ABSTRACT

Title of dissertation: DEVELOPING COMPUTATIONAL TOOLS FOR STUDYING CANCER METABOLISM AND GENOMICS

Rotem Katzir-Sheratzki, Doctor of Philosophy, 2021

Dissertation directed by: Professor Eytan Ruppin
Department of Computer Science

The interplay between different genomic and epigenomic alterations lead to different prognoses in cancer patients. Advances in high-throughput technologies, like gene expression profiling, next-generation sequencing, proteomics, and fluxomics, have enabled detailed molecular characterization of various tumors, yet studying this interplay is a complex computational problem.

Here we set to develop computational approaches to identify and study emerging challenges in cancer metabolism and genomics. We focus on three research questions, addressed by different computational approaches: (1) What is the set of metabolic interactions in cancer metabolism? To this end we generated a computational framework that quantitatively predicts synthetic dosage lethal (SDL) interactions in human metabolism, by developing a new algorithmic-modeling approach. SDLs offer a promising way to selectively kill cancer cells by targeting the SDL partners of activated oncogenes in tumors, which are often difficult to target directly. (2) What is the landscape of metabolic regulation in breast cancer? To this

end we established a new framework that utilizes different data types to perform multi-omics data integration and flux prediction, by incorporating machine learning techniques with Genome Scale Metabolic Modeling (GSMM). This enabled us to study the regulation of breast cancer cell line under different growth conditions, from multiple omics data. (3) What is the power of somatic mutations derived from RNA in estimating the tumor mutational burden? Here we develop a new tool to detect somatic mutations from RNA sequencing data without a matched-normal sample. To this end we developed a machine learning pipeline that takes as input a list of single nucleotide variants and classifies them as either somatic or germline, based on read-level features as well as position-specific variant statistics and common germline databases. We showed that detecting somatic mutations directly from RNA enables the identification of expressed mutations, and therefore represent a more relevant metric in estimating the tumor mutational burden, which is significantly associated with patient survival.

In sum, my work has been focused around developing computational methods to tackle different research questions in cancer metabolism and genomics, utilizing various types of omics data and a variety of computational approaches. These methods provide new solutions to some important computational challenges, and their applications help to generate promising leads for cancer research, and can be utilized in many future applications, analyzing novel and existing datasets.

DEVELOPING COMPUTATIONAL TOOLS FOR STUDYING METABOLISM AND GENOMICS

by

Rotem Katzir-Sheratzki

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2021

Advisory Committee:
Dr. Eytan Ruppin, Advisor, Co-Chair
Dr. James A. Reggia, Co-Chair
Dr. Najib El-Sayed
Dr. Hector Corrada Bravo
Dr. Max Leiserson

Preface

During the last years, my research was driven by my interest in the development of computational tools that harness different types of genomic data in order to study cancer metabolism and genomics. In this dissertation I present three computational approaches, designed to answer three different research questions. Each computational approach presented utilizes computational tools, designed to particularly answer the research question considering the relevant data in availability.

Initially, I was fascinated by the clinical potential of genetic interactions in cancer metabolism, and more specifically by a new concept called "synthetic dosage lethality" (SDL). I built an algorithmic approach to identify such interactions by employing Genome Scale Metabolic Modeling (GSMM) that enables perturbation simulations, and measured the effect of such perturbations on cell proliferation.

Completion of work described above, had stimulated my interest in studying the interactions between different data types, and understanding metabolic regulation. Therefore, I set to develop a computational approach to study metabolic regulation in cancer cells. By incorporating GSMMs, machine learning algorithms and data representation methods, I developed a pipeline to enable the integration of different genomic data types. This work enabled me to systematically chart the different layers of regulation in breast cancer cells, by predicting enzymes and pathways regulation levels, and laid a conceptual and computational basis for mapping metabolic regulation in additional cancers.

Finally, I was curious to study cancer immunity. Cancer is caused by the breakdown of the controls that regulate cells, often the result of mutations - changes in

the DNA/RNA sequence. Somatic mutations are the most common cause of cancer. Their analysis enabled the identification of driver mutations and contributed to cancer prognosis and interventions, through the development of biomarkers and targeted therapies. Recent studies indicate that a high tumor mutational burden (TMB) results in more neoantigens, increasing chances for T cell recognition, and has been recently approved by the FDA as a marker for immune checkpoint blockage therapy responses. To investigate the clinical utility of RNA-based mutations, I developed the first machine learning model that identifies the somatic mutations from RNA-seq of a given tumor sample, without its matched normal. This novel work allows the analysis of mutations from tumor RNA-sequencing alone, thus facilitating a profound investigation of mutations in numerous datasets, which was not feasible before. In the context of immunotherapy, our results demonstrate that estimating TMB from RNA is of a higher, or similar, predictive power, compare to TMB estimated from DNA.

In conclusion, the work presented in this thesis provides multiple computational approaches, designed to address different questions in cancer research. Working in close collaboration with different experimental labs on some of these projects provided me with a better understanding of the research questions, which motivated the computational approaches I developed to answer each question. I believe that much of this work can be used in future studies to advance the development of treatments and improve clinical decision-making for cancer patients.

To my parents.

Acknowledgments

This thesis summarizes seven significant years of my life. I wish to thank several people for their guidance, collaborations and support during these years.

First and foremost, I would like to express my profound gratitude and appreciation to my advisor Dr. Eytan Ruppin, for his advice, patience, compassion, and wisdom, even when I was on the other side of the world. Thank you for giving me the chance to work on challenging and interesting projects, and for being a role model as a scientist and as a person. I've learned a lot from you and I consider myself extremely fortunate for having you as my advisor.

Second, special gratitude to Dr. Keren Yizhak, for a wonderful mentorship since the first day of my Ph.D. journey, and for being my co-advisor in the last year and a half. Our work together was meaningful to me. Thank you for your trust and for letting me to develop as a scientist, while always being available to teach me and guide me. I feel fortunate for the privilege to work with you and to learn from you.

I would like to thank the members of the Ruppin Labs over the years for being such wonderful colleagues and friends, and especially to Sushant Patkar, Joo Sang Lee, Matthew Oberhardt, Welles Robinson and Erez Presi. Special thanks to Dr. Alejandro Shaffer, for guiding and inspiring me during the end of my Ph.D., and to Dr. Noam Auslander, for being a substantial part of this journey, as a supportive friend and as a great researcher.

I would like to thank all of my professors at the University of Maryland and to Tom Hurst, for being attentive and helpful.

Finally, I'm deeply and eternally grateful to my partner, my parents and my sister for their unconditional support, love and care. Thanks for always being there for me. My love and gratitude could not be expressed in words. I also want to thank my dog, Boten, for the joy she brings to my life.

Table of Contents

List of Figures

Chapter 1

Introduction

1.1 Modeling Cellular Metabolism

One of the ultimate goals of Computational Systems Biology is to build an *in silico model* of a living cell that included all its components and has a predictive value in simulating all cellular processes. While this goal has yet to be achieved, a first step in this direction was introduced by [1]. In this study the authors reconstructed a whole-cell computational model of the human pathogen *Mycoplasma genitalium* that includes all of its molecular components and their interactions. Even though *Mycoplasma genitalium* is a simple prokaryote with only 525 genes, this task has been very challenging, requiring an integrative approach combining diverse modeling approaches. A key difficulty is the lack of sufficient comprehensive knowledge on the pertaining biological processes and associated detailed kinetics. However, despite these difficulties, there is one domain where under simplifying assumptions, and due to two hundred years of biochemistry research, we were able to come closer towards realizing this *in silico* vision, and that is cellular metabolism [2]; Metabolism is by now the most studied and well known cellular process across many species, including humans. Over the last decade, recent strides in the computational study of metabolism have enabled its computational investigation on a genome-scale, demonstrating its value in predicting an array of cellular phenotypes. These advancements

have naturally began with the study of bacterial species [3, 4, 5, 6, 7, 8] then followed by eukaryotic and human modelling studies in an accelerating pace [9, 10, 11, 12, 13]. Cellular metabolism is defined as the set of biochemical reactions needed by biological cells to maintain life. These processes allow the cells to maintain their proper function, grow and respond to changes in the environment. Metabolism is often altered in disease, leading to an increased recognition of the importance of metabolic analysis in drug discovery and in understanding their mechanisms and modes-of-action [14]. Furthermore, metabolic processes involve the production of industrially important nutrients, resulting in a growing interest of metabolic biotechnological engineering applications [15]. Most of the chemical reactions within the cell are catalyzed by specific proteins called enzymes. There are two types of reactions: catabolic, that break down various substrates into metabolites, and anabolic, that collectively synthesize metabolites into amino acids, fatty acids, nucleic acids, and other needed building blocks. Reaction rate, or flux, is the rate of formation or consumption of metabolites in the reaction. The collection of these reactions forms highly complex metabolic networks. In general, the extreme complexity of cellular metabolism, involving thousands of cross-talking reactions, poses challenges for the field of metabolic modeling, requiring a system-level approach [3].

Traditionally, a Genome-Scale Metabolic Model (GSMM) reconstruction is a manual, bottom-up process, in which all the biochemical transformations taking place within a specific target organism or cell are identified and assembled into a structured metabolic network [16]. The network is represented mathematically by a stoichiometric matrix that comprises of the stoichiometric coefficients of the

network's reactions, and is accompanied by a detailed mapping of the genes and proteins to their catalyzed reactions [17]. Recent technological advancements have enabled the genome-wide quantification of genes, enzymes and metabolites levels, thus providing cues to an organism's metabolic state. However, despite this considerable progress, the most direct measure of activity in a metabolic network, the reactions flux rates, can be measured today for only a few dozens of reactions in central metabolism [18]. The analysis of GSMMs aims to bridge this gap and facilitate the prediction of the network's inner and outer (uptake and secretion) flux rates, thus characterizing the organism's metabolic state on a large-scale.

Ideally, one would like to use enzyme kinetics to characterize fully the mechanics of each reaction, in terms of how changes in metabolite concentrations affect local reaction rates. Namely, a kinetic model which is composed by a set of differential equations describing the change in metabolite concentration over time. However, a considerable amount of data and effort is required to parameterize even a small mechanistic model; the determination of such parameters is costly and time-consuming, and moreover much of the required information may be difficult or impossible to determine experimentally [19]. Example of a detailed small-scale kinetic model is of the human red blood cell [20] and a proposed overflow for the formulation of large-scale kinetic model was outlined [19, 21]. Instead of utilizing kinetic models, genome-scale metabolic modeling has applied constraint-based modeling (CBM) approach that relies on constraint-based analysis [22], which uses physicochemical constraints such as mass balance, energy balance, thermodynamics and flux limitations to describe the potential behavior of an organism. Such methods, however, ignore much of the

dynamic nature of the system and are unable to give insight into cellular substrate concentrations.

1.2 Constraint-based Modeling

A reconstruction of a genome-scale metabolic network (GSMM) is a process of identifying all the reactions that comprise a network, which relies on assembling various sources of information about all the biochemical reactions in the network. The reconstruction can be mathematically represented as an in-silico model for computing allowable network states under governing chemical and genetic constraints [3]. A fundamental step towards large scale human metabolic models has been taken in recent studies [23, 24] that reconstructed the global human metabolic network based on an extensive evaluation of genomic and bibliomic data. The reconstructed human model of [23] has been successfully used to predict disease co-morbidity [25] and tissue-specificity of disease genes [26], and for identifying diagnostic biomarkers for Inborn Errors of Metabolism (IEMs) [19].

Given a stoichiometric matrix [S] that encompasses all biochemical reactions and corresponding metabolites, a CBM model imposes mass-balance, thermodynamic and flux bound constraints to define a set of flux vectors that represent all possible steady-state solutions in the genome-scale metabolic network. The constraints imposed on the model are represented as a set of linear equations on the network's flux vector v

$$\mathbf{S} \cdot v = 0 \tag{1.1}$$

$$v_{min} \leq v \leq v_{max} \tag{1.2}$$

Where $v \in R_m$ (where m denotes the number of reactions in the network) denotes the predicted metabolic state and represents a feasible flux distribution through all reactions in the network. The steady state assumption represented in equation 1.1 assumes that there is no accumulation or depletion of metabolites in the metabolic network. Therefore, the production rate of each metabolite equal's its consumption rate and there is no concentration change. The thermodynamic constraints (i.e., under physiological conditions certain reactions are reversible while others are not) and flux capacity constraints (i.e., constraints on enzyme production rate) define bounds on the flux vector and are embedded in equation 1.2 (a negative lower bound corresponds to a reversible reaction and a zero lower bound corresponds to unidirectional reaction). Other constraints such as ones describing the available nutrients in the environment or a genetic perturbation may be also added. For example, in order to eliminate the activity of a gene ("knockout experiment") the minimal and maximal flux bounds of the corresponding reaction should be set to zero (i.e.: $0 \leq v_i \leq 0$). Similarly, to restrict the consumption of a metabolite from the environment the corresponding flux bounds should be set to zero.

The set of constraints that are imposed on the metabolic model defines a solution space of alternative feasible flux solutions that can be explored by different sampling and optimization techniques [27, 3]. The most frequently used optimization method is flux balance analysis (FBA), which assumes that a cell maximizes certain objective function (Figure 1.1). By searching for optimal steady-state solutions, it

narrows the set of feasible steady-state solutions. For micro-organisms, the most common objective function is of growth rate maximization (or biomass production) which is modeled by adding a pseudo reaction to the model (the biomass reaction), whose flux represents the cell's growth rate. The biomass reaction consumes essential biosynthetic precursors according to their contribution to the organism's dry cell weight [28]. Other objective functions can be postulated such as maximization of ATP production rate and minimization of nutrient uptakes. FBA is formulated as a Linear Programming (LP) problem, whose solution defines an optimal solution space, which is composed of alternate feasible steady-state solutions. Minimization of Metabolic Adjustment (MOMA) is a Quadratic Programming (QP) optimization method that searches for a flux distribution following gene knockout with a minimum Euclidean distance from the wild-type flux distribution [29]. An alternate approach to the same problem is called ROOM (Regulatory On/Off Minimization) which utilizes Mixed Integer Linear Programming (MILP) optimization in order to search for a flux distribution that minimizes the boolean regulatory changes between the wild-type and knockout strain fluxes [30].

1.3 Cancer Metabolism

Recent cancer genome studies have led to the identification of multiple cancer associated genes and pathways [32, 33]. It is clear now that cancer initiation and progression are controlled by a host of mutational events in these genes, combined together to support cancerous phenotypes. Furthermore, next-generation sequenc-

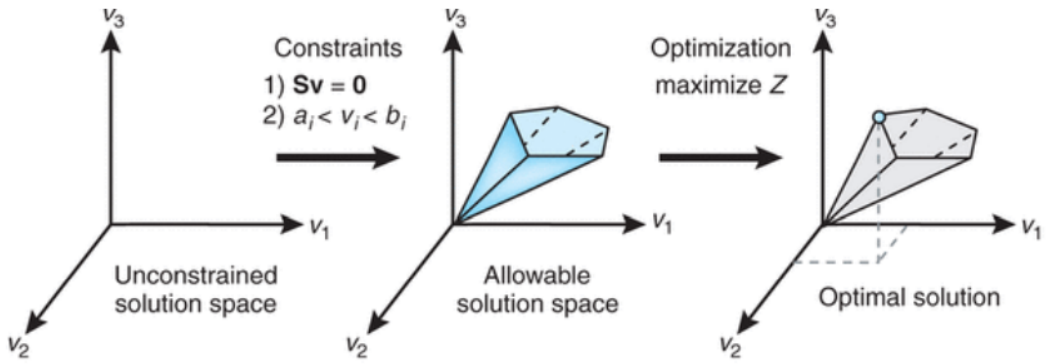


Figure 1.1: Without constraints, the flux distribution of a biological network may lie at any point in a solution space. When mass balance constraints imposed by the stoichiometric matrix S (labeled 1) and capacity constraints imposed by the lower and upper bounds (a and b) (labeled 2) are applied to a network, it defines an allowable solution space. The network may acquire any flux distribution within this space, but points outside this space are denied by the constraints. Through optimization of an objective function, FBA may identify a single optimal flux distribution that lies on the edge of the allowable solution space [31]

ing technologies have enabled the screening of numerous cancer types and subtypes, uncovering both inter and intra-tumor heterogeneity [34]. Despite this large diversity in dysregulated cellular processes, many key neoplastic events are converged to alter tumor cells metabolism. Indeed, cancer cells were found to have a metabolism that is remarkably different from the tissues from which they originated, due to their high demand for proteins, lipids, nucleotides and energy, all necessary for enhanced growth and proliferation [35]. This fundamental characteristic of cancer cells has led to the development of the first chemotherapy treatment, methotrexate, already in the early 1950's [36], in an attempt to target cancer cell proliferation. This drug is designed as an anti-metabolite that interferes with the use of folic acid by cancer cells, thus blocking DNA synthesis and halting cell growth. This common denominator amongst cancer cells together with additional accumulating evidences reviewed

below, have recently led to the recognition of altered tumor metabolism as one of the hallmarks of cancer [37].

Cellular metabolism is finely tuned by integrating signals from the intracellular and extracellular environments. The metabolic switch promoting deregulated growth is often triggered by mutations in signalling pathways that rest at the crux of anabolic and energetic homeostasis, such as HIF-1, PI3K/AKT, mTOR and AMPK [38, 39, 40, 41]. The mutated pathways result in constitutively active growth signals that induces cells to proliferate uncontrollably. In addition to the intracellular genetic modifications, the abnormal environmental conditions also play a major role in modifying cellular metabolism. Heterogeneity in oxygenation, PH levels and nutrient availability are combined with intrinsically altered tumor metabolism, optimizing for a continuous supply of building blocks and redox potential that allow cancer cells to survive and proliferate under strict selective pressure [42].

Recent years have significantly advanced our understanding of the genetic and molecular events underlying the metabolic functional phenotype of cancer cells. This has been achieved due to the considerable leap forward in omics measurement technologies, enabling the genome-wide characterization of different altered cellular processes. Accumulating data of gene sequences and gene methylation patterns, gene, protein and microRNA expression measurements, as well as metabolites levels, have revealed a comprehensive and complex picture of dysregulated cellular processes. Nonetheless, the entire metabolic network is comprised of more than a hundred different subsystems, spanning a few thousands of biochemical transformations. To comprehensively understand how the different cellular components

interact with each other, as well as figuring how the metabolic network responds to different genetic and environmental perturbations as a whole, computational tools come in hand. In particular, computer simulations enabling the investigation of the network's state under diverse conditions and on a genome-wide level are helpful for studying both normal and cancerous cellular metabolism and for advancing our ability to identify potential drug targets and biomarkers.

1.4 Multiomics data integration

The availability of high throughput multi-omics data, including transcriptomics, proteomics, phospho-proteomics and fluxomics, raises an emerging challenge of overlaying this data on top of the reconstructed metabolic networks, to more accurately infer the metabolic regulation reflected in the data. While much progress has been made on studying the regulation of metabolism in bacteria [43, 44, 45] and yeast [46] this question has not yet been studied in cancer cells. Using CBM as scaffolds for the analysis of high throughput omics-data suggests the possibility of inferring condition-dependent changes in the metabolic activity of an organism. Developing computational methods capable of predicting metabolic flux by integrating these data sources with a metabolic network is a major challenge of metabolic network modeling. Previous studies have already utilized GSMM to integrate high-throughout molecular datasets with a metabolic network in a qualitative manner: The methods developed by [47] and by [48] use gene expression data to identify genes that are absent or likely to be absent in certain contexts. While transcrip-

tomics and proteomics data provide important insight into hierarchical regulation of metabolic flux, phospho-proteomics may provide information on an additional level of regulation, called post-translational regulation. The latter denotes the effect of phosphorylation (the attachment of a phosphoryl group) that regulates protein function, subcellular localization, complex formation, degradation of proteins and therefore cell signaling networks. Currently, there are no GSMM methods that enables the integration of quantitative phospho-proteomics data with a metabolic network model to directly infer the metabolic fluxes themselves. In chapter 3, I provide the first chart of metabolic regulation in MCF7 breast cancer cells on a genome scale by integrating multi-omics data, and classifies the metabolic enzymes at three distinct regulation levels.

1.5 Synthetic Dosage Lethality and Genetic Interactions

Identification of proteins that interact to perform a common function is crucial to understanding the mechanisms of cellular processes. Both genetic and biochemical methods are used to uncover an interaction between two proteins. Synthetic lethal (SL) interactions, which occurs when the inhibition of two genes is lethal while the inhibition of each single gene is not [49, 50, 51, 52, 53, 54, 55, 56, 57], is a phenomenon offers a unique opportunity to develop selective anticancer drugs that will target a gene whose synthetic lethal (SL) partner is inactive only in the cancer cells [58, 59]. Synthetic dosage lethal (SDL) interactions, whereby the underexpression of one gene, combined with the overexpression of another gene is lethal, but not each event

individually [60] offer a promising way to kill cancer cells by inhibiting the activity of the partners of activated oncogenes in tumors [61, 62, 63, 64], as one of the hallmarks of cancer is over-expression of oncogenes. Screening technologies have been developed to detect SL/SDL-interactions in numerous model organisms [65] and in human cell lines [49, 50, 51, 52, 53, 54, 55, 56, 57]. However, as every pair of genes can potentially interact, the combinatorial search space consists of more than 500 million pairs, thus current experimental technologies are far from being able to address the challenge on a genome-scale. New bioinformatics approaches tried to address the challenge. While the model-based investigation of metabolic SLs GSMM is quite straightforward and has proven its value already for more than a decade (e.g., [66, 67]) and specifically in cancer [68, 69], SDLs have mostly been an uncharted land. In chapter 2 I present the first computational method for the identification of metabolic SDLs using GSMM. Our method does not only identify SDLs that are strictly lethal to the cell, but also those that have a significant effect on tumor growth or proliferation in clinical settings, depending on the measure of strength θ, assigned by the method. We further test and validate the predictive beneficial signal in cancer and show that the activation of SDLs is associated with smaller tumor sizes and longer patient survival.

1.6 Mutations in cancer

A mutation is a change in the DNA sequence of an organism. The human genome can harbor two types of mutations, germline and somatic. Germline mu-

tations are found in the first cell of the embryo and propagate to every cell of an individual. Conversely, somatic mutations are acquired by each cell lineage during development and post-natal life. For this reason, somatic mutations are found only in a portion of the cells of an individual, or even in a single cell. Somatic mutations are caused by several endogenous and exogenous factors. They can arise at very low frequencies due to the molecular instability of DNA bases and spontaneous reactions [70, 71]. Metabolic by-products such as reactive oxygen species (ROS), DNA replication before cell division as well as unwinding of the DNA double-helix during transcription all pose steady threats to DNA integrity [72, 73, 74]. Additionally, cells can be exposed to environmental mutagens such as tobacco smoke and aflatoxin or UV and ionising radiation [75, 76]. However, the accumulation of mutations is not always a gradual process as catastrophic events such as chromothripsis and kataegis can cause thousands of clustered chromosomal rearrangements or vast numbers of point mutations across a relatively short stretch of DNA [77, 78]. While the occurrence of an individual mutation is a stochastic event and can be caused by any of the factors outlined above, the type and context of mutations is not random. Each of the endogenous or exogenous factors and their associated repair mechanisms promote particular molecular reactions and thus, are associated with a spectrum of mutations. The first formalized approach to detect the association between different underlying mutational processes and their corresponding footprint was performed on the point mutation profiles of a large sequencing cohort of cancer patients [79, 80].

A primary hallmark of tumorigenesis is the accumulation of mutations in cancer cells [81]. The relationship between mutations and cancer emerged in the late

nineteenth and early twentieth century, when the German biologists David von Hansemann and Theodor Boveri observed abnormalities in mitotic divisions and corresponding chromosomal aberrations in cancerous epithelial cells [81, 82]. After the discovery of DNA as the inheritable substance and its structure in the forties and fifties, these mutagens were shown to cause chemical changes in the DNA, further strengthening the link between genetic alterations and the occurrence of cancer [83]. These mutations are found both in genes that drive cancer, and those that do not (passenger mutations) [34]. Regardless of driver status, these mutations provide a potential opportunity to specifically target tumor cells through the creation of tumor specific novel immunogenic peptides (neoantigens). These neoantigens are generated from peptides encoded by gene alterations that are present in tumor but not normal tissue, and therefore represent highly promising vaccine immunogens [84, 85]. Seminal studies have suggested the immunotherapeutic potential of neoantigens and have shown that: (a) mice and humans can mount T cell responses against mutated antigens [86, 87]; (b) mice can be tumor-protected by immunization with a single mutated peptide present in the tumor [88]; and (c) memory cytotoxic T lymphocyte (CTL) responses to mutated antigens are generated in patients who have unexpected long-term survival or have undergone effective immunotherapy [89, 90]. However, neoantigens also are almost exclusively personal, found uniquely in the tumor of each individual patient, and therefore have not been used for immunotherapy due to technical difficulties in their identification and testing [86]. Somatic mutation identification is traditionally performed on tumor and normal genomes/exomes [91, 92, 93, 94, 95, 33], comparing the DNA sequence from tumor samples with

their matched normal samples from the same individual. This allows subtraction of the germline variants shared by all cells in an individual, leaving only acquired somatic mutations. Due to the popularity of RNA-seq technology for gene expression profiling over microarray technology [96, 97, 98, 99], it has become routine for projects like The Cancer Genome Atlas (TCGA) to also sequence the tumor RNA, along with many more RNA-seq data that have been accumulated over the past few years. The majority of these RNA-seq data has been only studied for gene expression. Recently, Yitzhak et al. [100] have developed a tool to accurately detect somatic mutations from RNA-seq data. In chapter 4, I present the first model to identify somatic mutations from RNA-seq data without the matched normal. The clinical potential is huge, as the model allows for the first time to utilize a great amount of datasets available in the public domain that have RNA-seq data of the tumors of cancer patients, and it can potentially enable a better understanding of phenotype-genotype associations, as both the genetic (mutations) and expression levels are inferred from the same sample.

1.7 Immunotherapy

Up until recently, many doubts were raised regarding the ability of the immune system to control cancer, and the possibility of developing effective immunotherapy. However, evidences accumulated over the past years do not leave much room for skepticism. Studies have found that the abundance of CD8+ T-cells is one of the best predictors of overall survival in human cancers [101]. In addition, Adoptive T

cell therapy has shown to eradicate several types of blood cancers, and has been officially approved by the FDA as standard of care for some forms of non-Hodkin Lymphoma and acute lymphoblastic leukemia [102]. More recently, the development of checkpoint blockade (CPB) therapy such as anti-PD1 and anti-CTLA4 has resulted in long-lasting tumor responses in patients with a variety of cancers [103]. As a result, these drugs have been FDA-approved for many cancer types, including melanoma, non-small cell lung cancer, Urothelial carcinoma, Head and Neck squamous cell carcinoma and more [104].

CPB therapies were developed to overcome the dysfunction or exhaustion of T cells resulting from chronic antigen exposure and suppression by the tumor or cells in its microenvironment. However, it remains unclear why some patients respond to this therapy while others do not. Specifically, overall response rates with these agents as monotherapy are relatively low (15–20%), but some individuals can attain durable complete remissions. In specific cancer types (e.g., melanoma), front-line anti-PD-1 achieves higher response rates (40-45%), and the median progression free survival (PFS) is approximately six months [105, 106]. The combination of anti-PD1 and anti-CTLA4 (ipi/nivo), as reported in Phase I-III trials, is associated with even higher response rates (exceeding 60%) and improved PFS (median 12 months), but at the cost of high toxicity (grade 3 or 4 treatment-related adverse events 55+%) without improved overall survival (OS) [107, 108]. The variability of response to CPB highlights the need for identifying predictive biomarkers, leading to multiple attempts to develop such predictors.

Several biomarkers for CPB are already FDA-approved: The first has been

granted to patients with specific tumor types, including Non-small cell lung cancer and triple negative breast cancer, with high PDL-1 expression (50%). In these cases, a significant increase in PFS was observed. Additionally, in 2017 the FDA granted accelerated approval to pembrolizumab (anti-PD-1) for both adult and pediatric patients with microsatellite-instability-high (MSI-H) or defects in mismatch repair (dMMR) solid tumors, that have progressed following prior treatment. Overall response rate (ORR) in this group was 39.6%, and this was the first time the agency authorized a cancer treatment based on a genomic biomarker that was histology agnostic. Recently, an additional accelerated approval for anti-PD1 for the treatment of adult and pediatric with unresectable or metastatic solid tumors with tumor mutational burden-high (TMB-H) [$\geq$10 mutations/megabase (mut/Mb)] has been granted. However, even among these patients, the overall response rate stands on 29%.

Chapter 2

Metabolic Interactions in Cancer

⋆⋆ Published as "Synthetic dosage lethality in the human metabolic network is highly predictive of tumor growth and cancer patient", Proceedings of the National Academy of Sciences, 2015

2.1 Introduction

Synthetic lethality (SL) occurs when the combined loss of two non-essential genes renders a lethal phenotype [63]. SLs have been studied using experimental [109, 110] and computational approaches [111, 67, 112] to address various questions of cell function and evolution. The potential of SLs for cancer therapy has been recognized and accelerated the development of many SL screens [113, 114, 115, 116, 117]. See [118, 119, 120] for reviews of SLs applied in the context of cancer research.

Less studied are the so-called synthetic dosage lethality (SDL) interactions. An SDL is a genetic interaction between two genes where the underexpression of gene A (A↓) together with the overexpression of gene B (B↑) is lethal [121]. The observation that an interaction with an overexpressed gene can be lethal makes it particularly interesting for targeting cancer cells with (over-)expressed oncogenes. This is because many oncogenes that drive tumor growth are essential to cell function and thus difficult to target directly. Targeting the oncogenes' SDL partner, which

is a non-essential gene in normal cells, may nevertheless kill cancer cells. That SDLs can have important implications for cancer research, for instance to aid the design of new therapies, has also been recognized [118, 122, 60, 123]. Moreover, it has been shown that the overexpression of specific genes can be detrimental to cancer cell growth [124]. Recently, a data mining approach was used that identifies SLs and SDLs by analyzing large volumes of cancer genomic data [61]. Here we aim to complement data driven computational efforts with a biological network model approach to identify SDLs. This has recently become feasible in the realm of metabolism, with the advent of genome-scale metabolic modeling. We introduce a novel method that utilizes a constraint-based Genome-Scale Model of Metabolism [125, 126, 6, 127, 128] to predict metabolic SDLs. GSMMs have successfully resolved a wide range of research questions in model organisms [6, 4, 5, 3, 129, 130] and have been the basis for many computational studies of cancer [113, 114, 131, 132, 133, 134]. Furthermore, they have contributed to a systematic understanding of the underlying mechanisms leading to lethality and synthetic lethality [110, 111, 67, 112, 113] . A major advantage of a model based approach is that it can provide insights into the underlying network mechanisms causing SDLs. Furthermore, the modeling approach presented is general and can be used to identify SDLs in species and cell-types where omics data is missing.

We introduce IDLE, a computational approach for Identifying Dosage Lethality Effects in metabolism. IDLE predicts enzymatic SDLs from a GSMM with application to cancer. For each enzyme pair (A,B) in the human GSMM we systematically knock-out the enzyme flux through A combined with a stepwise flux

increase through enzyme B, and quantify the level of growth reduction. Pairs in which the growth is significantly more reduced than where either enzyme is perturbed individually are ranked as SDLs (A↓, B↑) with a corresponding value of "strength".

We demonstrate the predictive power of our approach in five different ways: First, by analyzing genome-wide experimental shRNA screens we show that A↓ in predicted SDLs (A↓, B↑) are indeed more likely essential in an overexpressed enzyme B↑ background than when B is not overexpressed. When A is underexpressed and B is overexpressed in a predicted SDL in a given tumor sample, we denote that SDL as 'active', that is, bearing potential functional effects on the tumor growth and the patient's survival. Secondly, we show that SDLs are less frequently active across cancer patients compared to randomly selected enzyme pairs, indicating that tumor cells select against the presence of SDLs to avoid cell death. Thirdly, we illustrate that the tumor size of breast cancer patients that have one or more active SDLs is significantly smaller than that of patients expressing randomly selected enzyme pairs. Fourth, we show that the predicted impeding effect of active SDLs on tumor growth correlates with a significantly longer patient survival time. These results become even more pronounced when one includes only highly ranked active SDLs (that show a stronger A↓, B↑ pattern at the transcriptional level) illustrating that our method successfully identifies the clinical impact of SDLs. Finally, we report that observed effects become stronger when more active SDLs are present in a given tumor, pointing to the cumulative effect of active SDLs in clinical tumors.

2.2 Results

2.2.1 Overview of IDLE Algorithm

The IDLE method (see 2.1 and methods for details) computes the effect on cell growth when an enzyme B increases its activity (we call this the reference GSMM), compared to its activity in a knockout (KO) GSMM, where additionally enzyme A is knocked out. The objective of IDLE is to find enzyme pairs (A, B) where this differential growth effect is marked, searching over the space of all possible pairs. For a given pair (A, B), we define a reference wild type GSMM and compute the maximum growth (biomass, µmax) with Flux Balance Analysis [135]. Similarly, µmax is computed for the KO GSMM, where reaction A is knocked-out. In both models, the maximum flux through B is computed without any constraint on µ (i.e., lower bound is zero, see 2.1 panel a and b for the reference and KO GSMM respectively). Now, the lower bound of the biomass reaction is stepwise increased (using n = 10 steps) towards µmax in both the reference (2.1 panel c) and KO (2.1 panel d) model. For each increase, the maximal allowable flux through reaction B is again computed. The increasing growth pressure may affect the allowable flux through reaction B, and if so, it must decrease. The basic idea behind IDLE is that this argument is reversible: if the growth requirement constrains the maximum allowable flux through B, then a further flux increase through B must decrease growth. This effect is quantified and expressed as a vector (2.1 panel e). The angle θ between the reference and KO vectors measures the difference between the effects on cellular growth of overexpressing enzyme B in the wild type (A, B↑) and after

a KO of enzyme A (A↓, B↑). If growth reduction is stronger in the KO situation (A↓, B↑) then we define θ positive and the enzymes (A, B) form an SDL. SDLs with the largest angle are predicted to have the maximum effect and are termed "high-impact" SDLs. We can therefore rank-order SDLs based on the computed angle θ (2.1 panel f).

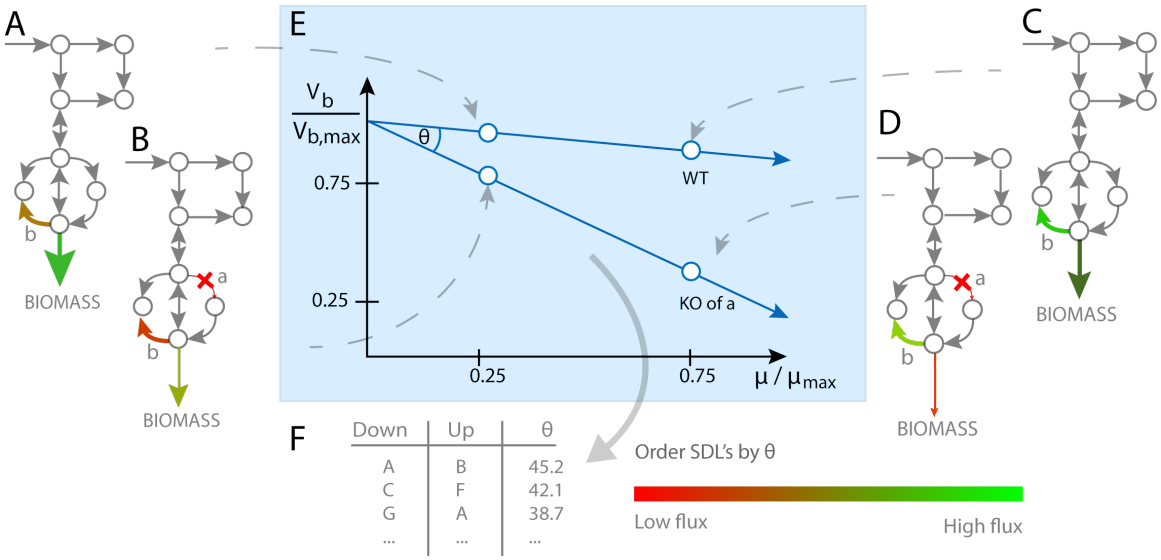


Figure 2.1: Conceptual overview of the IDLE method: (A) The maximum flux through enzyme B is computed when there is no biomass pressure (i.e., lower bound flux is zero). (B) This process is repeated for the KO model. (C) The biomass pressure is increased in stepwise fashion and the maximum flux through enzyme B is computed at each step. (D) This is repeated for the KO model. (E) The maximum relative flux of B (VB,max) is plotted at each biomass step (max) and the angle θ between the reference and KO vector is computed. (F) SDL pairs are ranked based on their growth impact, quantified by their angle θ.

2.2.2 The Metabolic SDL Network.

Our method discovered 12,447 SDL interactions. Reassuringly, the ranked list of SDLs significantly matches the top ranked metabolic SDLs identified by DAISY, an approach for data-driven inference of genetic interactions in cancer that is based on the discovery of underrepresented gene pairs in cancer genomic data

[61] (Wilcoxon rank-sum p ¡ 0.0038). SDLs are asymmetrical by definition, i.e., A↓, B↑ denotes a different interaction than A↑, B↓, and each may have a very different magnitude; in the first interaction enzyme A is the KO partner, while in the second interaction it is the overexpressed partner. Surprisingly, we discovered that six enzymes are major 'master' hubs, being the KO partners of many other over-activated B↑ in the SDL network (see subsection **??**). These major hubs (TPI, ENO, PGM, PYK, PGK and GAPD) all reside in the glycolysis pathway. Interestingly, when examining the hub partners, we observed that the B↑ partners are the same for 80% of the SDLs. The metabolic pathways that are enriched for these overexpressed partners are shown in SI1-Table S3. To better understand the putative mechanisms underlying the workings of these SDLs we conducted a further model-based analysis. First, we charted synthetic lethal (SL) interactions of the six master hubs, i.e., searched for genetic interacting pairs involving these six hub reactions where the growth reduction after their combined KO is larger compared to that observed after the single KOs. We were surprised to see that while these SDL hub reactions are highly sensitive to a synthetic dosage load (each being essential for 500 overexpressed partners), they have only very few SL partners (a list of these reactions and their pathways is shown in SI1-Table S4). Examining the SDL partners of the six central glycolytic hubs we find that they are quite distributed across the metabolic network in ten different pathways that are significantly enriched with the SDL partners (see SI1-Table S5). When further investigating these SDLs, we discovered that glycogen production is decreased by (on average) 60% when such SDLs are active compared to the wild-type and the knockout conditions. Interestingly, it

has recently been shown that glycogen metabolism and its initial accumulation is a key pathway induced by hypoxia and its activity is necessary for optimal glucose utilization in tumors [136].

2.2.3 SDL Is Predictive of in Vitro shRNA Essentiality Screens

We expect that a knockdown of enzyme A (A↓) will be lethal in a B↑ background in the case of an SDL (A↓B↑). To study this, we exploited gene essentiality at a genome-wide scale in cancer cell lines using experimental shRNA screens [137] and matched it with gene-expression profiles [138]. In a typical shRNA screen in a given cell line, each gene is individually knocked down by targeting its mRNA (both inhibiting and degrading it) by specific shRNAs that bind to it. Then, the effect of each individual gene knockdown on cell growth is measured from which scores are calculated that indicate gene essentiality (a $p = 0.05$ cutoff was used to consider a gene essential [137]). For each cancer cell line, we divided SDLs into two groups: group 1 consists of SDLs in which at least one of the B enzymes that form an SDL with enzyme A is overexpressed (B↑) and group 2 consists of SDLs where none of the B enzymes are overexpressed (see methods for definition of overexpression and for mapping genes to reactions). Then, the number of essential and non-essential A enzymes observed experimentally in the shRNA screen was compared between group 1 and group 2 in each cell line (one-tailed Fisher's exact test). Using a $p = 0.05$ cutoff we counted the number of cell lines in which enzymes A from group 1 are more frequently essential compared to these enzymes in group 2. The above

procedure was also repeated 5000 times for a set of random enzyme pairs of equal size. As expected, the number of cell lines in which essentiality of A in a B↑ background is enriched (group 1) is significantly higher for SDL than for random pairs (empirical p = 0.002).

2.2.4 Cancer Cells Select Against SDL

Cancer cells are expected to select against the negative effect that SDLs have on (tumor) growth. Thus, when the enzyme pair (A, B) is an SDL, underexpression of enzyme A and overexpression of enzyme B should occur less frequently than for random enzyme pairs. We analyzed a gene expression dataset of 7,362 patients from the TCGA cohort [139] and determined for each gene whether it is underexpressed (↓), overexpressed (↑), or unchanged compared with expression levels in normal tissue samples [140] (See methods for more information). We then computed for all SDLs the number of patients, F_{SDL}, with an active SDL (A↓, B↑) relative to those patients having only enzyme A underexpressed (A↓, B) or having only enzyme B overexpressed (A, B↑). This was repeated for 5,000 randomly constructed enzyme pair sets of equal size (F_{RANDOM}). As expected, F_{SDL} is significantly smaller than F_{RANDOM}, illustrating that an underexpression of A combined with an overexpression of B when A and B have an SDL relation occurs significantly less frequently than when the enzyme pair have no SDL relation (Figure 2.2). In fact, when the angle θ increases, the fraction of patients that have an active SDL approaches zero, testifying to the strong negative selection exerted on such SDLs.

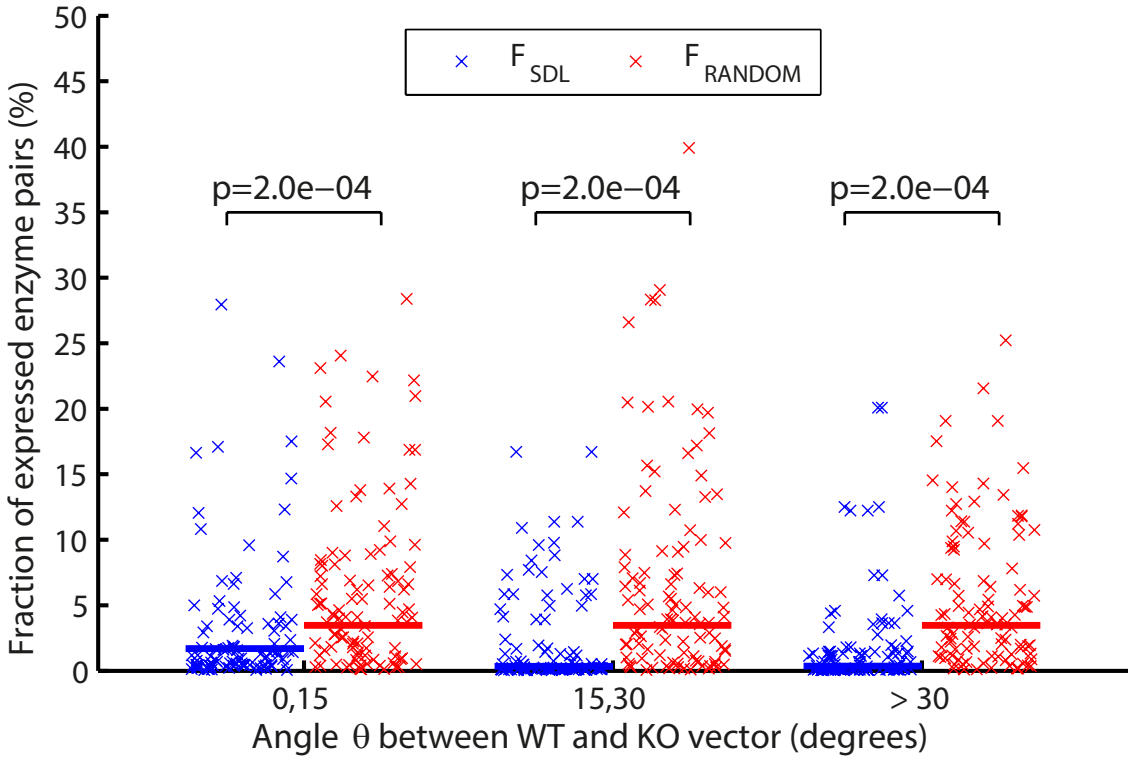


Figure 2.2: Percentage of active enzyme pairs (i.e., A, B) with A underexpressed and B overexpressed. When the angle θ increases, the fraction of active SDLs approaches zero. SDLs are significantly less frequently active than randomly chosen enzyme pairs. For all cutoffs, the P values obtain their maximum (1+1/5,000+1) significance.

2.2.5 SDL Correlates with Smaller BC Tumor Size.

Since SDL negatively affects growth in cancer cell lines we expect that the tumor size will be smaller for patients having at least one active SDL, compared to those who do not. To address this, we used a dataset where gene expression and matched tumor size data is available for 1587 breast cancer patients (41). We split the patients in this heterogeneous dataset based on the estrogen receptor (ER) sensitivity of their tumor (see SI1-Section S5 for key properties of the data set). We analyzed whether the tumor size of patients with an active SDL (A↓, B↑) is significantly smaller compared to patients that have one of the single effects, meaning only an under- (A↓, B) or overexpression (A, B↑) of enzyme A or B respectively. To investigate A↓, B↑ in relation to A↓, B we separated patients into two groups: patients that have B overexpressed (see methods for definition of overexpression) with varying underexpression of A (σ between 0 and 3 given the underlying gene expression distribution) and patients that have enzyme B not overexpressed with varying underexpression of A. When comparing A↓, B↑ with A, B↑ we also separated the patients into two groups: patients that have A underexpressed (see methods for definition of underexpression) with varying overexpression of B (σ between 0 and 3 given the underlying gene expression distribution) and patients that have enzyme A not underexpressed with varying overexpression of B. Finally, we created random enzyme pairs (n=5000) to serve as control for testing the specific effects of the SDLs. Statistical significance for all comparisons was computed with a signed Wilcoxon ranksum test, analogous to the signed Kaplan-Meier test defined in [61].

See SI1-Section S6 for a detailed procedure. As expected, we observed for ER+ breast cancer patients that patients with (at least one active) SDL have significantly smaller tumors compared to patients with only overexpression of B ($p < 4e\text{-}8$, 2.3). We found for ER- patients that the tumor sizes of patients with SDL are also significantly smaller compared to patients with only overexpression of B ($p < 5e\text{-}5$) as well as compared to those with only underexpression of A ($p < 7e\text{-}5$). Moreover, smaller tumors are observed for both ER- and ER+ patients with active SDLs compared to when patients have randomly selected enzyme pairs with the A↓, B↑ pattern active ($p < 2e\text{-}3$).

2.2.6 SDL Correlates with Increased Cancer Survival.

Since SDLs decrease breast cancer tumor size, we hypothesized that their presence also affects patient survival. For the breast cancer data, matched survival times were available such that we could correlate it to the level of SDL activation [141]. We hence performed a survival analysis analogous to the tumor size analysis described above. The significance of the results obtained for SDL were compared to the single effects and random pairs by a modified signed Kaplan-Meier test introduced in [61]. See methods for detailed procedure. As expected, we found that ER+ breast cancer patients with at least one active SDL have significantly better survival times compared to patients with only an underexpression of A ($p < 4e\text{-}03$, Figure 2.4 panel a and b). Patients that activate the highly ranked SDLs show the longest ER+ breast cancer survival times up to a median of over 12 years (Figure

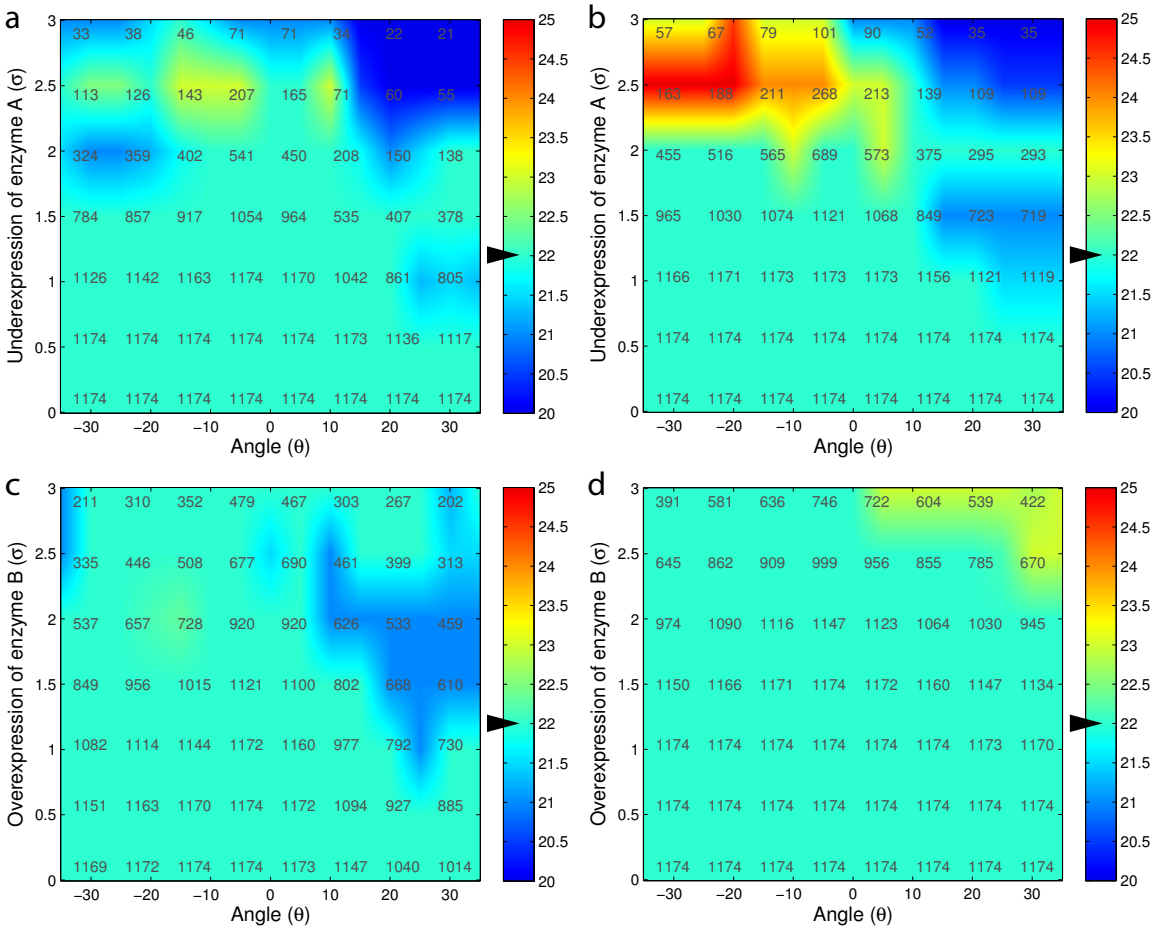


Figure 2.3: Median BC tumor size (in millimeters) for patients with ER^+ disease. Arrowheads denote the median tumor size for all patients with ER^+ BC (22 mm). The number of patients that express at least one enzyme pair are denoted inside the figures. (A) Patients with at least one active SDL (A↓, B↑) with constant overexpression of enzyme B. (B) Patients whose disease only underexpresses enzyme A (A↓, B) of the SDL. (C) Patients with at least one active SDL (A↓, B↑) with constant underexpression of enzyme A. (D) Patients whose disease only overexpresses enzyme B of the SDL (A, B↑).

2.4 panel a). In line with expectation, the survival time of patients with active SDL is significantly better compared to those having only enzyme B overexpressed (p ¡ 3e-4, Figure 2.4 panel c and d). Moreover, significant longer survival is also observed for patients with SDLs compared to those with random enzyme pairs with the A↓, B↑ pattern active (p < 1e-03). We refer to SI1-Section S7 for survival analysis of ER- patients. Since overexpression of enzyme B is generally not beneficial when enzyme A is not underexpressed, we wondered whether underexpressing enzyme B alone would be beneficial. SI-1-figure S5 indicates that this is not the case. In particular, severe underexpression of enzyme B correlates with increased tumor sizes (SI1-figure S4a,c) and decreased survival times (SI1- figure S4b,d) in both the ER+ and ER- breast cancer patients. SDLs predicted by IDLE are not expected to be specific for breast cancer. To examine their predictive power in another cancer type we analyzed a large cancer type-specific cohort of 921 patients diagnosed with serous epithelial ovarian cancer [142] with matched survival times. Indeed, the same observations were made as in the case of breast cancer ER+ patients, i.e., ovarian cancer patients with at least one active SDL have significantly better survival times compared to those having the single or random effects (p < 0.09 compared to A↓, B and p < 0.01 compared to all others, SI1-Figure S5). These results are even more apparent in the relapse free survival times of these patients (p < 0.02 compared to A↓, B and p < 9e-04 compared to all others, SI1-Figure S6).

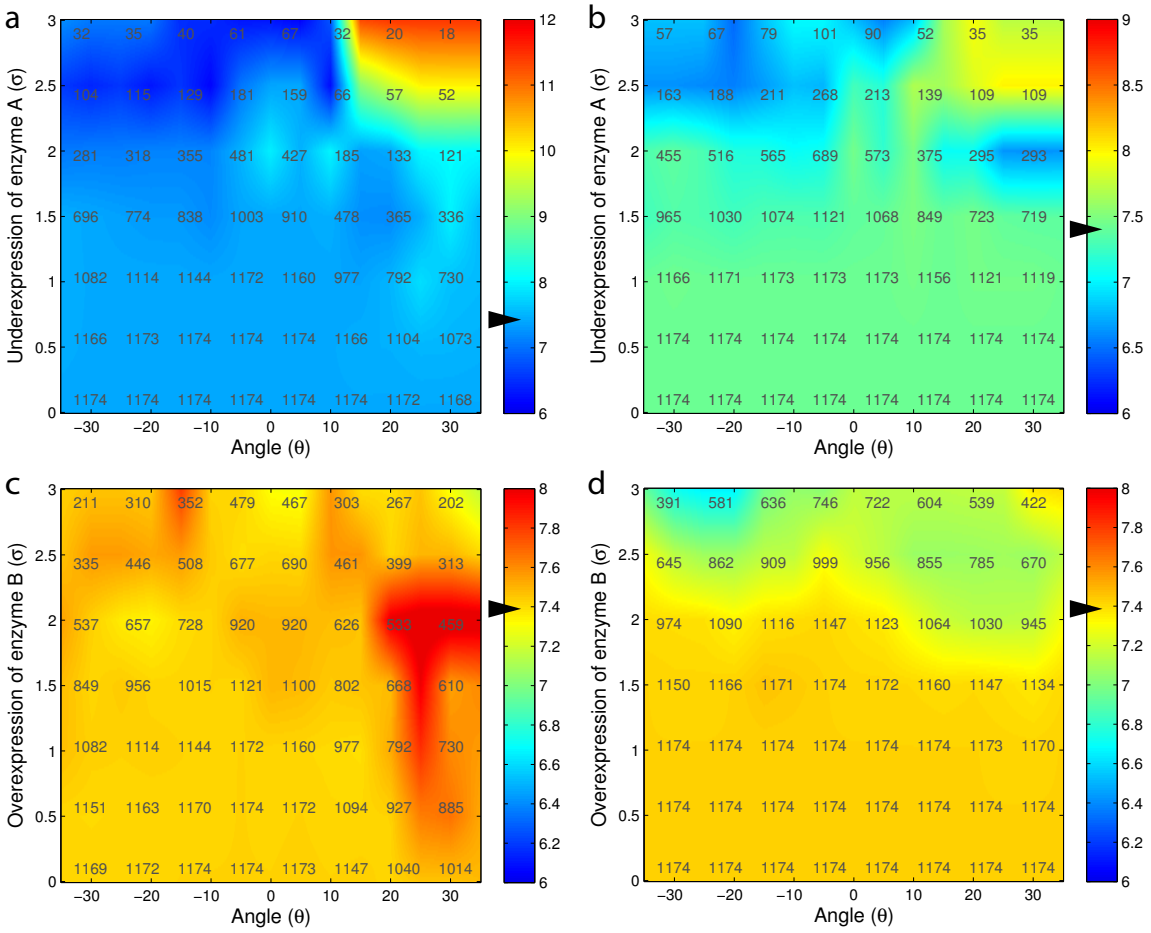


Figure 2.4: Median ER^+ BC survival time (in years). Arrowheads denote the median survival for all patients with ER^+ BC (7.4 y). The numbers of patients whose disease expresses at least one enzyme pair are denoted inside the figures. Note that the axis of figure a scales differently. (A) Patients with at least one active SDL (A↓ , B↑) with constant overexpression of enzyme B. (B) Patients whose disease only underexpresses enzyme A (A↓ , B) of the SDL. (C) Patients with at least one active SDL (A↓ , B↑) with constant underexpression of enzyme A. (D) Patients whose disease only overexpresses enzyme B of the SDL (A, B↑).

2.2.7 Cumulative Effect of SDLs in a Tumor Correlates to Better Survival

As SDL activity in a tumor correlates to survival prognosis, we asked if survival time would increase when patients have more SDLs active. We tested the presence of such a cumulative effect in the two largest cancer subtypes; the ER+ breast cancer (BC) patients (n=1174) and the serous epithelial ovarian cancer (OC) patients (n=921). Patients were categorized into three groups, those having 1-3, 4-8 or more than 8 active SDLs in their expression profiles (see methods for definition of over- and underexpression). The Kaplan-Meier survival curve in Figure **??** shows as expected better survival for patients with large number of active SDLs compared to those having only a few active SDLs. Indeed, a logrank test [143] revealed significantly improved survival times in both cancer types when the number of active SDLs increases ($p < 8e\text{-}03$ for BC ER+ and $p < 2e\text{-}03$ for OC and OC-RFS). The largest cumulative effect in the BC survival is related to SDLs being active with enzyme A as one of the major glycolytic hubs. Interestingly, the observed cumulative effect in OC is already present for patients that have 4-8 active (Figure **??** panel b). The under-activated enzymes A in these SDLs are enriched for pathways that utilize glutamine through glutamate metabolism, the TCA cycle and mitochondrial transport ($p < 0.001$, hypergeometric test). It has recently been shown that severe types of OC, such as the epithelial subtype we considered, are driven by glutamine metabolism, in contrast to BC tumors that depend on an over activity of glycolytic enzymes [144].

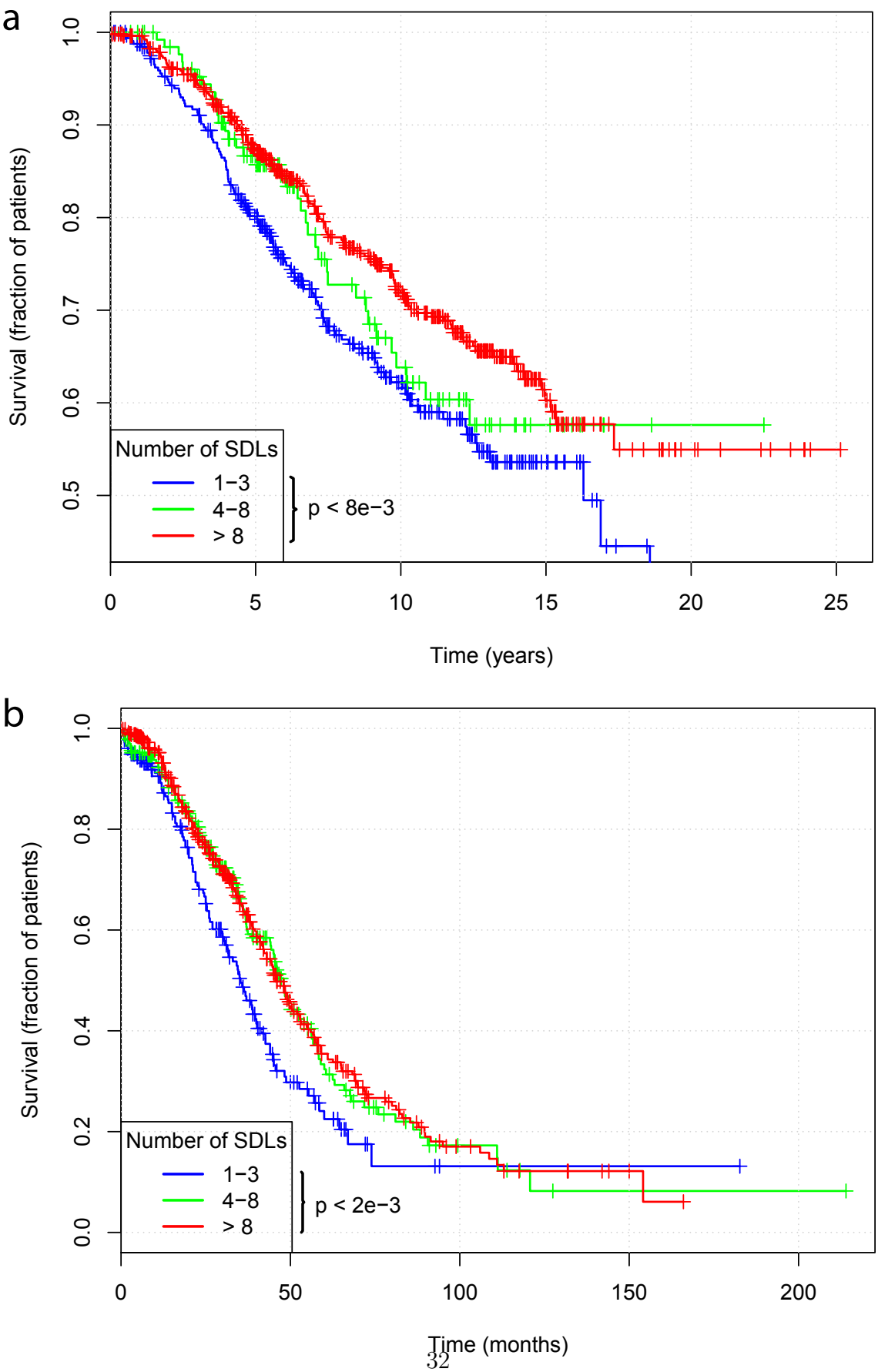


Figure 2.5: Kaplan–Meier survival curves for patient groups that have one to three, four to eight, or more than eight active SDLs. (A) Survival times for the patients with ER+ BC. (B) Survival times for patients with serous epithelial OC.

2.3 Discussion

We introduce the first computational method that captures enzymatic SDL effects in metabolic networks. Our method does not only identify SDLs that are strictly lethal to the cell, but also those that have a significant effect on tumor growth or proliferation in clinical settings (i.e., "synthetic dosage sick"). We show that our method is able to assign a measure of strength θ to each SDL, which correlates to its predictive power in an array of different tumor clinical attributes. It is therefore of interest to focus further research towards therapeutic interventions on the basis of "high-impact" pairs, which may have the largest beneficial effect on killing cancer cells. We show that SDLs are less frequently active than expected in cancer cells. This shows that rapidly expanding cancer cells select against interactions that reduce their growth rate. The activation of "highimpact" SDLs is associated with smaller tumor sizes and longer patient survival. The effect strongly depends on the extent to which SDLs are activated, but most SDLs we found do not require a complete enzyme knock-out to exert a functional predictive signal. Lastly, we demonstrated a cumulative effect of SDL presence; the more SDLs active in a tumor sample, the better this is for a patients' prognosis. This observation may shed light on targeting cancers that rely on glycolysis. Down-regulating glycolytic enzymes that are the major hubs in the SDL network is hence expected to have large inhibitory growth effect in tumor cells that overexpress many of the glycolytic SDL partners. As glycolysis is usually less active in normal cells and SDL partners of glycolytic hubs are less frequently overexpressed in normal cells compared to

cancer cells in the majority of tissue types (SI1-Section S2.2), targeting these glycolytic SDLs may be of therapeutic interest, especially when a large number of their partners are overexpressed. The current study, being the first of its kind, naturally focuses on harnessing the generic human metabolic model to identify a common core of SDLs that may be shared by many different cancer types. However, the IDLE approach is general and could be extended in the future to identify cancer type specific SDL interactions more precisely, by integrating patient- and tumor-specific omics data such as gene expression or proteomics. The results of our metabolic network modeling do not support the hypothesis that SDLs arise due to draining alternative compensatory pathways that compensate for the loss of the KO enzyme. This is because we do not find that the flux in such backup reactions of the major key glycolytic enzyme hubs is reduced following the over-expression of their SDL partners. Intriguingly, we do find that disrupted glycogen metabolism is predicted to be the major mechanism by which hundreds of SDLs of key glycolytic enzymes exert their growth inhibitory effects. Indeed, it has recently been shown that glycogen metabolism and its initial accumulation is key for optimal glucose utilization in tumors [136]. Thus, SDL relations do not arise via simple proximal interactions, but are likely to be the result of complex stoichiometric network relations that withdraw flux from biomass production through activation of other pathways. Our results testify to the potential contribution of model-based approaches to identify and uncover the mechanisms behind SDLs. Model-based SDL prediction via IDLE is widely applicable and not limited to cancer. It could be used to identify SDL networks in pathogenic bacteria or fungi, providing new antibiotic therapeutic leads. Other

possible applications include metabolic engineering to increase the yield of valuable metabolic byproducts. Specifically, this may be achieved by engineering an SDL effect to inhibit the production of undesired byproducts, or inversely, neutralizing the SDL effect to force an increased flux through desired pathways. Taken together, IDLE is expected to contribute to various research fields ranging from medical sciences to biotechnology.

2.4 Methods

2.4.1 The IDLE algorithm

The concept behind the IDLE algorithm is given in detail in Figure 2.6. IDLE compares fluxes in the reference genome-scale metabolic model (GSMM), in which no reaction is knocked-out (KO), with those in n KO GSMMs. Each KO GSMMs i is created by restraining the flux through reaction i to 0. For the reference GSMMs and each KO GSMMs, the maximum growth μ_{max} is computed with flux balance analysis (FBA) [135, 31]. KOs that reduce the biomass flux to zero are not considered for further analysis. The minimum growth rate is set to 0, and the maximum growth to µmax. The objective is then changed to compute the maximum flux through reaction B. This determines the starting point as indicated in Figure 2.6 (step 1). Then, the minimum growth rate is increased in steps of 10% until it equals μ_{max}. At each of these steps, again the maximum allowable flux through reaction B in the reference and KO models is computed Figure 2.6 (step 2).

Increasing the minimal growth rate may affect the allowable flux through reac-

tion B. If this is the case, the maximum allowable flux always decreases, because we add an extra constraint to the model, i.e. the maximum flux through B can already be reached when the biomass is unconstrained. Because initial flux and growth values may differ widely between the reference and KO model, we compare the relative effect of increasing the biomass flux. Therefore, the flux through reaction B and the biomass reaction are normalized by dividing all values by their maximum. Then, the maximum allowable flux is plotted against the growth rate and fitted by a straight line for both the reference and the KO models. Finally, the angle θ between both lines is calculated, which is a measure of growth rate reduction due to an increase of reaction B flux in the KO state (Figure 2.6, step 3).

Because the GSMM is a linear model we can deduce that an increased flux therefore leads to a lowered growth rate. The main reason that we constrain the growth rate and optimize the flux to simulate an up-regulated enzyme (and not vice versa) is that we can easily compare relative flux changes at linearly changing growth rates. We iterate over all possible reaction pairs ($\approx$2500) in the human model and therefore create over 6 million putative SDLs. We are only interested in those pairs with a significant difference between the reference and KO cell. Therefore, all pairs with $|\theta| < 2^\circ$ were removed, reducing the list to 12,447 putative SDLs. The effect of an increased enzyme flux on growth can be captures into four major types of enzyme relations, depending on whether growth is affected in the reference model, the KO model or both.

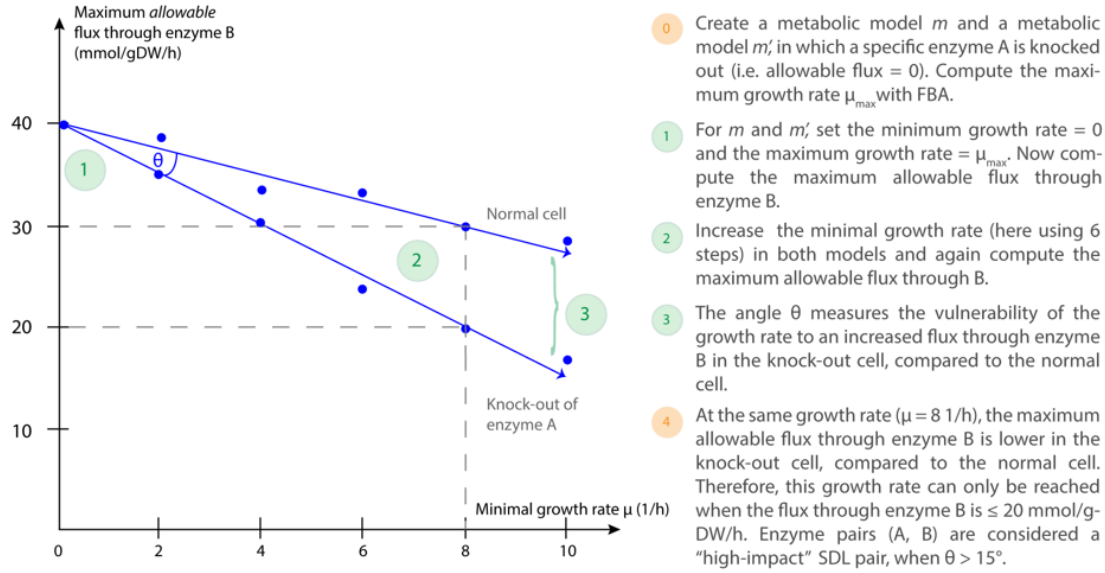


Figure 2.6: The IDLE method. IDLE measures the 'vulnerability' of the growth rate to a flux increase through enzyme B. This reference model m is compared with a model m' that computes this vulnerability when additionally enzyme A is knocked out. This difference can be quantified as the angle θ between the vectors in the m and m' models. To accommodate for differences in flux scaling, the computation is done using relative differences, which not shown here for clarity.

2.4.2 Over- /Under- expression

In all analyses, we defined an enzyme/gene to be under- (over)expressed when its expression was below (above) 0.5σ to 1.0σ from the mean in the gene expression distribution (see main text for references to gene expression datasets).

2.4.3 Mapping gene expression to enzymatic activity using the GPR associations

To map the gene expression to expression on the enzyme level, we used the boolean gene-protein- reaction (GPR) associations available in the H. sapiens recon1 [23] metabolic model, downloaded from the BIGG database [145]. These rules indicate which genes need to be expressed using the two boolean operators "and"

and "or". An example of such a rule is the following:

$$E1 = (g1 \; or \; g2) \; and \; g3 \tag{2.1}$$

which indicates that either gene 1 or gene 2 needs to be expressed (or both) in combination with gene 3 to encode the enzyme E1 and allow its activity. For the TCGA dataset with 7362 cancer patients with various tumor types and dataset matched with shRNA, presence/absence calls were assigned by the method described by [140]. Given the presence/absence of the gene transcripts, the GPR rules could be applied to determine the presence/absence of the enzyme.

For the breast- [141] and ovarian [142] cancer datasets, measured transcription levels were provided. In this case, we converted the boolean rules in a way that is commonly used in metabolic networks [146, 61].

- **OR** rules were converted to the maximum transcription level of either of the genes, i.e. (g1 or g2) was converted to max(g1, g2).
- **AND** rules were converted to the minimum transcription level of either of the genes, i.e. (g1 and g2) was converted to min(g1, g2).

To deal with different patients, having different tumor subtypes and expression levels, we first normalized the enzyme expression by subtracting the mean expression along all enzymes in each patient. Then, we computed the expression standard deviation σ_j for each enzyme j. Denote $E_{i,j}$ as the enzyme expression of enzyme j in patient i. Furthermore, denote E_j as the median expression of enzyme j across

all patients. Then, the SDL (A↓, B↑) is considered to be expressed at a cutoff in patient i, when $A_{i,j} \leq$A$_i$ - $\tau\sigma_j$ *and* B$_{i,k} \geq$B$_i$ + $\tau\sigma_k$, $\forall k \neq j$.

2.4.4 Computing the frequency of SDLs in cancer tissue (F$_{SDL}$)

Since SDL is detrimental to cell growth, cancer cells are expected to select against their activation. To test this hypothesis, we computed how often SDLs are activated in tumor cells, compared to randomly selected enzyme pairs. We used a dataset where gene expression was measured for 26 different tumor types collected from 7362 cancer patients. We followed the protocol from [140] to calculate for each gene in a tumor sample the Z-score and P-value to infer its underexpression (-1), overexpression (+1) or no alternation (0) compared to the level of expression in normal tissues. Comparisons were carried out between the exact same tissue types if 25 or more samples were available. Otherwise, all normal tissue samples irrespective of tissue type were used. To calculate Z-scores, normalized RSEM data was retrieved from the TCGA consortium and log2 transformation was applied on the normalized RSEM. Z-score is then calculated as Z = (expression in single tumor sample) – (mean expression in normal samples) / (standard deviation of expression of normal samples). We applied the false discovery rate method in R [144] to correct for multiple hypothesis testing. A cutoff of adjusted p-value 0.05 was used for defining under- or over-expression. Gene expression was then mapped to metabolic enzyme expression using the boolean GPR rules in the GSMM.

In equation 2.2 we compute n_1 , the number of SDLs (A↓, B↑) that are active,

relative to n_2 and n_3, the number of single under- (A↓, B) or over expressed (A, B↑) enzymes.

$$\text{f}=\frac{n_1}{n_1+n_2+n_3}\begin{cases} n_1 = A\downarrow, B\uparrow \\ n_2 = A\downarrow, B \\ n_3 = A, B\uparrow \end{cases} \tag{2.2}$$

We compute this fraction f for our list of SDLs: f_{sdl} and a list of randomly constructed enzyme pairs: f_{rand} of equal length. Let L be the number of pairs, then we computed the fraction

$$F = \frac{1}{L}\sum_{i=1} L1_{F_{sdl,i}<F_{rand,i}} \tag{2.3}$$

where 1 is an indicator function that returns 1 if the fraction F is smaller for the i^{th} SDL and 0 otherwise. The expression of an SDL does not occur frequently, meaning that n_1 is often 0. In that case, we cannot distinguish the expression of SDL from random expression. Therefore, we only consider cases where $n_1 > 0$, meaning that at least one patient activated the SDL or random pair. To analyze how the angle of our predicted SDL influences the result, we repeat this procedure for $N = 5000$ random iterations at different cutoffs. Finally, the empirical p-value is computed as:

$$p = \frac{|F < 0.5| + 1}{N + 1} \tag{2.4}$$

Notice that the reported empirical p-values of 2e-04 are the lowest possible, meaning that for all 5000 iterations, the median fraction of expressed pairs (A,B) was smaller for the SDLs.

2.4.5 Computing the tumor size and patient survival significance for SDLs

To test whether cancer patients with one or more SDLs active live significantly longer than patients expressing randomly selected enzyme pairs, we adopted a significance test analogously to the one presented in [61]. Based on the metabolic enzyme expression of each of the predicted SDLs we defined two groups of patients:

- $SDL^{+}group$: patients whose tumors under expressed enzyme A and over expressed enzyme B in the SDL (A↓, B↑);
- $SDL^{-}group$: patients whose tumors did not activate the SDL (A↓, B↑);

Enzymes were considered under (over) expressed when their expression was below (above) 1σ (OC) or 1.5σ (BC) the mean enzyme expression measured across all patients in the data set. For each SDL, a Kaplan-Meier (KM) [147] survival curve was plotted for the SDL^{+} and SDL^{-} group of patients. Then, we performed a logrank test [143] that returns a p-value denoting the significance in survival between the two groups considered. The logrank test takes censoring (patients did not die during the study) into account, which allows us to use both deceased and censored patients. In order to integrate the logrank p-values for all SDLs in

a later stage, we computed a signed KM-score. This KM-score is defined as sign * -ln(logrank p-value) and hence the more significant the logrank p-value is the higher the absolute KM-score will be. The sign of the KM-score is positive when the prognosis is better (measured by the median survival time) for the SDL^+ group than for the SDL^- group and negative otherwise. We repeated this analysis for 10,000 randomly selected enzyme pairs (A, B). The SDLs are expected to impede tumor growth and therefore lead to longer survival. Therefore, the median KM-score is expected to be positive. The randomly selected pairs are not expected to significantly affect tumor growth, and therefore their median KM-score is expected to be close to zero. Finally, we computed a p-value for the difference between the SDL and random group by applying a Wilcoxon ranksum test to their KM-scores. This score provides an integrated significance score for cancer survival associated with the SDLs, compared to random reaction pairs. The same approach was applied to test the survival difference between patients that express the SDLs and those that only under express enzyme A (A↓, B) or only over express enzyme B (A, B↑). For the tumor sizes the KM survival curves or logrank test are not applicable. To test whether tumor sizes were significantly smaller (larger) for SDLs compared to randomly selected enzyme pairs, we substituted the logrank test for a Wilcoxon ranksum test. Parallel to the signed KM test, we defined the sign positive when the median tumor sizes for the SDL patients were smaller and negative otherwise. Then results for all pairs were integrated in the same way as for the signed KM test. There is quite some redundancy in the SDLs we found. This can be explained in the following way: consider a linear metabolic pathway with three reactions, catalyzed

by enzymes A, B and C respectively. If enzyme D forms a SDL A-D, then B-D and C-D must also be SDLs and in fact they must have the same angle θ. This does not affect the heat maps we showed, because we only selected a patient once if she expressed at least one SDL. Now, we compare for each pair and we want to avoid counting a patient multiple times for related SDLs. To that end, we removed all duplicates, reducing the set from over 22.000 to little over 5000 "non-redundant" SDLs. Finally, to make our test more robust, we only included enzyme pairs (SDL, or random) that are expressed by at least 10 patients.

2.5 Supplementary Information

2.5.1 Six hub reactions

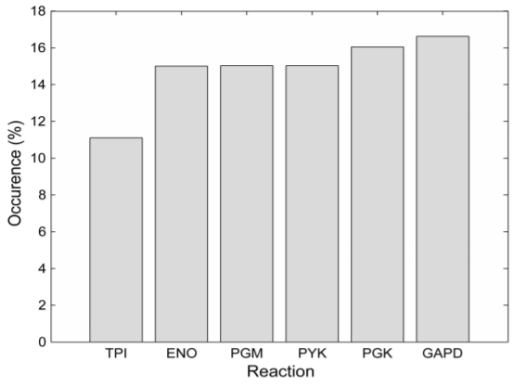


Figure S2 | Six reactions in the glycolysis pathway constitute major hubs in the SDL network. Together, they occur as the knock-out partner in nearly 90% of the SDLs with $\Theta > 15^{o}$.

2.5.2 Enriched metabolic pathways

Table S3: Metabolic pathways that are significantly (hypergeometric p-value) enriched with overexpressed SDL partners

Bile Acid Biosynthesis	0
D-alanine metabolism	0
Glyoxylate and Dicarboxylate Metabolism	0
Heme Degradation	0
Taurine and hypotaurine metabolism	0
Tetrahydrobiopterin	0
Transport, Peroxisomal	0
Cholesterol Metabolism	2.72E-14
Eicosanoid Metabolism	9.03E-08
Heme Biosynthesis	1.36E-06
Fatty acid oxidation, peroxisome	1.78E-06
Tryptophan metabolism	1.78E-06
Fatty acid elongation	6.76E-05
Tyrosine metabolism	0.0001353
R Group Synthesis	0.0002121
Arginine and Proline Metabolism	0.0002701
Pyrimidine Biosynthesis	0.0004913
CoA Biosynthesis	0.0011847
Vitamin D	0.0011847
N-Glycan Degradation	0.0015528
Cysteine Metabolism	0.0022291
Transport, Endoplasmic Reticular	0.0024682
Lysine Metabolism	0.0036013
beta-Alanine metabolism	0.0055741
CoA Catabolism	0.017605
Transport, Mitochondrial	0.0378463

2.5.3 Enriched metabolic pathways

Table S4: SL partners of the six major glycolytic hubs. Compared to the number of SDL partners, there are very few SL partners.

Reaction name	**Pathway**
phosphoglucomutase	Glycolysis/Gluconeogenesis
glycine hydroxymethyltransferase, reversible	Glycine, Serine, and Threonine Metabolism
phosphoglycerate dehydrogenase	Glycine, Serine, and Threonine Metabolism
ATP synthase (four protons for one ATP)	Oxidative Phosphorylation
ubiquinol-6 cytochrome c reductase, Complex III	Oxidative Phosphorylation
phosphoserine phosphatase (L-serine)	Glycine, Serine, and Threonine Metabolism
phosphoserine transaminase	Glycine, Serine, and Threonine Metabolism
cytochrome c oxidase, mitochondrial Complex IV	Oxidative Phosphorylation

2.5.4 Enriched metabolic pathways

Table S5: Pathway that are significantly enriched by dosage partners of the six glycolytic hubs. Significantly changed reactions were identified by sampling the flux space under three conditions (i.e. wild-type, KO, and SDL). Reactions were considered to be significantly different if the median of the sampled flux value under the SDL condition is 20% lower than in the other two conditions, or if the flux direction under the SDL condition was reversed. Using a hyper-geometric test 10 pathways were identified that are significantly affected by the SDL interaction.

Pathway	p-value
Bile Acid Biosynthesis	0
Eicosanoid Metabolism	0
Taurine and hypotaurine metabolism	0
Oxidative Phosphorylation	0.0003
beta-Alanine metabolism	0.0003
Transport, Peroxisomal	0.0006
Lysine Metabolism	0.001
Fatty acid elongation	0.0014
Fatty acid oxidation, peroxisome	0.0055
Glyoxylate and Dicarboxylate Metabolism	0.0057

2.5.5 ER- breast cancer survival times

Survival times for ER- patients show to our surprise a positive correlation with tumor size, quite the opposite from the ER+ patients. In general, patients ER- have poorer prognosis (median survival time 5.85 years) compared to ER+ patients (median survival time 7.42 years). In this case, although IDLE effectively finds the patients with the smallest tumor sizes they exhibit the worst survival time (figure 2.7 panel a). This also holds for patients whose tumors only under express enzyme A and do not over express the dosage enzyme B (figure 2.7 panel b). Only the SDL patients with an extreme underexpression of enzyme A (σ ¿ 3) show the negative correlation. ER- patients receive chemotherapy significantly more often than ER+ patients (56% on average, vs 11%). This probably affects the expression of oncogenes and tumor suppressors, which causes that more SDLs are active. Although this is beneficial for the tumor size of these patients, the fact that these patients needed chemotherapy could be reflected in their poor survival prognosis.

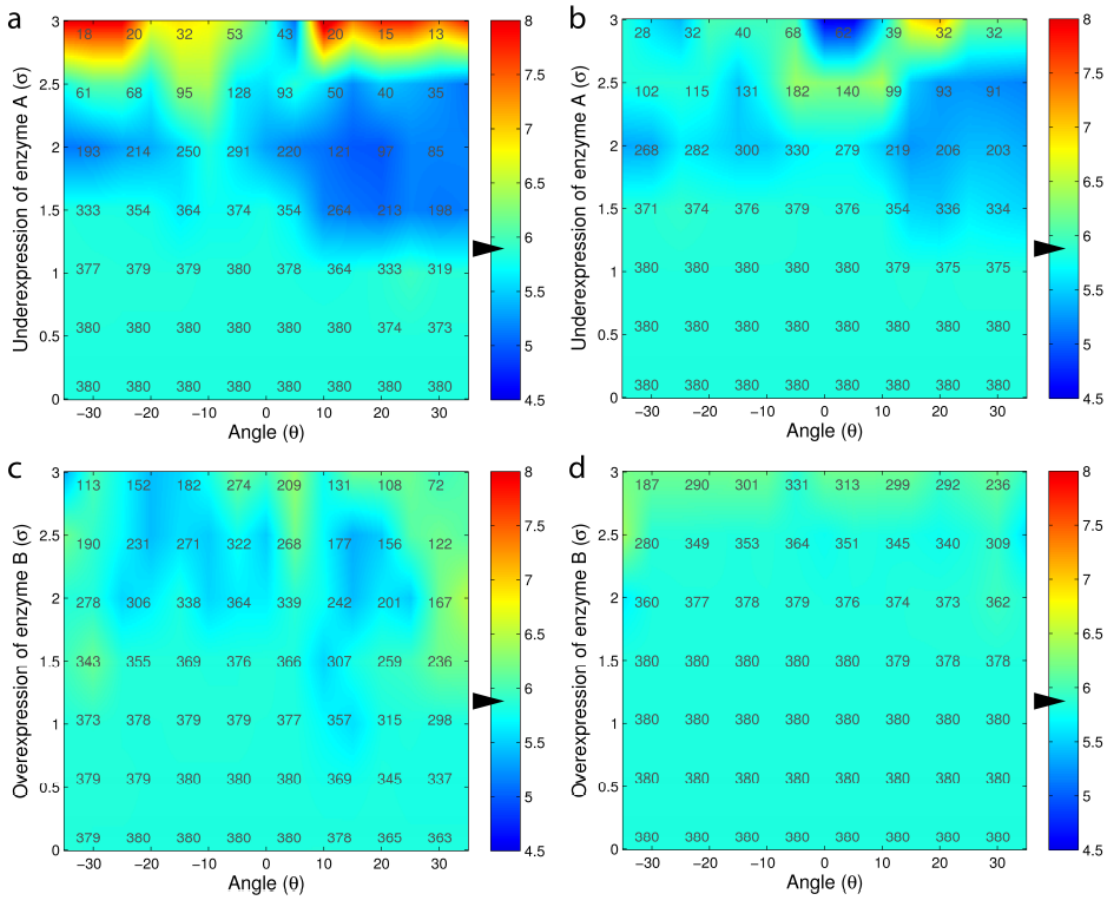


Figure 2.7: Median ER- breast cancer survival time (years). The arrowhead denotes the median survival time (5.8 years) for all ER- patients. a) Patients with at least one SDL (A↓, B↑) active, with enzyme B overexpressed at a constant (cutoff) value. b) Patients that only under express enzyme A (A↓, B). c) Patients with at least one SDL (A↓, B↑) active with A underexpressed at a constant (cutoff) value. d) Patients that only over express enzyme B (A, B↑).

Chapter 3

Studying the regulation of breast cancer metabolism from multi-omics data

⋆⋆ Published as "The landscape of tiered regulation of breast cancer cell metabolism", Scientific Reports 2019

3.1 Introduction

Cancer cells adapt their metabolism to facilitate biomass formation to support their rapid proliferation. Transcriptional regulation alone does not account for many of the metabolic alterations observed in cancer [148, 149], suggesting that post-transcriptional, post-translational and protein phosphorylation mechanisms may play an important role in modulating cancer metabolism and determining cancer cell phenotypes [150, 151, 152, 153]. Here we aim to chart the transcriptional, post-transcriptional and post-translational regulation of MCF7 breast cancer cell metabolism on a genome scale. This is performed via measurements of multi-omics data employing MCF7 breast cancer cells under three different in vitro growth conditions, and its analysis via an integration of this data within a genome scale metabolic model (GSMM) of human metabolism. Our approach is inspired by previous large-scale omics studies of the multi-level regulation of bacterial metabolism [44, 45, 43] and yeast [46], which have advanced our understanding of the organization and

regulation of metabolism in these organisms.

Genome scale metabolic modeling is an increasingly widely used computational framework for studying metabolism. Given the GSMM of a species alongside contextual information such as growth media and omics data, it has been shown that one can fairly reliably predict numerous metabolic phenotypes, including cells' growth rates, metabolite uptake and secretion rates and internal fluxes, gene essentiality, and more. Over the last few years, GSMMs have successfully served as a basis for many computational studies of cancer, e.g. [113, 114, 131, 132, 133, 134]. GSMMs have also been used to predict post-transcriptional regulation of metabolic enzymes in healthy tissues [26] but going beyond that to systematically analyze metabolic regulation in cancer is addressed here for the first time to the best of our knowledge.

3.2 Results

3.2.1 Data collection and preliminary model-free analysis

We collected omics measurements in MCF7, a breast cancer cell line, grown under three different conditions: (1) Minimum Essential Medium (MEM) with glucose and without glutamine (MEM-Gln), (2) MEM with glucose and glutamine (MEM) and (3) MEM with glucose, glutamine and supplemented with Oligomycin – an inhibitor of ATP synthase that inhibits cell respiration (MEM+Oli). The media were chosen because they reflect multiple stress conditions for the cell: one media (glutamine deprivation) is chosen because MCF7 cells rely on glutamine as the main

source of energy, and the other media (supplement of Oligomycin) is chosen because it emulates tumor hypoxic conditions.

The measurements were repeated twice under each condition at two time points - after 8 and 24 hours, resulting in overall 6×2 multi-omics datasets. Each such dataset includes the gene-expression of 1372 metabolic genes, proteomics for 486 metabolic enzymes (97% of the measured enzymes have gene expression values), phosphorylation values for 71 phosphorylation sites on metabolic enzymes, and flux measurements of 44 metabolic reactions (see methods). To obtain flux measurements, we fitted all the data obtained through spectrophotometric measurements and 13C assisted metabolomics experiments using our in-house developed software that simulates dynamics of metabolites 13C labeling, Isodyn [26, 154, 155, 156, 157]. Fitting the data allows determining the metabolic flux profiles of MCF7 breast cancer cells under three different growth conditions (see methods). Figure 3.1 summarizes the qualitative changes in the metabolites and their analysis using Isodyn. The analysis demonstrates a decrease in the fluxes of glycolysis, lactate production, pentose phosphate pathway (PPP) activity, tricarboxylic acid cycle (TCA) cycle utilization and fatty acid synthesis when the cells are at MEM-Gln growth condition compared to MEM. Moreover, increased pyruvate cycle, which is the conversion of pyruvate to oxaloacetate via pyruvate carboxylase followed by its conversion to malate and consequently back to pyruvate via malic enzyme, occurs mainly in MCF7 cells at MEM-Gln condition compared to the MEM growth condition. On the other hand, in the MEM+Oli growth condition, increased glycolysis, lactic acid fermentation and pyruvate cycle are observed compared to the MEM growth condition,

together with decreased TCA cycle activity, PPP and lipogenesis.

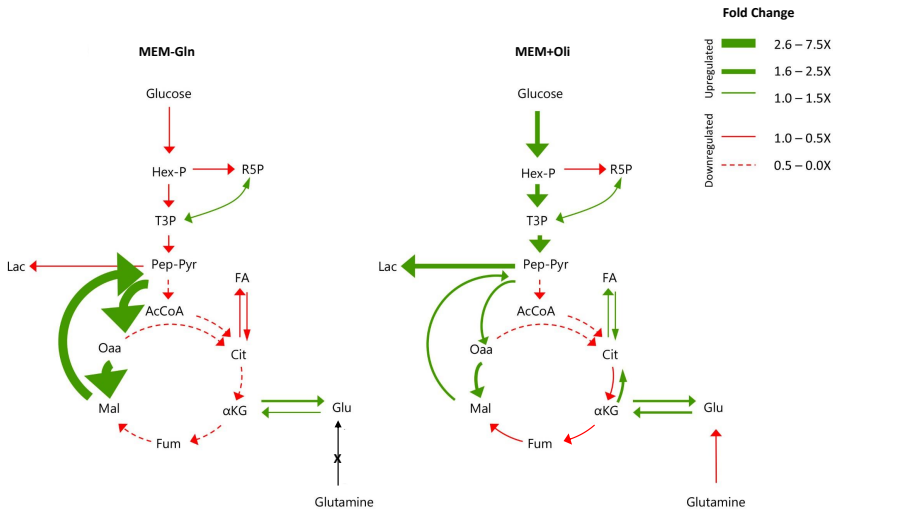


Figure 3.1: Metabolic flux map of MCF7 breast cancer cells under MEM-Gln or MEM+Oli growth conditions compared to MEM condition. The fluxes were estimated by using Isodyn software. In each growth condition, the calculated flux was normalized against the flux of MEM growth condition in order to calculate the net change.

To obtain a genome wide view of pathway-level differences in the transcriptional data across the different growth conditions, we first compared (using a t-test) the metabolic gene expression values between the different growth conditions to identify metabolic pathways that were significantly up or down regulated in any of these conditions compared to the others. We found that upon oligomycin treatment, carnitine shuttle pathway is downregulated compared to the other growth conditions, as well as the urea cycle/amino group metabolism pathway. On the other hand, fatty acid activation and C5-Branched dibasic acid metabolism (among other pathways) were found to be elevated upon such treatment - a full list of the significant growth condition-specific changes is provided in Supplementary Tables in subsections 3.4.1, 3.4.2, all p-values were FDR corrected for 0.05). A similar analysis

of the proteomics data revealed different results. While carnitine shuttle pathway activation was consistent with the gene expression analysis, the fatty acid pathways (activation, elongation and oxidation) were now found to be downregulated upon Oligomycin treatment. These results, consistent with previous observations both in yeast [158, 159] and in human [160, 161], point to the significant differences between the mRNA and protein levels of many metabolic enzymes and call for a systematic study of their potential functional regulatory implications.

3.2.2 Overview of the metabolic modeling based analysis

Our main goal in this study is to use the measured multi-omics data to systematically chart the different layers of metabolic regulation in breast cancer cells that orchestrate the actual metabolic flux across the network's reactions occurring in each growth condition. Ideally, measuring the actual fluxes in each condition directly via tracing experiments would be adequate, but obviously, this can currently be done only for a small number of fluxes that are mainly involving central cell metabolism. Hence, alternatively, we integrated the various omics data measured in each growth condition within a genome scale model of human metabolism [23] to infer the likely metabolic fluxes given these data in a genome wide manner. After an initial validation of these predictions, we proceeded to compare the flux predictions of the resulting reactions to the corresponding enzymes' omics data to identify their regulation. This is performed in a stepwise manner as follows (Figure 3.2):

1. **GSMM based identification of transcriptional and translational di-**

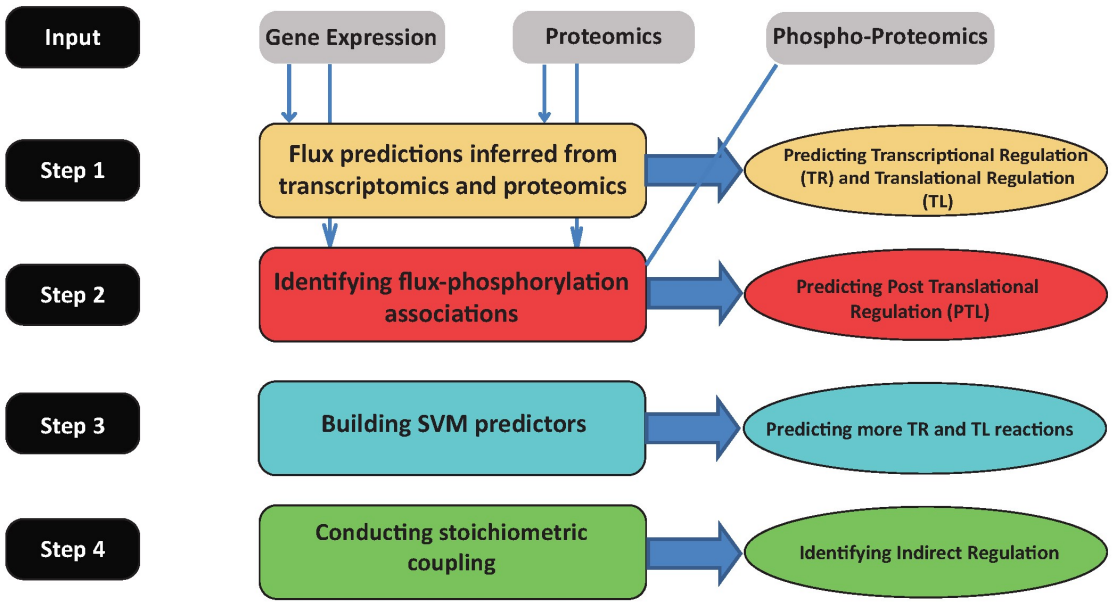


Figure 3.2: Systematic identification of reactions' regulation: Step 1: Using gene-expression and proteomics data to predict transcriptionally and translationally regulated reactions. Step 2: Using phospho-proteomic data to predict post-translationally regulated reactions. Step 3: Based on the results of step 1, build predictors of TR and TL regulation. Step 4: Identifying indirectly regulated reactions that are metabolically regulated via stoichiometric coupling.

rectly regulated reactions: We first identify reactions that are directly regulated – that is, reactions whose model-based predicted flux alterations across the different conditions studied can be accounted for by molecular alterations at any one of the levels measured: those include reactions that are primarily transcriptionally regulated and primarily translationally regulated. These assignments are done in a mutually exclusive manner, as follows: (1) transcriptionally regulated reactions (TR) are those reactions whose enzymes' gene expression levels match the predicted fluxes. (2) translationally regulated reactions (TL) are those reactions whose predicted flux levels do not match their gene expression levels, but they match the protein levels of their enzymes.

2. **GSMM based identification of post-translationally directly regu-**

lated reactions: Post-translationally regulated reactions' (PTL) assignments are given to the reactions where both the enzymes' gene expression and proteomics levels do not match the predicted flux levels but the predicted flux levels across the different growth condition can be significantly associated with changes in the phosphorylation levels of the enzymes.

3. **Building machine learning predictors of additional directly regulated reactions:** For the majority of the metabolic reactions, however, we did not find omics evidence testifying that they are directly regulated at any of these three levels. One major reason for that may be the limited scope of the proteomics and phospho-proteomics measurements. We, therefore, built machine learning based predictors of TR and TL regulation based on the reactions that have already been labeled as such via the model-based analysis in step (1). Then, we applied these predictors in a genome wide manner to further identify sets of reactions that are predicted to be TR or TL regulated (detailed below). We then performed various genome wide analyses to further test and validate the veracity of these predictions.

4. **Identifying stochiometrically coupled, indirectly regulated reactions:** Finally, even after this prediction step, a large set of reactions still remains unassigned and are labeled as indirectly regulated. A major likely source of such indirect regulation is metabolic regulation [162], which manifests itself in the stoichiometric coupling of the fluxes of different reactions across the metabolic network, and which we study further using the human

metabolic model.

Below we provide a detailed description of each of these four steps and the results they uncover.

3.2.3 Step 1: Identifying transcriptionally regulated (TR) and translationally regulated (TL) reactions

We first aimed to predict the fluxes of the reactions in each condition, to determine which reactions are directly regulated and at what level they are regulated. To this end, we used iMAT (the integrative Metabolic Analysis Tool) [26], a computational method that systematically predicts metabolic fluxes in a GSMM by incorporating omics data (transcriptomics and/or proteomics) that represent the activity level of the metabolic enzymes. iMAT considers the gene expression or protein levels as cues for the likelihood that the enzymes in question carry a metabolic flux in their associated reactions. It then leverages the GSMM to accumulate these cues into a global flux distribution that is stochiometrically consistent and maintains mass balance across the entire metabolic network (see methods).

To this end we first tested if the above described procedure yields flux predictions that are in accordance with those quantified with 13C Metabolic Flux Analysis (13C MFA). To this end, we combined both mRNA and protein expression measurements and used iMAT, a tool that extends upon the standard flux balance analysis (FBA) to predict the flux distribution that is the most likely given both types of data. Briefly, following a procedure already established and validated by [26], the

activity level of an enzyme was set according to the proteomics data when these data were available and according to the gene-expression otherwise, leaving the activity level unconstrained when large disparities existed between the gene expression and the proteomics data (see methods). Reassuringly, the accuracy of predicting the experimentally measured fluxes was significant across all growth conditions (Spearman correlation coefficient across all growth conditions = 0.42, p-values < 8.9671e-25, see Figure 3.3 for the correlations obtained at each of the three different growth conditions).

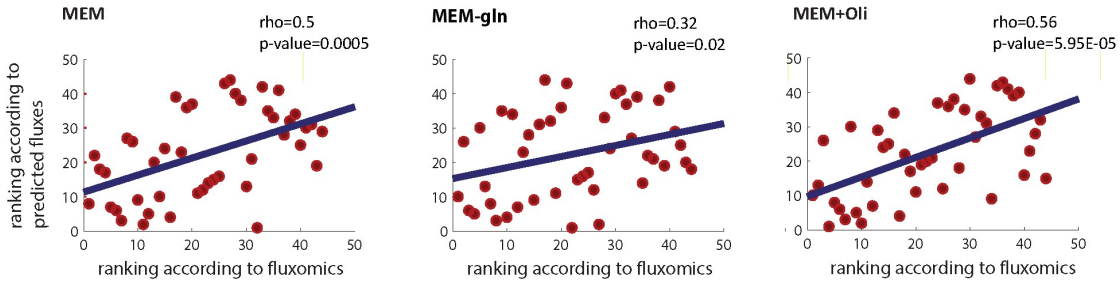


Figure 3.3: Scatter plot depicting the association between the measured and predicted fluxes in each of the three media conditions. Flux predictions were obtained by integrating the transcriptomics and proteomics data within the human metabolic model, as described in the main text.

Given these network wide flux predictions, we next set to identify the reactions that are transcriptionally regulated (TR). To this end we discretized the gene expression measurements and the predicted fluxes into three levels of activity: low (TR-low), moderate (TR-moderate) and high (TR-high). We then compared predicted flux level of each reaction to the discretized gene-expression level of the pertaining enzymes (see methods). Reactions whose predicted flux levels matched gene expression levels of their enzymes across the different measurements were considered to be TR. For the three conditions (MEM-Gln, MEM and MEM+Oli), 562,

550 and 556 reactions (approximately 28% of the model reactions) were identified as TR, respectively. Supporting these predictions, we found that the group of predicted TR reactions is enriched with transcription factor binding sites (using ENRICHR tool [163, 164], we calculated the enrichment according to several databases: Jaspar [165] and Transfar [166] (hyper-geometric p-value = 9.5892e-05), ChEA [167] (hyper-geometric p-value = 1.2819e-10) and ENCODE [168, 169] (hyper-geometric p-value = 0.0029)) (see methods).

To predict translational regulation (TL), we searched for reactions whose (discretized) predicted flux activity levels were different from the transcriptomic levels of their enzymes. Such transcriptomic/flux 'discordant' reactions whose activity levels were high (low) according to the gene expression of their enzymes but low (high) according to the flux predictions are considered to be post-transcriptionally down-(up-)regulated. The correlation between the proteomics data and the predicted fluxes for this subset of TL predicted reactions was high and significant (rho = 0.75, 0.6, 0.5, for the 3 growth conditions, all p-values <0.0071, Supplementary figure in subsection 3.4.3), as would be expected. It is important to note that in order to avoid circularity, this correlation was calculated in a cross-validation manner only for sub-group which was not constrained in the algorithm input. Among the reactions identified as post-transcriptionally regulated, we denoted the subset of reactions whose predicted flux state highly matches the proteomics (discretized) levels in a given growth condition as translationally (TL)-regulated. Among those, about 15 reactions are predicted to be TL-upregulated (the discretized flux/proteomics activity state is higher than the discretized transcriptomics state), and about 35 are predicted to be

TL-downregulated (the discretized flux/proteomics activity state is lower than the discretized transcriptomics state) (Table in subsection 3.4.5). The specific pathways that are predicted to be TR (high/low/moderate) and TL (up/down) regulated are listed in in Supplementary Table in subsection 3.4.4.

3.2.4 Step 2: Identifying post-translational (PTL) regulated reactions

To identify the reactions that are post-translationally (PTL) regulated, we used the fluxes predicted in the previous step as a reference point. That is, reactions whose predicted flux activity markedly differed both from their transcriptomics and proteomics expression levels (that are hence not predicted to be TR or TL regulated) may be post-translationally (PTL)-regulated. Overall, 34, 39, 42 such reactions have at least one measured phosphorylation site in MEM, MEM-Gln and MEM+Oli, respectively. We next inferred the impact of each of the measured phosphorylation sites on enzyme activity. The phosphorylation data included 56 metabolic enzymes phosphorylated at 71 different phosphorylation sites catalyzing 164 metabolic reactions. For each of the reactions, we computed the Spearman rank correlation between the predicted flux (computed via integrating the pertaining transcriptomics and proteomics data) and the corresponding site phosphorylation levels across all growth conditions and time points measured (subsections 3.4.5). 19 reactions manifested a significant p-value (<0.05) with a strong correlation (Spearman rho $>$—0.6—). These 19 reactions have 13 different phosphorylation sites (SI,

Fig. 4).

The functional impact of phosphorylation is currently known from the literature for only two of these enzymes: (1) phosphorylation of S1859 in carbamoyl-phosphate synthetase 2 (CAD) enhances its in vivo [170] activity, and (2) phosphorylation on S293 causes pyruvate dehydrogenase (PDHA1) enzyme inactivation [171]. Our predictions match both; for the CAD enzyme, we detected a high positive correlation (0.718) and for PDHA1 we obtained a strong negative correlation of 0.6. To test and validate these predictions in our cells further, we performed western blot experiments for both proteins (CAD and PDH together with their phosphorylated forms). We observed a marked phosphorylation of PDH in the predicted conditions for MEM-Gln and MEM+Oli compared to MEM growth condition, indicating its reduced activity under these conditions (Figure 3.4). This is additionally confirmed via flux measurements through 13C MFA. On the other hand, we observed a decreased phosphorylation at CAD protein, indicating a decrease at its activity at MEM-Gln and MEM+Oli conditions, as predicted (Figure 3.4).

3.2.5 Step 3: Genome wide prediction of TR and TL regulation of breast cancer metabolism

In the previous steps, we have identified about 500 reactions that are directly regulated at one of the three regulatory levels described above (TR, TL or PTL). In these reactions, the predicted flux changes were significantly associated with molecular alterations in the pertaining enzymes. However, this leaves a large number

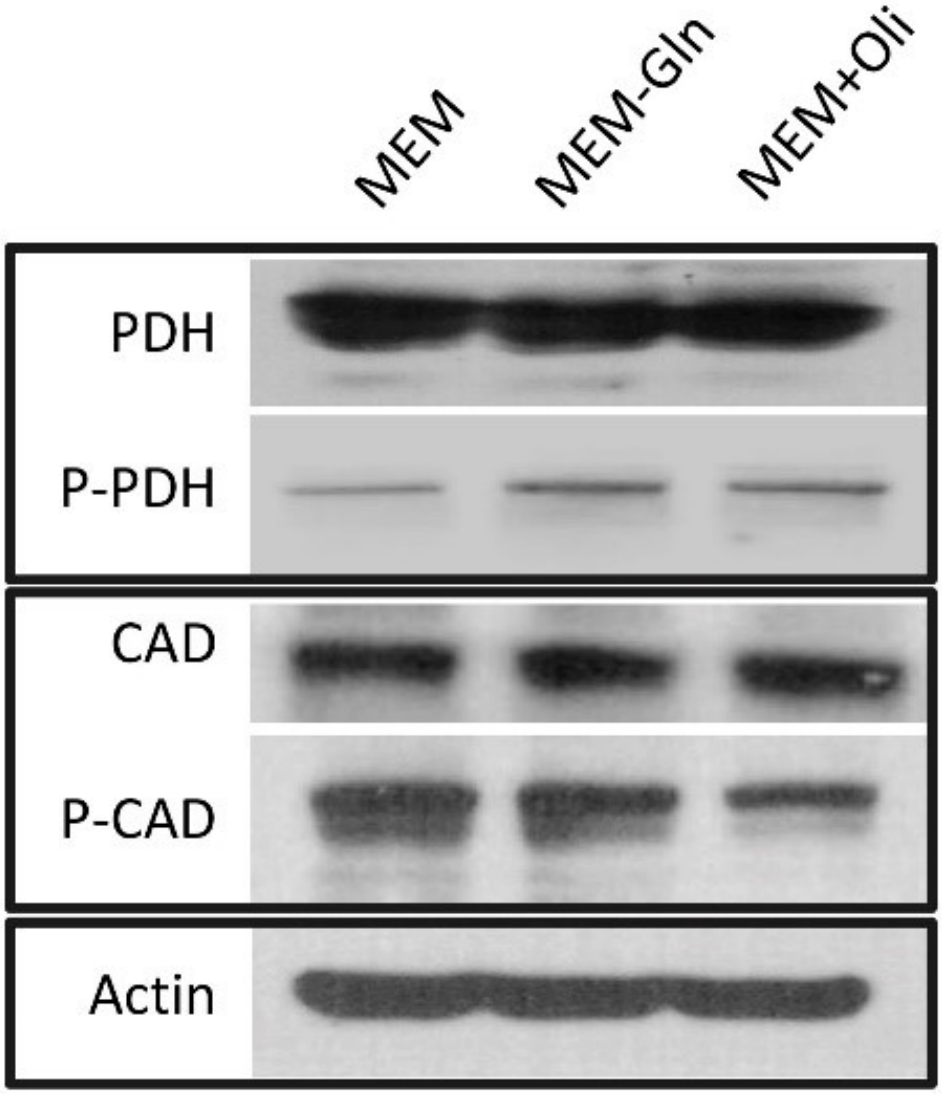
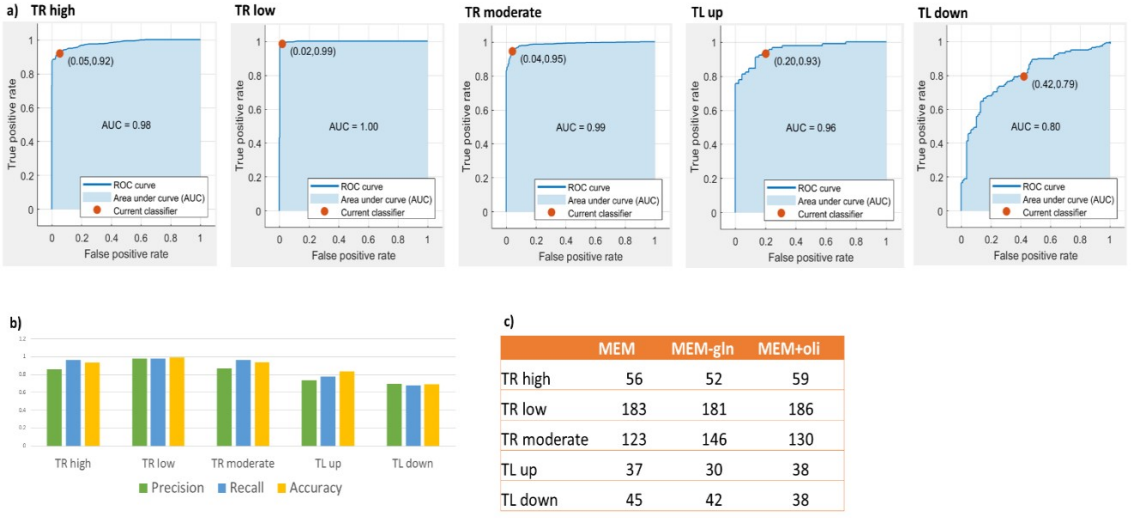


Figure 3.4: Phosphorylation of the indicated proteins (PDH and CAD) at MEM-Gln and MEM+Oli conditions were detected by western blot analysis.

of about 1450 reactions that were not assigned to any of these direct regulatory levels, which can be attributed to the limited scope of our measurements. In order to predict additional reactions that are likely to be directly regulated at TR or TL level, we built five Support Vector Machine (SVM) classifiers for five different direct regulation levels: TR-high, TR-low, TR-moderate, TL-up and TL-down. The goal of each classifier is to predict whether a reaction is regulated at one of these levels or not. The classifier was trained and evaluated using the reactions that have already been labeled as TR or TL regulated in the previous analysis at step (1), using a standard train and test 5-fold cross validation. The classifier input features included the gene expression, proteomics, predicted fluxes and metabolic network characteristics (reversibility information, number of participating metabolites, index of the relevant pathway, and more) of the given reactions, and the TR/TL labels already assigned in the previous steps (see methods). The accuracy of the classifier was measured by comparing the predicted labels against the known labels. The resulting classifiers achieved a high cross validation prediction accuracy (mean AUC >0.946 for all classifiers, all values are presented in Figure 3.5 panel a; recall and precision values are presented in Figure 3.5 panel b). Applying this to predict the direct regulation of the 1450 remaining reactions, 450 additional reactions were predicted to be regulated at exactly one of the TR/TL levels (in MEM, MEM-Gln and MEM+Oli, see Figure 3.5 panel c for their subdivision in each of the regulation groups). The predicted TR group is enriched with transcription factor binding sites (hyper-geometric p-value = 6.236e-119, see methods. Similarly, the predicted TL group has a significantly higher number of flux/proteomic states matches compared

to the randomly selected sets (empiric p-value = 0.04). It is important to note that the very small numbers of predicted PTL reactions did not enable us to build reliable predictors of regulation at this level. Interestingly, adding the new set of predicted reactions which are directly regulated to those reactions which are previously identified as directly regulated by model based integration uncovers a large number of new pathways that now become enriched in directly regulated reactions (in Supplementary Tables in subsections 3.4.7,3.4.8,3.4.9).



	MEM	MEM-gln	MEM+oli
TR high	56	52	59
TR low	183	181	186
TR moderate	123	146	130
TL up	37	30	38
TL down	45	42	38

Figure 3.5: (a) AUC curves for each of the direct regulation SVM classifiers; (b) mean precision and recall values for each of the SVM classifiers; (c) number of reactions that have been uniquely predicted to be directly regulated by one of the classifiers.

3.2.6 Step 4: Studying the reactions that are indirectly regulated via stoichiometric coupling

After the predictions we performed at step 1–3, around 1000 reactions yet remained not to be predicted as directly regulated, some of which are likely to be further identified as regulated with more extensive data. However, many of these

remaining unassigned reactions may still be truly indirectly regulated (IR) reactions where their flux may be primarily metabolically-regulated by changes in their substrate and product levels due to changes in the flux activities of other reactions in the metabolic network. That is, their flux may be stoichiometric coupling (SC-regulated) to the flux of other reactions in the metabolic network [172, 173, 174].

In the framework of MCA (Metabolic Control Analysis), it has been established that network structure is an important determinant of metabolic control [175]. Accordingly, a perturbation in enzyme abundance or activity can be propagated through reactions stoichiometry coupled to the reaction catalyzed by such enzyme. To study such dependencies on a genome-scale, we used flux sampling to quantify the pairwise stoichiometric couplings between all the metabolic reactions in the human network, identifying for each reaction how tightly its flux is coupled to the flux of each of the other reactions, in each of the different conditions (see Methods).

Remarkably, we found that the 1000 'unassigned' indirect reactions have significantly higher stoichiometric couplings to the TL and PTL directly regulated reactions than among themselves across the different growth conditions (using one sided Wilcoxon test, p-values = 6.9163e-158 and 2.945e-14, respectively). These findings point out that the regulation of cellular metabolism may be governed in a hierarchical manner where the flux of many indirectly regulated reactions is determined via stoichiometric coupling to the flux of others, directly regulated reactions. Finally, we found that the group of 1000 indirectly regulated reactions is highly enriched with bi-directional reactions (hyper-geometric p-value = 1.15e-28, 2.21e-32,

5.54e-32 for each condition, see Methods). This observation can be explained by metabolic control analysis (MCA) [176] theory: In the framework of MCA, enzyme activities catalyzing reversible reactions, which often are in rapid equilibrium, usually have low flux control coefficients and hence are poor targets of direct regulation. Indeed, the combination of the 'directional flexibility' of candidate SC-regulated reactions with their enhanced coupling to other directly-regulated reactions is likely to facilitate the formation of stoichiometrically feasible flux distributions across the metabolic network, providing a way for efficiently regulating the metabolic state with minimal cellular costs in terms of transcriptomics, proteomics and phosphoproteomics regulation.

3.2.7 Discussion

This study integrates transcriptomics, proteomics, phospho-proteomics and fluxomics data with metabolic modeling to provide the first chart of metabolic regulation in MCF7 breast cancer cells on genome scale. We classified the metabolic enzymes as those that are predicted to be directly regulated at three distinct levels (TR, TL, and PTL) and those that are predicted to be indirectly regulated, given the current coverage of omics data. As expected, we found that citric acid cycle is generally upregulated both on the transcription and translational level. Interestingly, while on the transcriptional level fatty acid oxidation was found to be generally down-regulated, it is up-regulated on the translational level. In addition, oxidative phosphorylation – another hallmark of cancer, was found to be up-regulated only on

the translational level (not including MEM+Oli medium). These findings further highlight the pivotal role of translational regulation in cancer and the importance of obtaining higher coverage of proteomic data, whenever possible.

Remarkably, we found that the flux of the indirectly regulated reactions is coupled to the flux of directly regulated ones. We also found that the indirectly regulated reactions are enriched with bi-directional reactions. These findings might open an opportunity for further research to determine an extent by which their activity levels are set by other reactions. Taken all together, these findings suggest that the regulation of breast cancer cell metabolism is controlled in a hierarchical manner where the direct regulation of about half of the reactions suffices to orchestrate the flux regulation through the whole metabolic network via flux coupling.

Like almost any other computational, genome scale investigation, our approach has quite a few limitations. First, the data itself, is still limited and noisy, and the coverage of different layers of omics data is uneven, due to obvious technical limitations. Second, guided by the data we collected, we focused here on studying post-translational modifications mediated by phosphorylation. However, obviously, post-translational modifications occur via a variety of additional mechanisms, including, e.g., acetylation, glycosylation and allosteric regulation [177, 178]. Consequently, the machine learning predictors built for predicting transcriptional regulation and post-transcriptional regulation, but not post-translational regulation. Fourthly, as we employ coarse discretization to overcome some of the noise in the data, we only identify regulatory alterations in reactions that are differentially active across the conditions of study. This limitation is partly mitigated, however, by

analyzing three very distinct metabolic states. Future work should aim to address these limitations by incorporating data sets covering more conditions, measuring a wider range of omics data with higher coverage, and ideally, move to perform such measurements in patients' tumor data. With the advent of omics technologies such data may become readily available soon and may be benefit from the conceptual and computational framework laid out in the current study.

Although we analyzed multiple layers of omics data, their coverage has been limited: while we had gene expression data for all 1372 metabolic genes, the coverage of our cutting-edge proteomics measurements provided data for only 486 metabolic enzymes and 71 of their phosphorylation sites. Flux measurements using 13C labeling are understandably even more limited in their scope, covering only central carbon metabolism. Aiming to make the best use of the available data and to obtain a genome-wide view of breast cancer cell metabolism, we used a modeling approach to integrate the data and infer the most likely genome-scale flux distributions. Additional work aiming to deal with the limited coverage problem was carried out via creating SVM predictors that used the known network properties together with measurements with high coverage and helped us extend the scope of the study to the utmost. With rapid advancement of high-throughput technology and accumulation of more comprehensive omics data across additional cellular conditions, the conceptual and computational framework exhibited here lays the methodological foundations for gradually obtaining a more comprehensive view of metabolic regulation in both breast cancer and other cancer types.

3.3 Materials and Methods

3.3.1 Genome-scale metabolic modeling (GSMM)

A metabolic network consisting of m metabolites and n reactions can be represented by a stoichiometric matrix S, where the entry S_{ij} represents the stoichiometric coefficient of metabolite i in reaction j [3]. A GSMM model imposes mass balance, directionality, and flux capacity constraints on the space of possible fluxes in the metabolic network's reactions through a set of linear equations:

$$S \cdot \ v = 0 \tag{3.1}$$

$$v_{min} \leq v \leq v_{max} \tag{3.2}$$

where v stands for the flux vector for all of the reactions in the model (i.e. the flux distribution). The exchange of metabolites with the environment is represented as a set of exchange (transport) reactions, enabling a pre-defined set of metabolites to be either taken up or secreted from the growth media. The steady-state assumption represented in equation 3.1 constrains the production rate of each metabolite to be equal to its consumption rate. Enzymatic directionality and flux capacity constraints define lower and upper bounds on the fluxes and are embedded in equation 3.2.

In the following, flux vectors satisfying these conditions will be referred to as feasible steady-state flux distributions.

3.3.2 Pathway enrichment analysis

Based on iMAT results, which was used to predict the regulation of the reactions in the metabolic model, a hypergeometric p-value was computed for each pathway in the model for being enriched with reactions that are regulated in each level. Data for reactions and their pathways were taken from BIGG database [145]. A correction for multiple hypotheses was done using false discovery rate method of 0.05.

3.3.3 Using iMAT with transcriptomics and proteomics as its input

We first employed a discrete representation of significantly high or low enzyme-expression levels across tissues. Gene expression and proteomics levels were discretized to highly (1), lowly (-1), or moderately (0) expressed, for each sample. This discretization was conducted as follows: the 1/3 of the proteomics with the highest values to be considered as highly expressed, and vice versa for lowly expressed. When proteomics data was not available, transcriptomics data was used (again – top 1/3 as lowly expressed, and vice versa). One could argue that the different levels of coverage between transcriptomics and proteomics could suggest using different thresholds for determining 'active' and 'inactive' genes in the respective analysis; To keep a systematic approach, here we opted to treat both data measurements in the same, uniform, way (but other approaches may be taken in the future. Lastly, in order to avoid direct effect of the coverage differences between proteomics and transcriptomics, we determined a moderate expression level for genes whose level

according to the gene expression was high (low) and according to the proteomics low (high), and left their corresponding enzymes/reactions unconstrained. In iMAT analysis, the discretized gene expression levels were incorporated into the metabolic model to predict a set of high and low activity reactions. Network integration was done by mapping the genes to the reactions according to the metabolic model (see methods), and by solving a constraint-based modeling (CBM) optimization problem to find a steady-state metabolic flux distribution. CBM models the cell as a network of metabolic reactions controlled by hundreds of genes and enables the prediction of feasible metabolic behavior under different genetic and environmental conditions, that are expressed as constraints in the network [9, 179]. By using the CBM approach, we assign permissible flux ranges to all the reactions in the network, in a way that satisfies the stoichiometric and thermodynamic constraints embedded in the model and maximizes the number of reactions whose activity is consistent with their expression state. iMAT's solution may not be unique as a space of alternative optimal solutions (in terms of its objective function) may exist. Therefore, we sampled 2,000 different flux distributions that are all consistent with the reactions' state of activity or inactivity defined in one of iMAT's optimal solutions. To address the potential degeneracy of the CBM solutions, we used the artificial-center-hit-and-run (ACHR) sampling approach [180] which is an efficient sampling approach for a linearly constrained space [181]. The mean flux distribution obtained over the 2,000 samples then serves as an approximation of the source metabolic state.

3.3.4 Gene to reaction mapping

To map the gene expression to expression on the reaction level, we used the boolean gene-protein-reaction (GPR) associations available in the H. sapiens recon1 metabolic model, downloaded from the BIGG database ([182]). These rules indicate which genes need to be expressed using the two Boolean operators "and" and "or". An example of such a rule is the following:

- R1 = (g1 or g2) and g3 (indicating that either gene 1 or gene 2 (or both) need to be expressed in combination with gene 3 to allow reaction 1 activity.
  - OR rules were converted to the maximum transcription level of either of the genes, i.e. (g1 or g2) was converted to max(g1, g2)
  - AND rules were converted to the minimum transcription level of either of the genes, i.e. (g1 and g2) was converted to min(g1, g2).

3.3.5 Bi-directional reactions

Bi-directional reactions are those that can potentially carry flux in both directions (this information is provided in the human GSMM model).

3.3.6 Identifying TR/TL reactions

We compared the discretized gene expression measurements to the activity levels of the predicted fluxes; we took 1/3 of the reactions with the highest flux values to be considered as highly active, and vice versa for lowly active reactions. The rest of the reactions considered to be moderately active. If the activity level

of a reaction matches the discretized value according to the gene expression, in at least 3 out of the 4 cell line replicates, the reaction is considered to be TR. For the rest of the reactions, if the activity level of a reaction matches the discretized value according to the proteomics, the reaction is considered to be TL.

3.3.7 Identifying PTL reaction

Among the reactions that haven't been classified as TR or TL in the way that mentioned above, we found the sub group of reactions that were associated with at least one phosphorylation site. Reactions whose predicted flux activity markedly differed from their transcriptomics or proteomics expression levels, and that were associated with at least one phosphorylation site in 3 of the 4 cell line replications, were predicted to be potentially post-translationally (PTL) regulated.

3.3.8 Finding transcription factor enrichment

First, we found the reactions that were predicted to be TR in all condition. Then, using the reaction-gene matrix, we found the list of genes that catalyze this group of reactions. Using ENRICHR tool [163, 164], we found how many of the genes have (at least one) TFs that bind to their promoter region, from exploring Jaspar [165], Transfar [166], ChEA [167] and ENCODE [168, 169] databases. Same for all model genes. These values were used in the hypergeometric calculation.

3.3.9 Support vector machine (SVM) classification

We built and trained five SVMs classifiers (representing 5 "classes" of regulation, as described in main text). We applied an SVM classifier with a quadratic kernel for each classifier, with the following features:

(1–4) gene expression measurements under 4 data points

(5–8) predicted fluxes under 4 data points

(9) A binary integer indicating if the reaction is reversible.

(10) An integer value associated with a unique metabolic pathway.

(11) The total number of metabolites participating in the reaction.

(12) The total number of substrates participating in the reaction.

(13) The total number of products participating in the reaction.

For the labels, we used the classification of the reactions from the previous steps (1 if it's regulated at that level, 0 otherwise). All SVM classifiers were trained on part of this data, and later tested on all data (mean recall and precision values presented in the text).

Cross-validation was performed by setting aside one fifth of the regulated-predicted reactions in the training set. The classifier was trained on the remaining four. The classifier's accuracy was measured by comparing the predicted labels against the known labels.

3.3.10 Computing pairwise flux correlations

For each growth condition, we found 2000 different flux distributions using flux balance analysis. Then, for each pair of reactions, we calculated the Spearman correlation between their flux values. For the coupling calculations, we used the absolute values of these correlations (as coupling between reactions can be either positive or negative).

3.3.11 Multiple hypotheses correction

Throughout our paper P-values were filtered by False Discovery Rate (FDR) to correct for multiple testing [183]. More specifically, first, all the p-values were sorted in increasing order, $P_1, P_2, .., P_n$. Next, we filtered p-values $p_i : p_i > \frac{i}{n} * 0.05$.

3.4 Supplementary Information

3.4.1 Enriched metabolic pathways

	MEM>MEM-gln	MEM>MEM+oli	MEM-gln>MEM	MEM-gln>MEM+oli	MEM+oli>MEM	MEM+oli>MEM-gln
Bile Acid Biosynthesis				0.002840541		
Biotin Metabolism	0.001100353					
C5-Branched dibasic acid metabolism					1.20E-05	0.000443096
Carnitine shuttle		7.37E-16		9.84E-16		
Citric Acid Cycle						0.003729332
Fatty acid activation					0.000317195	0.000281306
Fatty acid elongation						0.000186977
Galactose metabolism					0.003361259	
Glyoxylate and Dicarboxylate Metabolism			0.000283311			
IMP Biosynthesis					4.34E-05	
Inositol Phosphate Metabolism					0.000614716	0.00036079
Methionine Metabolism						0.001984337
Pentose Phosphate Pathway					0.000594431	0.002746309
Phenylalanine metabolism			0.003275101		0.000907837	
Salvage Pathway	0.001512486					
Starch and Sucrose Metabolism						0.003413082
Tetrahydrobiopterin				1.82E-03		
Transport, Extracellular			3.43E-21	0.003226941		
Transport, Mitochondrial		0.000735737	0.000365041	2.06E-15		
Tyrosine metabolism			0.000657783		9.71E-05	
Urea cycle/amino group metabolism		0.000115953		1.25E-03		
Vitamin A Metabolism			0.000808112			
Vitamin B6 Metabolism		0.00317668	0.000398338	1.71E-05		

Table 1(a): comparison between the different conditions on the transcription level. Presented are the pathways that had significant changes (FDR p-value < 0.05)

3.4.2 Enriched metabolic pathways

	MEM>MEM-gln	MEM>MEM+oli	MEM-gln>MEM	MEM-gln>MEM+oli	MEM+oli>MEM	MEM+oli>MEM-gln
Alanine and Aspartate Metabolism					0.002359855	
Bile Acid Biosynthesis				0.001300005		
Carnitine shuttle		3.23E-08		1.87E-08		
Cysteine Metabolism					0.000170908	
Fatty acid activation	9.98E-05	0.000585872				
Fatty acid elongation	1.85E-07	0.000202217				
Fatty acid oxidation, peroxisome	0.000304895	0.001239269				
Folate Metabolism					0.000973866	
Glycine, Serine, and Threonine Metabolism		0.0011368		0.001653511		
IMP Biosynthesis					0.000147021	0.000164663
Starch and Sucrose Metabolism					0.002084977	
Transport, Extracellular					9.04E-07	2.21E-05
Transport, Mitochondrial					2.96E-05	
Triacylglycerol Synthesis					0.00114011	
Urea cycle/amino group metabolism						0.000885848
Valine, Leucine, and Isoleucine Metabolism		9.44E-05		0.000465834		
Vitamin B6 Metabolism						0.000300163

Table 1(b): comparison between the different conditions on the translation level. Presented are the pathways that had significant changes (FDR p-value < 0.05)

3.4.3 Spearman Correlation Comparison

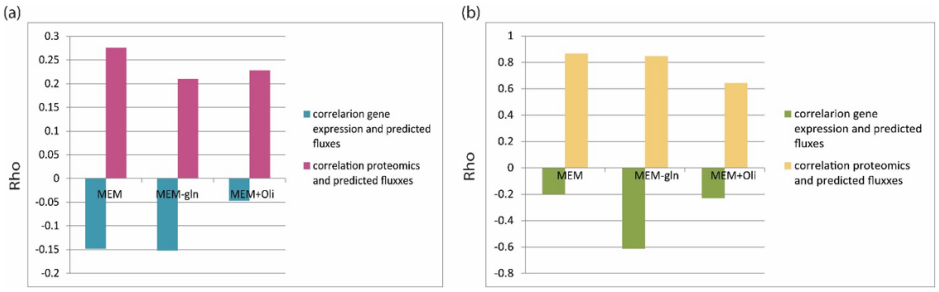


Figure 1: Comparison of the Spearman correlation between Gene expression and predicted fluxes, and between Proteomics data and predicted fluxes, for the subset of reactions that predicted to be (a) PTR and (b) TL.

3.4.4 Enriched metabolic pathways

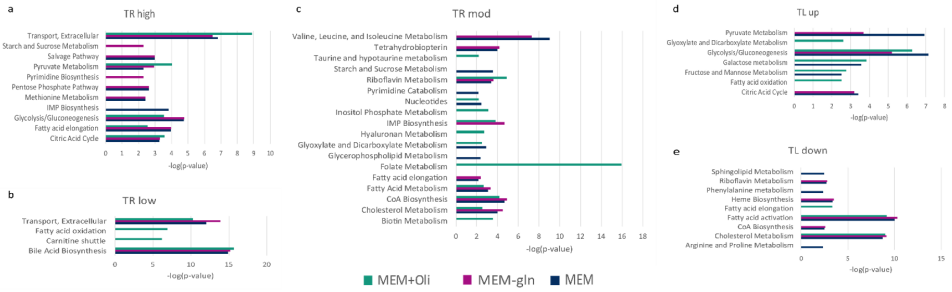


Figure 2: Pathways enriched with the reactions that are predicted as (a) TR high, (b) TR low, (c)TR moderate, (d) TL up and (e) TL down regulated (all p-values < 0.05 after FDR correction).

3.4.5 Spearman Correlation histogram

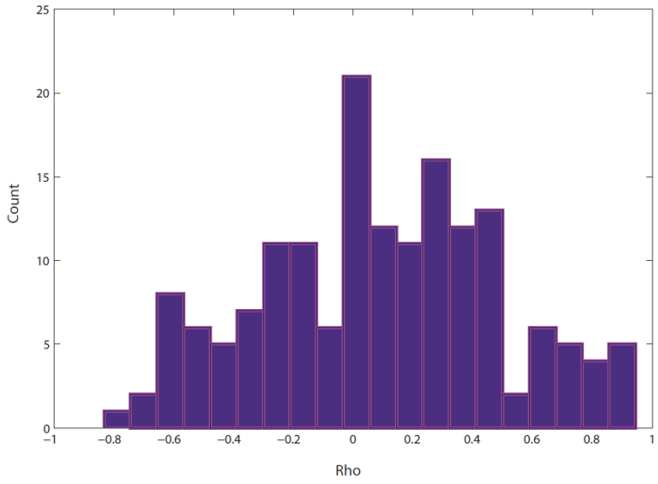


Figure 3: histogram of the Spearman correlation between phosphorylation and flux, for each reaction

3.4.6 Number of directly regulated reactions

	TR high	TR low	TR moderate	TL up-regulated	TL down-regulated	PTL up-regulated	PTL down-regulated
MEM	174	117	129	16	32	19	17
MEM-Gln	166	114	125	16	29	17	16
MEM+Oli	166	117	130	13	46	15	15

Table 4: Number of reactions that are directly regulated (based on measured data only), under each of the regulation categories, for each of the 3 growth conditions.

3.4.7 Enriched metabolic pathways

	TR-high	TR-low	TR-moderate	TL-up	TL-down
Aminosugar Metabolism			0.001420826		
Bile Acid Biosynthesis		3.22E-15			
Biotin Metabolism			0.008265235		
C5-Branched dibasic acid metabolism			0.000273157		
Cholesterol Metabolism	2.06E-10		0.006607643		
CoA Biosynthesis			3.48E-05		
Fatty Acid Metabolism			0.00503257		
Fatty acid activation			0.008265235		0.0006011
Fatty acid elongation	2.00E-08				
Fatty acid oxidation, peroxisome		0.000551			
Folate Metabolism			2.22E-16		
Galactose metabolism		0.003611			
Glycerophospholipid Metabolism			0.000154645		
Glycine, Serine, and Threonine Metabolism		7.15E-07			
Glycolysis/Gluconeogenesis	0.000411			0.000522	
Glyoxylate and Dicarboxylate Metabolism			0.001153372		
Heme Biosynthesis					9.40E-06
Histidine Metabolism		0.00027			
IMP Biosynthesis	0.000352				
Pyruvate Metabolism				3.55E-07	
ROS Detoxification	0.000119				
Riboflavin Metabolism			0.007724089		
Salvage Pathway	0.001622				
Steroid Metabolism		0.000282			
Tetrahydrobiopterin			0.00213589		
Transport, Extracellular	3.43E-08	2.79E-10		9.91E-05	
Transport, Mitochondrial			4.77E-09		
Tryptophan metabolism		0.000334			
Tyrosine metabolism		3.19E-06			

Table 5: (a) List of pathways enriched with reactions which are predicted to be regulated directly (both measured data and SVM predictions) by one of the regulation categories, and their corresponding p-value, under MEM growth condition (blank represents no enrichment).

3.4.8 Enriched metabolic pathways

	TR-high	TR-low	TR-moderate	TL-up	TL-down
Aminosugar Metabolism			0.002203192		
Bile Acid Biosynthesis		2.11E-15			
C5-Branched dibasic acid metabolism			0.000365947		
Carnitine shuttle			0.001944946		
Cholesterol Metabolism	2.24E-05				
CoA Biosynthesis			5.02E-05		
Fatty Acid Metabolism			0.006939117		
Fatty acid activation					0.0004393
Fatty acid elongation	1.07E-08				
Fatty acid oxidation, peroxisome		0.000515			
Folate Metabolism			1.11E-16		
Glycerophospholipid Metabolism			0.00032841		
Glycine, Serine, and Threonine Metabolism		5.69E-07			
Glycolysis/Gluconeogenesis				1.99E-05	
Glyoxylate and Dicarboxylate Metabolism			0.001709829		
Heme Biosynthesis					6.26E-06
Histidine Metabolism		0.000245			
IMP Biosynthesis	0.000261				
Pyrimidine Catabolism			0.005192029		
Pyruvate Metabolism	0.000529			1.12E-07	
ROS Detoxification	9.14E-05				
Salvage Pathway	0.00138				
Steroid Metabolism		0.000244			
Tetrahydrobiopterin			0.002658581		
Transport, Extracellular	8.79E-10	2.49E-10		0.000111	
Transport, Mitochondrial			5.81E-08		
Tryptophan metabolism		0.000293			
Tyrosine metabolism		2.51E-06			

Table 5 (b) List of pathways enriched with reactions which are predicted to be regulated directly (both measured data and SVM predictions) by one of the regulation categories, and their corresponding p-value, under MEM-Gln growth condition (blank represents no enrichment).

3.4.9 Enriched metabolic pathways

	TR-high	TR-low	moderate	TL-up	TL-down
Aminosugar Metabolism			0.000248865		
Bile Acid Biosynthesis		4.22E-15			
Carnitine shuttle			0.001313675		
Cholesterol Metabolism					5.50E-06
CoA Biosynthesis			4.07E-05		
Fatty Acid Metabolism			0.000698812		
Fatty acid activation	0.000619479				0.0008395
Fatty acid elongation	2.53E-07				
Fatty acid oxidation, peroxisome		0.000573239			
Folate Metabolism			6.66E-16		
Glycerophospholipid Metabolism			0.000214276		
Glycine, Serine, and Threonine Metabolism		8.17E-07			
Glycolysis/Gluconeogenesis				3.66E-05	
Glyoxylate and Dicarboxylate Metabolism			0.001366784		
Heme Biosynthesis					1.45E-05
Histidine Metabolism		0.000285485			
IMP Biosynthesis	0.00031143				
Pyruvate Metabolism	0.000684656			4.28E-06	
ROS Detoxification	0.00010683				
Riboflavin Metabolism			0.000309759		
Salvage Pathway	0.001518099				
Steroid Metabolism		0.000306961			0.001224317
Tetrahydrobiopterin			0.002346699		
Transport, Extracellular	2.20E-10	2.97E-11		5.04E-05	
Transport, Mitochondrial			1.42E-08		
Tryptophan metabolism		0.000361037			
Tyrosine metabolism		3.69E-06			0.003171941

Table 5 (c) List of pathways enriched with reactions which are predicted to be regulated directly (both measured data and SVM predictions) by one of the regulation categories, and their corresponding p-value, under MEM+Oli growth condition (blank represents no enrichment).

Chapter 4

Estimating tumor mutational burden from RNA-sequencing without matched-normal

4.1 Introduction

Somatic point mutations accumulate in the DNA of all dividing cells, both normal and neoplastic, and are the most common mechanism for altering gene function [184, 185, 186, 187]. Their detection in tumor samples is of high clinical value; first, when accumulated in specific genes termed "drivers", they may lead to the development of cancer. Identifying these mutations is therefore crucial for matching existing targeted therapies and for developing novel ones [188, 189, 190, 191]. In addition, somatic mutations are used for determining intra-tumor heterogeneity which is a major mechanism of therapeutic resistance [192], and for identifying mutational signatures, which have proven to be clinically useful biomarkers [193, 194]. Traditionally, detection of somatic point mutations is done using tumor and matched-normal whole exome or genome sequencing [195, 196, 197, 198, 199]. The latter is required for distinguishing between somatic mutations found exclusively in the tumor sample, and germline variants shared by all cells of an individual. Recently, several studies have developed a 'tumor-only' pipeline that detect somatic mutations without the matched-normal sample, at the cost of lower precision and recall

[200, 201, 202].

An additional extension to the traditional pipelines includes the detection of somatic mutations from RNA sequencing and a matched-normal DNA sample. In a recent publication, such a pipeline termed RNA-MuTect, was introduced, and showed that most of the mutations detected only in the RNA are filtered out by the pipeline, achieving and overall high precision. In addition, high sensitivity for mutations with sufficient detection power was observed, enabling the detection of most driver genes and mutational signatures [203].

In this study we took this approach one step further and developed RNA-MuTect-NMN, a pipeline for detecting somatic point mutations from RNA sequencing without a matched-normal sample. This is accomplished via a machine learning approach which utilizes a few dozens of features to classify single nucleotide variants as either somatic or germline. Our pipeline is trained and tested on the TCGA melanoma dataset achieving high precision and recall. In addition, it enables a reliable identification of both driver genes and mutational signatures. Finally, we applied our model to estimate the tumor mutational burden (TMB) which emerged as a proxy for neoantigen load. TMB is defined as the number of non-silent mutations found in a tumor DNA, and was found to be an independent marker of patient response to immune checkpoint inhibitor therapy (ICI), and for predicting patient survival, both in treated and treatment-naive patients [204, 205, 206, 207]. Here we show that estimating TMB from RNA which better reflects the set of expressed mutations, is either equivalent or superior to TMB estimated base on DNA. In addition, we show that this can be accomplished using a single RNA sample, further

emphasizing the potential clinical utility of our pipeline.

4.2 Results

4.2.1 Identifying somatic mutations from RNA-seq data without a matched normal sample

To develop a pipeline for detection of somatic point mutations from RNA-seq without a matched-normal sample, we leveraged RNA-seq and matched-normal DNA of 462 melanoma samples (SKCM) from the The Cancer Genome Atlas (TCGA) [208]. To obtain the ground truth of somatic and germline variants in these samples, we ran RNA-MuTect [203]; in short, RNA-MuTect works by first running MuTect [33] on RNA and matched-normal DNA, which classifies all variants into either germline or somatic. Since the set of somatic variants includes multiple noisy sites unique to the RNA, a series of filtering steps is then applied to yield the final set of true somatic mutations (Figure 4.1 panel A). As originally reported [203], focusing on the RNA mutations with sufficient detection power in the DNA, 90% were indeed found in the DNA, with a median of only 3 detected mutations per sample remained in the RNA set.

For each somatic and germline variant we collected a set of genomic features (Methods), such as the number of reference and alternate reads, variant classification type and MuTect likelihood score. In addition, we collected data on germline variants from dbSNP [209], gnomAD [210], 1000 genomes [68] and the Exome Sequencing project [211]. Finally, we utilized both DNA and RNA panel of normal

(PoN) which are based on ∼8000 TCGA and ∼6500 Genotype-Tissue Expression (GTEx) normal samples (Methods) [212]. These PoNs encode the distribution of alternate read counts across the entire sets of normal samples [213]. To test how well our features separate between somatic and germline variants, we performed a Wilcoxon rank sum test for each feature, and found that all features show a significant difference between these two types of variants (FDR corrected p-values $\leq$ 0.0111). However, when searching across a range of thresholds in each feature, we found that the Precision-Recall Area Under the Curve (PR-AUC) is very low ($\leq$ 0.08), as well as the F1 score ($\leq$ 0.16). This finding is a result of the substantial overlap between features' values in these two variant types, demonstrating the need for a more complex model. To this end we developed a machine learning framework named RNA-MuTect-NMN that gets as input a list of variants with their associated features, and classifies them as either somatic or germline. We first focused on an initial set of 100 samples. Each such sample contains the list of single nucleotide variants with their genomic features (Methods) and a somatic/germline label based on the RNA-MuTect pipeline, as described above (Figure 1A). We then trained random forest classifiers [214] in a 5-fold cross validation manner, such that in each iteration, 80 samples are used for training and 20 samples are used for validation. The median precision and recall achieved by our model on the validation sets are 0.82 and 0.83, respectively (Figure 4.1 panel B). To test our model performance, we used the remaining 362 samples and applied the following three step process: (1) we ran MuTect with tumor RNA-seq and without a matched-normal sample. In this step both somatic and germline variants are marked as true somatic mutations,

and a subset of sites are removed based on MuTect filtering scheme; (2) we then applied to this set of variants the 5 models built in the training step, and classified each variant as either somatic or germline, based on the majority vote; (3) finally, to remove any remaining RNA-specific noise, we applied the various RNA-MuTect filtering steps on the set of predicted somatic mutations. We have decided to run our pipeline as this three steps process due to a couple of time consuming steps implemented in the RNA-MuTect filtering pipeline, that would run more efficiently on the narrowed list of variants achieved after step 2. The final set of somatic and germline variants were then used to estimate the pipeline performance, showing a median precision and recall of 0.85 and 0.83, respectively (Figure 1B). Further investigating our results, we observed that a few samples achieved a relatively low precision. We found that this performance is due to their overall low number of somatic mutations in these samples (41 out of 46 samples with precision ≤ 0.6 had mutation count ≤ 50, Figure 4.1 panel C), and that the median precision on the remaining samples is 0.89. In addition, to circumvent the possibility that the high performance obtained by our model is a result of low purity levels which will in turn result with different allele fractions for somatic and germline variants, we examined the correlation between tumor purity and the obtained precision and recall levels. Indeed, we found this correlation to be insignificant (Spearman R = -0.0040, -0.0874, for precision and recall, respectively, P-value = N.S. for both). To better characterize our model we next examined which features are the most important using the feature importance score (methods). We found that the PoN DNA score 2 and PoN DNA score 1 features are the most important, followed by PoN RNA likelihood score

and gnomAD AF (Supplementary Table in subsection 4.4.1). Finally, we computed the Spearman correlation between the number of predicted somatic mutations, to that achieved by the DNA or RNA with a matched-normal DNA sample. In both cases, we found it to be highly significant ($R = 0.92$, P-value $\leq 4.15^{-151}$ for DNA and $R = 0.98$, P-value $\leq 8.7 * 10^{-286}$ for RNA, Figure 4.1 panels C-D, respectively).

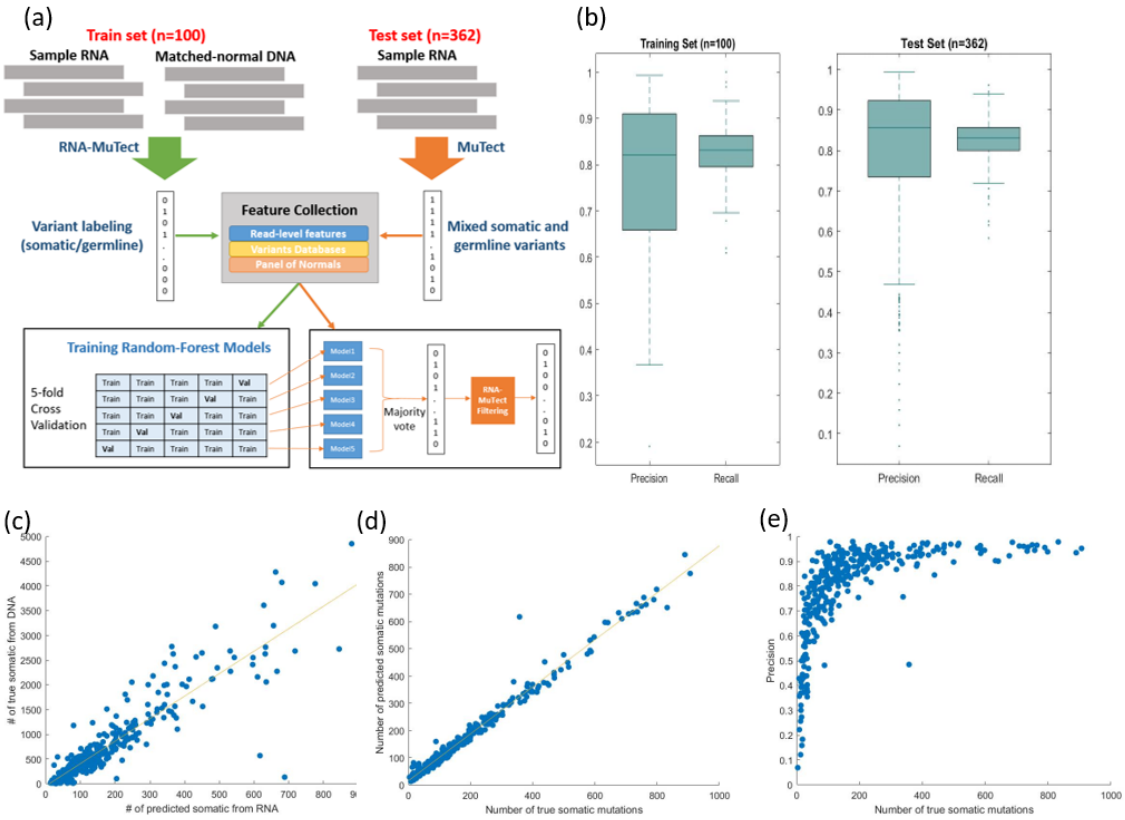


Figure 4.1: (a) An overview of the RNA-MuTect-NMN pipeline: In the training set (n=100, green arrows), RNA-MuTect is applied on tumor RNA and matched-normal DNA to obtain a list of variants labeled as somatic or germline. A random forest classifier is then trained with the collected set of features for each variant in a 5-fold cross validation manner. In the test set (n=362, orange arrows), MuTect is applied with tumor RNA and without a matched-normal sample, to yield a list of mixed somatic and germline variants. The five trained models are then applied to this set of variants in a majority vote manner. Finally, the predicted set of variants is further filtered by the various RNA-MuTect steps. (b) Precision and recall on training and validation sets. Box plots show median, 25^{th}, and 75^{th} percentiles. The whiskers extend to the most extreme data points not considered outliers, and the outliers are represented as dots. (c) Correlation between the number of predicted somatic mutations and true somatic mutations as determined by DNA. (d) Correlation between the number of predicted somatic mutations and true somatic mutations as determined by RNA. (e) Precision as the function of the number of true somatic mutations per sample.

4.2.2 Detecting mutational signatures and significantly mutated genes without a matched-normal sample

The overall high performance of RNA-MuTect-NMN enabled us to apply our standard analysis pipelines for identifying mutational signatures and significantly mutated genes. To this end we applied SignatureAnalyzer [215, 216] using the set of predicted somatic mutations, and identified 4 signatures (Figure 4.2 panel A): UV signature (COSMIC SBS7b, cosine similarity = 0.95) which is common in melanoma [185, 217], signature 5 (COSMIC SBS5, cosine similarity = 0.87) which is common in various cancer types, including melanoma, and a signature enriched with C>A mutations and was previously found only in ultraviolet light associated melanomas (SBS38, cosine similarity = 0.78). Importantly, the same three signatures were identified in the DNA (Supplementary Table in subsection 4.4.4). In addition, a signature enriched with T>G mutations was detected. This signature was not detected in the DNA but was detected in the RNA when somatic mutations were identified with a matched-normal DNA sample (Supplementary Table in subsection 4.4.3). Investigating this set we found that out of 552 mutations that are associated with this signature, 489 are mutations that were not found in the DNA mutations, suggesting a mechanism that is unique to RNA mutations.

Next, we identified significantly mutated genes by applying MutSig2CV [218] to the set of predicted somatic mutations. Out of 24 identified genes, 22 were found to be significantly mutated also when the matched-normal sample is taken into account (Figure 4.2 panel b), and only 2 were missed by our pipeline. Finally,

we examined our pipeline performance in identifying a set of 55 melanoma somatic driver genes based on the COSMIC database [219], 43 of them were found to carry at least one true somatic mutation in our dataset. Notably, we found that our pipeline achieves an even higher precision and recall on this set (median of 1 and 0.95, respectively), further demonstrating its high value.

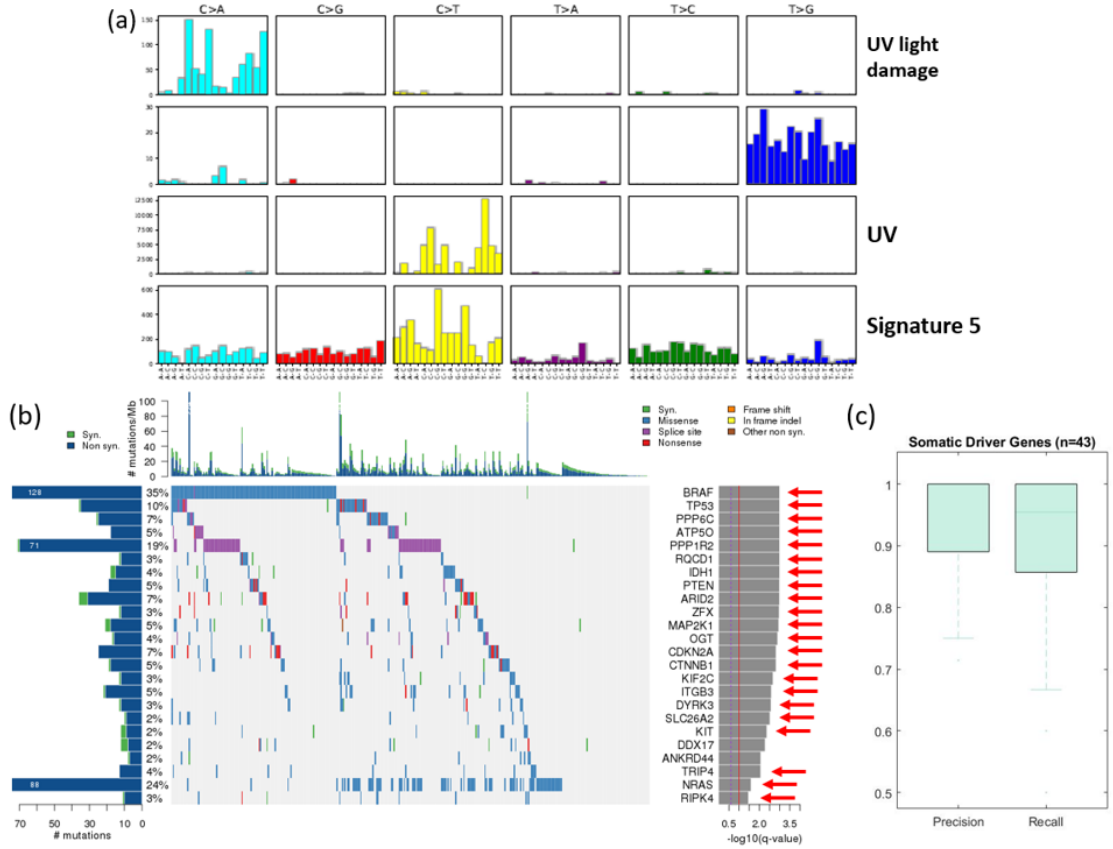


Figure 4.2: (a) Mutational signatures identified based on the set of predicted somatic mutations; (b) Co-mutation plot based on predicted somatic mutations in our test set. Overall frequencies, allele fractions, and significance levels of candidate cancer genes ($Q \leq 0.05$) identified by MutSig2C [218] are shown. Genes marked with a red arrow were also identified as significantly mutated based on the set of true somatic mutations. (c) Precision and recall on the set of know melanoma drivers. Box plots show median, $25^{th}, and 75^{th}$ percentiles. The whiskers extend to the most extreme data points not considered outliers, and the outliers are represented as dots.

4.2.3 TMB predicted by RNA-MuTect-NMN is associated with patient survival

The development of checkpoint blockade (CPB) therapy such as anti-PD1 and anti-CTLA4 has revolutionized cancer therapy and resulted in long-lasting tumor responses in patients with a variety of cancers [103]. As a result, these drugs have been FDA-approved for many cancer types, including melanoma, non-small cell lung cancer, Urothelial carcinoma, Head and Neck squamous cell carcinoma and more [104]. Recently, an accelerated approval for anti-PD1 for the treatment of adult and pediatric with tumor mutational burden-high (TMB-H, $\geq$ 10 mut/Mb) has been granted, making it a critical metric for the clinical decision process. Indeed, the TMB which is traditionally estimated via DNA sequencing has been found to be associated with patient survival, though to different extents depending on cancer type [220], prior and current treatment [221, 222, 223]. Here, based on the set of predicted somatic mutations from RNA sequencing alone, we estimated the tumor mutational burden as the number of non-silent somatic mutation in each sample. We then divided the patients into two groups with high- and low-TMB levels, using the median TMB as the cutoff value. We found that patients with high-TMB had a mild but significant increase in survival time as compared to those with low-TMB (log-rank P-value = 0.02, figure 4.3 panel a). Of note, performing the same analysis using the set of somatic mutations detected based on tumor and matched-normal DNA, similar results are obtained (logrank P-value = 0.01, Figure 4.3 panel b), further demonstrating the utility of our pipeline. In addition, we performed a

multivariate Cox proportional hazards regression analysis with patient age, tumor stage and TMB as the covariates, and found that TMB is the prognostic factor most associated with increased survival (HR = 0.59, 95% CI=0.36-0.96, P-value $\leq$ 0.03, Figure 4.3 panel e) . As discussed above, the extent of association between TMB and patient survival vary widely between different datasets according to cancer type and prior therapy. A recent publication by Valero et al. showed that among patients that were not treated with CPB, a very high TMB (top 10-20 percentile) is associated with poor survival [224]. Given that most of the patients in the TCGA cohort were not treated with CPB, we set to examine this observation in our data as well. Indeed, when we divide the patients into three groups with very high-, high- and low- TMB levels, using the top 10^{th} percentile for the very high group, and median for the remaining samples, we find that those with the highest TMB values have a poor survival (logrank P-value = 0.04 between high- and very high-TMB), and those with median high TMB have an improved survival as compared to those with low TMB (logrank p-value = 5.8*10-4, Figure 4.3 panel c). This result is robust to the selection of threshold for the very-high TMB group (top percentile between 10-18). Importantly, performing the same analysis based on DNA revealed the same trends, though with an inferior significance level (logrank P-value = 0.01, 0.04, respectively, Figure 4.3 panel d). Repeating the cox analysis while removing the top 10^{th} percentile, the association of TMB with survival is becoming even more significant (HR = 0.31, % CI=0.17-0.58, P-value $\leq$ 2*10-4, Figure 4.3 panel e). Overall, these results demonstrate that estimating TMB based on RNA alone is feasible and of a high predictive power.

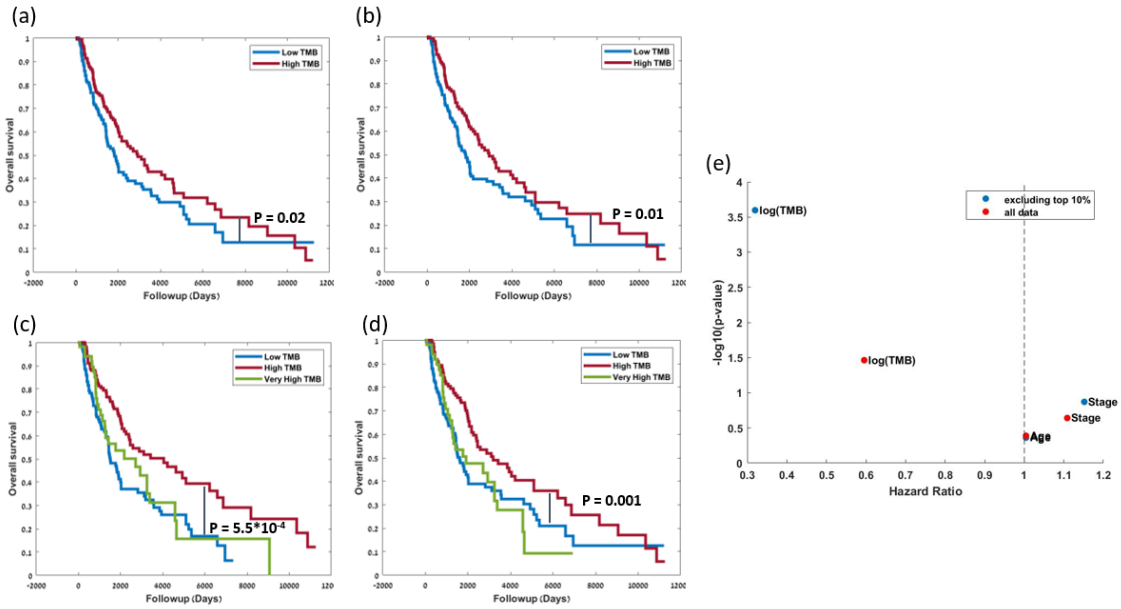


Figure 4.3: Kaplan–Meier survival curves for patient groups (a) patients with high vs. low TMB (estimated from the predicted RNA somatic mutations). The median is used to define the 'low TMB' and 'high TMB' subgroups. The P value is computed via a two-sided log-rank test. (b) same, for TMB estimated from DNA somatic mutations. (c) patients with very-high vs. high vs. low TMB (estimated from the predicted RNA somatic mutations). Subgroups were splitted by using the top 10^{th} percentile for the very high group, and median for the remaining samples. (d) same, for TMB estimated from DNA somatic mutations. (e) Hazard Ratio vs. –log10(p-value), obtained from multivariate Cox proportional hazards regression analysis. Red dots for results based on all samples, blue dots for results after excluding the top 10% of samples (very high TMB).

4.2.4 TMB estimation from RNA in patients treated with CPB

We next examined the prediction power of our model on an additional set of patients that were treated with nivolumab (anti-PD1), some are treatment naive and some have previously progressed on ipilimumab (anti-CTLA4) [225]. Raw RNA-sequencing data from 50 pre-therapy biopsies were obtained and aligned to the reference genome. Then, the set of mutations ¬obtained by MuTect using tumor RNA alone was further labeled by our model as either somatic or germline. To validate our calls we first applied SignatureAnalyzer and identified the set of mutational signatures that are active in these samples. Encouragingly, we found the UV signature (SBS7b), along with signatures SBS11 andSBS5 that were found by the authors based on DNA were also detected based on our predicted set of somatic mutations (cosine similarity = 0.86, 0.95 and 0.78, respectively, Figure 4.4 panel a). In addition, when applying MutSig2CV to identify significantly mutate genes, both NRAS and BRAF, known melanoma drivers, were found to be significantly mutated (Figure 4.4 panel b). Finally, we estimated the TMB based on the set of predicted somatic mutation. Interestingly, when considering the set of treatment naïve patients for which both DNA and RNA are available, no significant association between TMB and patient survival is found, based on neither DNA nor RNA. However, when considering the set of patients that were previously progressed on ipilimumab, a significant association between high TMB and poor survival is found (logrank P-value = 0.01, Figure 4.4 panel c). This is in similar to the trend that was reported by the authors using DNA (Figure 4.4 panel d). Overall, we again find

that estimating the TMB from tumor RNA alone is feasible and results with similar trends to those obtained with tumor and matched-normal DNA.

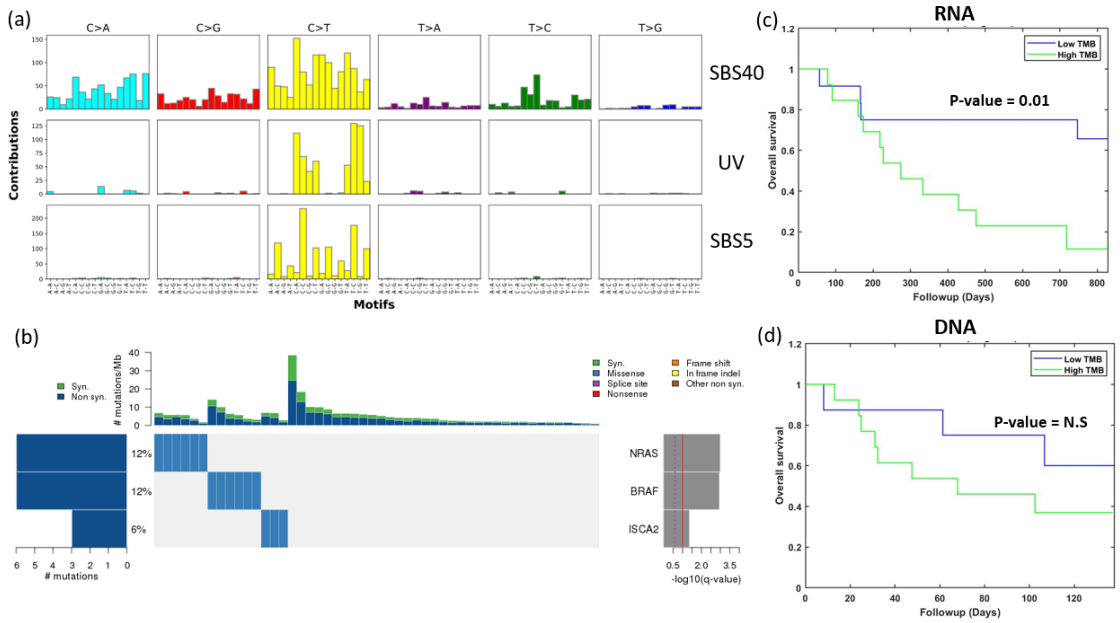


Figure 4.4: (a) Mutational signatures identified based on the set of predicted somatic mutations (using the RNA-seq data of 50 pre-therapy biopsies from [225]. (b) Co-mutation plot based on predicted somatic mutations. Overall frequencies, allele fractions, and significance levels of candidate cancer genes ($Q < 0.05$) identified by MutSig2C [38] are shown. (c) Kaplan–Meier survival curves for patient groups (a) patients with high vs. low TMB (estimated from the predicted RNA somatic mutations). The median is used to define the 'low TMB' and 'high TMB' subgroups. (d) same, for TMB estimated from DNA somatic mutations.

Acknowledgments

The results shown here are part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

4.3 Methods

4.3.1 Datasets

RNA-seq data of tumor from 462 Melanoma patients from the TCGA project [208] was utilized in this study. To infer the true labeling from RNA-mutect, RNA-seq from the adjacent matched normal tissue was used as well. For the survival analysis the clinical data of these patients was obtained. To compare between somatic mutations from RNA to somatic mutations from DNA, MuTect [33] was applied to the DNA data. We tested the model on an additional dataset, where RNA-seq data and clinical data for a cohort of 50 patients with advanced melanoma was obtained [225].

4.3.2 Somatic Mutation Calling

- RNA-somatic mutations were identified by applying RNA-MuTect [100] to the list of variants obtained from tumor RNA-seq and the variants obtained from DNA of the matched normal. This was also used as our labels for the model.

- DNA-somatic mutations were identified from tumor-normal paired alignments using MuTect [33], which identifies variants unique to the tumor sample by contrasting alignment pileups at each genomic position.

4.3.3 Feature Collection

The following features were collected and calculated for the random-forest classification model:

1) T_ref_count - count of reference alleles reads in tumor

2) T_alt_count - count of alternate alleles reads in tumor

3) T_lod_fstar - a statistic score assigned by MuTect; Log of (likelihood tumor event is real / likelihood event is sequencing error)

4) Tumor_f - allelic fraction of this candidate based on read counts

5-12) For each of the germline variants database (dbSNP, gnomAD, 1000Genome, ESP) two vectors were created:

1. The variant is present (1) or not (0) in each database (resulting with a 1X4 binary vector for each variant)

2. Allele Frequency (AF), when available, and when not available, the mean AF value was used (resulting with a 1X4 AF vector for each variant)

13) Variant_classification - (1) if the variant classification (assigned by Oncotator [226]) is one of the follows: 'IGR', 'Intron', 'RNA', 'lincRNA' , and (0) otherwise.

(14-31) Panel of Normals DNA (based on TCGA normal samples) – Each genomic position's histogram comprises eight bins used as features 14-21

14) total counts < 8 (insufficient coverage)

15) total counts ≥ 8 (and no alt count above the subsequent thresholds)

16) alt count ≥ 1 and alt fraction $\geq 0.1\%$

17) alt count ≥ 2 and alt fraction $\geq 0.3\%$

18) alt count ≥ 3 and alt fraction $\geq 1\%$

19) alt count ≥ 3 and alt fraction $\geq 3\%$

20) alt count ≥ 3 and alt fraction $\geq 20\%$

21) alt count ≥ 10 and alt fraction $\geq 20\%$

22) log-likelihood score (based on [213])

23-31) same for RNA (based on GTEx samples)

4.3.4 Panel of Normals (PoN)

The different PoN scores used as features in our model, are based on the method described in [213], where the idea is encoding the expected distribution of alternate allele read counts at every genomic position, based on a large panel normals ($\sim$8000 TCGA normal samples, in case of DNA, and $\sim$6500 GTEx samples in case of RNA).

4.3.5 Feature Importance

To calculate the feature importance, we used the built-in feature importance of scikit-learn, also known as GINI importance (or- mean decreased impurity). We obtained the feature importance scores for each of the 5 trained models, and final importance score for each feature was calculated as the average across all 5 models.

4.3.6 Significantly Mutated Genes

MutSig2CV [218] uses three tests to infer significantly mutated genes: abundance, which classifies whether a gene's observed mutation rate is significantly elevated

relative to its expected background mutation rate; clustering, which looks for genes harboring recurrently mutated loci; and conservation, which looks for genes whose mutations are significantly enriched in evolutionary conserved sites. Each of these tests returns a p-value for every gene, which are Fisher-combined and false discovery rate (FDR)-corrected via Benjamini-Hochberg. Genes were considered "significant" if their FDR value was below 0.05.

4.3.7 Statistical analysis

1. Multiple hypotheses correction. Throughout our paper P-values were filtered by False Discovery Rate (FDR) to correct for multiple testing [183].

2. Survival analysis. Survival analysis was performed using the Kaplan–Meier method [143] to generate survival curves, where the median values of TMB was used to split the patients into two groups, resulting with two survival curves. In the additional survival analysis, we used additional threshold of 10% for the group of very-high TMB samples, and median TMB for the rest of the samples. The log-rank test p-value was calculated to estimate the survival difference between the groups. Multivariable analysis was performed using Cox proportional hazards regression [227], where variables significant with univariate regression were included, namely, TMB, age and stage.

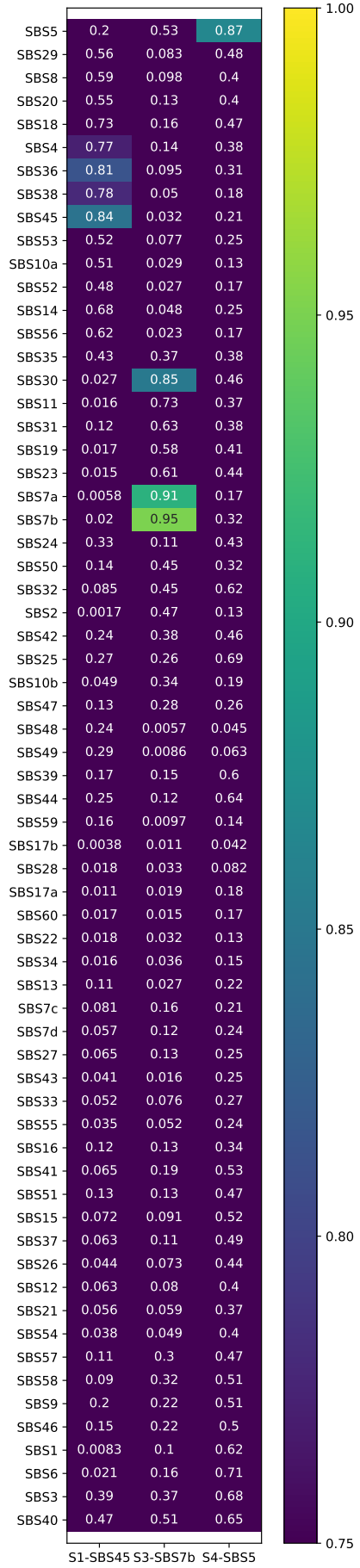
3. Distributions comparisons between germline and somatic groups was performed using the Wilcoxon rank-sum test [228].

4.4 Supplementary Information

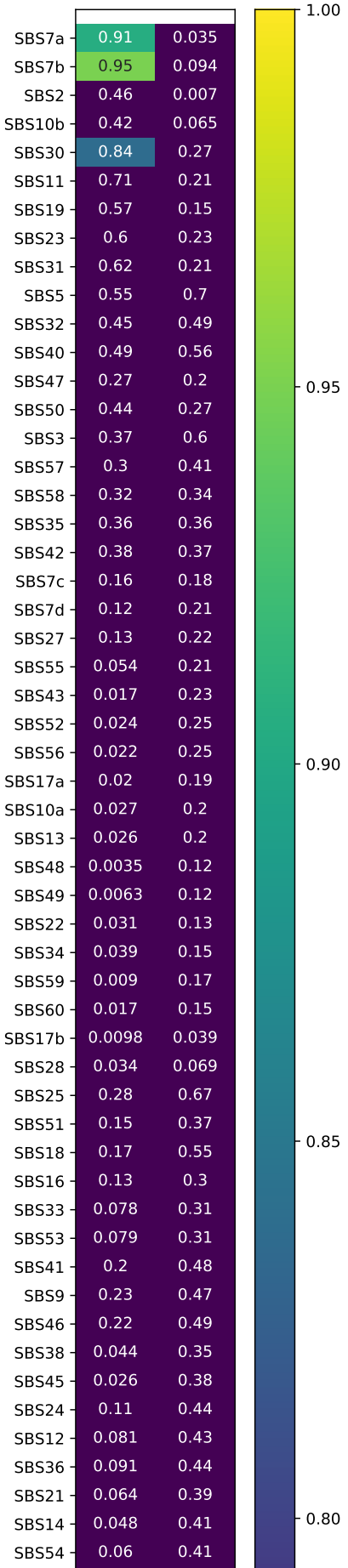
4.4.1 Feature Importance

Feature name	Feature importance score
pon_DNA_2	0.1678
pon_DNA_1	0.1077
log_like_RNA	0.0759
gnomad_AF	0.0705
log_like_DNA	0.0592
gnomad	0.0540
pon_DNA_8	0.0533
pon_DNA_3	0.0329
pon_DNA_5	0.0316
pon_DNA_4	0.0293
pon_RNA_7	0.0262
pon_RNA_8	0.0254
tumor_f	0.0241
t_alt_count	0.0235
t_lod_fstar	0.0233
pon_RNA_4	0.0224
pon_RNA_2	0.0191
esp_AF	0.0171
esp	0.0166
pon_DNA_7	0.0155
t_ref_count	0.0152
pon_RNA_3	0.0143
pon_DNA_6	0.0143
genome1000	0.0134
pon_RNA_1	0.0130
pon_RNA_6	0.0084
dbsnp	0.0081
dbsnp_AF	0.0072
pon_RNA_5	0.0064
classification_to_remove	0.0030
genome1000_AF	0.0016

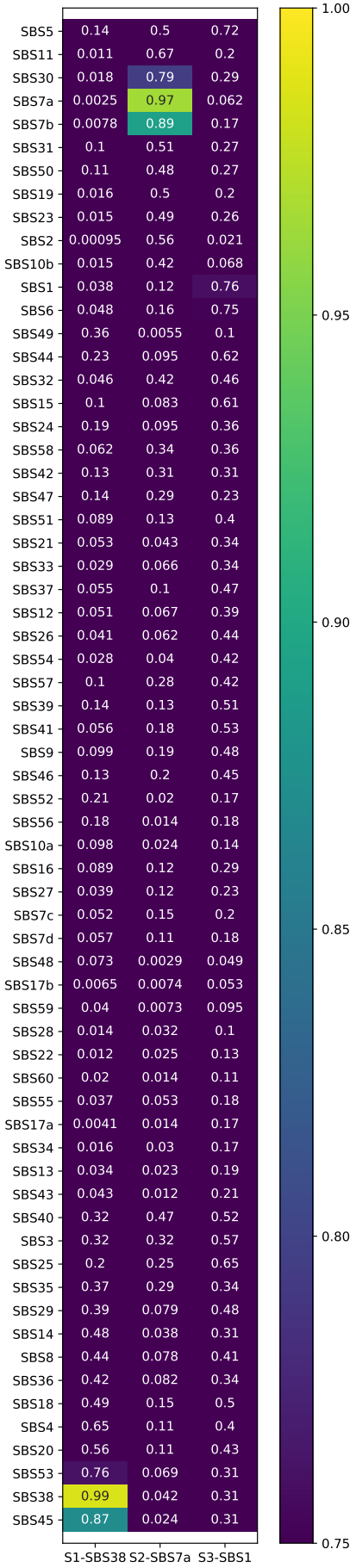
4.4.2 Mutational Signature (cosine similarity) - RNA predicted



4.4.3 Mutational Signature (cosine similarity) - RNA true



4.4.4 Mutational Signature (cosine similarity) - DNA



Chapter 5

Discussion

5.1 Summary and contributions

In this thesis I addressed emerging challenges in the field of GSMMs, focusing on the modeling of human metabolism, and also in the field of somatic mutation calling in cancer. This thesis begins with the presentation of a novel model-based method for the identification of synthetic dosage lethality interactions, that enables to selectively kill cancer cells. Next, in order to better study the metabolic regulation in breast cancer cell-line, I present an approach for multi-omics data integration and flux prediction, by establishing a conceptual framework that utilizes GSMM and machine learning techniques. Finally, to study somatic mutations in cancer, I developed the first model that enables the identification of somatic mutations in the tumor from RNA-seq data, without the matched normal.

Below I summarize these three studies, their limitations, and their future applications.

1. The IDLE (identifying dosage lethality effects) algorithm presents a novel approach to capture enzymatic SDL effects in metabolic networks. Previous GSMM-based methods to study interaction between enzymes were focused on synthetic lethal interactions, however, in order to target oncogenes, which are genes that over-expressed in the cancer cell, and cannot be targeted directly

because they are essential in the normal cell, an SDL interaction approach should be used.

To this end we came up with an algorithmic method, accomplished via a constraint-based modeling, that enables the prediction of cell growth under diverse genetic perturbations and constraints. Our approach enables the identification of SDL interactions with a range of lethal strength to the cell, and we show the correlation between this strength to different tumor clinical attributes, such as tumor size and patient survival, where the most lethal pairs have the largest effect on killing cancer cells. In addition, we show that the more SDLs active in a tumor sample, the better this is for a patients' prognosis. We also demonstrate that cancer cells selects against SDL interactions, such that SDLs are less frequently active than expected.

The IDLE approach presented here is general, however, it can be extended in the future to identify SDL interactions that are cancer type specific, by integrating omics data of the patients' tumor, such as gene expression or proteomics. In addition, the IDLE approach is not limited to cancer, and could be used to identify SDL networks in pathogenic bacteria or fungi, providing new antibiotic therapeutic leads. Other possible applications include metabolic engineering to increase the yield of valuable metabolic byproducts. Specifically, this may be achieved by engineering an SDL effect to inhibit the production of undesired byproducts, or inversely, neutralizing the SDL effect to force an increased flux through desired pathways.

Taken together, IDLE is expected to contribute to various research fields ranging from medical sciences to biotechnology.

2. In this part of the thesis, I describe a study that provides the first chart of metabolic regulation in MCF7 breast cancer cells on a genome scale, by integrating multi-omics data with the genome-scale metabolic model, along with machine learning techniques. In this study we classified the metabolic reactions as regulated at three distinct levels (transcriptionally-, translationally-, and post-translationally- regulated), and also characterized them as being either directly or indirectly regulated. We found that the flux of the indirectly regulated reactions is coupled to the flux of directly regulated ones, suggesting that the regulation of breast cancer cell metabolism is controlled in a hierarchical manner.

   The major limitation of this study is the data itself, being limited, noisy and having different levels of coverage for the different omics types. In addition, we focused on studying post-translational modifications mediated by phosphorylation, while post-translational modifications occur via a variety of additional mechanisms, including acetylation, glycosylation and allosteric regulation. Aiming to make the best use of the available data and to obtain a genome-wide view of breast cancer cell metabolism, we (1) employ coarse discretization to overcome some of the noise in the data, and only identify regulatory alterations in reactions that are differentially active across the conditions of study, and (2) build SVM predictors that use the known network

properties together with measurements with high coverage and help us to deal with the limited coverage and to extend the scope of the study to the utmost.

Future work should aim to address these limitations by measuring a wider range of omics data with higher coverage; with the rapid advancement of high-throughput technology and the accumulation of more comprehensive omics data across additional cellular conditions, such data may become readily available soon and may be benefit from the conceptual and computational framework laid out in the current study.

3. In this study we introduce RNA-MuTect-NMN, the first computational method that identifies somatic mutations from RNA-seq data without a matched-normal sample. The pipeline is based on the RNA-MuTect method [203] which is designed to detect somatic mutations from tumor RNA-seq and matched-normal DNA. To extend it to a 'tumor-only' mode we built a random forest classification model that distinguishes between somatic and germline variants using various features, including mutation specific ones, and those derived from large panels and databases of normal individuals. Our model was trained on a subset of the TCGA melanoma dataset, and achieved high precision and recall (85% and 83%, respectively), when applied to an independent test set of additional >350 melanoma samples. Additionally, we show that estimating the tumor mutational burden from RNA rather than from DNA is feasible, and that the exact same trends as those estimated using tumor DNA with a matched-normal sample are observed. As previously shown, we find that

in melanoma patients that were not treated with checkpoint blockade (CPB), very high TMB is associated with poor survival [224], while median high is associated with improved survival as compared to patients with low TMB. In addition, in treated patients that were previously progressed on anti-CTLA4, we find that high TMB is significantly associated with poor survival compared to low TMB. These results are in concordance with the original findings [225]. Importantly, the model built in this study is based on melanoma samples, however, the RNA-MuTect-NMN approach is generic and can be easily applied for any cancer type, given a sufficient number of samples with RNA-seq of the tumor, along with tumor and matched normal DNA for validation. Moreover, melanoma is a highly mutated cancer with a sufficient number of somatic mutations that can be used for model training, and where the fraction of germline contamination predicted by our model is negligible. Hence, the performance of our approach should be further tested on lowly mutated cancers where significantly less somatic mutations are available for training, and where the fraction of germline contamination can become substantial. These limitations can be potentially addressed by down-sampling of the germline group, or by combining multiple datasets together.

We believe that the motivation for using RNA-MuTect-NMN is three-fold:

(a) For future studies, it diminishes the need for collecting and sequencing matched-normal samples, thus significantly reducing sequencing cost, especially for large cohort analysis.

(b) It enables the analysis of RNA-seq data in retrospective studies where RNA was originally sequenced for expression-based analyses.

(c) It enables a combined analysis where both genetic and phenotypic data can be inferred from the exact same sample. This is especially crucial in cancer where different regions of a tumor from which DNA and RNA are extracted may be significantly different due to tumor heterogeneity. These applications can significantly increase the number of samples analyzed and thus aid biomarker and drug target discovery.

5.2 Future challenges in the modeling human metabolism

While there has been a remarkable progress in the last years in the genome-scale modeling of human metabolism, additional challenges lies ahead, including the utilization of richer datasets from both cell-lines and clinical samples, the modeling of cell environment including its interactions with surrounding cells, and studying the potential of emergent drug resistance to metabolic drugs. Generally, GSMMs lend themselves naturally to the early stages of drug development, most notably the determination of new targets for target-based screens. However, as direct drivers of innovation and as scaffolds for interpretation of complex large-scale datasets, the potential of GSMMs in drug development is yet largely untapped.

5.2.1 Integrating additional omics data sources

The GSMM framework is a platform for omics data integration that can be of significant value. Nonetheless, transcriptomics and proteomics have been the main data source for deciphering metabolic phenotypes while other data sources have been rarely used. New technologies for next-generation sequencing (NGS) has enabled a systematic cataloguing of human genomes through national and international genomics projects. This is most prominent in cancer through resources such as The Cancer Genome Atlas and the International Cancer Genomics Consortium. These databases are examples for comprehensive resources where mutational signatures and potentially new therapeutic targets across cancer types have been identified [79, 144]. By focusing on the subset of mutated metabolic enzymes and evaluating their effect on protein function, one can potentially use these datasets to model multiple human cells and identify their unique metabolic vulnerabilities. A first step in this direction has been taken by Nam et al. [129]. In this study the authors integrated genetic mutation data from more than 1,700 cancer genomes along with their gene expression levels. Predicted flux changes between normal and cancer cells were then evaluated by simulating loss-of-function mutations in metabolic enzymes, leading to the identification of 15 onco-metabolites, including the well-studied succinate and fumarate.

Apart from genomics, metabolomics is an additional accumulating data resource for studying human metabolic disorders. For instance, metabolomic profiles of cancer cells have been widely used for the past several years to distinguish between

different cell lines and tumor types both in vitro and in vivo [229, 230]. Furthermore, cancer-associated mutations in certain metabolic genes were found to induce an abnormal accumulation of onco-metabolites [231]. The ability to both integrate and predict metabolite concentrations at the genome-scale level is therefore of major importance in studying human metabolism. While the information on extracellular metabolites has been used to constrain a given GSMM [133, 232], the prediction and/or integration of intracellular metabolite levels require the usage of thermodynamic information and the knowledge of the kinetic parameters of the network [233, 234], which are largely unknown. The utilization of metabolomic data for analyzing GSMMs therefore calls for new, more sophisticated methodologies, designed to address these emerging challenges.

5.2.2 Modeling cancer cells environment and interactions

While many studies have focused on growing cancer cells in vitro and out of their tumorigenic context, it is now widely accepted that the tumor microenvironment plays an important role in defining and reprogramming cancer cell metabolism [235]. The computational study of cell and tissue interactions via GSMMs has already been demonstrated in both microorganisms and human tissues [236, 237, 238], but has not been explored in the context of cancer cells and supporting cells in their environment. Modeling the dynamic exchange of material between these different cells can bring us closer to a more accurate modeling of tumors in vivo and reveal metabolically related phenotypes that could not have been discovered by the

modeling of each cancer cell alone.

5.2.3 Studying the emergence of resistance to metabolic drug targets

GSMMs can be utilized in the context of resistance in bacteria and cancer, to identify promiscuous functions of existing metabolic enzymes, thus revealing alternative pathways capable of bypassing the targeted reaction(s). Furthermore, this approach can be used to identify gain-of function enzyme mutations and increase our understanding of enzymes' catalytic side activities. Promiscuous functions of metabolic enzymes have already been studied by GSMM of Escherichia coli, both revealing fundamental features of these enzymes [129], as well as identifying novel metabolic pathways that produce precursors for cell growth under diverse environmental conditions [130].

In summary, GSMM is a stepping stone for whole-cell modelling, and this vision, that was already firstly realized by [1] in bacteria, should inspire us to aim at modeling the entire cellular dynamics of different human cells.

5.3 Future challenges in somatic mutation calling

Variant calling algorithms have been evolving and improving in the past years. The underlying models are getting more and more complex in order to describe the physical process of NGS experiments and to model different types of artifacts. However, it is still very challenging to precisely detect somatic Single nucleotide variants

(SNVs) due to low variant allele frequencies (VAFs), sequencing artifacts and lower than desired coverage. Low VAFs in tumor samples are caused by several reasons including tumor-normal cross contaminations, tumor ploidy, sub-clonality (also called intra-tumor heterogeneity), and local copy-number variation in the cancer genome. In addition, the performance of a particular caller varies dataset by dataset. Previous studies also showed that the output of different somatic callers for a given dataset is highly divergent and the calling results show a very low level of concordance across callers [239, 240, 241, 242, 243, 244, 245]. Due to discrepancies among callers, finding a single best caller for various datasets is considered impractical [244].

Below are some of the open challenges in the field of mutations calling and somatic mutations.

5.3.1 Ensemble of callers

Ensemble approaches have been used to combine prediction results generated by multiple somatic variant callers. The idea is based on the "wisdom of crowds"; since the patterns of statistical models used in different classifiers do not necessarily overlap, the complementary information about these patterns could potentially be harnessed to improve overall performance. To get a good ensemble, it is generally believed that the base learners should be as accurate as possible, and as diverse as possible. Thus, there are two major questions raised regarding how to construct a feasible and effective ensemble approach. First, how to select a reasonable number

of component callers with higher accuracy while maintaining the diversity of the callers [246, 247]. Second, how to combine the results from individual callers to determine whether a variant should be called or not. Most ensemble approaches for somatic calling belong to two categories: The simple approach is combining the predictions from multiple callers by simple fixed rules, such as majority voting [248] or consensus approaches [249, 250]. The more complex approaches employ machine learning (ML) methods, which treat prediction results or metrics of individual callers as input features. These inputs are then combined with other genomic features and used to train a classifier that is then applied on an unknown new dataset to predict variants. These ML-based methods include stacking [251], Bayesian approach [252], decision trees [244, 253] and deep learning [254, 255]. Consensus approaches with a fixed rule are easily implemented and can save tremendous training and prediction time. In contrast to consensus approaches, ML-based ensemble approaches can leverage the information from the training sets with known truth, which may provide potentially better performance than fixed combination rules. However, a downside of the ML-based ensemble approaches is higher computational complexity that may be very sensitive to the training dataset.

Two significant concerns for current existing ensemble approaches still exist:

1) Due to insufficient real "ground-truth" somatic variants and evolving software, the caller selection from previous studies may be out of date and not ideal for current studies.

2) The replicability and reproducibility of ML-based ensemble methods have not been thoroughly examined.

5.3.2 Benchmarking studies

Although most variant callers were published with benchmarking results against other mainstream pipelines of their time, the claimed performance may not be replicated on independent datasets. A number of independent studies to benchmark and compare various somatic variant callers have been published [256, 257, 258, 240, 241, 239, 259], but inconsistent performance data and contradicting rankings of the variant callers were reported. The inconsistency of benchmarking results is due to two reasons. First, most variant callers need to be fine-tuned to achieve the expected accuracy on a naive dataset, yet the optimal parameter values are unknown to the tester. In this case, applying the default values seems a reasonable solution and indeed a common practice in benchmarking studies. For example, Cai et al. [259] applied default settings in comparing four tumor-normal callers. Sandmann et al. [257] used default settings except for VAF threshold. Kroigard et al. [239] applied default settings for when benchmarking on exome-sequencing data and adjusted parameters for targeted sequencing data. Second, some variant callers were original designed for certain types of applications and then published without extensive validation on a wide range of datasets, so their performance may drop in some occasions.

To conclude, variant calling algorithms have been evolving and improving in the past years. The underlying models are getting more and more complex in order to describe the physical process of NGS experiments and to model different types of artifacts. Somatic mutations identification holds great potential in cancer

treatment; systematic sequencing studies of larger numbers of tumors from a wide variety of cancer types will yield further insights into the development of human cancer, providing new opportunities for molecular diagnosis and therapeutics.

Bibliography

[1] Jonathan R. Karr. A whole-cell computational model predicts phenotype from genotype. *Cell*, 150(2):389–401.

[2] L. Kuepfer. Towards whole-body systems physiology. *Mol Syst Biol*, page 6.

[3] N.D. Price, J.L. Reed, and B.O. Palsson. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nature Reviews Microbiolgy*, 2(11):886–97.

[4] A.M. Feist and B.O. Palsson. The growing scope of applications of genome-scale metabolic reconstructions using escherichia coli. *Nat Biotech*, 26(6):659–667.

[5] M.A. Oberhardt, B.O. Palsson, and J.A. Papin. Applications of genome-scale metabolic reconstructions. *Molecular Systems Biology*, page 5.

[6] N.D. Price. Genome-scale microbial in silico models: the constraints-based approach. *Trends Biotechnol*, 21(4):162–9.

[7] M.A. Oberhardt, K. Yizhak, and E. Ruppin. Metabolically re-modeling the drug pipeline. *Current Opinion in Pharmacology*, 13(5):778–785.

[8] K.R. Patil, M. Åkesson, and J. Nielsen. Use of genome-scale microbial models for metabolic engineering. *Current Opinion in Biotechnology*, 15(1):64–69.

[9] A. Bordbar. Constraint-based models predict metabolic and associated cellular functions. *Nat Rev Genet*, 15(2):107–120.

[10] A. Bordbar and B.O. Palsson. Using the reconstructed genome-scale human metabolic network to study physiology and pathology. *Journal of Internal Medicine*, 271(2):131–141.

[11] A. Mardinoglu and J. Nielsen. Systems medicine and metabolic modelling. *Journal of Internal Medicine*, 271(2):142–154.

[12] M.J. Herrgard. A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nat Biotech*, 26:1155 – 1160.

[13] C.G. Oliveira Dal'Molin and L.K. Nielsen. Plant genome-scale metabolic reconstruction and modelling. *Current Opinion in Biotechnology*, 24(2):271–277.

[14] T. Töpel. Ramedis: the rare metabolic diseases database. *Applied Bioinformatics*, 5(2):115–118.

[15] M. Durot, P.-Y. Bourguignon, and V. Schachter. Genome-scale models of bacterial metabolism: reconstruction and applications. *FEMS Microbiology Reviews*, 33(1):164–190.

[16] I. Thiele and B.O. Palsson. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protocols*, 5(1):93–121.

[17] J.D. Orth, I. Thiele, and B.O. Palsson. What is flux balance analysis? *Nat Biotech*, 28(3):245–248.

[18] J. Niklas, K. Schneider, and E. Heinzle. Metabolic flux analysis in eukaryotes. *Current Opinion in Biotechnology*, 21(1):63–69.

[19] T. Shlomi, M.N. Cabili, and E. Ruppin. Predicting metabolic biomarkers of human inborn errors of metabolism. *Molecular Systems Biology*, 5:263.

[20] n Jamshidi. Dynamic simulation of the human red blood cell metabolic network. *Bioinformatics*, 17(3):286–287.

[21] N. Jamshidi and B.O. Palsson. Formulating genome-scale kinetic models in the postgenome era. *Molecular Systems Biology*, page 4.

[22] M.W. Covert and B.O. Palsson. Constraints-based models: regulation of gene expression reduces the steady-state solution space. *Journal of theoretical biology*, 221(3):309–25.

[23] N.C. Duarte. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences*, 104(6):1777–1782.

[24] H. Ma. The edinburgh human metabolic network reconstruction and its functional analysis. *Molecular Systems Biology*, 3(135).

[25] D.S. Lee. The implications of human metabolic network topology for disease comorbidity. *Proceedings of the National Academy of Sciences*, 105(29):9880–5.

[26] T. Shlomi. Network-based prediction of human tissue-specific metabolism. *Nature Biotechnology*, 26(9):1003–10.

[27] B.O. Palsson. *Systems Biology : properties of reconstructed networks*. Cambridge University Press, New York.

[28] A. Varma and B. Palsson, and metabolic capabilities of escherichia coli. ii. *Optimal growth patterns. J Theor Biol*, 165:503– 522.

[29] D. Segre, D. Vitkup, and G.M. Church. Analysis of optimality in natural and perturbed metabolic networks. *Proceedings of the National Academy of Sciences*, 99(23):15112–15117.

[30] T. Shlomi, O. Berkman, and E. Ruppin. Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proceedings of the National Academy of Sciences*, 102(21):7695–700.

[31] J.D. Orth, I. Thiele, and B.O. Palsson. What is flux balance analysis? *Nat Biotech*, 28(3):245–248.

[32] M.S. Lawrence. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, 505(7484):495–501.

[33] K. Cibulskis, M.S. Lawrence, S.L. Carter, A. Sivachenko, and D. Jaffe. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, 31:213–219.

[34] Michael S. Lawrence. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–218.

[35] M.G. Vander Heiden, L.C. Cantley, and C.B. Thompson. Understanding the warburg effect: the metabolic requirements of cell proliferation. *Science*, 324:1029–1033.

[36] M.C. Li, R. Hertz, and D.B. Spencer. Effect of methotrexate therapy upon choriocarcinoma and chorioadenoma. *Experimental Biology and Medicine*, 93(2):361–366.

[37] D. Hanahan and R.A. Weinberg. Hallmarks of cancer: The next generation. *Cell*, 144(5):646–674.

[38] G.L. Semenza. Hif-1: upstream and downstream of cancer metabolism. *Current Opinion in Genetics Development*, 20(1):51–56.

[39] R.J. Shaw, L.C. Cantley, and Ras. Pi(3)k and mtor signalling controls tumour cell growth. *Nature*, 441(7092):424–430.

[40] D.A. Guertin and D.M. Sabatini. Defining the role of mtor in cancer. *Cancer Cell*, 12(1):9–22.

[41] D.R. Wise. Myc regulates a transcriptional program that stimulates mitochondrial glutaminolysis and leads to glutamine addiction. *Proceedings of the National Academy of Sciences*, 105:18782–18787.

[42] R.A. Cairns, I.S. Harris, and T.W. Mak. Regulation of cancer cell metabolism. *Nature Reviews Cancer*, 11(2):85–95.

[43] Eva Yus. Impact of genome reduction on bacterial metabolism and its regulation. *Science*, 26(5957):1263–1268.

[44] Marc Güell. Transcriptome complexity in a genome-reduced bacterium. *Science*, 326(5957):1268–1271.

[45] Sebastian Kühner. Proteome organization in a genome-reduced bacterium. *Science*, 326(5957):1235–1240.

[46] A.P.E.A. Oliveira. Regulation of yeast central metabolism by enzyme phosphorylation. *Molecular Systems Biology*, 8, 1.

[47] M. Åkesson, J. Förster, and J. Nielsen. Integration of gene expression data into genome-scale metabolic models. *Metabolic Engineering*, 6(4):285–293 27.

[48] S.A. Becker and B.O. Palsson. Context-specific metabolic networks are consistent with experiments. *PLoS Computational Biology*, 4(5):1000082.

[49] Elizabeth Iorns, Christopher J. Lord, Nicholas Turner, and Alan Ashworth. Utilizing rna interference to enhance cancer drug discovery. *Nature reviews. Drug discovery*, 6(7):556–568.

[50] Rachel Brough, Jessica R. Frankum, Sara Costa-Cabral, Christopher J. Lord, and Alan Ashworth. Searching for synthetic lethality in cancer. *Current Opinion in Genetics and Development*, 21(1):34–41.

[51] Archana Bommi-Reddy, Ingrid Almeciga, Jacqueline Sawyer, Christoph Geisen, Wenliang Li, Ed Harlow, William G. Kaelin, and Dorre Grueneberg. Kinase requirements in human cells: Iii. altered kinase requirements in vhl-/- cancer cells detected in a pilot synthetic lethal screen. *Proceedings of the National Academy of Sciences*, 105(43):16489,.

[52] Caponigro G. Stransky N. Venkatesan K. Margolin A.A. Kim S. Wilson C.J. Lehár J. Kryukov G.V. Sonkin D. Barretina, J. and A. Reddy. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–307.

[53] Edelman E.J. Heidorn S.J. Greenman C.D. Dastur A. Lau K.W. Greninger P. Thompson I.R. Luo X. Soares J. Garnett, M.J. and Q. Liu. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, 483(7391).

[54] Sarah A. Martin, Afshan McCarthy, Louise J. Barber, Darren J. Burgess, Suzanne Parry, Christopher J. Lord, and Alan Ashworth. Methotrexate in duces oxidative dna damage and is selectively lethal to tumour cells with defects in the dna mismatch repair gene msh2. *EMBO Molecular Medicine*, 1(6-7):323–337.

[55] Nicholas C. Turner, Christopher J. Lord, Elizabeth Iorns, Rachel Brough, Sally Swift, Richard Elliott, Sydonia Rayter, Andrew N. Tutt, and Alan Ashworth. A synthetic lethal sirna screen identifying genes mediating sensitivity to a parp inhibitor. *The EMBO Journal*, 27(9):1368–1377.

[56] C.J. Lord and A. Ashworth. Mechanisms of resistance to therapies targeting brca-mutant cancers. *Nature Medicine*, 19(11):1381–1388.

[57] A systematic mammalian genetic inter action map reveals pathways underlying ricin susceptibility. *Cell*, 152(4).

[58] Alan Ashworth, Christopher J. Lord, and Jorge S. Reis-Filho. Genetic interac tions in cancer progression and treatment. *Cell*, 145(1):30–38.

[59] H Hartwell, Philippe Szankasi, Christopher J. Roberts, Andrew W.Mur ray, and Stephen H. Friend. Integrating genetic approaches into the discovery of anticancer drugs. *Science*, 278(5340):1064–1068.

[60] Babu V. Sajesh, Brent J. Guppy, and Kirk J. McManus. Synthetic genetic targeting of genome instability in cancer. *Cancers*, 5(3):739–61.

[61] Livnat Jerby-Arnon, Nadja Pfetzer, YedaelY Waldman, Lynn McGarry, Daniel James, Emma Shanks, Brinton Seashore-Ludlow, Adam Weinstock, Tamar Geiger, PaulA Clemons, Eyal Gottlieb, and Eytan Ruppin. Predicting cancer-specific vulnerability via data-driven detection of synthetic lethal ity. *Cell*, 158(5):1199–1209.

[62] Alan Ashworth, Christopher J. Lord, and Jorge S. Reis-Filho. Genetic interac tions in cancer progression and treatment. *Cell*, 145(1):30–38.

[63] Leland H. Hartwell, Philippe Szankasi, Christopher J. Roberts, Andrew W.Mur ray, and Stephen H. Friend. Integrating genetic approaches into the discovery of anticancer drugs. *Science*, 278(5340):1064–1068.

[64] I.B.Weinstein Cancer. Addiction to oncogenes–the achilles heal of cancer. *Science*, 297(5578):63–64.

[65] The genetic landscape of a cell. *Science*, 327(5964):425–431.

[66] DeLuna A. Church G.M. Segre, D. and R. Kishony. Modular epistasis in yeast metabolism.

[67] D. Deutscher, I. Meilijson, M. Kupiec, and E. Ruppin. Multiple knockout analysis of genetic robustness in the yeast metabolic network. *Nature Genetics*, 38(9):993–998.

[68] Meilijson I. Kupiec M. Deutscher, D. and E. Ruppin. *Multiple knockout analysis of genetic robustness in the yeast metabolic network*.

[69] Zheng L. Folger O. Rajagopalan K.N. MacKenzie E.D. Jerby L. Micaroni M. Chaneton B. Adam J. Hedley A. Frezza, C. and G. Kalna. *Haem oxygenase is synthetically lethal with the tumour suppressor fumarate hydratase*.

[70] T. Lindahl. Instability and decay of the primary structure of dna. *Nature*, 362:709–715,.

[71] T. Lindahl and B. Nyberg. Rate of depurination of native deoxyribonucleic acid. *Biochemistry*, 11:3610–3618,.

[72] J. Cadet and J.R. Wagner. Dna base damage by reactive oxygen species, oxidizing agents, and uv radiation. *Cold Spring Harb Perspect Biol*, 5.

[73] D. Branzei and M. Foiani. The dna damage response during dna replication. *Current Opinion Cell Biology*, 17:568–575,.

[74] G. D'Alessandro and F. Fagagna. Transcription and dna damage: Holding hands or crossing swords? *Journal of Molecular Biology*, 429:3215–3229,.

[75] L.A. Loeb and C.C. Harris. Advances in chemical carcinogenesis: a historical review and prospective. *Cancer Research*, 68:6863–6872,.

[76] R.P. Rastogi, Kumar Richa, Tyagi A., M. B., and R.P. Sinha. Molecular mechanisms of ultraviolet radiation-induced dna damage and repair. *Journal of Nucleic Acids*, 592980.

[77] P.J. Stephens. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell*, 144:27–40,.

[78] S. Nik-Zainal. Mutational processes molding the genomes of 21 breast cancers. *Cell*, 149:979–993,.

[79] L.B. Alexandrov. Signatures of mutational processes in human cancer. *Nature*, 500:415–421,.

[80] L.B. Alexandrov, S. Nik-Zainal, D.C. Wedge, P.J. Campbell, and M.R. Stratton. Deciphering signatures of mutational processes operative in human cancer. *Cell Reports*, 3:246–259,.

[81] M.R. Stratton, P.J. Campbell, and P.A. Futreal. The cancer genome. *Nature*, 458:719–724,.

[82] D. Hansemann and T. Bovieri. Ueber asymmetrische zelltheilung in epithelkrebsen und deren biologische bedeutung. *Archiv für pathologische Anatomie und Physiologie und für klinische Medicin*, 119:110.

[83] L.A. Loeb and C.C. Harris. Advances in chemical carcinogenesis: a historical review and prospective. *Cancer Research*, 68:6863–6872,.

[84] Nir Hacohen. Getting personal with neoantigen-based therapeutic cancer vaccines. *Cancer Immunology Research*, 1(1):11–15.

[85] Bianca Heemskerk, Pia Kvistborg, and Ton N.M. Schumacher. The cancer antigenome. *The EMBO Journal*, 32(2):194–203.

[86] Marialuisa Sensi and Andrea Anichini. Unique tumor antigens: Evidence for immune control of genome integrity and immunogenic targets for t cell–mediated patient-specific immunotherapy. *Clinical Cancer Research*, 12(17):5023–5032.

[87] Giorgio Parmiani. Universal and stemness-related tumor antigens: potential use in cancer immunotherapy. *Clinical Cancer Research*, 13(19):5675–5679.

[88] Ofer Mandelboim. Regression of established murine carcinoma metastases following vaccination with tumour-associated antigen peptides. *Nature Medicine*, 1(11):1179–1183.

[89] Volker Lennerz. The response of autologous t cells to a human melanoma is dominated by mutated neoantigens. *Proceedings of the National Academy of Sciences*, 102(44):16013–16018.

[90] Juhua Zhou. Persistence of multiple tumor-specific t-cell clones is associated with complete tumor regression in a melanoma patient receiving adoptive cell transfer therapy. *Journal of Immunotherapy*, 1997:28 1.

[91] D.C. Koboldt, Q. Zhang, D.E. Larson, D. Shen, and M.D. McLellan. Varscan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, 22:568–576.

[92] C.T. Saunders, W.S. Wong, S. Swamy, J. Becq, and L.J. Murray. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics*, 28:1811–1817.

[93] D.E. Larson, C.C. Harris, K. Chen, D.C. Koboldt, and T.E. Abbott. Somaticsniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics*, 28:311–317.

[94] D.C. Koboldt, K. Chen, T. Wylie, D.E. Larson, and M.D. McLellan. Varscan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, 25:2283–2285.

[95] R. Goya, M.G. Sun, R.D. Morin, G. Leung, and G. Ha. Snvmix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics*, 26:730–736.

[96] Z. Wang, M. Gerstein, and M. Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10:57–63.

[97] J. Shendure. The beginning of the end for microarrays? *Nature Methods*, 5:585–587.

[98] Y. Guo, Q. Sheng, J. Li, F. Ye, D.C. Samuels, and Y. Shyr. Large scale comparison of gene expression levels by microarrays and rnaseq using tcga data. *PLoS One*, 8:71462.

[99] Y. Guo, C.I. Li, F. Ye, and Y. Shyr. Evaluation of read count based rnaseq analysis methods. *BMC Genomics*, 14(Suppl. 8):2.

[100] Keren Yizhak. Rna sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. *Science*, 364(6444).

[101] T.F. Gajewski, H. Schreiber, and Y.-X. Fu. Innate and adaptive immune cells in the tumor microenvironment. *Nature Immunology*, 14:1014.

[102] S.A. Rosenberg and J.N. Kochenderfer. Personalized cell transfer immunotherapy for b-cell malignancies and solid cancers. *Molecular Therapy*, 19:1928–1930.

[103] D.M. Pardoll. The blockade of immune checkpoints in cancer immunotherapy. *Nature Reviews Cancer*, 12:252.

[104] S.C. Wei, C.R. Duffy, and J.P. Allison. Fundamental mechanisms of immune checkpoint blockade therapy. *Cancer Discovery*, 8:1069 – 1086.

[105] C. Robert, Long G. V, B. Brady, C. Dutriaux, M. Maio, and L. Mortier. Nivolumab in previously untreated melanoma without braf mutation. *New England Journal of Medicine*, 372:320–330.

[106] C. Robert, J. Schachter, Long G. V, A. Arance, J.J. Grob, and L. Mortier. Pembrolizumab versus ipilimumab in advanced melanoma. *New England Journal of Medicine*, 372:2521–2532.

[107] M.A. Postow, J. Chesney, A.C. Pavlick, C. Robert, K. Grossmann, and D. McDermott. Nivolumab and ipilimumab versus ipilimumab in untreated melanoma. *New England Journal of Medicine*, 372:2006–2017.

[108] J.D. Wolchok, H. Kluger, M.K. Callahan, M.A. Postow, N.A. Rizvi, and A.M. Lesokhin. Nivolumab plus ipilimumab in advanced melanoma. *New England Journal of Medicine*, 369:122–133.

[109] M. Costanzo. The genetic landscape of a cell. *Science*, 327(5964):425–431.

[110] B. Szappanos. An integrated approach to characterize genetic interaction networks in yeast metabolism. *Nature Genetics*, 43(7):656–662.

[111] P.F. Suthers, A. Zomorrodi, and C.D. Maranas. Genome-scale gene/reaction essentiality and synthetic lethality analysis. *Molecular Systems Biology*, 5(301).

[112] D. Segrè, A. Deluna, G.M. Church, and R. Kishony. Modular epistasis in yeast metabolism. *Nature Genetics*, 37(1):77–83.

[113] O. Folger. Predicting selective drug targets in cancer through metabolic networks. *Molecular Systems Biology*, 7(501).

[114] C. Frezza. Haem oxygenase is synthetically lethal with the tumour suppressor fumaratehydratase. *Nature*, 477(7363):225–228.

[115] C. Laufer, B. Fischer, M. Billmann, W. Huber, and M. Boutros. Mapping genetic interactions in human cancer cells with rnai and multiparametric phenotyping. *Nature Methods*, 10(5):427–431.

[116] J. Luo. A genome-wide rnai screen identifies multiple synthetic lethal interactions with the ras oncogene. *Cell*, 137(5):835–848.

[117] X. Lu, P.R. Kensche, M.A. Huynen, and R.A. Notebaart. Genome evolution predicts genetic interactions in protein complexes and reveals cancer drug targets. *Nature Communications*, 4(2124).

[118] W.G. Kaelin, Jr. The concept of synthetic lethality in the context of anticancer therapy. *Nature Reviews Cancer*, 5(9):689–698.

[119] F.L. Rehman, C.J. Lord, and A. Ashworth. Synthetic lethal approaches to breast cancer therapy. *Nature Reviews Clinical Oncology*, 7(12):718–724.

[120] D.A. Chan and A.J. Giaccia. Harnessing synthetic lethal interactions in anticancer drug discovery. *Nature reviews Drug discovery*, 10(5):351–364.

[121] E.S. Kroll, K.M. Hyland, P. Hieter, and J.J. Li. Establishing genetic interactions by a synthetic dosage lethality phenotype. *Genetics*, 143(1):95–102.

[122] Y. Bian. Synthetic genetic array screen identifies pp2a as a therapeutic target in mad2-overexpressing tumors. *Proceedings of the National Academy of Sciences*, 111(4):1628–1633.

[123] R. Sopko. Mapping pathways and phenotypes by systematic gene overexpression. *Molecular Cell*, 21(3):319–330.

[124] A. Wagner. Computational evaluation of cellular metabolic costs successfully predicts genes whose expression is deleterious. *Proceedings of the National Academy of Sciences*, 110(47):19166–19171.

[125] M.L. Mo, N. Jamshidi, and B.Ø. Palsson. A genome-scale, constraint-based approach to systems biology of human metabolism. *Molecular Biosystems*, 3(9):598–603.

[126] J. Schellenberger. Quantitative prediction of cellular metabolism with constraint-based models: The cobra toolbox v2.0. *Nature Protocols*, 6(9):1290–1307.

[127] I. Thiele. A community-driven global reconstruction of human metabolism. *Nature Biotechnology*, 31(5):419–425.

[128] M. Terzer, N.D. Maynard, M.W. Covert, and J. Stelling. Genome-scale metabolic networks. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 1(3):285–297.

[129] H. Nam. Network context and selection in the evolution to enzyme specificity. *Science*, 337(6098):1101–1104.

[130] R.A. Notebaart. Network-level architecture and the evolutionary potential of underground metabolism. *Proceedings of the National Academy of Sciences*, 111(32):11762–11767.

[131] L. Jerby. Metabolic associations of reduced proliferation and oxidative stress in advanced breast cancer. *Cancer Research*, 72(22):5712–5720.

[132] L. Jerby and E. Ruppin. Metabolic associations of reduced proliferation and oxidative stress in advanced breast cancer. *Cancer research*, 72(22):5712–5720.

[133] R. Agren. Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using init. *PLOS Computational Biology*, 8(5).

[134] F. Gatto, I. Nookaew, and J. Nielsen. Chromosome 3p loss of heterozygosity is associated with a unique metabolic network in clear cell renal carcinoma. *Proceedings of the National Academy of Sciences*, 111(9).

[135] A. Varma and B.O. Palsson. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type escherichia coli w3110. *Applied and Environmental Microbiology*, 60(10):3724–3731.

[136] E. Favaro. Glucose utilization via glycogen phosphorylase sustains proliferation and prevents premature senescence in cancer cells. *Cell Metabolism*, 16(6):751–764.

[137] R. Marcotte. Essential gene profiles in breast, pancreatic, and ovarian cancer cells. *Cancer Discovery*, 2(2):172–189.

[138] J. Barretina. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607.

[139] L. Chin, W.C. Hahn, G. Getz, and M. Meyerson. Making sense of cancer genomic data. *Genes Development*, 25(6):534–555.

[140] X. Lu, W. Megchelenbrink, R.A. Notebaart, and M.A. Huynen. Predicting human genetic interactions from cancer genome evolution. *PLoS One*, 10(5).

[141] C. Curtis. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352.

[142] B. Gyorffy, A. Lánczky, and Z. Szállási. Implementing an online tool for genome-wide validation of survival-associated biomarkers in ovarian-cancer using microarray data from 1287 patients. *Endocrine-related Cancer*, 19(2):197–208.

[143] J.M. Bland and D.G. Altman. The logrank test. *BMJ*, 328(7447).

[144] L. Yang. Metabolic shifts toward glutamine regulate tumor growth, invasion and bioenergetics in ovarian cancer. *Molecular Systems Biology*, 10(5):20134892.

[145] J.E.A. Schellenberger. Bigg: a biochemical genetic and genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics*, 11(1):213.

[146] S. Rossell, M.A. Huynen, and R.A. Notebaart. Inferring metabolic states in uncharacterized environments using gene-expression measurements. *PLoS Computational Biology*, 9(3):1002988.

[147] J.M. Bland and D.G. Altman. Survival probabilities (the kaplan-meier method. *Bmj*, 317(7172):1572–1580.

[148] Y. Audic and R.S. Hartley. Post-transcriptional regulation in cancer. *Biology of the Cell*, 96(7):479–498.

[149] B. Ell and Y. Kang. Transcriptional control of cancer metastasis. *Trends in Cell Biology*, 23(12):603–611.

[150] Deng X. Ruvolo, P.P. and W.S. May. Phosphorylation of bcl2 and regulation of apoptosis. *Leukemia*, 15(4):515.

[151] Bodenmiller B. Uotila A. Stahl M. Wanka S. Gerrits B. Aebersold R. Huber, A. and R. Loewith. Characterization of the rapamycin-sensitive phosphoproteome reveals that sch9 is a central coordinator of protein synthesis. *Genes Development*, 23(16):1929–1943.

[152] Muñoz J. Braam S.R. Pinkse M.W. Linding R. Heck A.J. Mummery C.L. Van Hoof, D. and J. Krijgsveld. Phosphorylation dynamics during early differentiation of human embryonic stem cells. *Cell Stem Cell*, 5(2):214–226.

[153] Sgarbi G. Solaini, G. and A. Baracca. Oxidative phosphorylation in cancer cells. *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, 1807(6):534–542.

[154] Marin S. Lee P.W. Selivanov, V.A. and M. Cascante. Software for dynamic analysis of tracer-based metabolomic data: estimation of metabolic fluxes and their statistical analysis. *Bioinformatics*, 22(22):2806–2812.

[155] Meshalkina L.E. Solovjeva O.N. Kuchel P.W. Ramos-Montoya A. Kochetov G.A. Lee P.W. Selivanov, V.A. and M. Cascante. Rapid simulation and analysis of isotopomer distributions using constraints based on enzyme mechanisms: an example from ht29 cancer cells. *Bioinformatics*, 21(17):3558–3564.

[156] Puigjaner J. Sillero A. Centelles J.J. Ramos-Montoya A. Lee P.W.N. Selivanov, V.A. and M. Cascante. An optimized algorithm for flux estimation from isotopomer distribution in glucose metabolites. *Bioinformatics*, 20(18):3387–3397.

[157] I.M.E.A. Mas. Compartmentation of glycogen metabolism revealed from 13c isotopologue distributions. *BMC Systems Biology*, 5(1):175.

[158] S.P.E.A. Gygi. Correlation between protein and mrna abundance in yeast. *Molecular and Cellular Biology*, 19(3):1720–1730.

[159] P.-J.E.A. Lahtvee. Absolute quantification of protein and mrna abundances demonstrate variability in gene-specific translation efficiency in yeast. *Cell Systems*, 4(5):495–504.

[160] F.E.A. Edfors. Gene-specific correlation of rna and protein levels in human cells and tissues. *Molecular Systems Biology*, 12, 10.

[161] G.E.A. Chen. Discordant protein and mrna expression in lung adenocarcinomas. *Molecular Cellular Proteomics*, 1(4):304–313.

[162] J.F.E.A. Moxley. Linking high-resolution metabolic flux phenotypes and transcriptional regulation in yeast modulated by the global regulator gcn4p. *Proceedings of the National Academy of Sciences*, 106(16):6477–6482.

[163] E.Y.E.A. Chen. Enrichr: interactive and collaborative html5 gene list enrichment analysis tool. *BMC Bioinformatics*, 14(1):128.

[164] M.V.E.A. Kuleshov. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, 44(W1):90– 97.

[165] A.E.A. Sandelin. Jaspar: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research*, 32:91– 94.

[166] V.E.A. Matys. Transfac: transcriptional regulation, from patterns to profiles. *Nucleic Acids Research*, 31(1):374–378.

[167] A.E.A. Lachmann. Chea: transcription factor regulation inferred from integrating genome-wide chip-x experiments. *Bioinformatics*, 26(19):2438–2444.

[168] K.R.E.A. Rosenbloom. Encode whole-genome data in the ucsc genome browser: update 2012. *Nucleic Acids Research*, 40(D1):912– 917.

[169] E.P. Consortium. An integrated encyclopedia of dna elements in the human genome. *Nature*, 57(489):7414.

[170] I.E.A. Ben-Sahra. Stimulation of de novo pyrimidine synthesis by growth signaling through mtor and s6k1. *Science*, 339(6125):1323–1328.

[171] L.G.A.M.S.P. Korotchkina. Mutagenesis studies of the phosphorylation sites of recombinant human pyruvate dehydrogenase. *Site-specific regulation., Journal of Biological Chemistry*, 270(24):14297–14304.

[172] C.B.P.A.M.J.L. Pál. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nature Genetics*, 37(12):1372.

[173] J.G.E.A. Bundy. Evaluation of predicted network modules in yeast metabolism using nmr-based metabolite profiling. *Genome Research*, 17(4):510–519.

[174] R.A.E.A. Notebaart. Co-regulation of metabolic genes is better explained by flux coupling than by network distance. *PLoS Computational Biology*, 4(1):26.

[175] E.W.L.A.C.W. Klipp. Inferring dynamic properties of biochemical reaction networks from structural knowledge. *Genome Informatics*, 15(1):125–137.

[176] M.E.A. Cascante. Metabolic control analysis in drug discovery and disease. *Nature Biotechnology*, 20(3):243.

[177] M.A.P.B. Strumillo. Towards the computational design of protein post-translational regulation. *Bioorganic Medicinal Chemistry*, 23(12):2877–2882.

[178] M. Audagnotto and M. Dal Peraro. Protein post-translational modifications: In silico prediction tools and molecular modeling. *Computational and Structural Biotechnology Journal*, 15:307–319.

[179] I.M.E.A. Mas. Stoichiometric gene-to-reaction associations enhance model-driven analysis performance: Metabolic response to chronic exposure to aldrin in prostate cancer. *BMC Genomics*, 20(1):1–12.

[180] S.A.E.A. Becker. Quantitative prediction of cellular metabolism with constraint-based models: the cobra toolbox. *Nature Protocols*, 2(3):727.

[181] D.E.A.R.L.S. Kaufman. Direction choice for accelerated convergence in hit-and-run sampling. *Operations Research*, 46(1):84–95.

[182] J.R.E.A. Wiśniewski. Universal sample preparation method for proteome analysis. *Nature Methods*, 6(5):359.

[183] Y.A.Y.H. Benjamini. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological*, 57(1):289–300.

[184] J. Vijg. *Somatic mutations, genome mosaicism, cancer and aging*. Curr Opin Genet Dev.

[185] I. Martincorena. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science*, 348(6237):880–886.

[186] M.A. Lodato, M.B. Woodworth, and S. Lee. Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science*, 350(6256):94–98,.

[187] F. Blokzijl. Tissue-specific mutation accumulation in human adult stem cells during life. *Nature*, 538(7624):260–264.

[188] M.S. Lawrence. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–218.

[189] T. Davoli. Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell*, 155(4):948–962.

[190] L.A. Garraway and E.S. Lander. Lessons from the cancer genome. *Cell*, 153(1):17–37.

[191] B. Vogelstein. Cancer genome landscapes. *Science*, 339(6127):1546–1558.

[192] T.P. Dentro, S. C., I. Leshchiner, K. Haase, M. Tarabichi, J. Wintersinger, A.G. Deshwar, and Yang. Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. *Cell*, 184(8):2239–2254 ,.

[193] J. Ma, J. Setton, and N. Lee. The therapeutic significance of mutational signatures from dna repair deficiency in cancer. *Nat Commun*, 9:3292,.

[194] A. Vanderstichele, P. Busschaert, S. Olbrecht, D. Lambrechts, and I. Vergote. Genomic signatures as predictive biomarkers of homologous recombination deficiency in ovarian cancer. *Eur J Cancer*, (v;86:5-14).

[195] K. Cibulskis. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol*, 31(3):213–219.

[196] D.C. Koboldt. Varscan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res*, 22:568–576.

[197] S. Kim. Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods*, 15:591–594,.

[198] A. McKenna. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome Res*, 20:1297–1303,.

[199] Z. Lai. Vardict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res*, 44:108– 108.

[200] S. Hiltemann, G. Jenster, J. Trapman, P. Spek, and A. Stubbs. Discriminating somatic and germline mutations in tumor dna samples without matching normals. *Genome Res*, 25(9):1382–1390,.

[201] J.K. Teer, Y. Zhang, and L. Chen. Evaluating somatic tumor mutation detection without matched normal samples. *Hum Genomics*, 11:22,.

[202] J.X. Sun. A computational approach to distinguish somatic vs. germline origin of genomic alterations from deep sequencing of cancer specimens without a matched normal. *PLoS Comput. Biol*, 14(2):1005965.

[203] K. Yizhak. Rna sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. *Science*, 364(6444).

[204] P. Riviere. High tumor mutational burden correlates with longer survival in immunotherapy-naïve patients with diverse cancers. *Mol. Cancer Ther*, 19(10):2139–2145.

[205] A.M. Goodman. Tumor mutational burden as an independent predictor of response to immunotherapy in diverse cancers. *Mol. Cancer Ther*, 16(11):2598–2608.

[206] I. Bonta. Correlation between tumor mutation burden and response to immunotherapy.

[207] H.-X. Wu. Tumor mutational and indel burden: a systematic pan-cancer evaluation as prognostic biomarkers. *Ann. Transl. Med*, 7(22).

[208] R. Akbani. Genomic classification of cutaneous melanoma. *Cell*, 161(7).

[209] K. Sherry, Ward S.T., M., and Sirotkin. dbsnp—database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res*, pages 9, 677–679,.

[210] K.J. Karczewski, L.C. Francioli, and G. Tiao. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581:434–443,.

[211] E.V. Server. Nhlbi go exome sequencing project (esp. Online]. Available:.

[212] B. Lonsdale, Thomas J., Salvatore J., Phillips M., Lo R., Shad E., Hasz S., Walters R., Garcia G., Young F., N., and Foster. The genotype-tissue expression (gtex) project. *Nat. Genet*, 45(6):580–585 ,.

[213] H.J. Ellrott, Bailey K., Saksena M.H., Covington G., Kandoth K.R., Stewart C., Hess C., Ma J., Chiotti S., McLellan K.E., M., and Sofia. Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. *Cell Syst*, 6(3):271–281 ,.

[214] T.K. Ho. Random decision forests. In *Proc. 3rd Int. Conf. Doc. Anal. Recognit*, volume 1, page 278–282. IEEE.

[215] L.B. Alexandrov, J. Kim, and N.J. Haradhvala. The repertoire of mutational signatures in human cancer. *Nature*, 578:94–101,.

[216] J. Kim, K. Mouw, and P. Polak. Somatic ercc2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat Genet*, 48:600–606,.

[217] N. Saini. The impact of environmental and endogenous damage on somatic mutation load in human skin fibroblasts. *PLoS Genet*, 12(10):1006385.

[218] M. Lawrence, P. Stojanov, and C. Mermel. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, 505:495–501,.

[219] P. Tate, Bamford J.G., Jubb S., Sondka H.C., Beare Z., Bindal D.M., Boutselakis N., Cole H., Creatore C.G., Dawson C., E., and Fish. Cosmic: the catalogue of somatic mutations in cancer. *Nucleic Acids Res*, pages 47 1 , 941– 947 ,.

[220] N.T. McGrail, Pilié D.J., Rashid P.G., Voorwerk N.U., Slagter L., Kok M., Jonasch M., Khasraw E., Heimberger M., Lim A.B., B., and Ueno. High tumor mutation burden fails to predict immune checkpoint blockade response across all cancer types. *Ann. Oncol*, 32(5):661–672 ,.

[221] P. Klempner, Fabrizio S.J., Bane D., Reinhart S., Peoples M., Ali T., Sokol S.M., Frampton E.S., Schrock G., Anhorn A.B., R., and Reddy. Tumor mutational burden as a predictive biomarker for response to immune checkpoint inhibitors: a review of current evidence. *Oncologist*, pages 25 1 , 147 ,.

[222] E. Litchfield, Reading K., Puttick J.L., Thakkar C., Abbosh K., Bentham C., Watkins R., Rosenthal T.B., Biswas R., Rowan D., A., and Lim. Meta-analysis of tumor-and t cell-intrinsic mechanisms of sensitization to checkpoint inhibition. *Cell*, 184(3):596–614 ,.

[223] T.J. Samstein, Lee R.M., Shoushtari C.H., Hellmann A.N., Shen M.D., Janjigian R., Barron Y.Y., Zehir D.A., Jordan A., Omuro E.J., A., and Kaley. Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nat. Genet*, 51(2):202–206 ,.

[224] J.J. Valero, Lee C., Hoen M., Wang D., Nadeem J., Patel Z., Postow N., Shoushtari M.A., Plitas A.N., Balachandran G., V.P., and Smith. The association between tumor mutational burden and prognosis is dependent on treatment context. *Nat. Genet*, 53(1):11–15 ,.

[225] N. Riaz. Tumor and microenvironment evolution during immunotherapy with nivolumab. *Cell*, 171(4):934–949.

[226] Lichtenstein L. Gupta M. Lawrence M.S. Pugh-T.J. Saksena G. Meyerson M. Ramos, A.H. and G. Getz. Oncotator: cancer variant annotation tool. *Human mutation*, 2015.

[227] D.R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 1972.

[228] Wilcoxon F. Individual comparisons by ranking methods. *Biometrics*, 1945.

[229] C.L. Florian. Characteristic metabolic profiles revealed by 1h nmr spectroscopy for three types of human brain and nervous system tumours. *NMR in Biomedicine*, 8(6):253–264.

[230] A.R. Tate. Towards a method for automated classification of 1h mrs spectra from brain tumours. *NMR in Biomedicine*, 11(4-5):177–191.

[231] M. Yang, T. Soga, and P.J. Pollard. Oncometabolites: linking altered metabolism with cancer. *The Journal of clinical investigation*, 123(9):3652–3658.

[232] B.J. Schmidt. Gim3e: condition-specific models of cellular metabolism developed from metabolomics and expression data. *Bioinformatics*, 29(22):2900–2908.

[233] K. Yizhak. Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model. *Bioinformatics*, 26(12):255–260.

[234] C. Cotten and J. Reed. Mechanistic analysis of multi-omics datasets to generate kinetic parameters for constraint-based metabolic models. *BMC Bioinformatics*, 14(1):32.

[235] A. Morandi and P. Chiarugi. Metabolic implication of tumor:stroma crosstalk in breast cancer. *Journal of Molecular Medicine*, 92(2):117–126.

[236] A. Bordbar. A multi-tissue type genome-scale metabolic network for analysis of whole-body systems physiology. *BMC Systems Biology*, 5(1):180.

[237] S. Freilich. Competitive and cooperative metabolic interactions in bacterial communities. *Nat Commun*, 2:589.

[238] A.R. Zomorrodi and C.D. Maranas. Optcom: A multi-level optimization framework for the metabolic modeling and analysis of microbial communities. *PLoS Comput Biol*, 8(2):1002363.

[239] A.B. Kroigard, M. Thomassen, A.V. Laenkholm, T.A. Kruse, and M.J. Larsen. Evaluation of nine somatic variant callers for detection of somatic mutations in exome and targeted deep sequencing data. *PLoS ONE*, 11:0151664.

[240] N.D. Roberts. A comparative analysis of algorithms for somatic snv detection in cancer. *Bioinformatics*, 29:2223–2230.

[241] Q. Wang. Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. *Genome Med*, 5:91.

[242] S.Y. Kim and T.P. Speed. Comparing somatic mutation-callers: beyond venn diagrams. *BMC Bioinform*, 14:189.

[243] C. Xu. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Comput. Struct. Biotechnol. J*, 16:15–24.

[244] I. Anzar, A. Sverchkova, R. Stratford, and T. Clancy. Neomutate: an ensemble machine learning framework for the prediction of somatic mutations in cancer. *BMC Med. Genomics*, 12:63.

[245] J. O'Rawe. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med*, 5:28.

[246] L.K. Hansen and P. Salamon. Neural network ensembles. *IEEE Trans. Pattern Anal*, 12:993–1001.

[247] G. Brown, J. Wyatt, R. Harris, and X. Yao. Diversity creation methods: a survey and categorisation. *Inf. Fusion*, 6:5–20.

[248] A. D. et al. Ewing. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nature Methods*, 12:623–630.

[249] D.L. Goode. A simple consensus approach improves somatic mutation prediction accuracy. *Genome Med*, 5:90.

[250] M. Callari. Intersect-then-combine approach: improving the performance of somatic variant calling in whole exome sequencing data using multiple aligners and callers. *Genome Med*, 9:35.

[251] S.Y. Kim, L. Jacob, and T.P. Speed. Combining calls from multiple somatic mutation-callers. *BMC Bioinform*, 15:154.

[252] B.L. Cantarel. Baysic: a bayesian method for combining sets of genome variants with improved specificity and sensitivity. *BMC Bioinform*, 15:104.

[253] L.T. Fang. An ensemble approach to accurately detect somatic mutations using somaticseq. *Genome Biol*, 16:197.

[254] B.J. Ainscough. A deep learning approach to automate refinement of somatic variant calling from cancer sequencing data. *Nat. Genet*, 50:1735–1743.

[255] S.M.E. Sahraeian. Deep convolutional neural networks for accurate somatic mutation detection. *Nat. Commun*, 10:1041.

[256] Xu H., DiCarlo J., Satya R.V., Peng Q., and Wang Y. Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. *BMC Genomics*, 15(1).

[257] *Evaluating variant calling tools for non-matched next-generation sequencing data*. Sci Rep.

[258] Spencer D.H., Tyagi M., Vallania F., Bredemeyer A.J., Pfeifer J.D., and Mitra R.D. Performance of common analysis methods for detecting low-frequency single nucleotide variants in targeted next-generation sequence data. *J Mol Diagn*, 16(1):75–88.

[259] Cai L., Yuan W., Zhang Z., He L., and Chou K.-C. In-depth comparison of somatic point mutation callers based on different tumor next-generation sequencing depth data. *Sci Rep*, 6(36540).