, it is still beneficial to use the TCA cycle. In such cases, it contrasts with the hypoxic case where it is beneficial to dispose of NADH, since the enzymatic cost of running the TCA cycle is lower per ATP produced than that of OXPHOS. As long as the NADH can be oxidized, the TCA cycle can therefore

complete more cycles as an alternative to running full oxidation in OXPHOS. In enzyme-limited cases, it is therefore optimal to minimize the NADH/$FADH_2$ ratio since there at least in the model is no mechanism to oxidize the $FADH_2$ other than complex II in the electron transport chain.

The use of glutamine by cancer cells to feed the TCA cycle instead of pyruvate derived from glucose is well known in literature, and is often termed "glutamine addiction" [24], [149]–[151]. However, the reason why cancer cells prefer glutamine as substrate is not well understood, which the model could potentially help with. We therefore investigated the high uptake of glutamine in the model (Fig. 33). When glutamate (mainly derived from glutamine) is used to feed the TCA cycle, only part of the cycle is used. We compared the use of glutamate as substrate to the cycle with lactate, since lactate is what will in such cases be exported from glycolysis. While lactate produces 5 NADH per round of the cycle (one extra when converting lactate to pyruvate), only 2 NADH are produced when using glutamate. Under enzyme-limited conditions, the TCA cycle can therefore be run more cycles with glutamine and increase the total ATP production. As part of the process, aspartate is produced and exported by the model.

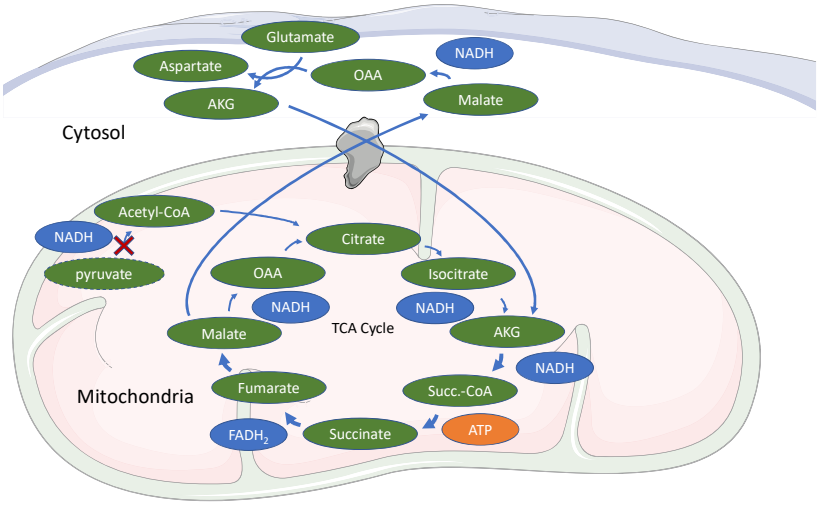

Fig. 33: The pathway used when glutamine feeds the TCA cycle. *Using glutamine to feed the TCA cycle leads to less production of NADH for each round in the cycle. When OXPHOS is limited by its high enzyme usage, the TCA cycle flux can therefore be increased, producing more ATP. The arrow thicknesses indicate the amount of flux that passes through the reactions, where thick arrows represent higher flux.*

While glutamine import is well documented, aspartate export is not, and was not observed in a cell culture experiment of the NCI-60 cell lines [152]. However, there are several pathways that can convert aspartate to other metabolites. For example, it can be converted to lactate (via fumarate, malate, and pyruvate) without extra ATP cost or altering the redox balance (except for shifts between NADH and NADPH), which may explain why aspartate export is not observed. In addition, aspartate can be exported to dihydroorotate, which was reported for hypoxic conditions [150].

Turning the attention to proline export in cancers, which has been observed in both NCI60 and HEK293 cell lines [25], [152], there are two potential pathways to consider (Fig. 34). Under normal conditions, the PYCR enzymes (PYCR3 or other enzymes of the same family) is responsible for proline synthesis, while the PRODH is used in the opposite direction. However, an intriguing option is the possibility that PRODH could be run in

reverse, which is predicted as favorable in the model. We have not found any evidence of this behavior in the literature, and since it is less favorable from a thermodynamic perspective compared to PYCR, it is unclear if it is favorable enough to carry flux. However, it has been shown that complex II, which catalyzes a similar reaction, can be run in reverse in rats [153]. Running PRODH in reverse would be favorable in hypoxia, since it would enable the running of complex I without oxygen consumption, and thereby enable production of extra ATP. PYCR is shown by to be favorable at enzyme-limited conditions by the model, since it provides means to dispose of NADH.

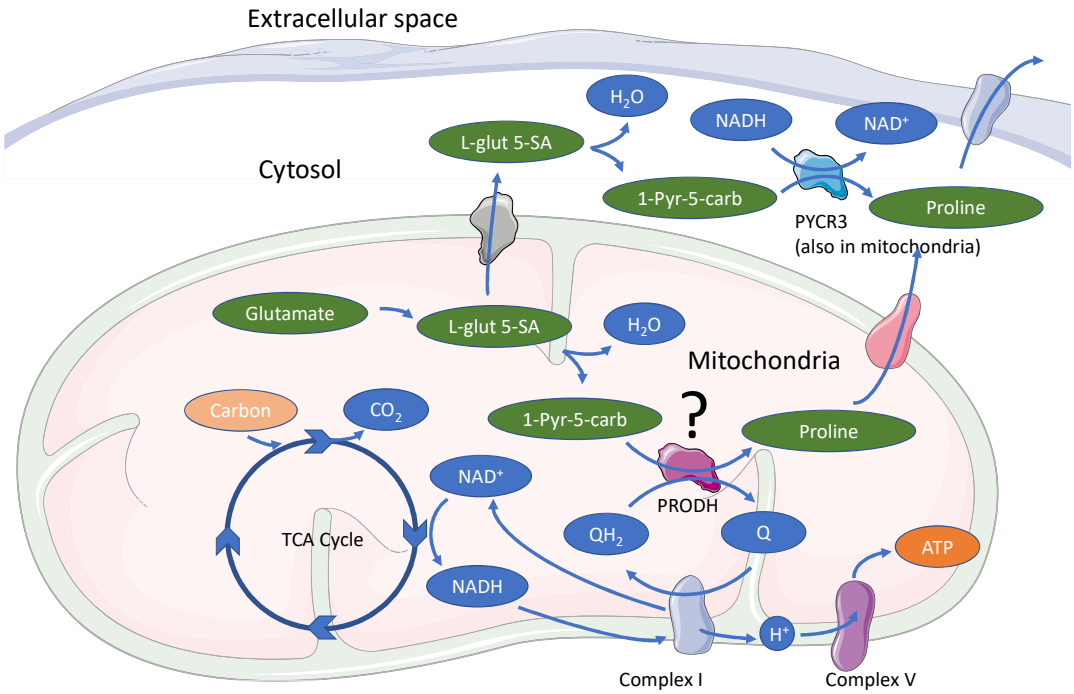

Fig. 34: The potential mechanism for running PRODH in reverse. *Proline synthesis can assist in increasing the flux through the TCA cycle, leading to an increase of ATP production, by the following mechanisms: 1) By oxidation of NADH through any of the PYCR enzymes, since less enzymatic capacity would need to be spent on OXPHOS in situations where enzyme usage is limiting. 2) By running PRODH in reverse, if possible. This would enable increased flux through complex I in hypoxic conditions, since PRODH can be used instead of using oxygen for converting ubiquinol (QH_2) into ubiquinone (Q). Consequently, more ATP can be produced in complex V since complex I pumps more protons out of the mitochondrial matrix.*

To quantify the metabolism of each amino acid in different conditions, we ran simulations where the model was given a single substrate of a limited amount and maximized ATP production (Table 2). Under hypoxia and no enzyme constraints, PRODH in reverse was beneficial for many amino acids, and many amino acids gave a higher ATP yield than lactate. With PRODH in reverse blocked, the positive effect for amino acids was lost except for serine, threonine and glycine, which still gave a higher ATP yield compared to lactate. With a model where enzyme usage constraints were limiting for ATP production, the overall picture changed, and glutamine (and glutamate) became the substrates with the highest yield. As described above, this can be understood from an increased flux in the TCA cycle due to better possibilities to dispose the NADH (Fig 35).

Table 2: Simulation of ATP production from lactate and amino acids under different conditions. *The ATP production (mmol/gDW/h) is maximized given a maximal uptake of 5 mmol/gDW/h of a single substrate and varying oxygen availability. "O2-limited": The oxygen uptake is limited to 5 mmol/gDW/h, which is not enough to fully oxidize any of the substates. "O2-limited, no PRODH": In addition to "O2-limited" the reverse PRODH reaction is blocked. "No O2, no PRODH": The O2 uptake constraint is set to zero and the reverse PRODH reaction is blocked. "Enzyme Lim.": The total available enzyme usage pool is constrained to a low value (0.001 g/gDW). PRODH is not active in such conditions. Green background corresponds to a higher ATP production compared to lactate, white to identical, and red to a lower flux.*

Substrate	Low O_2	Low O_2, no PRODH	No O_2, no PRODH	Enzyme lim.
lactate	5	5	0	0.0938
aspartate	6.5	5	0	0.0937
glutamine	10	5	0	0.1042
glycine	5.7	5.7	0	0.0929
proline	4.7	4.7	0	0.0943
serine	9	9	3.33	0.0945
threonine	9	9	3.33	0.0931
alanine	5	5	0	0.0932
arginine	12.3	5	0	0.0942
asparagine	6.5	5	0	0.0940
cysteine	5	5	0	0.0927
glutamate	10	5	0	0.1020
histidine	11.1	5	0	0.0923
isoleucine	5	5	0	0.0921
leucine	5	5	0	0.0917
lysine	4.8	4.8	0	0.0918
methionine	5	5	0	0.0830
phenylalanine	5	5	0	0.0900
tryptophan	3.8	3.8	0	0.0877
tyrosine	5	5	0	0.0915
valine	5	5	0	0.0917

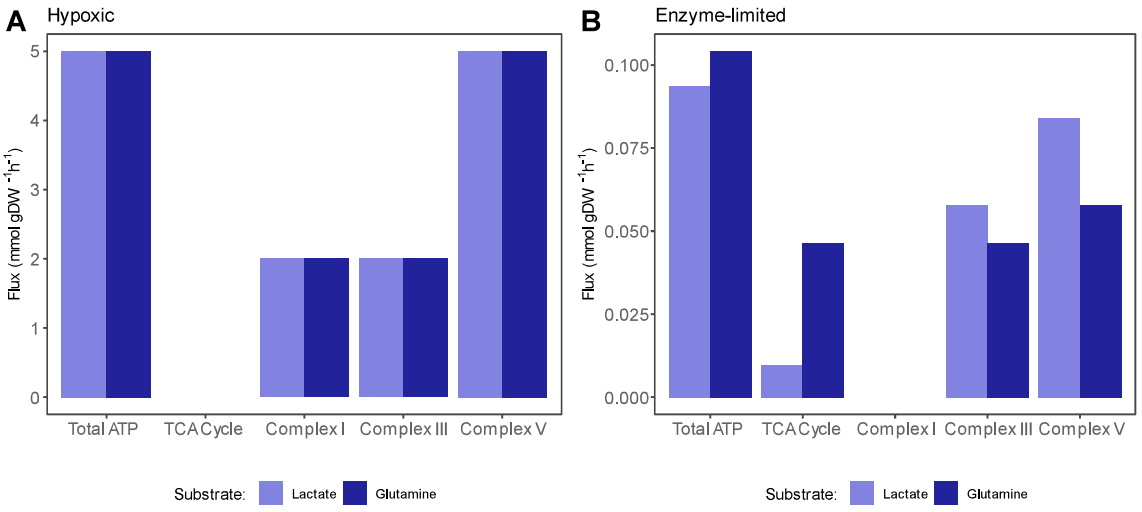

Fig. 35: ATP production from the substrates glutamine and lactate. *A. Hypoxic conditions (reverse PRODH reaction blocked). B. Enzyme-limited conditions.*

7.4. Evaluation of metabolic collaboration between cell types in the TME

It is commonly believed that stromal cells in the TME support cancer cells by providing them with resources. For example, it has been proposed that cancer-associated fibroblasts (CAFs) in the TME supply cancer cells with pyruvate and lactate [28], [154]. Furthermore, macrophages the TME could consume dead cell and debris and convert this mass to resources useful for cancer cell growth. To investigate such collaboration scenarios, we constructed a combined model of three cell types: cancer cells, fibroblasts, and other cells, where the latter represents cells that do not collaborate with the cancer cells (for example immune cells) (Fig. 36A-B). While the tumor cells need to grow to enable tumor growth, the fibroblasts and other cells are expected to be recruited to the tumor. The fibroblasts however need to produce the extracellular matrix (ECM) for the tumor.

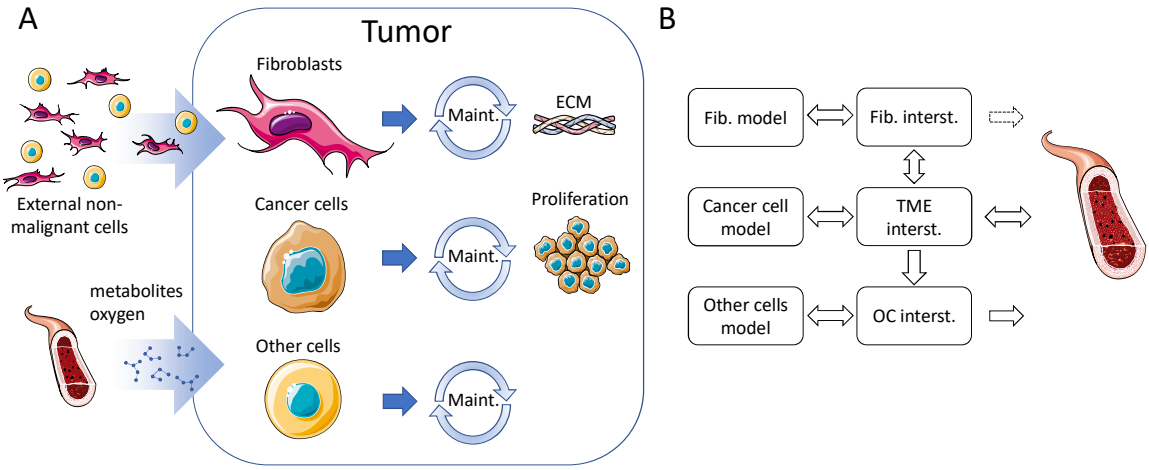

Fig. 36: Modeling setup for cell type collaboration in the TME. *A. Modeling setup. The model consists of interconnected cell types: Cancer cells, fibroblasts, and other cells (for example immune cells). Each cell type is represented by a separate model of the type that was used in the simulations in Fig. 30, with NGAM included. The model is fed with metabolites from blood as defined by the diffusion model. Fibroblasts and other cells are assumed to be recruited from outside the tumor, and thus their biomass production is not included in the tumor growth. However, the tumor cells need to grow, and the fibroblasts need to build the extracellular matrix for the tumor. B. Communication between compartments in the combined model. While the fibroblasts can send metabolites back to the cancer cells, the other cells cannot since they are not expected to collaborate with the cancer cells.*

The combined model can be parameterized regarding the cell type mixture in the tumor and the portion of the tumor that the ECM constitutes, and we worked with 7 different configurations in the simulations (Table 3). The ECM was fixed to 80% collagen (represented by collagen I) and 20% glycosaminoglycans (represented by heparan sulfate).

Table 3: Models used in the simulations. *The m0 model contains only tumor cells and is identical to the model used in Fig. 30. The ECM fraction represents the weight fraction of the total objective that the ECM constitutes.*

Model	Cancer cell fraction	Fibroblast cell fraction	Other cells fraction	ECM fraction
m0	1	0	0	0
m1	0.6	0.2	0.2	0.01
m2	0.6	0.2	0.2	0.25
m3	0.6	0.2	0.2	0.5
m4	0.75	0.05	0.2	0.25
m5	0.45	0.35	0.2	0.25
m6	0.9699	0.0001	0.03	0.00001

As a direct consequence of the first law of thermodynamics, fibroblasts cannot increase the total available amount of accessible energy in the TME. However, it is possible for the fibroblasts to lend enzymatic capacity to the cancer cells, and thereby increase the growth rate. Indeed, the models predict such a behavior in some cases (Fig. 37A). Depending on the size of the extra burden from generation of ECM, the combined model can grow faster than the model where only cancer cells are present in cases where resources are plentiful. To investigate which metabolites contributed to the increase in growth, we designed an iterative algorithm that extracted all metabolites that were exported from the fibroblasts and imported into the cancer cells, identifying in total 233 such potential collaboration metabolites (Fig. 37B). However, many of these metabolites are unrealistic and lack support in the literature (for example ATP). We therefore limited the allowed collaboration metabolites to those proposed in literature, namely lactate, pyruvate, free fatty acids, glutamine, ketone bodies, and alanine [154], [155]. With the transport of metabolites from fibroblasts to cancer cells limited to these metabolites, we could no longer observe any growth advantage compared to if the same transport is completely shut down (not shown, see **paper V** for details).

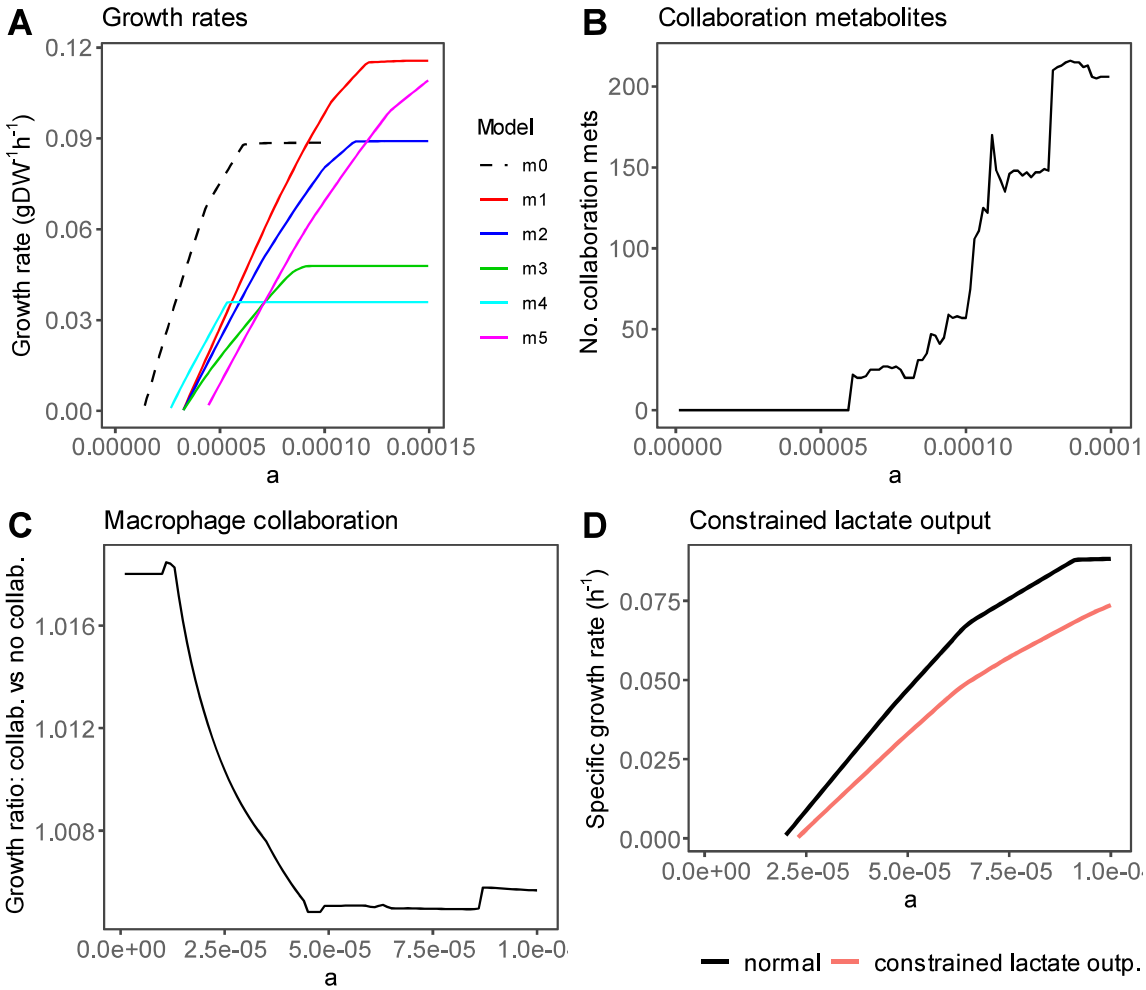

Fig. 37: Tumor model simulations. *A. Growth simulations with different model setups. B. The number metabolites sent from the fibroblasts to the tumor cells (collaboration metabolites) during growth for the m2 model. C. Increase in growth rate by adding 10% of the biomass back as input for the tumor, which simulates the macrophage activity to clean up dead cells. D. Change in growth rate from limiting the lactate output to half of the maximum uptake rate of glucose.*

Macrophages can scavenge dead cells and debris and convert it into usable metabolites, which could be beneficial for growth. To simulate such a scenario, we assumed that 10% of the tumor cells die, and that the materials those cells consist of can be used as input for growth. Since the macrophage function is complicated, we didn't explicitly model the macrophages, but rather extracted the relevant parts of the biomass equation and used 10% of those metabolites as input for growth. The ATP costs for growth were omitted since that part is consumed in the growth process, and we also omitted any costs, including cell maintenance, related to the operation of the macrophages. The growth advantage of the extra input of metabolites was small (Fig. 37C), which can be attributed to the lack of oxygen or enzymatic capacity for the different values of *a*. The major limiting factor for growth is ATP production, and for most materials found in a cell, oxygen and the use of OXPHOS is required for ATP generation.

In another type of collaboration often proposed in literature, oxygenated cancer cells consume lactate instead of glucose and thereby increase the available amount of glucose in the TME [156], [157]. Sonveaux et al [156] presented experimental evidence that the necrotic regions of tumors were enlarged when blocking lactate uptake in mice. However, I reason that although there will be an increase in glucose availability, there will also be a corresponding decrease in oxygen availability, which likely cancels out any such positive effect. I hypothesize that the effect could instead be related to the pH in tumors – inhibition of the lactate transporter MCT1 (which was done in the study) will likely reduce lactate uptake in the entire body. The lactate levels will then likely be higher in the blood and tumor-adjacent tissue, leading to smaller lactate gradients and lower diffusion rate from the tumor into blood and adjacent tissue. For a cell to survive, the internal pH needs to stay within a narrow range [151], [157]–[159], and it has been proposed that a higher external pH requires a larger ATP maintenance to sustain the internal pH [158], [160]. A reduction in lactate diffusion may then set an upper limitation to cellular lactate production. A reduction of the maximum lactate output to half of the glucose uptake bound led to a large growth reduction and an increase of the necrotic range (Fig. 37D). Mammalian cells have also been shown take up lactate if available when not under stress [161], for example to regulate the pH, and cancer cells may simply share this behavior with healthy cells. In this alternative hypothesis, the lactate uptake by oxygenated cancer cells may not be directly beneficial for growth in the hypoxic parts of the tumor.

7.5. Summary

In this chapter, I have demonstrated another use case for GEMs – simulations of metabolism using enzyme usage constraints. A diffusion model was developed to estimate the metabolite uptake constraints, which enabled simulation of the whole range of hypoxia in tumors. I here looked at the optimal behavior in the TME, assuming the cancer cells will express the required enzymes. The model recapitulated known behaviors of metabolism in tumors, showcasing the usefulness of both the diffusion model and enzyme usage constraints. Specifically, the model gave a plausible explanation to the phenomenon known as "glutamine addiction", where the use of glutamine instead of pyruvate as input to the TCA cycle leads to a higher ATP production in enzyme-limited conditions. In addition, the model predicted interesting behaviors regarding proline export through PYCR1 and PRODH (in reverse) , which matches experimental observations [25], [152]. PYCR1, one of the PYCR enzymes, is commonly overexpressed across different cancers [162]. While the role of the reverse PRODH is unclear, PYCR enzymes are predicted by the model to

help in disposing of NADH under enzyme usage limitation conditions. I also investigated metabolic collaboration scenarios between stromal cells and cancer cells in the TME. I conclude that the growth benefit for cancer cells from such scenarios is likely small, or potentially non-existent.

8. Conclusions

In this thesis, I have investigated the usefulness of genome-scale metabolic models for understanding the metabolic behavior of cancers. I have used two different strategies for this purpose: 1) use omics data to generate context-specific models and compare the active networks between different cell types or conditions and 2) perform advanced simulations including enzyme usage constraints to try to understand the behavior of cells.

The use of single cell RNA-Seq data for generating context-specific models has been central in this thesis. I have concluded that it is nearly impossible to generate complete context-specific models from the data of a single cell, while it is plausible to pool data from cells in a cell population to generate a model. However, with current technologies, typically thousands of cells are needed per cell population, and the needed pool size should be estimated for each dataset. I have also showed that the available metabolic network differs substantially across cell types, which motivates the study of their metabolism individually. While context-specific models can be generated from other types of data, scRNA-Seq clearly offers an advantage in that multiple cell types can be studied from complex organs. Although this is to some extent also possible via FACS-sorting followed by bulk RNA-Seq, the advantage of scRNA-Seq is that the method is not dependent on surface markers for defining cell populations, that the cell populations doesn't have to be defined in advance, and the high availability of public single-cell datasets.

While investigating the properties of single-cell RNA-Seq, I discovered that there is a bias across genes in scRNA-Seq data that can be estimated using copies per UMI. Under supervision by Prof. Lior Pachter I developed the BUTTERFLY method, which can be used to reduce the bias. While the method is useful, the most important part of the work was the realization that the problem, which we call "the pooled amplification paradox", exists in single-cell data and that it gives rise to for example batch effects between datasets.

In **paper V**, I performed metabolic simulations on cancer with enzyme usage constraints based on a diffusion model for setting the metabolite uptake constraints. As part of the work, I also developed a method for adding enzyme usage constraints to models, called GECKO Light. I used the model to explain metabolic behaviors such as glutamine addiction and protein secretion and showed that the metabolic collaborations between stromal cells and cancer cells are likely not important.

My three main contributions to the field are the development of a series of methods useful for genome-scale modeling, especially for investigating the tumor microenvironment, an improvement of the quantification of single-cell RNA-Seq data, and interesting discoveries around metabolism in the TME. I hope and believe that these findings and methods will prove useful both for myself in my future research but also for both the single-cell and modeling community as well as in cancer research, where I predict that mathematical modeling will become increasingly important.

9. Future perspectives

In this thesis, I have in addition to single-cell RNA-Seq studies developed methods for genome-scale metabolic modeling and applied some of them to explore the metabolism in tumors. Both the use of context-specific GEMs generated from single-cell RNA-Seq and the use of enzyme usage constraints for the study of human health and disease are in their infancy, but show great promise for the future. For example, I find it fascinating that the model, with just the reaction network and constraints on metabolite availability and enzyme usage, still can reproduce metabolic behaviors of tumors that are poorly understood. With further development in the field, I am convinced that it will be possible to get new mechanistic insights of human metabolism. So far, I did not combine the generation of context-specific GEMs from scRNA-Seq with enzyme-constrained models, which could be used to increase the accuracy of the simulations further, especially when performing simulations with multiple cell types. Likewise, I mainly focused on cancer, but there are many diseases where metabolism plays a role. For example, diabetes and Alzheimer's disease could be cases where different modeling approaches involving single-cell RNA-Seq and enzyme usage constraints could be applied.

The human body has through evolution been exposed to starvation, and the metabolism is likely therefore optimized to preserve energy. Likewise, there is sometimes a need for high energy production, which can be vital for example in muscles when exposed to danger or in expanding T cells when responding to a pathogen. Therefore, I think that the metabolic programs we encounter in the body are to a large extent optimized for a combination of these two aspects, and how much of each depends on the urgency for energy. Aerobic glycolysis, which is active in for example the Warburg effect, is one such example. This strategy potentially makes it possible to increase ATP production in a cell without a net loss of energy for the organism, as long as other cells with lower ATP needs can take up the lactate and use it (resulting in a lower ATP production in those cells). This could be compared to other pathways suggested by the model, that for example oxidate NADH without gain in ATP, which are less realistic since they lead to a net loss of energy for the organism. It seems plausible that the reason why aerobic glycolysis is used during high ATP need is because it uses less enzymatic capacity per ATP produced. The enzyme allocation needed per reaction flux plays an important part in this analysis, and together with understanding the metabolic needs of the cell and the urgency, this information holds promise to explain many behaviors in human cells. Software that adds enzyme usage constraints to metabolic models, such as GECKO Light, are an important piece of this puzzle. Unfortunately, the gene associations are still missing for many reactions in Human1, and in addition, k_{cat} values are also not available for many enzymes. However, the continuous improvements of models and k_{cat} databases improves the prediction of enzyme usage costs over time. In addition, efforts such as DLKcat [163], where unknown k_{cat} values are predicted using a deep learning approach based on substrates and protein sequence, will likely play an important role.

To narrow down the possible behaviors of different cell types in the human body, omics data plays an important role. In particular, single-cell RNA-Seq data holds promise to increase the understanding of individual cell populations. While I in this thesis only investigated cell populations identified by clustering, different approaches for defining populations are possible. One example is to define cell populations using a sliding window through a cell continuum, which may be useful for identifying metabolic switches. With

the recent arrival of spatial transcriptomics [164], even more possibilities arise. The context-specific model can then be connected to for example surrounding cells and be constrained according to an estimated metabolite availability based on spatial location.

I have in this thesis used scRNA-Seq to determine if a reaction is present or not in a cell type. While useful, such a method fails to detect many metabolic differences between cell types, for example when a pathway is overexpressed in one cell type compared to another, but still exists in both. An alternative approach could be to generate context-specific difference models by using p values from differential expression as input to a MILP, used much in the same way as in tINIT. Another alternative approach could be to penalize flux through reactions where there is poor evidence of enzyme presence, in a similar way as in the COMPASS method [104]. Single-cell RNA-Seq data contains much information, and it is evident that a simple thresholding approach per cluster and gene only uses a small part of that information content. It is therefore likely possible to develop new methods that extract more information about the metabolism in cell populations from scRNA-Seq data.

As part of this thesis, I investigated and developed a correction method for amplification biases across genes in scRNA-Seq data. I have recently learned that false positives that arise from alignment add biases to the amplification measurements, making them less reliable, especially for pseudo-alignment tools that map reads to the transcriptome only. However, we can still observe large differences in amplification across data aligned using STAR (**paper I**), suggesting that either similar problems exist with full alignment tools or that the true PCR amplification biases across genes are substantial. While RNA-Seq has existed for more than a decade, it still seems that the quantification problem is not fully solved. This is also apparent from our results in **paper I**, where the technical variation between samples is high. Tools such as BUTTERFLY, that strive to improve quantification, will likely be important in future RNA-Seq pipelines in combination with methods that effectively filter out falsely aligned reads.

Mathematical modeling is a powerful tool, and I strongly believe it will become increasingly important in future efforts to understand the metabolism in human health and disease. To understand the motivations behind metabolism with human reasoning and measurements alone is hard, and without modeling, the progress will be slower, especially when trying to understand complex behaviors. This thesis is a step in the direction of making complex modeling of human metabolism more available to scientists, and thereby ultimately assist in the effort to understand human metabolism. I foresee that in the future, metabolic modeling will be seen as a key component in studies concerning metabolism.

10. References

[1] A. Sánchez López de Nava and A. Raja, “Physiology, Metabolism,” in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2022. Accessed: Mar. 28, 2022. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK546690/

[2] I. Martínez-Reyes and N. S. Chandel, “Mitochondrial TCA cycle metabolites control physiology and disease,” *Nat Commun*, vol. 11, no. 1, Art. no. 1, Jan. 2020, doi: 10.1038/s41467-019-13668-3.

[3] Y. Chaban, E. J. Boekema, and N. V. Dudkina, “Structures of mitochondrial oxidative phosphorylation supercomplexes and mechanisms for their stabilisation,” *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, vol. 1837, no. 4, pp. 418–426, Apr. 2014, doi: 10.1016/j.bbabio.2013.10.004.

[4] S. A. Mookerjee, A. A. Gerencser, D. G. Nicholls, and M. D. Brand, “Quantifying intracellular rates of glycolytic and oxidative ATP production and consumption using extracellular flux measurements,” *J Biol Chem*, vol. 292, no. 17, pp. 7189–7207, Apr. 2017, doi: 10.1074/jbc.M116.774471.

[5] I. Marchiq and J. Pouysségur, “Hypoxia, cancer metabolism and the therapeutic benefit of targeting lactate/H+ symporters,” *J Mol Med (Berl)*, vol. 94, pp. 155–171, 2016, doi: 10.1007/s00109-015-1307-x.

[6] A. Nilsson, E. Björnson, M. Flockhart, F. J. Larsen, and J. Nielsen, “Complex I is bypassed during high intensity exercise,” *Nat Commun*, vol. 10, no. 1, Art. no. 1, Nov. 2019, doi: 10.1038/s41467-019-12934-8.

[7] B. Jiang, “Aerobic glycolysis and high level of lactate in cancer metabolism and microenvironment,” *Genes Dis*, vol. 4, no. 1, pp. 25–27, Feb. 2017, doi: 10.1016/j.gendis.2017.02.003.

[8] K. Vermeulen, D. R. Van Bockstaele, and Z. N. Berneman, “The cell cycle: a review of regulation, deregulation and therapeutic targets in cancer,” *Cell Proliferation*, vol. 36, no. 3, pp. 131–149, 2003, doi: 10.1046/j.1365-2184.2003.00266.x.

[9] J. Zheng, “Energy metabolism of cancer: Glycolysis versus oxidative phosphorylation (Review),” *Oncology Letters*, vol. 4, no. 6, pp. 1151–1157, Dec. 2012, doi: 10.3892/ol.2012.928.

[10] N. Goyal, M. Padhiary, I. A. Karimi, and Z. Zhou, “Flux measurements and maintenance energy for carbon dioxide utilization by Methanococcus maripaludis,” *Microb Cell Fact*, vol. 14, p. 146, Sep. 2015, doi: 10.1186/s12934-015-0336-z.

[11] Y. Shieh, M. Eklund, G. F. Sawaya, W. C. Black, B. S. Kramer, and L. J. Esserman, “Population-based screening for cancer: hope and hype,” *Nat Rev Clin Oncol*, vol. 13, no. 9, Art. no. 9, Sep. 2016, doi: 10.1038/nrclinonc.2016.50.

[12] S. T. Mayne, M. C. Playdon, and C. L. Rock, “Diet, nutrition, and cancer: past, present and future,” *Nat Rev Clin Oncol*, vol. 13, no. 8, Art. no. 8, Aug. 2016, doi: 10.1038/nrclinonc.2016.24.

[13] C. Sawyers, “Targeted cancer therapy,” *Nature*, vol. 432, no. 7015, pp. 294–297, Nov. 2004, doi: 10.1038/nature03095.

[14] D. Hanahan and R. A. Weinberg, “The Hallmarks of Cancer,” *Cell*, vol. 100, no. 1, pp. 57–70, Jan. 2000, doi: 10.1016/S0092-8674(00)81683-9.

[15] D. Hanahan and R. A. Weinberg, “Hallmarks of Cancer: The Next Generation,” *Cell*, vol. 144, no. 5, pp. 646–674, Mar. 2011, doi: 10.1016/j.cell.2011.02.013.

[16] J. C. Forster, W. M. Harriss-Phillips, M. J. Douglass, and E. Bezak, “A review of the development of tumor vasculature and its effects on the tumor microenvironment,” *Hypoxia*, vol. 5, p. 21, 2017, doi: 10.2147/HP.S133231.

[17] J. A. Nagy, S.-H. Chang, A. M. Dvorak, and H. F. Dvorak, “Why are tumour blood vessels abnormal and why is it important to know?,” *Br J Cancer*, vol. 100, no. 6, pp. 865–869, Mar. 2009, doi: 10.1038/sj.bjc.6604929.
[18] R. K. Jain, J. D. Martin, V. P. Chauhan, and D. G. Duda, “8 - Tumor Microenvironment: Vascular and Extravascular Compartment,” in *Abeloff’s Clinical Oncology (Sixth Edition)*, J. E. Niederhuber, J. O. Armitage, M. B. Kastan, J. H. Doroshow, and J. E. Tepper, Eds. Philadelphia: Elsevier, 2020, pp. 108-126.e7. doi: 10.1016/B978-0-323-47674-4.00008-6.
[19] M. Sefidgar, M. Soltani, K. Raahemifar, H. Bazmara, S. M. M. Nayinian, and M. Bazargan, “Effect of tumor shape, size, and tissue transport properties on drug delivery to solid tumors,” *Journal of Biological Engineering*, vol. 8, no. 1, p. 12, Jun. 2014, doi: 10.1186/1754-1611-8-12.
[20] M. Busk, J. Overgaard, and M. R. Horsman, “Imaging of Tumor Hypoxia for Radiotherapy: Current Status and Future Directions,” *Seminars in Nuclear Medicine*, vol. 50, no. 6, pp. 562–583, Nov. 2020, doi: 10.1053/j.semnuclmed.2020.05.003.
[21] A. Fortunato, A. Boddy, D. Mallo, A. Aktipis, C. C. Maley, and J. W. Pepper, “Natural Selection in Cancer Biology: From Molecular Snowflakes to Trait Hallmarks,” *Cold Spring Harb Perspect Med*, vol. 7, no. 2, p. a029652, Feb. 2017, doi: 10.1101/cshperspect.a029652.
[22] M. V. Liberti and J. W. Locasale, “The Warburg Effect: How Does it Benefit Cancer Cells?,” *Trends Biochem Sci*, vol. 41, no. 3, pp. 211–218, Mar. 2016, doi: 10.1016/j.tibs.2015.12.001.
[23] E. L. Lieu, T. Nguyen, S. Rhyne, and J. Kim, “Amino acids in cancer,” *Exp Mol Med*, vol. 52, no. 1, Art. no. 1, Jan. 2020, doi: 10.1038/s12276-020-0375-3.
[24] D. R. Wise and C. B. Thompson, “Glutamine Addiction: A New Therapeutic Target in Cancer,” *Trends Biochem Sci*, vol. 35, no. 8, pp. 427–433, Aug. 2010, doi: 10.1016/j.tibs.2010.05.003.
[25] A. Nilsson *et al.*, “Quantitative analysis of amino acid metabolism in liver cancer links glutamate excretion to nucleotide synthesis,” *PNAS*, Apr. 2020, doi: 10.1073/pnas.1919250117.
[26] R. Nilsson *et al.*, “Metabolic enzyme expression highlights a key role for MTHFD2 and the mitochondrial folate pathway in cancer,” *Nat Commun*, vol. 5, no. 1, p. 3128, Jan. 2014, doi: 10.1038/ncomms4128.
[27] R. Baghban *et al.*, “Tumor microenvironment complexity and therapeutic implications at a glance,” *Cell Communication and Signaling*, vol. 18, no. 1, p. 59, Apr. 2020, doi: 10.1186/s12964-020-0530-4.
[28] U. E. Martinez-Outschoorn, M. P. Lisanti, and F. Sotgia, “Catabolic cancer-associated fibroblasts transfer energy and biomass to anabolic cancer cells, fueling tumor growth,” *Seminars in Cancer Biology*, vol. 25, pp. 47–60, Apr. 2014, doi: 10.1016/j.semcancer.2014.01.005.
[29] S. Pavlides *et al.*, “The reverse Warburg effect: Aerobic glycolysis in cancer associated fibroblasts and the tumor stroma,” *Cell Cycle*, vol. 8, no. 23, pp. 3984–4001, Dec. 2009, doi: 10.4161/cc.8.23.10238.
[30] A. Avagliano *et al.*, “Metabolic Reprogramming of Cancer Associated Fibroblasts: The Slavery of Stromal Fibroblasts,” *Biomed Res Int*, vol. 2018, p. 6075403, 2018, doi: 10.1155/2018/6075403.
[31] J. Zhou, Z. Tang, S. Gao, C. Li, Y. Feng, and X. Zhou, “Tumor-Associated Macrophages: Recent Insights and Therapies,” *Frontiers in Oncology*, vol. 10, 2020,

Accessed: Apr. 01, 2022. [Online]. Available: https://www.frontiersin.org/article/10.3389/fonc.2020.00188
[32] J. S. Edwards and B. O. Palsson, "Systems Properties of the Haemophilus influenzaeRd Metabolic Genotype *," *Journal of Biological Chemistry*, vol. 274, no. 25, pp. 17410–17416, Jun. 1999, doi: 10.1074/jbc.274.25.17410.
[33] H. Lu *et al.*, "A consensus S. cerevisiae metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism," *Nat Commun*, vol. 10, no. 1, Art. no. 1, Aug. 2019, doi: 10.1038/s41467-019-11581-3.
[34] J. L. Robinson *et al.*, "An atlas of human metabolism," *Sci. Signal.*, vol. 13, no. 624, Mar. 2020, doi: 10.1126/scisignal.aaz1482.
[35] H. Lopes and I. Rocha, "Genome-scale modeling of yeast: chronology, applications and critical perspectives," *FEMS Yeast Research*, vol. 17, no. 5, p. fox050, Aug. 2017, doi: 10.1093/femsyr/fox050.
[36] B. Papp, B. Szappanos, and R. A. Notebaart, "Use of Genome-Scale Metabolic Models in Evolutionary Systems Biology," in *Yeast Systems Biology: Methods and Protocols*, J. I. Castrillo and S. G. Oliver, Eds. Totowa, NJ: Humana Press, 2011, pp. 483–497. doi: 10.1007/978-1-61779-173-4_27.
[37] J. D. Orth, I. Thiele, and B. Ø. Palsson, "What is flux balance analysis?," *Nature Biotechnology*, vol. 28, no. 3, Art. no. 3, Mar. 2010, doi: 10.1038/nbt.1614.
[38] Gurobi Optimization, LLC, "Gurobi Optimizer Reference Manual." 2022. [Online]. Available: https://www.gurobi.com
[39] B. J. Sánchez, C. Zhang, A. Nilsson, P.-J. Lahtvee, E. J. Kerkhoven, and J. Nielsen, "Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints," *Molecular Systems Biology*, vol. 13, no. 8, p. 935, Aug. 2017, doi: 10.15252/msb.20167411.
[40] A. Chang *et al.*, "BRENDA, the ELIXIR core data resource in 2021: new developments and updates," *Nucleic Acids Research*, vol. 49, no. D1, pp. D498–D508, Jan. 2021, doi: 10.1093/nar/gkaa1025.
[41] B. Srinivasan, "A guide to the Michaelis–Menten equation: steady state and beyond," *The FEBS Journal*, vol. n/a, no. n/a, doi: 10.1111/febs.16124.
[42] H. Wang *et al.*, "Genome-scale metabolic network reconstruction of model animals as a platform for translational research," *PNAS*, vol. 118, no. 30, Jul. 2021, doi: 10.1073/pnas.2102344118.
[43] I. Domenzain *et al.*, "Reconstruction of a catalogue of genome-scale metabolic models with enzymatic constraints using GECKO 2.0," *bioRxiv*, p. 2021.03.05.433259, Mar. 2021, doi: 10.1101/2021.03.05.433259.
[44] A. Schultz and A. A. Qutub, "Reconstruction of Tissue-Specific Metabolic Networks Using CORDA," *PLOS Computational Biology*, vol. 12, no. 3, p. e1004808, Mar. 2016, doi: 10.1371/journal.pcbi.1004808.
[45] S. A. Becker and B. O. Palsson, "Context-Specific Metabolic Networks Are Consistent with Experiments," *PLOS Computational Biology*, vol. 4, no. 5, p. e1000082, maj 2008, doi: 10.1371/journal.pcbi.1000082.
[46] R. Agren, A. Mardinoglu, A. Asplund, C. Kampf, M. Uhlen, and J. Nielsen, "Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling," *Molecular Systems Biology*, vol. 10, no. 3, p. 721, Mar. 2014, doi: 10.1002/msb.145122.
[47] M. Karas, D. Bachmann, U. Bahr, and F. Hillenkamp, "Matrix-assisted ultraviolet laser desorption of non-volatile compounds," *International Journal of Mass*

Spectrometry and Ion Processes, vol. 78, pp. 53–68, Sep. 1987, doi: 10.1016/0168-1176(87)87041-6.
[48] R. Aebersold and M. Mann, "Mass spectrometry-based proteomics," *Nature*, vol. 422, no. 6928, Art. no. 6928, Mar. 2003, doi: 10.1038/nature01511.
[49] M. Uhlén *et al.*, "Tissue-based map of the human proteome," *Science*, vol. 347, no. 6220, p. 1260419, Jan. 2015, doi: 10.1126/science.1260419.
[50] M. Labib and S. O. Kelley, "Single-cell analysis targeting the proteome," *Nat Rev Chem*, vol. 4, no. 3, Art. no. 3, Mar. 2020, doi: 10.1038/s41570-020-0162-7.
[51] R. Lowe, N. Shirley, M. Bleackley, S. Dolan, and T. Shafee, "Transcriptomics technologies," *PLoS Comput Biol*, vol. 13, no. 5, p. e1005457, May 2017, doi: 10.1371/journal.pcbi.1005457.
[52] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," *Nat Rev Genet*, vol. 10, no. 1, pp. 57–63, Jan. 2009, doi: 10.1038/nrg2484.
[53] M. Sultan *et al.*, "A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome," *Science*, vol. 321, no. 5891, pp. 956–960, Aug. 2008, doi: 10.1126/science.1160342.
[54] Q. Pan, O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe, "Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing," *Nat Genet*, vol. 40, no. 12, Art. no. 12, Dec. 2008, doi: 10.1038/ng.259.
[55] S. Aldridge and S. A. Teichmann, "Single cell transcriptomics comes of age," *Nat Commun*, vol. 11, no. 1, Art. no. 1, Aug. 2020, doi: 10.1038/s41467-020-18158-5.
[56] F. Edfors *et al.*, "Gene-specific correlation of RNA and protein levels in human cells and tissues," *Mol Syst Biol*, vol. 12, no. 10, p. 883, Oct. 2016, doi: 10.15252/msb.20167144.
[57] S. Sidoli, K. Kulej, and B. A. Garcia, "Why proteomics is not the new genomics and the future of mass spectrometry in cell biology," *Journal of Cell Biology*, vol. 216, no. 1, pp. 21–24, Dec. 2016, doi: 10.1083/jcb.201612010.
[58] A. Adil, V. Kumar, A. T. Jan, and M. Asger, "Single-Cell Transcriptomics: Current Methods and Challenges in Data Acquisition and Analysis," *Frontiers in Neuroscience*, vol. 15, 2021, Accessed: Apr. 01, 2022. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnins.2021.591122
[59] J. Picot, C. L. Guerin, C. Le Van Kim, and C. M. Boulanger, "Flow cytometry: retrospective, fundamentals and recent instrumentation," *Cytotechnology*, vol. 64, no. 2, pp. 109–130, Mar. 2012, doi: 10.1007/s10616-011-9415-0.
[60] D. Seiter *et al.*, "Quantity and location of the tumor cells in a biopsy specimen," *Journal of Nuclear Medicine*, vol. 59, no. supplement 1, pp. 248–248, May 2018.
[61] M. Griffith, J. R. Walker, N. C. Spies, B. J. Ainscough, and O. L. Griffith, "Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud," *PLoS Comput Biol*, vol. 11, no. 8, Aug. 2015, doi: 10.1371/journal.pcbi.1004393.
[62] S. Martín-Alonso, E. Frutos-Beltrán, and L. Menéndez-Arias, "Reverse Transcriptase: From Transcriptomics to Genome Editing," *Trends in Biotechnology*, vol. 39, no. 2, pp. 194–210, Feb. 2021, doi: 10.1016/j.tibtech.2020.06.008.
[63] "The polymerase chain reaction: An overview and development of diagnostic PCR protocols at the LCDC," *Can J Infect Dis*, vol. 2, no. 2, pp. 89–91, 1991.
[64] D. Risso, K. Schwartz, G. Sherlock, and S. Dudoit, "GC-Content Normalization for RNA-Seq Data," *BMC Bioinformatics*, vol. 12, no. 1, p. 480, Dec. 2011, doi: 10.1186/1471-2105-12-480.

[65] W. E. Johnson, C. Li, and A. Rabinovic, “Adjusting batch effects in microarray expression data using empirical Bayes methods,” *Biostatistics*, vol. 8, no. 1, pp. 118–127, Jan. 2007, doi: 10.1093/biostatistics/kxj037.
[66] J. T. Leek, W. E. Johnson, H. S. Parker, A. E. Jaffe, and J. D. Storey, “The sva package for removing batch effects and other unwanted variation in high-throughput experiments,” *Bioinformatics*, vol. 28, no. 6, pp. 882–883, Mar. 2012, doi: 10.1093/bioinformatics/bts034.
[67] H. T. N. Tran *et al.*, “A benchmark of batch-effect correction methods for single-cell RNA sequencing data,” *Genome Biology*, vol. 21, no. 1, p. 12, Jan. 2020, doi: 10.1186/s13059-019-1850-9.
[68] G. K. Marinov *et al.*, “From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing,” *Genome Res.*, vol. 24, no. 3, pp. 496–510, Mar. 2014, doi: 10.1101/gr.161034.113.
[69] E. Tunnacliffe and J. R. Chubb, “What Is a Transcriptional Burst?,” *Trends in Genetics*, vol. 36, no. 4, pp. 288–297, Apr. 2020, doi: 10.1016/j.tig.2020.01.003.
[70] T. Kivioja *et al.*, “Counting absolute numbers of molecules using unique molecular identifiers,” *Nat Methods*, vol. 9, no. 1, Art. no. 1, Jan. 2012, doi: 10.1038/nmeth.1778.
[71] Q. H. Nguyen, N. Pervolarakis, K. Nee, and K. Kessenbrock, “Experimental Considerations for Single-Cell RNA Sequencing Approaches,” *Front Cell Dev Biol*, vol. 6, Sep. 2018, doi: 10.3389/fcell.2018.00108.
[72] “Chromium Automated Single Cell 3′ Reagent Kits v3.1 User Guide • Rev B.” 10x Genomics. Accessed: Aug. 12, 2020. [Online]. Available: https://assets.ctfassets.net/an68im79xiti/72l3yNeOYCOi3sNIAeDPQj/ceadc80df6f50cbb8d41485563145460/CG000286_ChromiumSingleCell3__v3.1_Automation_UG_RevB.pdf
[73] G. X. Y. Zheng *et al.*, “Massively parallel digital transcriptional profiling of single cells,” *Nature Communications*, vol. 8, p. 14049, Jan. 2017, doi: 10.1038/ncomms14049.
[74] E. Z. Macosko *et al.*, “Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets,” *Cell*, vol. 161, no. 5, pp. 1202–1214, May 2015, doi: 10.1016/j.cell.2015.05.002.
[75] S. Picelli, O. R. Faridani, Å. K. Björklund, G. Winberg, S. Sagasser, and R. Sandberg, “Full-length RNA-seq from single cells using Smart-seq2,” *Nat Protoc*, vol. 9, no. 1, pp. 171–181, Jan. 2014, doi: 10.1038/nprot.2014.006.
[76] M. Hagemann-Jensen *et al.*, “Single-cell RNA counting at allele and isoform resolution using Smart-seq3,” *Nature Biotechnology*, vol. 38, no. 6, Art. no. 6, Jun. 2020, doi: 10.1038/s41587-020-0497-0.
[77] P. J. A. Cock, C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice, “The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants,” *Nucleic Acids Res*, vol. 38, no. 6, pp. 1767–1771, Apr. 2010, doi: 10.1093/nar/gkp1137.
[78] A. Dobin *et al.*, “STAR: ultrafast universal RNA-seq aligner,” *Bioinformatics*, vol. 29, no. 1, pp. 15–21, Jan. 2013, doi: 10.1093/bioinformatics/bts635.
[79] D. Kim, J. M. Paggi, C. Park, C. Bennett, and S. L. Salzberg, “Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype,” *Nat Biotechnol*, vol. 37, no. 8, Art. no. 8, Aug. 2019, doi: 10.1038/s41587-019-0201-4.

[80] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter, “Near-optimal probabilistic RNA-seq quantification,” *nbt*, vol. 34, no. 5, pp. 525–527, May 2016, doi: 10.1038/nbt.3519.
[81] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford, “Salmon provides fast and bias-aware quantification of transcript expression,” *Nature Methods*, vol. 14, no. 4, Art. no. 4, Apr. 2017, doi: 10.1038/nmeth.4197.
[82] Y. Du, Q. Huang, C. Arisdakessian, and L. X. Garmire, “Evaluation of STAR and Kallisto on Single Cell RNA-Seq Data Alignment,” *G3: Genes, Genomes, Genetics*, vol. 10, no. 5, pp. 1775–1783, May 2020, doi: 10.1534/g3.120.401160.
[83] P. Melsted *et al.*, “Modular, efficient and constant-memory single-cell RNA-seq preprocessing,” *Nature Biotechnology*, pp. 1–6, Apr. 2021, doi: 10.1038/s41587-021-00870-2.
[84] A. Srivastava, L. Malik, T. Smith, I. Sudbery, and R. Patro, “Alevin efficiently estimates accurate gene abundances from dscRNA-seq data,” *Genome Biology*, vol. 20, no. 1, p. 65, Mar. 2019, doi: 10.1186/s13059-019-1670-y.
[85] B. Kaminow, D. Yunusov, and A. Dobin, “STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data.” bioRxiv, p. 2021.05.05.442755, May 05, 2021. doi: 10.1101/2021.05.05.442755.
[86] M. I. Love, W. Huber, and S. Anders, “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2,” *Genome Biology*, vol. 15, p. 550, Dec. 2014, doi: 10.1186/s13059-014-0550-8.
[87] V. Svensson, “Droplet scRNA-seq is not zero-inflated,” *Nat Biotechnol*, pp. 1–4, Jan. 2020, doi: 10.1038/s41587-019-0379-5.
[88] F. W. Townes, S. C. Hicks, M. J. Aryee, and R. A. Irizarry, “Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model,” *Genome Biology*, vol. 20, no. 1, p. 295, Dec. 2019, doi: 10.1186/s13059-019-1861-6.
[89] P. Qiu, “Embracing the dropouts in single-cell RNA-seq analysis,” *Nat Commun*, vol. 11, no. 1, Art. no. 1, Mar. 2020, doi: 10.1038/s41467-020-14976-9.
[90] T. Stuart *et al.*, “Comprehensive Integration of Single-Cell Data,” *Cell*, vol. 177, no. 7, pp. 1888-1902.e21, Jun. 2019, doi: 10.1016/j.cell.2019.05.031.
[91] C. Hafemeister and R. Satija, “Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression,” *Genome Biology*, vol. 20, no. 1, p. 296, Dec. 2019, doi: 10.1186/s13059-019-1874-1.
[92] L. McInnes, J. Healy, N. Saul, and L. Großberger, “UMAP: Uniform Manifold Approximation and Projection,” *Journal of Open Source Software*, vol. 3, no. 29, p. 861, Sep. 2018, doi: 10.21105/joss.00861.
[93] M. D. Young and S. Behjati, “SoupX removes ambient RNA contamination from droplet based single-cell RNA sequencing data,” *bioRxiv*, p. 303727, Feb. 2020, doi: 10.1101/303727.
[94] C. S. McGinnis, L. M. Murrow, and Z. J. Gartner, “DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors,” *Cell Systems*, vol. 8, no. 4, pp. 329-337.e4, Apr. 2019, doi: 10.1016/j.cels.2019.03.003.
[95] A. A. AlJanahi, M. Danielsen, and C. E. Dunbar, “An Introduction to the Analysis of Single-Cell RNA-Sequencing Data,” *Mol Ther Methods Clin Dev*, vol. 10, pp. 189–196, Aug. 2018, doi: 10.1016/j.omtm.2018.07.003.
[96] T. Ilicic *et al.*, “Classification of low quality cells from single-cell RNA-seq data,” *Genome Biol*, vol. 17, Feb. 2016, doi: 10.1186/s13059-016-0888-1.

[97] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, “Mapping and quantifying mammalian transcriptomes by RNA-Seq,” *Nat. Methods*, vol. 5, no. 7, pp. 621–628, Jul. 2008, doi: 10.1038/nmeth.1226.
[98] M. D. Robinson and A. Oshlack, “A scaling normalization method for differential expression analysis of RNA-seq data,” *Genome Biology*, vol. 11, p. R25, Mar. 2010, doi: 10.1186/gb-2010-11-3-r25.
[99] J. K. Pickrell *et al.*, “Understanding mechanisms underlying human gene expression variation with RNA sequencing,” *Nature*, vol. 464, no. 7289, pp. 768–772, Apr. 2010, doi: 10.1038/nature08872.
[100] A. T. L. Lun, K. Bach, and J. C. Marioni, “Pooling across cells to normalize single-cell RNA sequencing data with many zero counts,” *Genome Biology*, vol. 17, no. 1, p. 75, Apr. 2016, doi: 10.1186/s13059-016-0947-7.
[101] A. T. L. Lun, D. J. McCarthy, and J. C. Marioni, “A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor,” *F1000Research*, vol. 5, p. 2122, Oct. 2016, doi: 10.12688/f1000research.9501.2.
[102] V. Y. Kiselev, T. S. Andrews, and M. Hemberg, “Challenges in unsupervised clustering of single-cell RNA-seq data,” *Nature Reviews Genetics*, vol. 20, no. 5, Art. no. 5, May 2019, doi: 10.1038/s41576-018-0088-9.
[103] N. Alghamdi *et al.*, “A graph neural network model to estimate cell-wise metabolic flux using single-cell RNA-seq data,” *Genome Res.*, vol. 31, no. 10, pp. 1867–1884, Oct. 2021, doi: 10.1101/gr.271205.120.
[104] A. Wagner *et al.*, “Metabolic modeling of single Th17 cells reveals regulators of autoimmunity,” *Cell*, vol. 184, no. 16, pp. 4168-4185.e21, Aug. 2021, doi: 10.1016/j.cell.2021.05.045.
[105] L. S. Yilmaz, X. Li, S. Nanda, B. Fox, F. Schroeder, and A. J. M. Walhout, “Modeling tissue-relevant Caenorhabditis elegans metabolism at network, pathway, reaction, and metabolite levels,” *Molecular Systems Biology*, vol. 16, no. 10, p. e9649, Oct. 2020, doi: 10.15252/msb.20209649.
[106] Y. Zhang, M. S. Kim, E. Nguyen, and D. M. Taylor, “Modeling metabolic variation with single-cell expression data,” *bioRxiv*, p. 2020.01.28.923680, Jan. 2020, doi: 10.1101/2020.01.28.923680.
[107] M. A. Oberhardt and E. P. Gianchandani, “Genome-scale modeling and human disease: an overview,” *Frontiers in Physiology*, vol. 5, 2015, Accessed: Apr. 04, 2022. [Online]. Available: https://www.frontiersin.org/article/10.3389/fphys.2014.00527
[108] C. B. Steen, C. L. Liu, A. A. Alizadeh, and A. M. Newman, “Profiling Cell Type Abundance and Expression in Bulk Tissues with CIBERSORTx,” *Methods Mol Biol*, vol. 2117, pp. 135–157, 2020, doi: 10.1007/978-1-0716-0301-7_7.
[109] G. E. Hoffman and E. E. Schadt, “variancePartition: interpreting drivers of variation in complex gene expression studies,” *BMC Bioinformatics*, vol. 17, no. 1, p. 483, Nov. 2016, doi: 10.1186/s12859-016-1323-z.
[110] F. Avila Cobos, J. Alquicira-Hernandez, J. E. Powell, P. Mestdagh, and K. De Preter, “Benchmarking of cell type deconvolution pipelines for transcriptomics data,” *Nat Commun*, vol. 11, no. 1, Art. no. 1, Nov. 2020, doi: 10.1038/s41467-020-19015-1.
[111] P. A. C. ’t Hoen *et al.*, “Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories,” *Nature Biotechnology*, vol. 31, no. 11, Art. no. 11, Nov. 2013, doi: 10.1038/nbt.2702.

[112] Y. Zhang, G. Parmigiani, and W. E. Johnson, “ComBat-seq: batch effect adjustment for RNA-seq count data,” *NAR Genomics and Bioinformatics*, vol. 2, no. 3, p. lqaa078, Sep. 2020, doi: 10.1093/nargab/lqaa078.
[113] D. Lambrechts *et al.*, “Phenotype molding of stromal cells in the lung tumor microenvironment,” *Nature Medicine*, vol. 24, no. 8, pp. 1277–1289, Aug. 2018, doi: 10.1038/s41591-018-0096-5.
[114] Š. Konjar and M. Veldhoen, “Dynamic Metabolic State of Tissue Resident CD8 T Cells,” *Frontiers in Immunology*, vol. 10, 2019, doi: 10.3389/fimmu.2019.01683.
[115] J. Ding *et al.*, “Systematic comparison of single-cell and single-nucleus RNA-sequencing methods,” *Nature Biotechnology*, pp. 1–10, Apr. 2020, doi: 10.1038/s41587-020-0465-8.
[116] Y. Benjamini and T. P. Speed, “Summarizing and correcting the GC content bias in high-throughput sequencing,” *Nucleic Acids Res*, vol. 40, no. 10, p. e72, May 2012, doi: 10.1093/nar/gks001.
[117] I. J. Good and G. H. Toulmin, “The number of new species, and the increase in population coverage, when a sample is increased.,” *Biometrika*, vol. 43, no. 1–2, pp. 45–63, Jun. 1956, doi: 10.1093/biomet/43.1-2.45.
[118] R. A. Fisher, A. S. Corbet, and C. B. Williams, “The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population,” *Journal of Animal Ecology*, vol. 12, no. 1, pp. 42–58, 1943, doi: 10.2307/1411.
[119] T. Daley and A. D. Smith, “Predicting the molecular complexity of sequencing libraries,” *Nat Methods*, vol. 10, no. 4, pp. 325–327, Apr. 2013, doi: 10.1038/nmeth.2375.
[120] C. Deng, T. Daley, and A. D. Smith, “Applications of species accumulation curves in large-scale biological data analysis,” *Quant Biol*, vol. 3, no. 3, pp. 135–144, Sep. 2015, doi: 10.1007/s40484-015-0049-7.
[121] C. Deng, T. Daley, P. Calabrese, J. Ren, and A. D. Smith, “Estimating the number of species to attain sufficient representation in a random sample,” *arXiv:1607.02804 [stat]*, May 2018, Accessed: Apr. 21, 2020. [Online]. Available: http://arxiv.org/abs/1607.02804
[122] A. Srivastava *et al.*, “Alignment and mapping methodology influence transcript abundance estimation,” *Genome Biology*, vol. 21, no. 1, p. 239, Sep. 2020, doi: 10.1186/s13059-020-02151-8.
[123] M. Ghandi *et al.*, “Next-generation characterization of the Cancer Cell Line Encyclopedia,” *Nature*, vol. 569, no. 7757, Art. no. 7757, May 2019, doi: 10.1038/s41586-019-1186-3.
[124] R. M. Meyers *et al.*, “Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells,” *Nat Genet*, vol. 49, no. 12, Art. no. 12, Dec. 2017, doi: 10.1038/ng.3984.
[125] L. J. Carithers *et al.*, “A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project,” *Biopreserv Biobank*, vol. 13, no. 5, pp. 311–319, Oct. 2015, doi: 10.1089/bio.2015.0032.
[126] J. W. Squair *et al.*, “Confronting false discoveries in single-cell differential expression,” *Nat Commun*, vol. 12, no. 1, Art. no. 1, Sep. 2021, doi: 10.1038/s41467-021-25960-2.
[127] A. S. Booeshaghi *et al.*, “Isoform cell type specificity in the mouse primary motor cortex,” Mar. 2020. doi: 10.1101/2020.03.05.977991.

[128] L. J. Falomir-Lockhart, G. F. Cavazzutti, E. Giménez, and A. M. Toscani, “Fatty Acid Signaling Mechanisms in Neural Cells: Fatty Acid Receptors,” *Frontiers in Cellular Neuroscience*, vol. 13, 2019, Accessed: Mar. 06, 2022. [Online]. Available: https://www.frontiersin.org/article/10.3389/fncel.2019.00162
[129] R. Moretti and P. Caruso, “The Controversial Role of Homocysteine in Neurology: From Labs to Clinical Practice,” *International Journal of Molecular Sciences*, vol. 20, no. 1, Art. no. 1, Jan. 2019, doi: 10.3390/ijms20010231.
[130] N. Kim *et al.*, “Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma,” *Nat Commun*, vol. 11, no. 1, p. 2285, May 2020, doi: 10.1038/s41467-020-16164-1.
[131] V. Petrova, M. Annicchiarico-Petruzzelli, G. Melino, and I. Amelio, “The hypoxic tumour microenvironment,” *Oncogenesis*, vol. 7, no. 1, Art. no. 1, Jan. 2018, doi: 10.1038/s41389-017-0011-9.
[132] J. Fu, M. Yu, W. Xu, and S. Yu, “Research Progress of Bile Acids in Cancer,” *Frontiers in Oncology*, vol. 11, 2022, Accessed: Mar. 07, 2022. [Online]. Available: https://www.frontiersin.org/article/10.3389/fonc.2021.778258
[133] C. Frezza *et al.*, “Haem oxygenase is synthetically lethal with the tumour suppressor fumarate hydratase,” *Nature*, vol. 477, no. 7363, Art. no. 7363, Sep. 2011, doi: 10.1038/nature10363.
[134] P. S. Bekiaris and S. Klamt, “Automatic construction of metabolic models with enzyme constraints,” *BMC Bioinformatics*, vol. 21, no. 1, p. 19, Jan. 2020, doi: 10.1186/s12859-019-3329-9.
[135] S. Opdam, A. Richelle, B. Kellman, S. Li, D. C. Zielinski, and N. E. Lewis, “A Systematic Evaluation of Methods for Tailoring Genome-Scale Metabolic Models,” *Cell Syst*, vol. 4, no. 3, pp. 318-329.e6, Mar. 2017, doi: 10.1016/j.cels.2017.01.010.
[136] D. G. Kilburn, M. D. Lilly, and F. C. Webb, “The Energetics of Mammalian Cell Growth,” *Journal of Cell Science*, vol. 4, no. 3, pp. 645–654, May 1969.
[137] S. Harada *et al.*, “Reliability of plasma polar metabolite concentrations in a large-scale cohort study using capillary electrophoresis-mass spectrometry,” *PLOS ONE*, vol. 13, no. 1, p. e0191230, Jan. 2018, doi: 10.1371/journal.pone.0191230.
[138] E. N. Hoogenboezem and C. L. Duvall, “Harnessing Albumin as a Carrier for Cancer Therapies,” *Adv Drug Deliv Rev*, vol. 130, pp. 73–89, May 2018, doi: 10.1016/j.addr.2018.07.011.
[139] O. Quehenberger *et al.*, “Lipidomics reveals a remarkable diversity of lipids in human plasma1[S],” *Journal of Lipid Research*, vol. 51, no. 11, pp. 3299–3305, Nov. 2010, doi: 10.1194/jlr.M009449.
[140] O. Siggaard-andersen, I. H. Gøthgen, P. D. Wimberley, and N. Fogh-andersen, “The oxygen status of the arterial blood revised: Relevant oxygen parameters for monitoring the arterial oxygen availability,” *Scandinavian Journal of Clinical and Laboratory Investigation*, vol. 50, no. sup203, pp. 17–28, Jan. 1990, doi: 10.3109/00365519009087488.
[141] X. Zhang, C.-G. Li, C.-H. Ye, and M.-L. Liu, “Determination of Molecular Self-Diffusion Coefficient Using Multiple Spin-Echo NMR Spectroscopy with Removal of Convection and Background Gradient Artifacts,” *Anal. Chem.*, vol. 73, no. 15, pp. 3528–3534, Aug. 2001, doi: 10.1021/ac0101104.
[142] S. R. Chary and R. K. Jain, “Direct measurement of interstitial convection and diffusion of albumin in normal and neoplastic tissues by fluorescence photobleaching,” *PNAS*, vol. 86, no. 14, pp. 5385–5389, Jul. 1989, doi: 10.1073/pnas.86.14.5385.

[143] T. K. Goldstick, V. T. Ciuryla, and L. Zuckerman, “Diffusion of oxygen in plasma and blood,” *Adv Exp Med Biol*, vol. 75, pp. 183–190, 1976, doi: 10.1007/978-1-4684-3273-2_23.
[144] D. P. Valencia and F. J. González, “Estimation of diffusion coefficients by using a linear correlation between the diffusion coefficient and molecular weight,” *Journal of Electroanalytical Chemistry*, vol. 681, pp. 121–126, Aug. 2012, doi: 10.1016/j.jelechem.2012.06.013.
[145] J. Fan *et al.*, “Glutamine-driven oxidative phosphorylation is a major ATP source in transformed mammalian cells in both normoxia and hypoxia,” *Molecular Systems Biology*, vol. 9, no. 1, p. 712, Jan. 2013, doi: 10.1038/msb.2013.65.
[146] E. P. Seidlitz, M. K. Sharma, Z. Saikali, M. Ghert, and G. Singh, “Cancer cell lines release glutamate into the extracellular environment,” *Clin Exp Metastasis*, vol. 26, no. 7, pp. 781–787, 2009, doi: 10.1007/s10585-009-9277-4.
[147] T. Takano, J. H.-C. Lin, G. Arcuino, Q. Gao, J. Yang, and M. Nedergaard, “Glutamate release promotes growth of malignant gliomas,” *Nat Med*, vol. 7, no. 9, pp. 1010–1015, Sep. 2001, doi: 10.1038/nm0901-1010.
[148] F. Lange, J. Hörnschemeyer, and T. Kirschstein, “Glutamatergic Mechanisms in Glioblastoma and Tumor-Associated Epilepsy,” *Cells*, vol. 10, no. 5, p. 1226, May 2021, doi: 10.3390/cells10051226.
[149] C. Chinopoulos and T. N. Seyfried, “Mitochondrial Substrate-Level Phosphorylation as Energy Source for Glioblastoma: Review and Hypothesis,” *ASN Neuro*, vol. 10, p. 1759091418818261, Jan. 2018, doi: 10.1177/1759091418818261.
[150] Y. Wang *et al.*, “Coordinative metabolism of glutamine carbon and nitrogen in proliferating cancer cells under hypoxia,” *Nat Commun*, vol. 10, no. 1, p. 201, Jan. 2019, doi: 10.1038/s41467-018-08033-9.
[151] C. Corbet and O. Feron, “Tumour acidosis: from the passenger to the driver’s seat,” *Nat Rev Cancer*, vol. 17, no. 10, Art. no. 10, Oct. 2017, doi: 10.1038/nrc.2017.77.
[152] M. Jain *et al.*, “Metabolite profiling identifies a key role for glycine in rapid cancer cell proliferation,” *Science*, vol. 336, no. 6084, pp. 1040–1044, May 2012, doi: 10.1126/science.1218595.
[153] C. L. Quinlan, A. L. Orr, I. V. Perevoshchikova, J. R. Treberg, B. A. Ackrell, and M. D. Brand, “Mitochondrial Complex II Can Generate Reactive Oxygen Species at High Rates in Both the Forward and Reverse Reactions*,” *Journal of Biological Chemistry*, vol. 287, no. 32, pp. 27255–27264, Aug. 2012, doi: 10.1074/jbc.M112.374629.
[154] J. G. Jung and A. Le, “Targeting Metabolic Cross Talk Between Cancer Cells and Cancer-Associated Fibroblasts,” in *The Heterogeneity of Cancer Metabolism*, A. Le, Ed. Cham: Springer International Publishing, 2021, pp. 205–214. doi: 10.1007/978-3-030-65768-0_15.
[155] C. M. Sousa *et al.*, “Pancreatic stellate cells support tumour metabolism through autophagic alanine secretion,” *Nature*, vol. 536, no. 7617, pp. 479–483, Aug. 2016, doi: 10.1038/nature19084.
[156] P. Sonveaux *et al.*, “Targeting lactate-fueled respiration selectively kills hypoxic tumor cells in mice,” *J Clin Invest*, vol. 118, no. 12, pp. 3930–3942, Dec. 2008, doi: 10.1172/JCI36843.
[157] K. G. de la Cruz-López, L. J. Castro-Muñoz, D. O. Reyes-Hernández, A. García-Carrancá, and J. Manzo-Merino, “Lactate in the Regulation of Tumor Microenvironment and Therapeutic Approaches,” *Front Oncol*, vol. 9, p. 1143, Nov. 2019, doi: 10.3389/fonc.2019.01143.

[158] P. Swietach, R. D. Vaughan-Jones, A. L. Harris, and A. Hulikova, “The chemistry, physiology and pathology of pH in cancer,” *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 369, no. 1638, p. 20130099, Mar. 2014, doi: 10.1098/rstb.2013.0099.
[159] V. Estrella *et al.*, “Acidity Generated by the Tumor Microenvironment Drives Local Invasion,” *Cancer Res*, vol. 73, no. 5, pp. 1524–1535, Mar. 2013, doi: 10.1158/0008-5472.CAN-12-2796.
[160] R. D. Vaughan-Jones and M. L. Wu, “Extracellular H+ inactivation of Na(+)-H+ exchange in the sheep cardiac Purkinje fibre,” *J Physiol*, vol. 428, pp. 441–466, Sep. 1990, doi: 10.1113/jphysiol.1990.sp018221.
[161] S. Hui *et al.*, “Glucose feeds the TCA cycle via circulating lactate,” *Nature*, vol. 551, no. 7678, pp. 115–118, Nov. 2017, doi: 10.1038/nature24057.
[162] L. Burke *et al.*, “The Janus-like role of proline metabolism in cancer,” *Cell Death Discov.*, vol. 6, no. 1, pp. 1–17, Oct. 2020, doi: 10.1038/s41420-020-00341-8.
[163] F. Li *et al.*, “Deep learning based kcat prediction enables improved enzyme constrained model reconstruction.” bioRxiv, p. 2021.08.06.455417, Aug. 08, 2021. doi: 10.1101/2021.08.06.455417.
[164] A. Rao, D. Barkley, G. S. França, and I. Yanai, “Exploring tissue architecture using spatial transcriptomics,” *Nature*, vol. 596, no. 7871, Art. no. 7871, Aug. 2021, doi: 10.1038/s41586-021-03634-9.