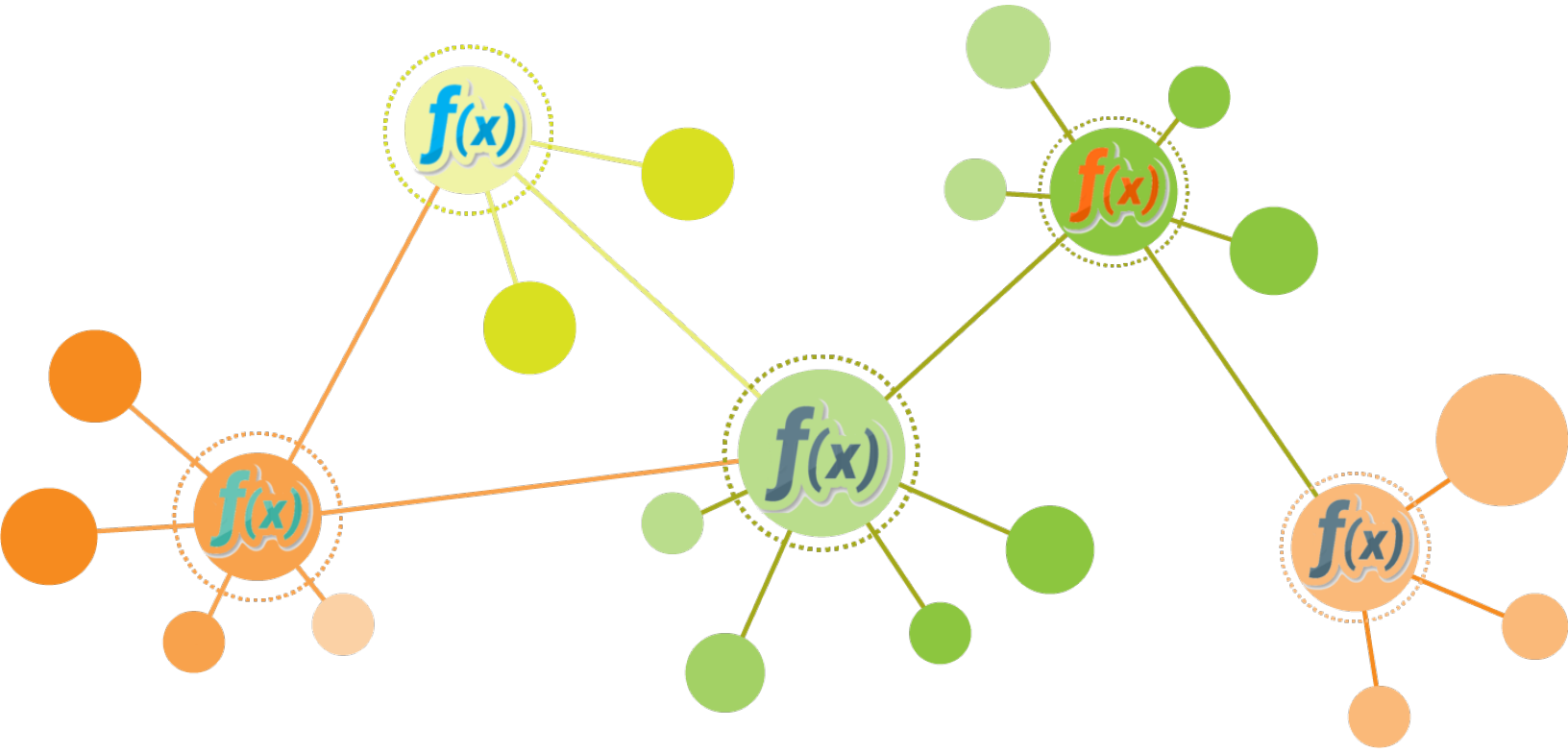
Modeling Functional Modules Using Statistical and Machine Learning Methods



Cankut CUBUK
PhD Thesis

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Modeling Functional Modules Using Statistical and Machine Learning Methods

Cankut Çubuk

Department of Biotechnology
Universitat Politècnica de València

Supervisor: Dr. Joaquín Dopazo
Tutor: Dr. Joaquín Cañizares Sales

This dissertation is submitted for the degree of
Doctor of Philosophy
March 2020

The supreme guide in life is science/knowledge.

Mustafa Kemal Atatürk

Acknowledgements

There are many people who contributed to my thesis directly and indirectly. This includes friends and collaborators who gave me valuable ideas and helped. All my colleagues and people who I met during this long journey were very friendly from very beginning to end and let me feel myself at my home with full of La Paella weekends. Thanks to all these people.

First and foremost, I would like to thank Dr. Joaquin Dopazo who believed in me and gave me the opportunity to work with him in his group and let me develop my academic career under his supervision with full support.

Thanks to doctoral school of UPV, its academic commission, my thesis tutor Dr. Joaquin Canizares and Dr. María Belén Picó who was the programme coordinator, I never felt as a foreign student and could overcome bureaucratic issues.

Dr. José Carbonell Caballero and Dr. Francisco García García were very patient and helpful when I joined the group of Dr. Dopazo as an intern in 2012. Thanks to their contribution I could start to feel as a bioinformatician and loved more and more what I was doing. While I was working at CIPF, we had a small working group called Voltron with the members of Dr. Alicia Amadoz, Dr. Marta Hidalgo, Dr. Jose Carbonell and Francisco Salavert. For me, it was the first cross-functional team working experience which was full with of fun, learning and developing methods. I really appreciated linguistic support of Asunción Gallego for the abstract of this thesis in Spanish and Valencian languages.

I also would like to thank members of MLPM project and especially to the project coordinator Dr. Karsten Borgwardt. It was a great international project experience where I could extend my social and professional network. I really appreciated the given secondment opportunities at Pharmatics Ltd with Dr. Felix Agakov and ETH Zürich with Dr. Karsten Borgwardt. The economical support of all national and EU fundings during this scientific career development episode was very essential and I am glad for receiving these contributions.

For sure, family support is not unignorable. Even from very far away, they always let me feel that I am not alone. At the beginning of this journey, I met with a precious French lady, Helene Spera, who entered my life as a flatmate. Since then, we lived in 4 different flats, in 3 cities in 2 countries. She is still my flatmate, plus now the soulmate and mother of our baby Ada Iris. I thank her for being perfect mate and make my life meaningful.

To the next generation of my family.

For now Ada Iris and Yade Deren ÇUBUK :)

Abstract

Understanding the aspects of the cell functionality that account for disease or drug action mechanisms is the main challenge for precision medicine. In spite of the increasing availability of genomic and transcriptomic data, there is still a gap between the detection of perturbations in gene expression and the understanding of their contribution to the molecular mechanisms that ultimately account for the phenotype studied. Over the last decade, different computational and mathematical models have been proposed for pathway analysis. However, they are not taking into account the dynamic mechanisms contained by pathways as represented in their layout and the interactions between genes and proteins. In this thesis, I present two slightly different mathematical models to integrate human transcriptomic data with prior knowledge of signalling and metabolic pathways to estimate the Mechanistic Pathway Activities (MPAs). MPAs are continuous and individual level values that can be used with machine learning and statistical methods to determine biomarkers for the early diagnosis and subtype classification of the diseases, and also to suggest potential therapeutic targets for individualized therapeutic interventions.

The overall objective is, developing new and advanced systems biology approaches to propose functional hypotheses that help us to understand and interpret the complex mechanism of the diseases. These mechanisms are crucial for robust personalized drug treatments and predict clinical outcomes. First, I contributed to the development of a method which is designed to extract elementary sub-pathways from a signalling pathway and to estimate their activity. Second, this algorithm adapted to metabolic modules and it is implemented as a webtool. Third, the method used to reveal a pan-cancer metabolic landscape. In this study, I analyzed the metabolic module profile of 25 different cancer types and the method is also validated using different computational and experimental approaches. Each method developed in this thesis was benchmarked against the existing similar methods, evaluated for their sensitivity and specificity, experimentally validated when it is possible and used to predict clinical outcomes of different cancer types. The research described in this thesis and the results obtained were published in different systems biology and cancer-related peer-reviewed journals and also in national newspapers.

Resumen

La comprensión de los aspectos de la funcionalidad de las células que cuentan para los mecanismos de las enfermedades es el mayor reto de la medicina personalizada. A pesar de la disponibilidad creciente de los datos de genómica y transcriptómica, sigue existiendo una notable brecha entre la detección de las perturbaciones en la expresión de genes y la comprensión de su contribución en los mecanismos moleculares que últimamente tienen relación importante con el fenotipo estudiado. A lo largo de la última década, distintos modelos computacionales y matemáticos se han propuesto para el análisis de las rutas. Sin embargo, estos modelos no toman en cuenta los mecanismos dinámicos de las rutas como la estructura y las interacciones entre genes y proteínas. En esta tesis doctoral, presento dos modelos matemáticos ligeramente distintos, para integrar los datos transcriptómicos masivos de humano con un conocimiento previo de de las rutas de señalización y metabólicas para estimar las actividades mecánicas que están detrás de esas rutas (MPAs). Las MPAs son variables continuas con valores de nivel individual que pueden ser usadas con los modelos de aprendizaje de máquinas y métodos estadísticos para determinar los biomarcadores que podemos usar para los diagnósticos tempranos y la clasificación de subtipos de enfermedades, además de poder sugerir las dianas terapéuticas potenciales para las intervenciones individualizadas.

El objetivo global es desarrollar nuevos y avanzados enfoques de la biología de sistemas para proponer unas hipótesis funcionales que nos ayuden a entender e interpretar los mecanismos complejos de las enfermedades. Estos mecanismos son cruciales para mejorar los tratamientos personalizados y predecir los resultados clínicos. En primer lugar, contribuí al desarrollo de un método que está diseñado para extraer las subrutas elementales desde la ruta de señalización con sus actividades estimadas. Posteriormente, este algoritmo se ha adaptado a los módulos metabólicos y se ha implementado como una herramienta web. Finalmente , el método ha revelado un panorama metabólico para una lista completa de diferentes tipos de cánceres. En este estudio, analicé el perfil metabólico de 25 tipos de cáncer distintos y se validó el método usando varios enfoques computacionales y experimentales. Cada método desarrollado en esta tesis ha sido enfrentado a otros métodos similares existentes, evaluados por sus sensibilidades y especificidades, experimentalmente validados cuando fue posible y usados para predecir resultados clínicos de varios tipos de cánceres. La investigación descrita en esta tesis y los resultados obtenidos fueron publicados en distintas revistas arbitradas que están relacionadas con el cáncer y biología de sistemas, y también en los periódicos nacionales.

Resum

La comprensió dels aspectes de la funcionalitat de les cèl·lules que compten per als mecanismes de les malalties és el major repte de la medicina personalitzada. Malgrat la disponibilitat creixent de les dades de genòmica i transcriptómica, continua existint una notable bretxa entre la detecció de les pertorbacions en l'expressió de gens i la comprensió de la seua contribució en els mecanismes moleculars que últimament tenen relació important amb el fenotip estudiat. Al llarg de l'última dècada, diferents models computacionals i matemàtics s'han proposat per a l'anàlisi de les rutes. No obstant això, aquests models no tenen en compte els mecanismes dinàmics de les rutes com l'estructura i les interaccions entre gens i proteïnes. En aquesta tesi doctoral, presente dos models matemàtics lleugerament diferents, per a integrar les dades transcriptómicos massius d'humà amb un coneixement previ de de les rutes de senyalització i metabòliques per a estimar les activitats mecàniques que estan darrere d'aqueixes rutes (MPAs). Les MPAs són variables contínues amb valors de nivell individual que poden ser usades amb els models d'aprenentatge de màquines i mètodes estadístics per a determinar els biomarcadores que podem usar per als diagnòstics primerencs i la classificació de subtipus de malalties, a més de poder suggerir les dianes terapèutiques potencials per a les intervencions individualitzades.

L'objectiu global és desenvolupar nous i avançats enfocaments de la biologia de sistemes per a proposar unes hipòtesis funcionals que ens ajuden a entendre i interpretar els mecanismes complexos de les malalties. Aquests mecanismes són crucials per a millorar els tractaments personalitzats i predir els resultats clínics. En primer lloc, vaig contribuir al desenvolupament d'un mètode que està dissenyat per a extraure les subrutas elementals des de la ruta de senyalització amb les seues activitats estimades. Posteriorment, aquest algorisme s'ha adaptat als mòduls metabòlics i s'ha implementat com una eina web. Finalment, el mètode ha revelat un panorama metabòlic per a una llista completa de diferents tipus de càncers. En aquest estudi, vaig analitzar el perfil metabòlic de 25 tipus de càncer diferents i es va validar el mètode usant diversos enfocaments computacionals i experimentals. Cada mètode desenvolupat en aquesta tesi ha sigut enfrontat a altres mètodes similars existents, avaluats per les seues sensibilitats i especificitats, experimentalment validats quan va ser possible i usats per a predir resultats clínics de diversos tipus de càncers. La investigació descrita en aquesta tesi i els resultats obtinguts van ser publicats en diferents revistes arbitrades que estan relacionades amb el càncer i biologia de sistemes, i també en els periòdics nacionals.

Contents

List of figures

List of tables

Glossary

MPA Mechanistic Pathway Activities
PCR Polymerase Chain Reaction
NGS Next-Generation Sequencing
STP Signal Transduction Pathway
PTM Post-Translational Modifications
GRP Gene Regulation Pathways
DEG Differentially Expressed Gene
GSEA Gene Set Enrichment Analysis
ORA Over-Representation Analysis
TPA Topology-based Pathway Analysis
GPR Gene-Protein-Reaction
HMA Human Metabolic Network
FBA Flux Balance Analysis
PPI Protein-Protein Interactions
MoA Mechanisms of Action
GO Gene Ontology
TPR True Positive Rate
FPR False Positive Rate
TCGA The Cancer Genome Atlas
GCC Genome Characterization Center
TMM Trimmed Mean of M-values normalization method
CBM Constraint-Based Models
FDR False Discovery Rate
CH Carbohydrate
AA Amino Acid
LP Lipid
NT Nucleotide
PCA Principal Component Analysis
ICCG International Cancer Genome Consortium
WT Wild Type
KD Knock Down
AIC Akaike Information Criterion
K-M Kaplan-Meier
BLCA Bladder Urothelial Carcinoma
BRCA Breast Invasive Carcinoma

CESC Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma

CLLE Chronic Lymphocytic Leukemia

COAD Colon Adenocarcinoma

GBM Glioblastoma Multiforme

HNSC Head and Neck Squamous Cell Carcinoma

KIRC Kidney Renal Clear Cell Carcinoma

KIRP Kidney Renal Papillary Cell Carcinoma

LAML Acute Myeloid Leukemia

LGG Brain Lower Grade Glioma

LIHC Liver Hepatocellular Carcinoma

LUAD Lung Adenocarcinoma

LUSC Lung Squamous Cell Carcinoma

MALY Malignant Lymphoma

OV Ovarian Serous Cystadenocarcinoma

PAAD Pancreatic Adenocarcinoma

PACA Pancreatic Cancer

PAEN Pancreatic Endocrine Neoplasms

PRAD Prostate Adenocarcinoma

READ Rectum Adenocarcinoma

SKCM Skin Cutaneous melanoma

STAD Gastric Adenocarcinoma

THCA Thyroid Papillary Carcinoma

UCEC Uterine Corpus Endometrial Carcinoma

Chapter 1

Introduction

Chapter 1
INTRODUCTION

1.1 Systems Biology

In the late 1860s, the nucleic acid was discovered by Dr Friedrich Miescher [1]. Almost one century after this discovery, two more important contributions to the molecular biology and genetics field were done. In 1953, Watson and Crick characterized the three-dimensional structure of DNA, and then, in 1958, Francis Crick explained how the DNA is converted into the functional components of the cells. This process is called Central Dogma and it describes the flow of genetic information from DNA to RNA and RNA to protein. This is the point where the systems biology was born to determine and explain biological systems. Frederick Sanger, the Nobel Prize winner and a scientist who pioneered the first sequencing technique in the 1970s stated that "*A knowledge of sequences could contribute much to our understanding of living matter*" [2]. It was known that the discovery of new genes in the genome, the quantification of their functional units and their characterization were very important. And this could be done through the base sequence content of biological molecules. Thus, the sequencing technique proposed by Sanger led to the development of other methods; first *Polymerase Chain Reaction* (PCR), then *Microarray* and *Next-Generation Sequencing* (NGS).

The biological systems are operated through dynamic interactions among genes and their products, regulatory circuits and metabolic networks. Over the last decades, the molecular techniques that are mentioned in the previous paragraph were widely used to generate a large amount of quantitative expression data for the genes and proteins which act as members of a big networking system. Such data is used to uncover the physical and chemical interactions of functional units and their causal relationships to build the structured map of cellular mechanisms by mathematical modelling. These mechanisms are called biological pathways and networks. Interestingly, it is not clear when they were pronounced for the first time, which could be assigned as a major milestone in the systems biology area. The complexity of these cellular systems is the biggest challenge in the life sciences and their better understanding can let us elucidate phenotypic traits and especially to discover the disease maps; from their initiation to progression and to their treatment.

The systems biology is an integrated approach to decipher the complexity of biological systems and it is based on computational and mathematical modelling. In a unique sentence, it is rationally defined as "*Systems biology is based on the understanding that the whole is greater than the sum of the parts*" [3]. Nowadays, this field is mainly saturated by the reference network construction and the context-specific network modelling using the genomic and transcriptomic data as constraints on top of the existing reference models. The sink of systems biology mainly fills the bucket of databases with the curated and new pathways. Therefore, "*how the dynamic*

mechanisms of these pathways can be used for the efficient treatment of complex diseases and especially for the decision making systems of the personalized medicine applications" are the most interesting questions that remain to be answered.

The research described in this PhD thesis is developed under a systems biology perspective to elucidate the biological mechanisms of complex diseases and to determine their dynamic behaviours by integrating the genomic and transcriptomic data over the curated pathways. The method developed in this thesis can be considered as a next-generation systems biology approach which considers the activity of mechanisms (*whole*) as responsible from diseases rather than the activity of their components (*the sum of the parts*).

1.2 Biological pathways and networks

Pathways and networks are the indivisible parts of systems biology. There are some small conceptual differences between these two terms which will be defined under this section.

1.2.1 Pathways

A pathway is a schematic representation of the dynamic cellular processes (flow) which indicate the different types of interactions among the molecules in a cell. Each pathway covers a small fraction of the genome and these fractions are generally bounded by the biological concepts that are also used to categorize the pathways. To indicate the direction of flow, pathways are drawn as directed graphs. Each graph, $G=(V,E)$, contains 2 main elements that are vertex/node (V) and edge/arrow (E), to represent proteins and directed interactions, respectively. Based on the pathway type, the annotation of pathway nodes can vary as; proteins, genes, enzymes and metabolites. The biological pathways can be divided into three main categories; signalling pathways, metabolic pathways and genetic pathways.

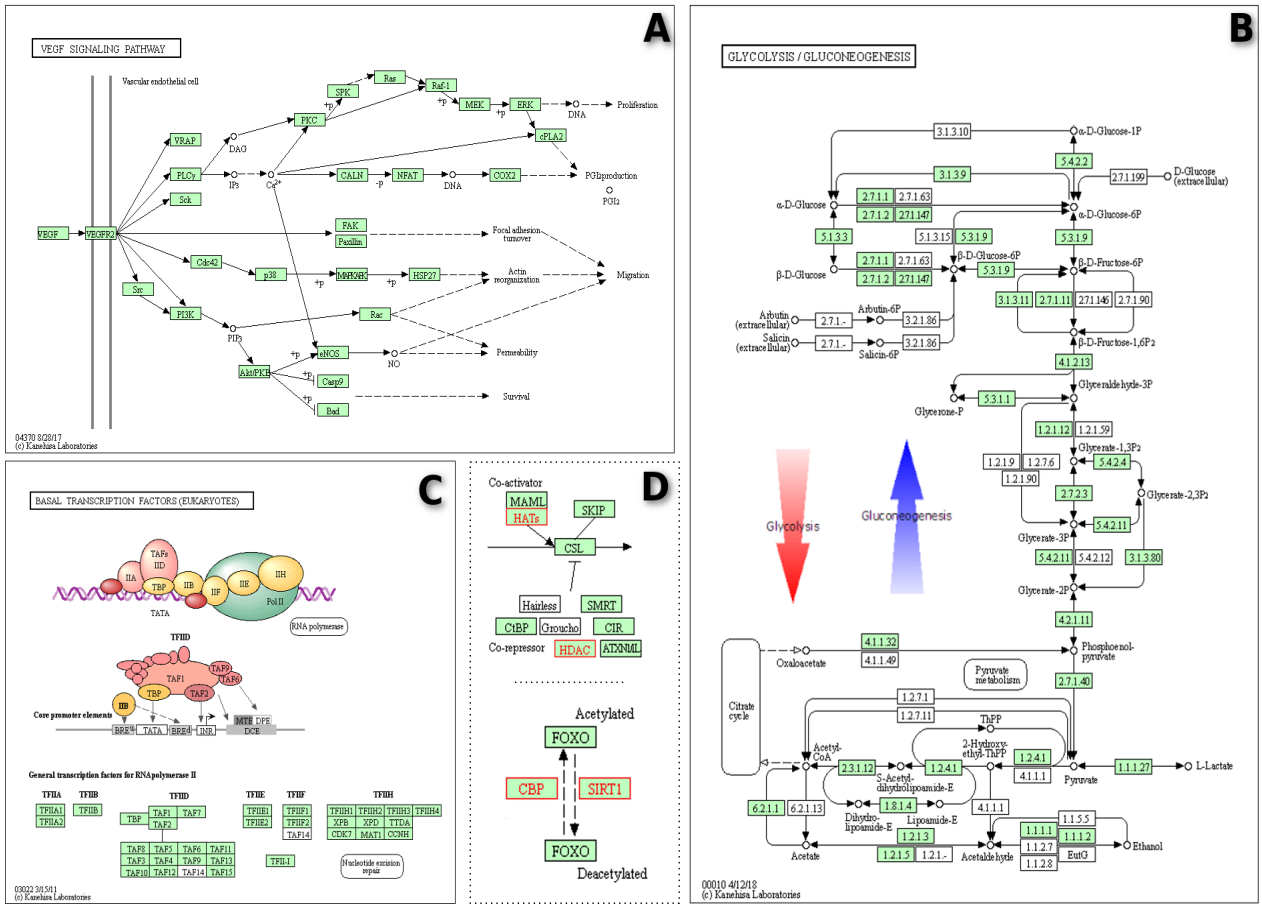


Figure 1.1: Examples for human signalling, metabolic and genetic pathways taken from KEGG database. A) VEGF signalling pathway (hsa04370). B) Glycolysis/Gluconeogenesis (hsa00010). C) Basal transcription factors (hsa03022). D) Edge (PTMs) representations in signalling pathways.

Signalling pathways are also known as Signal Transduction Pathways (STPs), transmit the specific information carried by the extracellular signalling molecules (growth factors, hormones, cytokines, metabolites, neurotransmitters, etc.) from cell's exterior to its interior (see Figure 1.1 A). Cells can capture the physical or chemical stimulants by their cell membrane receptors and convert them into the signals to initiate intracellular signalling cascades. The signal generated by the receptor must firstly be transferred into the cytoplasm and then transmitted over the intracellular protein cascade until it arrives at the effector protein or protein complex which is specialized to trigger a specific cellular function (to elicit specific cellular responses). These three signalling stages are called reception, transduction and response and they are all together aimed to maintain and control the flow of celular information. The proteins in a signalling pathway are represented as nodes and the post-translational modifications are represented as edges. The Post-Translational Modifications (PTMs) (phosphorylation, dephosphorylation, ubiquitination, glycosylation, etc.) modulate protein functions by series of complex biological reactions, however, on the contrary, their consequences on the proteins are more simple. PTMs can inhibit or activate their target proteins. Thus, the edges in the signalling pathways are drawn using *t* and *delta* arrows, inhibition and activation, respectively (see Figure 1.1 D, top). PTMs can be reversible or irreversible (like

acetylation and deacetylation of FOXO protein, by CBP and SIRT1 proteins, see Figure 1.1 D). Since the different mechanisms of reversible PTMs are catalyzed by different proteins, they can be dissected in two different irreversible modifications. And this is a general practice applied by most of the signalling pathway databases (see Figure 1.1 D, bottom) which also helps to reduce the loops. Compared to metabolic and genetic pathways, the signalling pathways contain very few feedback loops.

Metabolic pathways are the sequence of biochemical reactions that occur in the cells. These reactions are catalyzed by the enzymes and each reaction converts the reactants into the products with the help of intermediating cofactors. Inputs and outputs of metabolic reactions are called metabolites and their different bioactive forms are generally called compounds. Generally, the metabolites in a metabolic pathway are represented as circular nodes and the reactions as rectangular nodes. The edges show the direction of a reaction. In some cases, for visualization purposes, the reaction nodes can be melted over the edges and not shown. The enzymatic reactions can conduct two different processes; while the *anabolic process* builds the molecules needed by an organism such as hormones and it is an energy-consuming action, the *catabolic process* breaks down the large molecules into the smaller ones and it usually releases energy (the energy used by the anabolic process). The metabolic reactions which have the capacity of controlling these two processes known as reversible, otherwise, they are classified as irreversible reactions. The reactions of glycolysis and gluconeogenesis are the best examples for the reversible reactions. Actually, the topology of these two processes are exactly same while glycolysis is a catabolic process that breaks down glucose into small molecules to generate ATP which is the main source of energy for the most cellular processes, and gluconeogenesis is an anabolic process that is the exact opposite of glycolysis and used by the cells when there is not enough demand for glucose and ATP (see Figure 1.1 B).

Each reversible reaction in the metabolism is considered as a feedback loop and they are the smallest loops that can be found in any pathway. Thus, the metabolic pathways are more complex biological systems when they are compared with the other pathway types. Because of the highly interconnected nature of metabolic pathways, it is difficult to divide metabolic flow processes into stages. Moreover, their functional interpretation is more challenging compared to signal transduction. The research objective of this PhD thesis is focused on developing a new method that can be used both for cell signalling and metabolism by means of some minor algorithmic adaptations.

Genetic pathways are also called Gene Regulation Pathways (GRP). They demonstrate how the expression levels of mRNA and proteins are directed by some other sequence-specific DNA-binding proteins that are classified as transcription factors (see Figure 1.1 C). This is a self-controlled cellular system that is used by cells to adapt to their microenvironment and it prevents abundant production of the proteins. In simple terms, we can say that the gene regulation pathways turn genes on and off. Because of the time constraints of my PhD study, the genetic

pathways are not studied and for that reason, there is no more information given about GRPs in this thesis.

1.2.2 Pathway databases

Most of the databases contain similar pathway information. The pathway descriptions are mostly stored in systems biology specific markup language formats. Biopax, SBML and SBGN are the standard formats with different levels and used by different databases. Some databases have their own formats; like KGML from the KEGG database. The well known and prominent pathway databases are KEGG, Reactome, ASCN, SignaLink, and WikiPathways [4–8]. Comprehensive list of the pathways can be found at the studies of meta-databases that compile all pathway databases in one. OmniPath, Pathway Commons and PathCards are examples to these meta-databases [9–11].

1.2.3 Pathway analysis methods

From the late ’90s to 2015, DNA array technologies used widely. In parallel to the usage of microarrays, in the early 2000s, the databases for the functional gene sets were started to developed rapidly [4, 12]. Microarrays were used to identify the Differentially Expressed Genes (DEGs) between different conditions. However, it was most interesting to know which functional classes were regulated by these DEGs. This commonly asked question was addressed first with Over-Representation Analysis (ORA) or enrichment analysis. The ORA identifies relevant pathways by comparing the number of DEGs found in a pathway against a background gene list and in many cases this background is the number of all DEGs found in the genome. Fisher’s exact is the most commonly used statistical test to determine whether a significant number of DEGs belongs to a set of genes (gene ontology, pathways, etc.). The results of ORA can be biased by some factors; the number of genes in a pathway which varies a lot between pathways and the thresholds (e.g. $p<0.05$, fold-change>2) used to define the DEGs. To deal with the threshold limitations of ORA, Gene Set Enrichment Analysis (GSEA) was proposed and it is considered as the second generation pathway analysis method that uses the sorted list of the gene rankings. GSEA is a method proposed by Subramanian et al., 2003 [13] and since than GSEA with different ranking metrics were developed. The comprehensive list of these metrics is given by Zyla et. al., 2017 [14] with their benchmark results. Nevertheless, there is no any GSEA method that can deal with the interactions that are presented in a pathway. Thus, the new methods which use the gene content of pathways and concurrently with pathways’ topology became an emerging issue in the systems biology field. The first Topology-based Pathway Analysis (TPA) was introduced by Draghici el. al, 2007 [15]. TPAs are categorized as the third generation pathway analysis. They are one step closer to the reality of biological systems and could be used to simulate the effect of

genetic perturbations. Within the period of developments from ORA to TPA, as well as the analysis methods, the pathways were also improved; they were drawn with more detailed and curated information. Thus, this advanced pathway knowledge increased our understanding of how the pathways regulate cellular functions. First, a pathway is composed of several sub-mechanisms which are connected between them and each of them can be independently activated or deactivated. These mechanisms overlap partially and they are differentiated by at least one node (gene or protein). But even this kind of very small differences between sub-mechanisms can have very different biological implications. These sub-mechanisms are called sub-pathways or circuits. The topology-based sub-pathway activity analysis is the fourth generation pathway analysis and known as Mechanistic Pathway Activity (MPA). For the first time, when it was proposed by Sebastián-León et al., 2013 [16], the method was compatible with only the Affymetrix microarray platform. While the ORA and GSEA methods are able to analyze all kinds of pathways, TPA and MPA can only analyze the signalling pathways. The second version of MPA is developed within this PhD thesis project as platform-independent (PCR, Microarray, RNA-Seq, etc.) and compatible with both signalling and metabolic pathways. The details of this method are discussed in chapter 2 and 3.

1.2.4 Networks

A network compiles all canonical pathways under a single roof to reconstruct a more complex mathematical model. Thus, the biological networks have the size of genome-scale and demonstrate the crosstalks between the pathways. Networks get topological structure from pathways and additionally they contain more complete biochemical information such as stoichiometric coefficients and Gene-Protein-Reaction (GPR) rules. Genome-scale networks are more successfully constructed for metabolism rather than cell signalling and extensively used in the bioengineering of genetically modified microorganisms to increase the efficiency of production [17].

The BiGG, SEED, Metabolic Atlas and Biomodels are the most known resources that store the networks in XML like files such as SBML [18–21]. Human Metabolic Network (HMA) series are named with the prefix of “*Recon”* and “*HRM”*. The latest HMA was published in March 2018 and contains 13,543 metabolic reactions involving 4,140 unique metabolites [22]. Flux Balance Analysis (FBA) and its derivatives are the methods that are used to analyze metabolic networks and mostly for constraint-based modelling [23, 24]. FBA solves the series of mathematical problems by forcing the solution to get the maximum or minimum amount of the objective function under the following assumption; the production and the consumption rates of each metabolite are steady-state. The objective function can be the biomass or the production (or consumption) rate of a specific metabolite. The biomass and the amount of the targeted metabolite can be easily

measured for the microorganisms, however, this is a challenging issue for mammalian cells. For that reason, FBA methods are more successfully applied and advanced for microorganisms than animal or plant cells [25]. On the other hand, setting a correct objective function for animal cells and especially for cancer cells is still not very easy.

The protocol to construct a metabolic network from scratch is well defined by Thiele et al., 2010 [26] and it seems to be more straightforward than the signalling and regulatory network construction [27]. While metabolic pathways clearly separated from STPs and GRPs, there is a big intersection between these two non-metabolic pathway types which makes their network representation quite similar; as a relevant example, we can give the map of Protein-Protein Interactions (PPI). Apart from the compilation of all known STPs and GRPs in one network, PPI networks can also contain interactions based on causal inferences, co-expression and correlation analysis of quantified network entities. However, with this kind of analysis, the directionality of interactions can not be defined robustly [28]. Therefore, PPIs are constructed as undirected graphs to make the network edges more homogenous. PPIs can be found on the following databases; IntAct, MINT and BioGRID [29–31]. A PPI can contain up to 700.000 interactions including low-confidence level interactions. The visualization and analysis of big monolithic networks are close to being impossible because of their node sizes and the crossing edges [32]. The complexity of enormous networks suggests that the network can be divided into clusters based on the interaction strength of proteins. From the highly interconnected networks, different functional elementary modules can be attributed and these operational entities can be identified by network clustering algorithms. Short random walks, edge betweenness and greedy optimization of modularity are some of the community detection methods that have unquestionable success and widely used in this field. The extensive collection of all other methods can be found under *igraph* network analysis package. The set of detected network communities can be analyzed using enrichment methods [33–35].

1.3 Thesis outline and overview

In spite of the increasing availability of genomic and transcriptomic data, there is still a gap between the detection of perturbations in gene expression and the understanding of their contribution to the molecular mechanisms that ultimately account for the phenotype studied. Alterations in cell metabolism and signalling are behind the initiation and progression of many diseases, including cancer. The present work aims to develop a method which uses the patient's transcriptomic data to calculate the activity level of the metabolic and signalling sub-pathways that are the key contributors to human diseases. The method developed within this thesis work provides an individual level output which is very rarely happening in the systems biology field. This prominent feature of the method which has been developed in this study lets us develop machine

learning and statistical approaches to determine biomarkers for the early diagnosis and subtype classification of diseases, and also to suggest potential therapeutic targets for individualized therapeutic interventions. This thesis consists of new method development, its application, its benchmarking with similar methods and its validation in the context of cell signalling and metabolic pathways. Thus, this thesis organized into three core chapters. Each chapter starts by giving a short overview of the objectives to be achieved. Although each chapter covers a different research question, the overall objective is, developing new and advanced systems biology approach that can help us to understand and interpret the complex mechanism of the diseases for robust personalized drug treatments. Finally, all the results are exposed and discussed. The following chapters are given in this thesis for the objectives of this thesis;

1) A model of mechanistic pathway activity

2) Metabolizer web tool for differential metabolic activity analysis and discovery of therapeutic targets using summarized metabolic pathway models

3) A pan-cancer metabolic landscape based on gene expression integration into pathway modules

Chapter 2

A model of mechanistic pathway activity

Chapter 2 is adapted from the following publications: “Hidalgo MR, **Cubuk C**, Amadoz A, et al. (2017). High throughput estimation of functional cell activities reveals disease mechanisms and predicts relevant clinical outcomes. Oncotarget, 8(3):5160–5178. DOI: 10.18632/oncotarget.14107”, and “Amadoz A, Hidalgo M, **Cubuk C**, et al. (2018). A comparison of mechanistic signalling pathway activity analysis methods. Briefings in Bioinformatics, bby040, https://doi.org/10.1093/bib/bby040”. All the results and figures that appear here are derived from the work of the PhD student in collaboration with the other authors.

Chapter 2
A MODEL OF MECHANISTIC PATHWAY ACTIVITY

2.1 Overview and objectives

Although the most phenotypic traits (including disease and drug response) are the consequence of the combination of multiple altered genes (multigenic), the broad majority of biomarkers in use are based on single-gene perturbations (expression change, mutation, etc.) Obviously, the determination of the status of a single gene is technically easier than multiple gene measurements. However, regardless of their extensive clinical utility, single-gene biomarkers not always have mechanistic links to the fundamental cellular processes responsible for disease progression or therapeutic response. Such processes are better understood as pathological alterations in the normal operation of functional modules caused by different combinations of gene perturbations rather than by alterations of a unique gene [36]. Of particular interest are signalling pathways, a type of functional module known to play a key role in cancer origin and progression, as well as in other diseases. Consequently, analysis of the activity of signalling pathways should provide a more informative insight into cellular function. Actually, the recent demonstration that the inferred activity of the c-Jun N-terminal kinase pathway shows a significantly higher association with neuroblastoma patients' mortality than the activity of their genes (including MICN, the conventional neuroblastoma biomarker) [37] that constitutes a neat confirmation of this concept. In a similar example, drug sensitivity is shown to be better predicted using probabilistic signalling pathway models than directly using gene activity values [38]. However, conventional methods for pathway analysis, even the most sophisticated ones are based on pathway topology and can only detect the existence of a significant level of gene activity within the pathway [39]. On the other hand, these methods ignore the obvious fact that many pathways are multifunctional and often trigger opposite functions (e.g. Figure 2.1; depending the receptor and the effector proteins involved in the transduction of the signal, the apoptosis pathway may trigger survival or cell death). Thus, the independent analysis of each mechanism in a pathway is becoming increasingly important. Moreover, whether the level of gene activity detected by conventional methods actually triggers cell functionalities or not and, if so, what genes are the ultimate responsible for the resulting cell activity is something that must be determined a posteriori, usually by heuristic methods. Thus, pathway activity analysis emerges as an alternative way of defining a new class of mechanistic biomarkers, whose activity is related to the molecular mechanisms that account for disease progression or drug response. However, capturing the aspects of the activity of the pathway that is really related to cell function is not trivial. This requires an appropriate description of the elementary sub-pathways and an adequate computation of the individual contributions of gene activities to the actual activity of the sub-pathway.

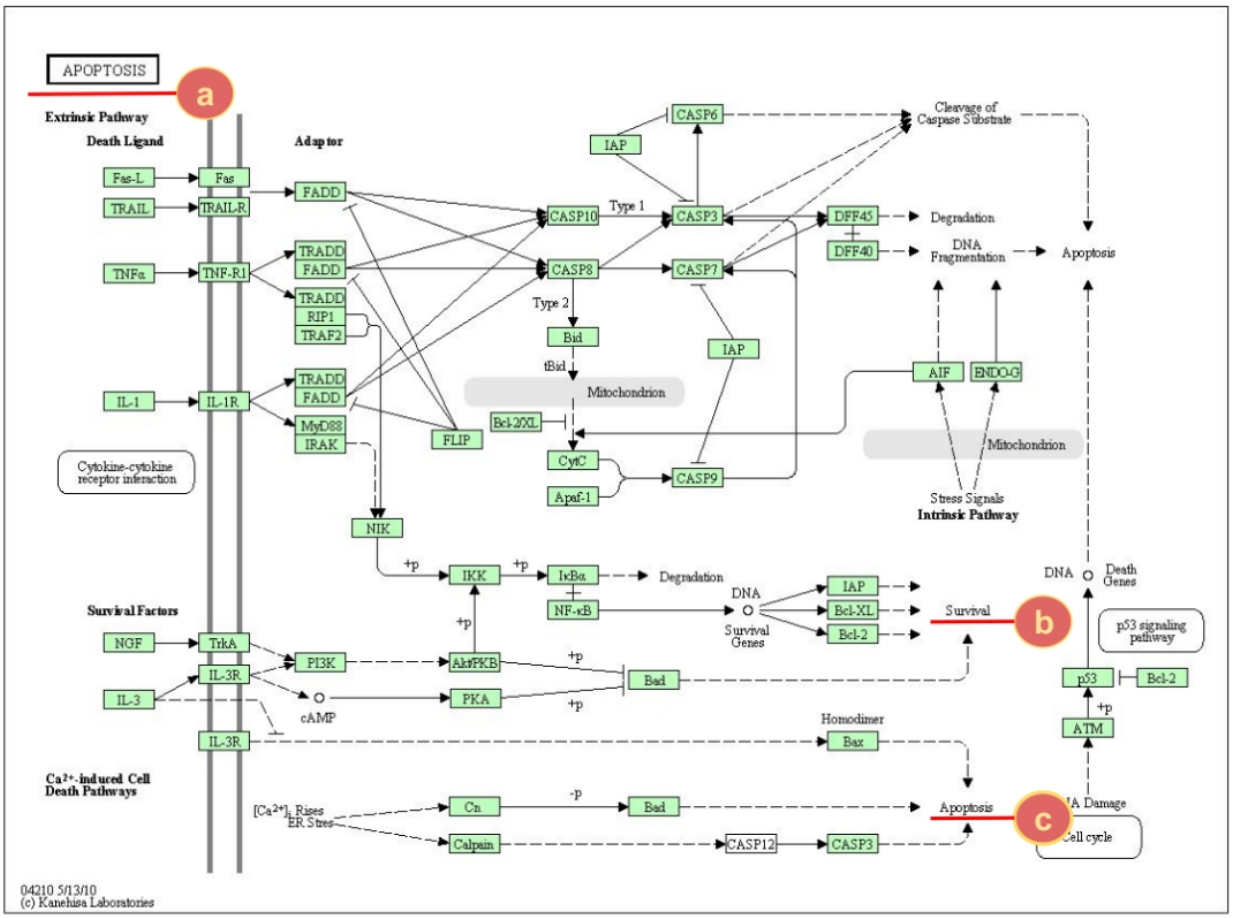


Figure 2.1: An example of multifunctional pathway with opposite functions. Human apoptosis pathway (hsa04010, 2010) taken from KEGG. Apoptosis pathway (a) can trigger opposite functions (b and c, survival and apoptosis, respectively) through different circuits and effector proteins.

Different approaches of computing activity scores for diverse sub-pathway definitions using gene expression values [16, 40–42], or even gene mutations [43], have been proposed by others (see Table 2.1). However, in most of them, the sub-pathway definition has not a functional annotation. Because it is either disconnected or only collaterally related, to the functional consequences of pathway activity. This chapter introduces a new method, which is called HiPathia, to estimate the activity within a pathway that uses biological knowledge of cell signalling to recode individual gene expression values (and/or gene mutations) into the activity scores of the pathway that ultimately account for cell functionalities at the same ranges of these scores. Specifically, it estimates the level of activity of stimulus-response sub-pathways (also signalling circuits thereinafter) within signalling pathways, which trigger cell responses (e.g. proliferation, cell death, etc.). The activity values of these canonical circuits connected to the activation (or deactivation) of cell functionalities can be considered to be multigenic mechanistic biomarkers that can easily be related to phenotypes and provide direct clues to understand disease mechanisms and drug mechanisms of action (MoA). Therefore, HiPathia is designated as a novel Mechanistic Pathway Activity (MPA) tool for cell signalling concept. Users of this tool can reveal the dynamic cellular functions of the phenotype of interest by using only their omics data. HiPathia contains pre-computed signalling circuits that are ready to use the proposed network propagation algorithm. It is implemented as a web server and R package, available on http://hipathia.babelomics.org and https://bioconductor.org/packages/hipathia, respectively.

Method	Date	Code	Pathway modelled	Circuit definition	Scoring method	Activation / Inhibition
HiPathia	2017	Web application; R package	KEGG	Receptor-to-effector circuits	Propagation algorithm	Yes
TAPPA	2007	ToPASeq R package	KEGG	All possible circuits	Scores of co-expression that explain the compared conditions	No
PWEA	2010	ToPASeq R package	User-defined pathways	All possible circuits	Mutual influence among gene expression within the circuit	No
CLIPPER	2013	Web application; ToPASeq R package	KEGG; Reactome	All possible circuits	Weighted sum of GE	No
PRS	2012	ToPASeq R package	KEGG	Trees of associated DE genes	Topologically weighted sum of DE	No
DEGraph	2012	ToPASeq R package	KEGG; User-defined pathways	All possible circuits	Multivariate two-sample tests of means of DE genes within a subgraph	No
DEAP	2013	Python code	KEGG	Receptor-to-effector linear circuits	Running sum of discretized DE	Yes
SubSPIA	2015	R code	KEGG	Minimal spanning trees (MST)	DE genes used to define the MST	Yes
MinePath	2016	Web application	KEGG	All possible circuits	Discretized GE values with logical operators	Yes

GE = Gene expression; DE = Differentially expressed

Table 2.1: List of mechanistic pathway activity methods compared in this chapter.

2.2 Materials and methods

2.2.1 Modelling strategy of the pathways

The concept of pathway activity first requires a neat description of the relationships between proteins within the pathway, which can be taken from different pathway repositories [44]. The well known canonical pathways, such as the ones used in this study are quite similar between public repositories. Here, KEGG signalling pathway definitions [45] are used, but any other repository could be used instead, as Reactome [5] or others. The pathway activity concept also requires a way to estimate the activation status of each protein, which accounts for the intensity of the signal that they can transmit along the pathway. A total of 60 KEGG pathways (Table 6 in Hidalgo et al. 2017, the short link: bit.ly/2k5Lhb0), which include the main KEGG categories related to signalling, such as; signal transduction pathways, signalling molecules and interaction pathways, cell growth and death, cell communication, endocrine system and immune system, as well as some other related pathways are used in this modelling framework. This selection of pathways includes a total of 2212 gene products that participate in 3379 protein nodes. It must be noted that any gene product can participate in more than one node (even in different pathways) and a node can contain more than one gene product.

The pathways are directed networks in which nodes (composed by one or more proteins) relate to each other by edges. In KEGG pathways, the relation edges are consist of different types of protein interactions (mainly the post-translational modifications) that include phosphorylation, dephosphorylation, ubiquitination, deubiquitination, glycosylation, methylation, etc. Generally, each interaction includes a specific label to indicate the influence type of modification. These post-translational related influence types are *inhibition* and *activation* (see Figure 2.2). Another edge type is the *Association* which has two subcategories as *groups* and *binding-associations*. These kinds of edges represent the necessary cooperation of different genes in order to make the transmission of the signal possible. Thus, rather than computing the signal propagation, they are used for the reconstruction of nodes. To this end, the pathway topology is remodelled in such a way that these edges are removed and new nodes representing the new combinations are introduced. These new nodes are called complex nodes, in contrast to the original nodes which are called plain nodes. Complex nodes are created in two ways [46]:

- ***Binding-associations:*** are defined by undirected edges between the nodes of KEGG pathways to describe the transient associations. Thus, the original plain nodes that let the signal reaches to the effector node are preserved but also a new complex node which represents the association is created. Figure 2.2 (A) illustrates this modelling.

- ***Groups:*** are defined as separated entities that are consisting of different combinations of already defined nodes. Groups have their own interactions defined in the KGML files, and therefore a complex node with the specified interactions is created based on the information given in the pathway files. Figure 2.2 (B) illustrates this modelling.

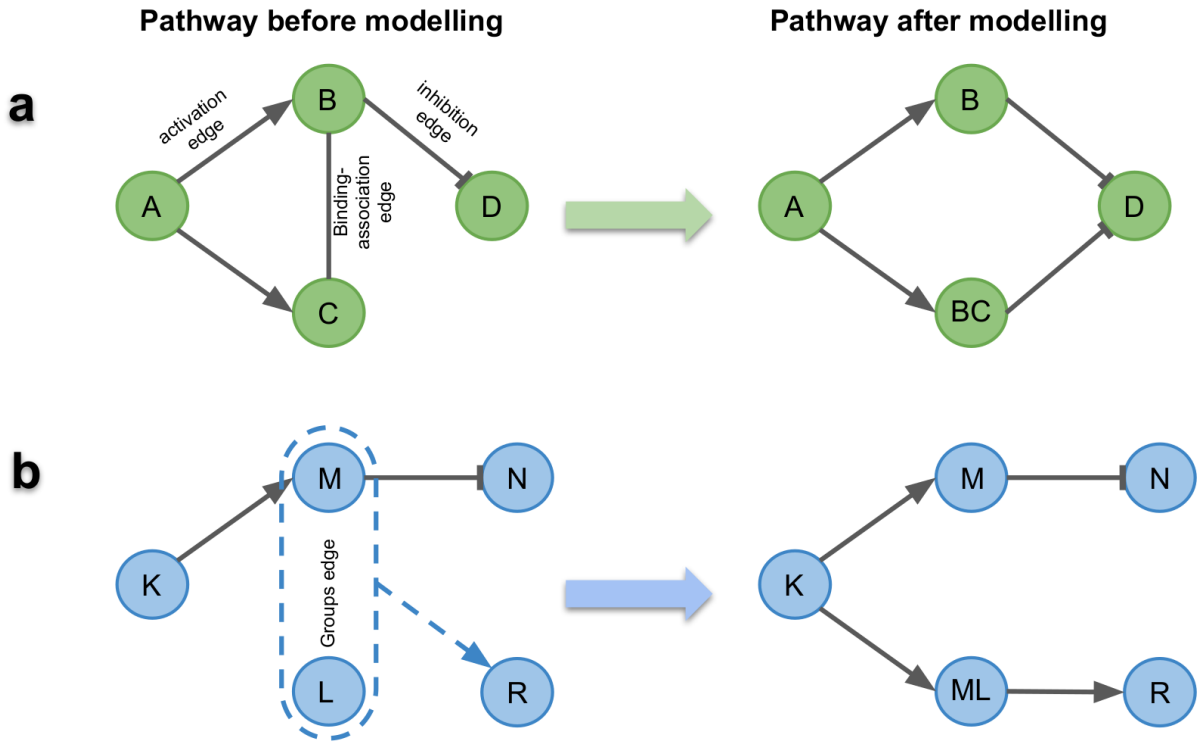


Figure 2.2: Pathway (node end edge) modelling framework. a) Node A activates node B, B inhibits node D and node C can be functional if it binds to B. Thus, a new complex node by the association rule of binding-associations, node BC, is introduced and now node A also activates node BC and BC inhibits node D. b) When ML complex node is constructed by the association rule of groups, it can be activated by node K and then its active form can activate node R.

In order to transmit the signal along the pathway, a protein needs: first, to be present and functional, and second, to be activated/inhibited by other protein(s). Preferably, the amount and activity of proteins should be quantified by proteomic, phosphoproteomic and chemoproteomic experiments [47], however, the production of these types of data is still relatively complex [48]. Instead, the extensively used systems biology approaches are taking the presence of the mRNA (gene expression) corresponding to the protein as a proxy for the presence of the protein [16, 40–42, 48, 49]. Therefore, the presence of the mRNAs corresponding to the proteins present in the pathway is quantified as a normalized gene expression value between 0 and 1. Second, a value of signal intensity transmitted through a protein is computed as taking into account; the level of expression of the corresponding mRNA and the intensity of the signal arriving in it. The net value of the signal transmitted across the pathway corresponds to the signal values transmitted by the last proteins of the pathway that ultimately trigger the cell functions. Figure 2.3 illustrates a toy model

for the quantification of the protein nodes and a circuit with the signal intensity transmitted through these nodes.

2.2.2 Decomposing pathways into circuits

Pathways are represented by directed graphs, which connect input (receptor) nodes to output (effector) nodes. The signal arrives in a receptor node and is transmitted along the pathway following the direction of the interactions until it reaches an output node that triggers an action within the cell. Thus, from different receptor nodes, the signal may follow different routes along the pathway to reach different output nodes. Within this modelling context, a canonical circuit is defined as any possible route the signal can traverse to be transmitted from a particular receptor to a specific output node (see Figure 2.4, left). The output node is the ultimate responsible elements of the circuit to carry out the signal propagation and is in charge of functional events in the cell. Then, from a functional viewpoint, an effector circuit can be defined as a higher-level signalling entity composed by the collection of all the canonical circuits ending in a unique output (effector) node (see Figure 2.4, centre). When applied to effector circuits, the method returns the joint intensity of the signal arriving at the corresponding effector node. A total of 6101 canonical circuits and 1038 effector circuits can be defined in the 60 pathways modelled.

2.2.3 Estimating the value of protein node activation

The methodology proposed, as proposed herein, accepts a matrix of gene expression values (m_{genes} x $n_{samples}$) as the input data, which are the proxies of protein amounts, and consequently, of potential protein activation values [16, 40–42, 48, 49]. The matrix is first normalized and then transformed into the node values matrix (k_{nodes} x $n_{samples}$ where $m \neq k$ and generally $k > m$ because one gene can participate in various nodes) using the information of node composition taken from KEGG database. This transformed matrix includes the normalized values of inferred activity of each node for each sample. The expression matrix is normalized into the values between 0 – 1 interval by subtracting the minimum and dividing by the maximum value of each gene.

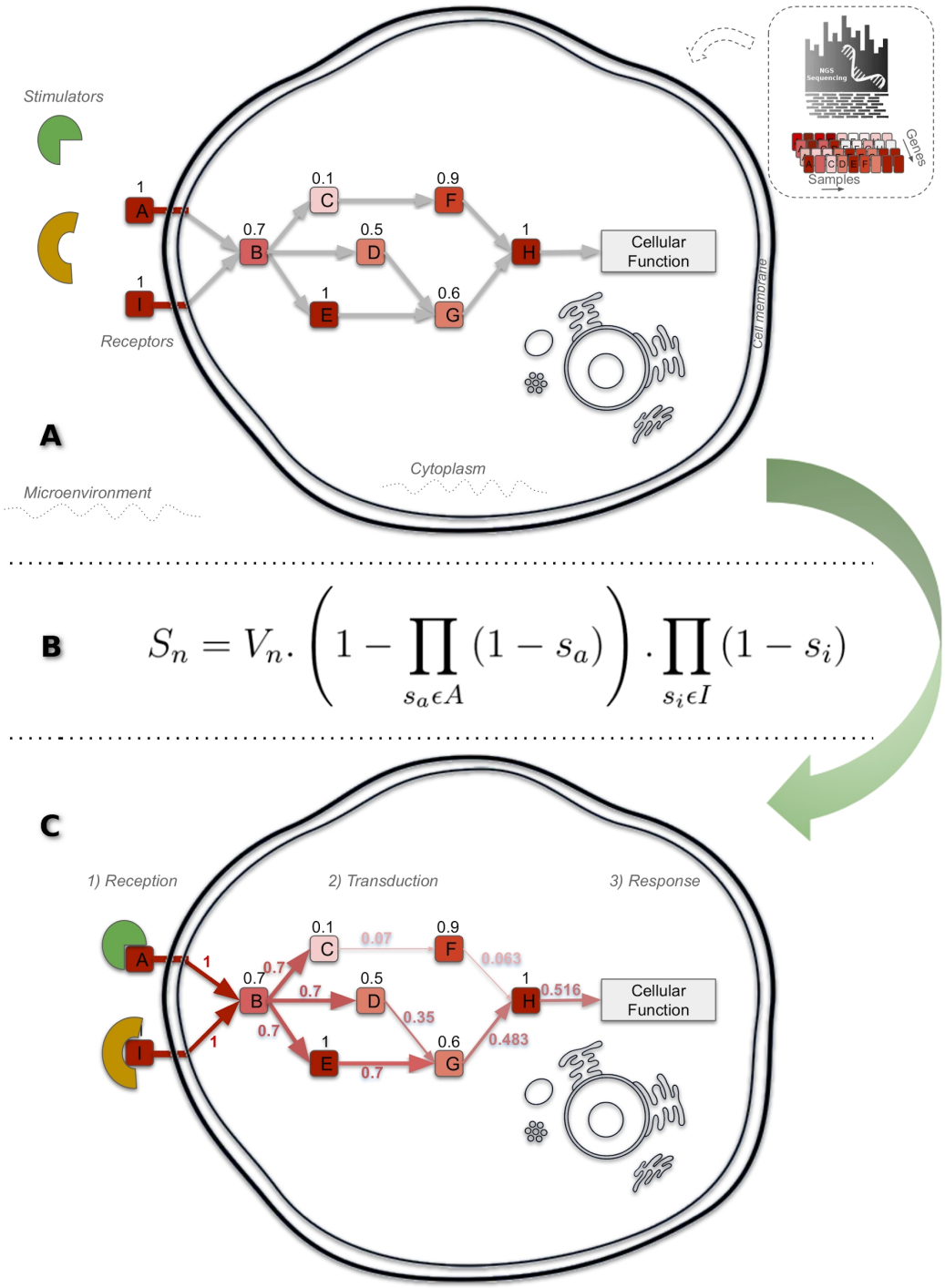


Figure 2.3: A toy model of a circuit with the propagation of cellular signal. A) Gene expression data used as proxies of protein node values. B) Equation of the iterative algorithm that is used at each node for the computation of signal intensity. C) Illustration of the signal propagation. The signal is transmitted across the signal transduction circuit in three main steps; reception, transduction and response. The signal leaving the effector node, node H, is considered as the quantity of the cellular function.

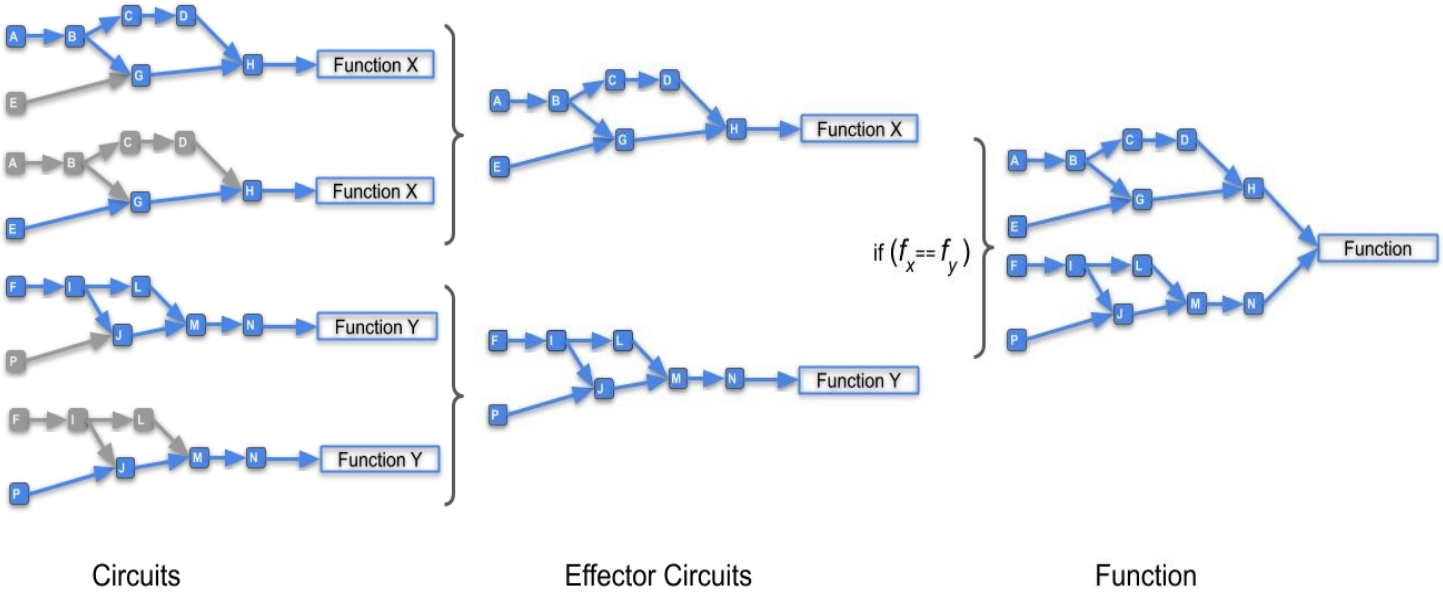


Figure 2.4: Schema that illustrates the relationship between circuits, effector circuits and functions. Left: signalling circuits, which are canonical sub-pathways that transmit signals from a unique receptor to a unique effector node. Center: effector circuits that represent the combined activity of all the signals that converge into a unique effector node. Right: functional activity that represents the combined effect of the signal received by all the effectors that trigger a particular cell function or similar functions (see also section 2.2.5).

As it is mentioned above, in the modelling strategy section, there are two different kinds of nodes: plain and complex nodes. The normalized value of each node is computed according to the type of node;

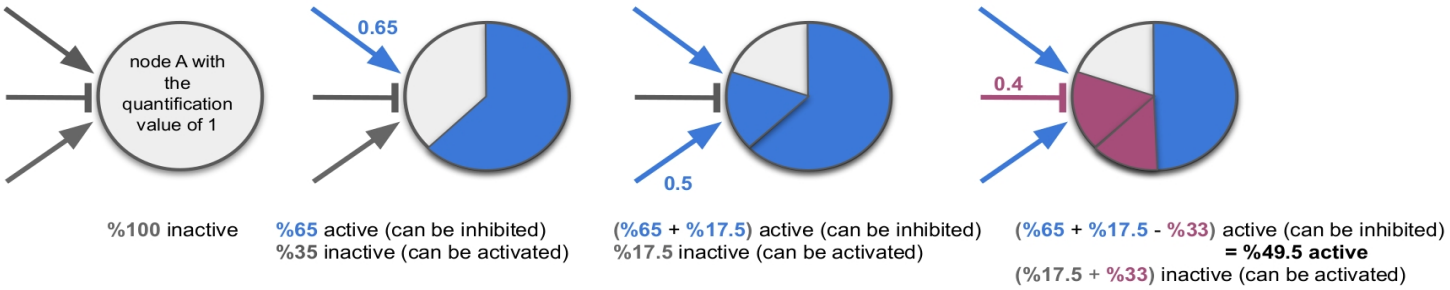
- **Plain nodes:** They represent an original KEGG node, which may contain one or several proteins. To compute the plain node value, a summarization of the values of the proteins included in the node is applied, generally taking the percentile 90. If no protein is associated with the node, it takes the value of 0.5 (imputing the missing values with a default value) in order to not interfere in the signal propagation.

- **Complex nodes:** They include more than one node, as a result of modelling of the groups or binding-association relations. To compute their node value, the value of each included plain node is computed separately, and then the minimum value among them is taken. In this way, it captures the idea that all the proteins in the complex are necessary to transmit the signal through the complex node.

Since the matrix of gene expression values is normalized between 0 and 1, node values will be in the same interval.

2.2.4 Computing the circuit activity

The computation of the signal intensity for a circuit is performed by means of an iterative algorithm starting from the receptor nodes of this circuit. In order to initialize the circuit signal, it is assumed that a signal value of 1 that stimulates the receptor nodes of any circuit always present (see Figure 2.3, A). This assumption substitutes for barely measured molecular signals (oxidative stress, hormone level, metabolite concentration, etc.) that are captured from the microenvironment of a cell by its receptors. Then, for each node n of the network, the signal value is propagated along the nodes according to the following rule (see Figure 2.3, B for the equation): where S_n is the signal intensity for the current node n, v_n is its normalized value, A is the total number of activator signals (S_a), arriving in the current node from activation edges, I is the total number of inhibitory signals (S_i) arriving in the node from inhibition edges. The algorithm to compute the transmission of the signal along the network is a recursive method and conceptually quite similar to sum-product message-passing algorithm for directed trees [50]. By this algorithm, at each node, the signal is being updated and the recursive visiting of the graph nodes continues until the difference between the previous and updated signal values becomes smaller than the given threshold. This threshold is a value very close to zero (default=10^{-6}) and allows for more precise estimation of circuit activities by converging the loops. It is also a decision criterion to terminate the recursive visits on a graph. Many MPA methods simply cannot handle loops and artificially disconnect them or even remove them from the calculations [40–42, 46, 51–53]. Figure 2.3 and Figure 2.5 represent the computation of intensity of signal transmission in a node and also demonstrates a simple scenario of how the signal is transmitted across a circuit.



$$S_n = v_n \cdot (1 - \prod_{s_i \in A}(1 - s_i)) \cdot \prod_{s_j \in I}(1 - s_j)$$

$$S_n = 1.\,(1 - ((1 - 0.65)\,.(1 - 0.5)).\,(1 - 0.4)$$

$$S_n = 0.495$$

Figure 2.5: Illustration of the signal propagation over the protein node. The figure illustrates a protein node which receives two activating and one inhibiting signals with the strength of 0.65, 0.5 and 0.4, respectively. In this example, to make the calculations more apparent, the node value is given as 1. In this case, the 0 value means no expression and 1 indicates the highest amount of the expression among the samples for the given node. We assume that the protein node is %100 inactive until the first activating signal is received. Activating signals can activate the inactive parts and inhibiting signals can inhibit the active parts, and vice versa when the first signal is inhibition. When a protein node receives activating and inhibiting signals, the order of the calculation does not matter while the main mathematical operation between activation and inhibition is multiplication. Finally, the signal intensity of this node is 0.495.

2.2.5 Circuits for functional analysis

Effector proteins that are located at the end of each circuit trigger specific cellular functions in the cell. The magnitude of cellular functions is correlated with the strength of the signal received by these proteins. Thus, once the circuit activities are calculated, from the matrix of circuits a matrix of functions can be inferred, which contains an intensity value of the signal for each molecular function and for each sample. Different effector circuits of different pathways may end in the same effector protein, thus they trigger the same cellular function. Also, different effector nodes may trigger the same function. Therefore, the signal intensity, S_f, received by a particular function ***f***, is summarized from the intensity signal values of all the circuits ending in an effector node related to ***f*** by taking the mean of all signal intensity values related to the function ***f***. Figure 2.4 illustrates how effector circuits are composed of different signalling circuits and how functions can be triggered by several effector circuits. Since the KGML files do not contain the functional annotations of effector proteins that are shown on KEGG pathway layouts, the external databases

were used for the annotation of effectors to make the process automatized and to minimize manual data entry by the users. In this study, Gene Ontology (GO) [12] terms corresponding to the biological process ontology (http://geneontology.org/docs/download-ontology/, February 16, 2016 release) and molecular function keywords of Uniprot [54] (http://www.uniprot.org/help/keywords, the release of September 21, 2015) are used to define the functions triggered by the effector proteins. Different functional annotations can be also used for functional analysis. Here, we selected these two external annotations, because they were covering the high percentage of the effector protein annotations.

2.2.6 Case examples for the application of HiPathia

HiPathia was successfully applied in different research scenarios which include the analysis of functions triggered in cancer cells and their predictive power in patient survival [55–58], the discovery of the cellular processes triggered by death and during the post-mortem interval [59], and the explanation of molecular mechanisms behind the obesity [60]. In all the studies cited in this section, the PhD student of this dissertation was listed as a co-author. These reference studies contain a detailed functional interpretation of the results obtained by the method proposed. I believe that the utility of the activity of signalling circuits as highly reliable mechanistic biomarkers was demonstrated well enough within these references. On the other hand, similar applications, analyses and functional interpretations will be discussed in the following chapters in the context of metabolic circuits (modules). Thus, in this chapter, I will only focus on the benchmarking of the proposed method with other similar methods.

2.2.7 Comparison with other available methods for defining and scoring circuit activity

Although HiPathia is a unique method, it was needed to check the accuracy of this method by benchmarking it with other proposals that define circuits and calculate their activity scores. The nine methods listed in Table 2.1 were satisfying three basic conditions needed for a fair comparison: they can be applied to RNA-Seq data, they have a common definition of the pathway (KEGG pathways constituted the unique common pathway definition) and there is software available for running them. Since circuit definition and scoring methods used will potentially have an impact on the relative performances of the methods, a comparative benchmark has been carried out to study their relative performance. The relative performance of the methods compared was derived from the estimation of their ratio of true and false negatives.

2.2.8 Sensitivity and specificity of the methods

To assess the sensitivity (true positive rate - TPR) of the methods compared, the 12 cancer types listed in Table 2.1 were used. In any of the 12 normal versus tumour comparison (e.g. BRCA Tumor vs BRCA Normal), it was expected to detect the changes of activity in the 14 KEGG pathways (Table 2.3) that belong to the cancer pathways category. As different methods have different circuit definitions, the comparison of the methods was carried out at the level of the whole pathway definition rather than the numbers of significant circuits. This means that a pathway was considered as altered when at least one circuit in this pathway was found significantly activated by means of a Wilcoxon test with Bonferroni correction. For each method, TPR was estimated as the number of altered cancer pathways (containing one or more differentially activated circuits) divided by the total number of cancer pathways given in Table 2.3.

To estimate the specificity (false positive rate - FPR) of the methods compared, two groups that composed by identical individuals of the same cancer types for the same cancer pathways were used. As the individuals compared belong to the same phenotype of the same cancer type (e.g. BRCA Tumor vs BRCA Tumor), any differentially activated circuit would be a false positive detection. For each comparison, randomly selected two data sets of 25 tumour samples were used. Comparisons between both data sets were repeated 100 times per method and cancer type. For each of these repetitions, the sampling was done independently to ensure the sample variation between the iterations. Similarly to the case of TPR, the FPR of any method is calculated as the number of cancer KEGG pathways in which the method finds one or more circuits significantly activated, then divided by the total number of pathways analyzed (14 for HiPathia and DEAP and 13 for the rest of methods, because PPAR signalling pathway [hsa03320] was not implemented in them). Again, Wilcoxon test with Bonferroni correction was used to assess significantly activated circuits.

Finally, TPR and FPR distributions of each method were compared using the same statistical assessment as mentioned above, to detect significant differences among them.

TCGA Identifier	Cancer	Primary tumour	Normal adjacent tissue
BLCA	Bladder Urothelial Carcinoma	301	17
BRCA	Breast invasive carcinoma	1057	113
COAD	Colon adenocarcinoma	451	41
HNSC	Head and Neck squamous cell carcinoma	480	42
KIRC	Kidney renal clear cell carcinoma	526	72
KIRP	Kidney renal papillary cell carcinoma	222	32
LIHC	Liver hepatocellular carcinoma	294	48
LUAD	Lung adenocarcinoma	486	55
LUSC	Lung squamous cell carcinoma	428	45
PRAD	Prostate adenocarcinoma	379	52
THCA	Thyroid carcinoma	500	58
UCEC	Uterine Corpus Endometrial Carcinoma	516	23

Table 2.2: Cancers types used in this chapter with the number of samples sequenced of both tumour biopsy and normal adjacent tissue.

KEGG ID	Pathway name
hsa04010	MAPK signalling pathway
hsa04310	Wnt signalling pathway
hsa04350	TGF-beta signalling pathway
hsa04370	VEGF signalling pathway
hsa04630	Jak-STAT signalling pathway
hsa04024	cAMP signalling pathway
hsa04151	PI3K-Akt signalling pathway
hsa04150	mTOR signalling pathway
hsa04110	Cell cycle
hsa04210	Apoptosis
hsa04115	p53 signalling pathway
hsa04510	Focal adhesion
hsa04520	Adherens junction
hsa03320	PPAR signalling pathway

Table 2.3: Fourteen KEGG pathways belonging to the subcategory of ‘Pathways in cancer’ were used to detect changes when cancers versus control comparisons were done.

2.2.8 Data source and processing

A large The Cancer Genome Atlas (TCGA) data set of RNA-Seq counts for 12 cancer types analyzed was used for the benchmark provided in this chapter. Only the cancer types in which the RNA-Seq counts for healthy control and cancer samples were available by the time of this analysis were used for the benchmark analysis (Table 2.2). The data were downloaded from the ICGC data portal (https://dcc.icgc.org/releases/release_20/Projects) and preprocessed as given below.

Since TCGA cancer data has different origins and underwent different management processes, non-biological experimental variations (batch effect) associated with Genome Characterization Center (GCC) and plate ID must be removed from the RNA-Seq data. The COMBAT method [61] was used to remove the batch effect. Then, we applied the Trimmed Mean of M-values normalization method (TMM) method [62] for data normalization. TMM is a very efficient normalization method that corrects a well-known artefact derived from RNA-Seq technology: the RNA-composition bias. When comparing two different samples, the number of read counts of an equally expressed gene may vary depending on the level of expression of the other genes due to the fact that the library depth is fixed. The read counts of a gene represent the proportion of the gene with respect to the total RNA production of the sample, but this proportion is not a quantitative number which can be compared if the total RNA production is different between samples. TMM normalization estimates the ratio of RNA production between samples with a weighted trimmed mean of the log expression ratios (trimmed mean of M values or TMM). Then it uses this estimation to modify the observed library size of a sample to a comparable library size which follows the proportion of RNA production between the samples. The resulting normalized values were used as input of the circuit activity methods.

2.3 Results

2.3.1 Estimation of the sensitivity and specificity of the MPA methods

As it is explained in the section of materials and methods, the tumour samples of each cancer type were compared with their corresponding healthy tissue samples and within this strategy, the expectation was to find TPRs through the number of significant cancer-associated KEGG pathways. In Figure 2.6a, the violin plots show for any method the mean TPR in the central line. This figure shows how only HiPathia was able to detect the changes in the activity of circuits that are belonging to all the cancer pathways analyzed across the 12 cancer types. Other groups of three methods (TAPPA, DEGraph and subSPIA), with a significantly different performance (P-value $= 3 \times 10^{-4}$), was able to detect between 50% and 75% of the alterations in the cancer pathways used here. The rest of the methods could detect differential activity in less

than 50% of the cancer pathways.

To check whether the high sensitivity of HiPathia, TAPPA, DEGraph and subSPIA is real or is only the consequence of low specificity, we calculated their FPRs. To achieve this, data sets of identical samples were compared and significant differences in circuit activity found by a particular method in the comparisons are considered as false positives. A total of 10,800 comparisons (100 times × 9 methods × 12 cancer types) between pairs of data sets of 25 samples each, randomly sampled among cancer samples, were performed. The FPR was computed as the mean of the number of significant cancer pathways divided by the total number of cancer pathways per method and cancer. Figure 2.6b shows how most of the methods have a low FPR, except PWEA, which displays a high ratio of false positives (over 30%). Note that this figure, on its y-axis, displays 1-FPR for better visualization purpose. The best performer is SubSPIA (P-value = 0.006), which, together with CLIPPER and HiPathia methods, showed the highest specificity (P-value = 0.001). Here the p-values were obtained by the statistical comparison of FPRs of methods or cluster of methods.

2.4 Discussion

Since the early 2000s, different functional profiling methods have been proposed for the interpretation of the omics data, such as over-representation methods (ORA) or gene-set enrichment methods (GSEA) [63–66], that focus on the collective activity of genes within biologically relevant entities such as pathways. However, given that most pathways have multifunctional entities and these methods, even in their most sophisticated versions do not consider the internal structure of the pathway. They are simply illustrative and fail to provide real mechanistic information on the specific outcomes of a cell. Especially nowadays, the high throughput transcriptomic experiments are affordable and provide a wealth of precise data that must be interpreted in light of their biological consequences and implications. On the other hand, this increasing availability of the omics data lets us design detailed and well-curated pathways. By this means, it also opens a new door to the development of MPA models which are essential to make the functional interpretations as much as close to the real. Models of MPA fill the gap between the conventional approaches based on single-gene biomarkers, or functional enrichment methods, and the more realistic, model-based approaches. These methods use the biological knowledge available on the relevant biological modules (such as signalling pathways) to explain how their perturbations ultimately cause diseases or responses to treatments. Therefore, MPA methods provide a link between gene perturbations (measured as gene expression changes) and disease mechanisms or drug MoAs [67, 68].

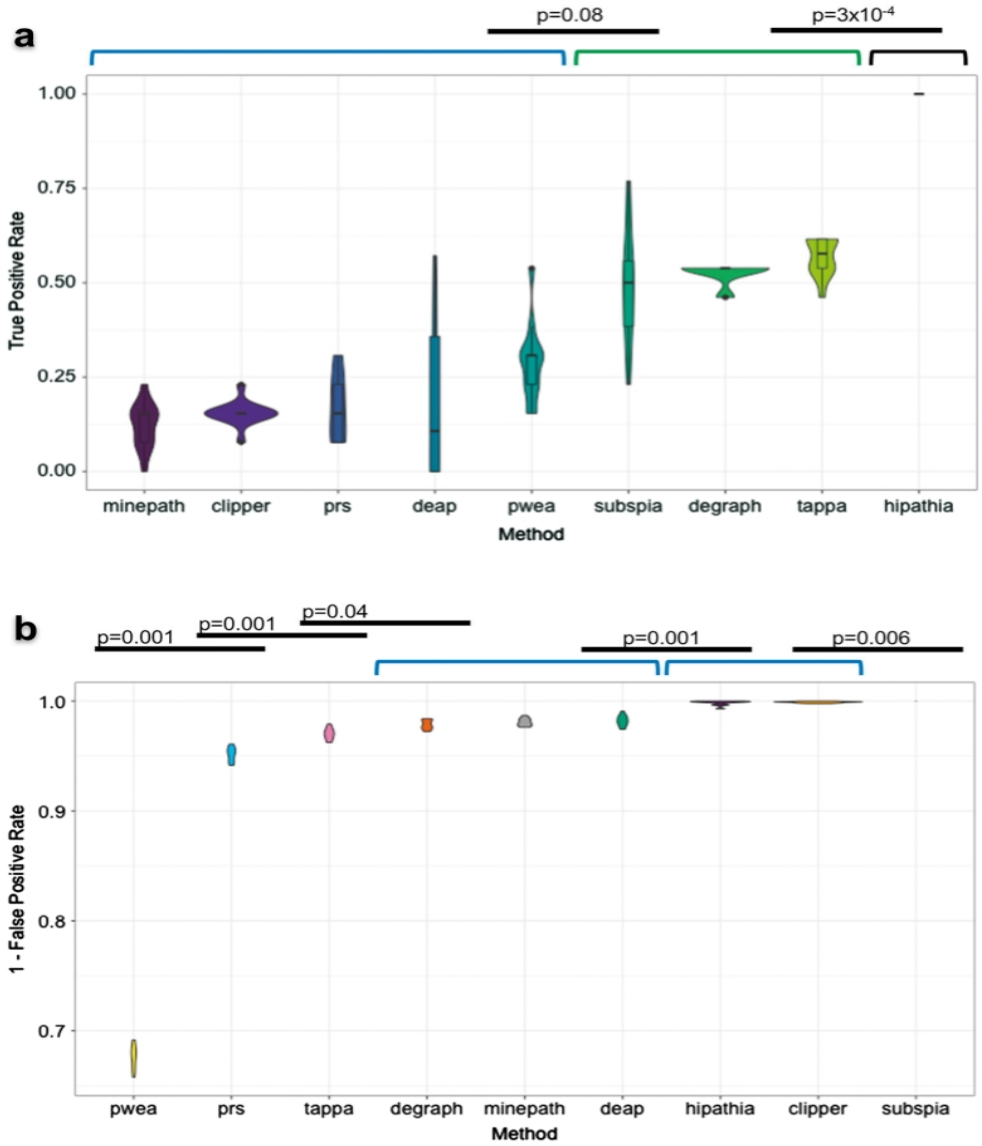


Figure 2.6: Simultaneous comparison of sensitivities and specificities of the different MPA methods. A Wilcoxon test with Bonferroni correction was used to compare successive FPR or TPR distributions to detect significant differences among them. Black lines with the p-values denote significant differences between consecutive methods. Brackets define groups of methods with no significant differences in their performances. A) TPR or sensitivity was computed as the number of significant cancer pathways found, when cancer samples are compared with samples of the tissue of reference, divided by the total number of cancer pathways per method and cancer. Violin plots obtained using 12 cancer types show for any method the mean TPR in the central line. The figure shows the methods ranked by TPR value. B) FPR or specificity was computed as the mean of the number of significant cancer pathways found, when cancer samples are compared with cancer samples, divided by the total number of KEGG cancer pathways along 100 bootstraps (see section 2.2.8), per method and cancer. Violin plots show average values and distributions of FPR for each method. The figure shows the methods ranked by FPR value and the y-axis represents 1 - FPR.

To evaluate the performance of the different pathway definition and scoring strategies used by the different MPA methods to capture biological information that accounts for cell behaviour (such as signal transduction circuit activities) and relate them to phenotypic conditions, we have produced a benchmarking of nine MPA methods that could be compared in similar conditions. Both together, Figure 2.6 offers a summarized view of both specificity and sensitivity of the methods analyzed in this chapter. Whereas most of the MPA methods show an excellent specificity (except PWEA), the differences in terms of sensitivity are more pronounced. They could be clustered in four groups according to their relative sensitivity and specificity. The first group of highly sensitive and specific methods would include HiPathia, the only one able to detect differences in the activities of all the cancer pathways across all the cancer types analyzed while maintaining a high specificity as well. The second group of methods with medium sensitivity but still high specificity, which includes TAPPA, DEGraph and SubSPIA, detects changes in only 75%–50% of the cases (P-value = 0.08). The third group of methods, with low sensitivity but still high specificity, which includes CLIPPER, DEAP, MinePath and PRS, shows poorer performance, detecting changes in circuit activities in less than 25% of the cases. Finally, the conceptually most different method, PWEA, does not only present a low specificity but also a low sensitivity.

It is difficult to attribute the relative performance of the methods to a unique factor and it rather seems to be a combination of several of them. Apparently, the use of a receptor-to-effector definition of the circuit and the distinction between activations and deactivations seem to be important factors that differentiate HiPathia from the rest of methods in terms of sensitivity. MinePath and DEAP also use activation/inhibition information to calculate the score and DEAP uses receptor-to-effector circuit definitions, but in both cases, the scoring algorithm uses discretized values of differential gene expression, which seem to reduce drastically the sensitivity. The most representative feature of the second group of methods seems to be the use of differential gene expression or co-expression to obtain scores for circuits. These circuits that can effectively separate the conditions compared are chosen as differentially activated. In the third group, showing poorer sensitivity (below 25%), the discretization of differential gene expression values seems to represent a hurdle for obtaining a better sensitivity for two of the methods (DEAP and MinePath). The case of CLIPPER and PRS is probably related to a combination between the scoring strategy and the circuit definition. Finally, the PWEA presents, in addition, a low specificity. Probably, the PWEA case is a combination of the circuit definition and a scoring algorithm, based on mutual influence among genes, which is not capturing properly the underlying biology of the pathway activity. Moreover, all the methods in their original publications demonstrated to be more sensitive than the conventional ORA and GSEA methods [40–42, 52, 53, 57, 69–71].

Receptor-to-effector subpathways are relevant circuit definitions from a biological standpoint, as they represent the possible routes taken from the beginning of a pathway, where the signal is originated, to its end, where a function is triggered. Within the context of signalling pathways, such circuit definitions effectively model signal transduction events. MPA methods

implementing these circuit definitions model more realistically biological events and consequently produce better results. In addition, an interesting feature of the methods that use receptor-to-effector circuits is that the changes in the activity of such circuits can be easily associated with cell functionalities triggered by the effector protein [57]. Contrarily, given the fact that many genes and subnetworks can be shared by several pathways, pathway definitions based in subnetworks are, in principle, more prone to false positives. Because most of the methods only accept KEGG pathways as input, it is not possible to test potential biases of the different methods under different pathway definitions. As a more extensive approach, in future this benchmark can be done in the following scenario; testing each MPA algorithm with all different circuit definitions to find the best combination of circuit activity algorithm and circuit definition.

Chapter 3

Metabolizer web tool for differential metabolic activity analysis and discovery of therapeutic targets using summarized metabolic pathway models

Chapter 3 is adapted from the following publication: “**Cubuk C**, Hidalgo MR, Amadoz A, et al. (2019). Differential metabolic activity and discovery of therapeutic targets using summarized metabolic pathway models. npj Systems Biology and Applications, 5(7). DOI: 10.1038/s41540-019-0087-2”.

Chapter 3
METABOLIZER WEB TOOL

3.1 Overview and objectives

Most rare diseases have an identified genetic origin and multigenic nature. Thus, the rare diseases are often better understood as failures of functional modules caused by different combinations of perturbed gene activities rather than by the failure of a unique gene [36]. In fact, an increasing corpus of recent evidence suggests that the activity of well-defined functional modules, like pathways, provide a better prediction of complex phenotypes, such as patient survival [37, 57], drug effect [38], etc., than the activity of their constituent genes. In particular, the importance of metabolism in cancer [72] and other diseases [73] makes of metabolic pathways an essential asset to understand disease mechanisms and drug modes of action (MoA) and search for new therapeutic targets.

Genes encode proteins and proteins operate cell function. Therefore, the thousands of genes expressed in a particular cell determine what that cell can do. Gene expression changes have been used to understand pathway activity in different manners. Initially, conventional gene enrichment [74] and gene set enrichment analysis (GSEA) [64] were used to detect pathway activity from changes in gene expression profiles [75]. However, these methods provided an excessively simplistic view on the activity of complex functional modules that ignored the intricate network of relationships among their components. Other methods took advantage of network structures to gain understanding in mechanisms of action [76] using massive transcriptomic data on massive cell perturbation repositories [77]. In the previous chapter, Chapter 2, we discussed the advantages of mechanistic models over the enrichment methods for the pathway analysis. Mechanistic models focus into the elementary components of the pathways associated with functional responses of the cell [42, 57] and in this way they provide a more accurate picture of the cell activity [78]. Specifically, in the context of metabolic pathways, constraint-based models (CBM) have been applied to find the relationship between different aspects of the metabolism and the phenotype [79]. Using transcriptomic data, CBM allows the analysis of human metabolism at an unprecedented level of complexity [80, 81]. However, as many mathematical models, CBM presents some problems, such as their dependence on initial conditions or the arbitrariness of some assumptions, along with difficulties of convergence to unique solutions [79, 82]. Moreover, with limited exceptions [83],19 most of the software that implements CBM models only run in commercial platforms, such as MatLab and working with them require skills beyond the experience of experimental researchers.

Because of the highly interconnected nature of metabolic pathways, they considered as complex biological processes and less dissectable. However, metabolic modules have been defined to provide a comprehensive curated summary of the main aspects of metabolic activity and

account for the production of the main classes of metabolites (nucleotides, carbohydrates, lipids, and amino acids) [84]. This chapter presents a simple model that accounts for the activity of metabolic modules [84] taking into account the complex relationships among their components and the integrity of the chain of biochemical reactions that must occur to guarantee the transformation of simple to complex metabolites or vice versa. The likelihood of such reactions to occur is inferred from gene expression values within the context of metabolic modules. The model has been used in a pan-cancer study that has demonstrated high precision in detecting cancer vulnerabilities [85]. In order to make these models accessible and easily usable to the biomedical community, I have developed Metabolizer, an interactive and intuitive web tool for the interpretation of the consequences that changes in gene expression levels within metabolic modules can have over cell metabolite production. A case study of Metabolizer, pan-cancer analysis of metabolic modules, is given and discussed in chapter 4 rather than here. Thus, this chapter consists of the method and tool description, proof of concept, and comparison of Metabolizer to other methods.

3.2 Materials and methods

This section describes the webtool and also provides validation of the method. To increase the readability, the material and methods part was organized under two subsections as *implementation of Metabolizer web server* and *evaluating the predictive power of Metabolizer.*

3.2.1 Implementation of Metabolizer web server

Under this section, you will find the details of the method, technical details and functionalities of the web tool.

3.2.1.1 Estimating the metabolic activity of a KEGG module

Pathway modules [84] are used to depict the complex interactions among proteins carrying out the reactions that account for the main metabolic transformations in the cell. Here, a total of 95 modules were used, that comprise a total of 446 reactions and 553 genes. The comprehensive list of modules with their detailed information (module ID, main metabolic category, sub metabolic category, description/name, KGML file, pathways, KEGG link of module, start/end metabolite) can be found at the supplementary tables of Cubuk et al. Cancer Research., 2018 [85] and Cubuk et al., npj SBA, 2019 [86]. The pathway modules (Table 3.8) were downloaded through REST-style KEGG API from the KEGG MODULE (http://www.genome.jp/kegg/module.html) database in plain text format files that include information of the metabolites, genes and reactions. Metabolic pathways were downloaded from KEGG PATHWAY database in KGML format files. Then, each KEGG module is made up of reaction nodes (composed by one or several isoenzymes or

enzymatic complexes [49], which are connected by edges in a graph that describes the sequence of reactions that transforms simple metabolites into complex metabolites, or vice-versa. The potential catalytic activity level of a KEGG module can be derived from the potential catalytic activities of all the reaction nodes, assuming all the intermediate metabolites are present and available (equivalent to setting all their values to 1, assuming that all of them are present). However, if metabolite measurements are available, their normalized values can easily be integrated into this modelling framework. Under this modelling framework, the potential for the catalytic activity of a reaction node is inferred from the presence of the constituent proteins. However, given the difficulty of obtaining direct measurements of protein levels, an extensively used proxy for protein presence is the observation of the corresponding mRNA within the context of module [16, 40, 42, 46, 48, 49, 57]. Therefore, the presence of the mRNAs corresponding to the proteins present in the pathway is quantified as a normalized gene expression value between 0 and 1. To estimate the potential activity of the reaction node two scenarios are considered [87]: isoenzymes, where the activity is produced if at least one of them is present ($Expression_{Isoenzyme1}$ OR $Expression_{Isoenzyme2}$ OR …) and enzymatic complexes, where the activity occurs only if all the enzymes are present ($Expression_{EnzymeA}$ AND $Expression_{EnzymeB}$ AND …). For example, in Figure 3.1, the last reaction transforming isocitrate into 2-oxoglutarate is catalysed by either an enzymatic complex or two alternative isoenzymes, represented as “(R01899 AND R00268) OR R00267 OR R00709”, which may be estimated from the normalized gene expression values of the mRNAs corresponding to proteins R01899, R00268, R00267 and R00709 as:

$V = max\{min\{E_{R01899}, E_{R00268}\}, E_{R00267}, E_{R00709}\}$ where V is the activity of the node and E_p is the 90th percentile of normalized expression of the gene corresponding to the enzyme p. This approach is very similar to computing the Gene-Protein-Reaction (GPR) rules in metabolic networks [87]. It is worth noting that some enzymes can participate in more than one node (even in different modules), and thereby contribute to different reactions.

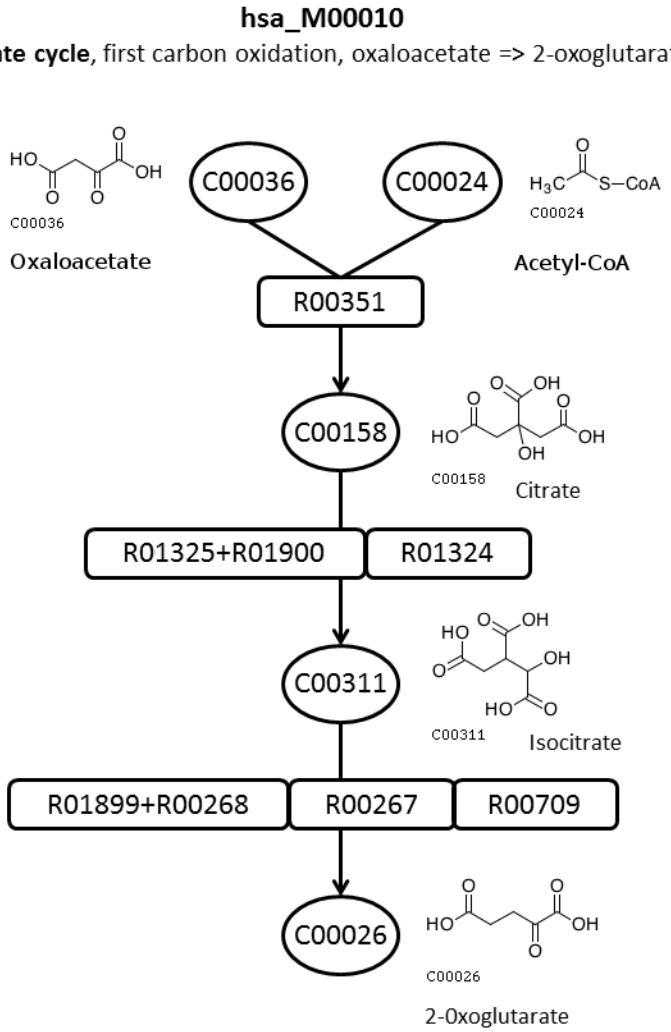


Figure 3.1: The metabolic module Citrate cycle, first carbon oxidation, oxaloacetate => 2-oxoglutarate module (M00010).

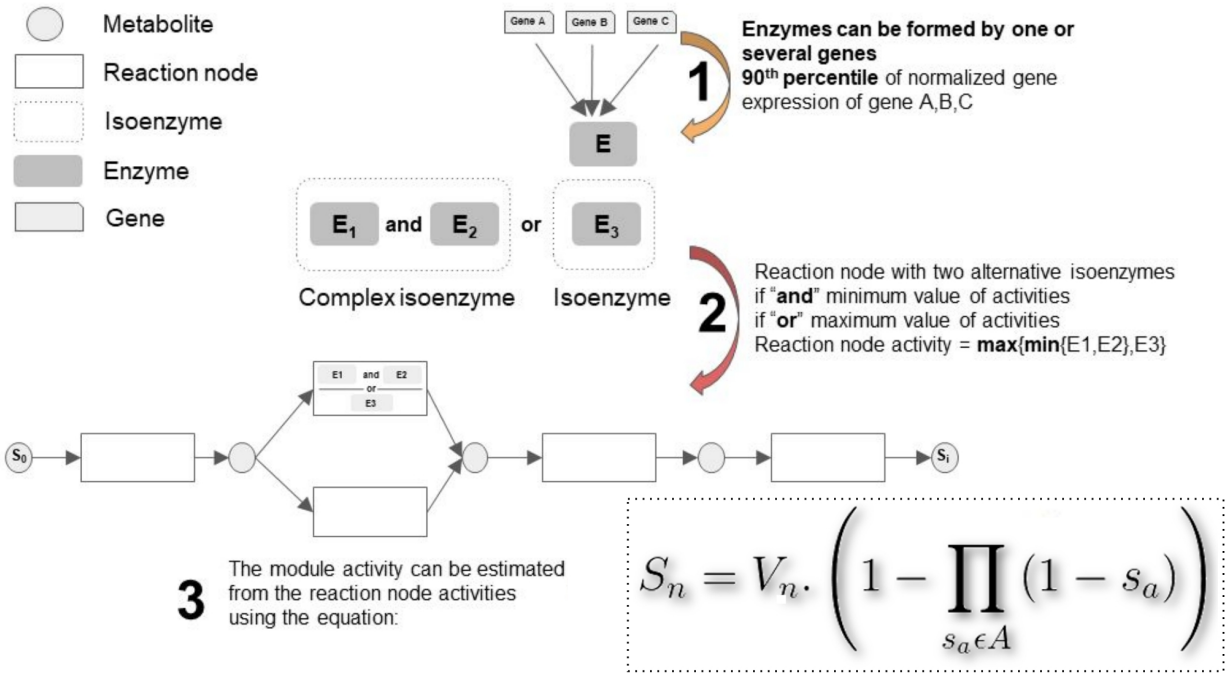


Figure 3.2: Procedure used to estimate reaction node activity and module activity from the constituent gene expression activities. 1) If an enzyme is composed for more than one gene, the enzyme activity value is obtained as the 90th percentile of normalized gene expression (to avoid possible outliers or artefacts). 2) When an isozyme is a complex, composed by more than one enzyme, the complex isozyme activity is obtained as the minimum activity value of all the involved enzymes (the limiting reaction capability). When a reaction node is composed of more than one isozyme, then, the maximum isoenzyme activity value is taken as the activity of the reaction node. 3) Finally, the module activity is inferred from the corresponding node activities.

Then, the contribution of potential catalytic activities of all individual nodes to the whole module metabolic activity can be derived by using a recursive method that sequentially traverses the module from the simpler to the more complex metabolite. Assuming a value of 1 for the initial node of the module, the potential catalytic activity of the subsequent nodes is calculated by the equation given in Figure 3.2. In the formula, S_n is the catalytic activity of the current node n, V_n is the value of catalytic node activity which is inferred from normalized gene expression values of the current node n, A_n is the set of edges arriving in the node n that, within this modelling framework, accounts for the flux of metabolites produced by the corresponding reactions in other nodes with activity values s_a.

The resulting integrity value of the whole sequence of reactions represented in the module is summarized by the value of catalytic activity propagated until the last node, which carries out the last the transformations of the chain of reactions that ultimately produces the final metabolite [85]. This method is an adapted version of the propagation algorithm on graphs successfully used to estimate cell signalling activities in cancer [57]. Here, in metabolism, there are only reactions instead of gene interactions such as activations and inhibitions that appear in signalling. Then, the formula accounts for the integrity of the chain of reactions that connect the initial to the final

metabolite. Figure 3.2 outlines the procedure. In brief, this formula is equivalent to the activation part of the given formula in chapter 2, under the concept of cell signalling. Despite the similarity of these formulas given in this chapter and chapter 2, the formulas are explained slightly different based on the differences of biological mechanisms and components that exist between the cell signalling and metabolic pathway.

3.2.1.2 Differential metabolic module activity

The significant alterations in module activities across the compared conditions can be calculated in a similar way to differentially expressed genes using linear models, t-test or Wilcoxon test. Here, Wilcoxon test is used to assess the significance of the observed changes of metabolic module activity when samples of two conditions are compared. Since many modules are simultaneously tested, multiple testing effects need to be corrected. FDR method (known as Benjamini & Hochberg procedure) [88] is used for this purpose.

3.2.1.3 Build a predictive model

The class prediction functionality includes two sub-functionalities: training process, where the predictor is built using a training set (the model is stored for future use), and classification process, where the predictor can be used for class prediction purposes. In order to build a predictor, a training set composed by samples belonging to two or more classes is required. The selection of samples that properly represent the variability of the classes is critical for the generatability of the predictor. Additionally, in order to avoid the production of suboptimal and biases models, the following rules are set; sample size of each class needs to be higher than 10 and sample sizes between different classes need to be balanced. The minimum allowed ratio between the sample sizes of the two classes is 0.66.

Two powerful prediction algorithms, Random Forest (RF) [89], as implemented in randomForest package in R (https://cran.r-project.org/web/packages/randomForest/), and support vector machines (SVM), [90] as implemented in e1071 package (https://cran.r-project.org/web/packages/e1071/), can be chosen to train the predictor. There is no specific reason behind the selection of these two supervised learning algorithms in the Metabolizer rather than their fast and easy parameter and model optimisation. These algorithms use the profiles of metabolic module activities of the two or more groups of samples compared. The accuracy obtained by the predictor is assessed by k-fold cross-validation and the area under the receiver operating characteristic (ROC) curve. Once a model has been trained, the predictor can be saved and can be used in a second phase to classify unknown samples. Thus, using the option “Test existing model” in the Metabolizer web interface, a list of samples can be uploaded and the

proper predictor can be selected from the list of saved predictors. The predictor chosen will return a table with the probabilities of belonging to any of the classes for each sample.

3.2.1.4 Prediction of the impact of KOs in metabolism

The method proposed can be used not only to derive metabolic module activity profiles in real conditions but also in simulated conditions. Therefore, knockouts (KOs) or over-expressions, alone or in combinations, can easily be simulated by changing the values of the targeted genes to 0 or 1 (or to any other low or high value between 0 and 1), respectively. Then, the simulated condition is compared to the original condition and a fold change threshold of 2 (that can be modified by the user) can be used to detect the most relevant changes in module metabolic activity. Since only two individual conditions (before and after KO) are compared a conventional test cannot be applied here. In addition to individual gene interventions, the effect of drugs with known targets (as described in DrugBank [91]) over the different metabolic modules can be studied. It is possible to simulate the effect of drugs alone, in combinations, or combined with gene KOs or over-expressions. Since it is common that genes participate in more than one pathway and drugs often affect more than one gene, the results can contain more than one module and pathway. Obviously, off-target effects not described in DrugBank cannot be included in the predictions. However, Metabolizer would allow conjecturing new off-target effects by checking inconsistencies between the expected metabolic module activities from the prediction and the real ones observed upon the application of the drug. This fact reinforces the utility of comprehensive holistic modelling approaches like the one presented here. The single gene intervention strategy implemented here is similar to the one used in the PathAct web tool [43] in the context of signalling pathway genes.

3.2.1.5 Automatic detection of optimal therapeutic targets

The Knockout option of Metabolizer implements the Auto Knockout functionality to find the optimal KO to revert a condition. Within this modelling framework a gene KO is easy to simulate. Simply, the expression value is multiplied by 0.01. This is a randomly selected constant that simulates the KO effect by decreasing the expression value. Zero value is not used to allow propagation in the network. Once the KO is applied, the method recalculates the module activity profiles. It is worth noting that a gene can participate in more than one module and that, depending on the location of the KO gene in the topology of the module, the KO can have a drastic or an irrelevant effect on the module activity.

Then, if two groups of samples are provided, metabolizer finds the KO intervention that makes samples of one of the classes resemble more to samples of the other class at the level of metabolic module activity profiles. This functionality has been designed to compare diseased to

healthy conditions, or similar scenarios, and find the KO intervention that produces the maximum reversion from the disease to the healthy condition. Firstly, a class predictor is built, using RF, that best discriminate among the two classes compared. Since only 553 genes participate in the modules, for each sample all the possible gene KOs can be carried out. For each simulated KO, the metabolic values are recalculated and then the probability of a sample of belonging to the opposite class is calculated. All metabolic profiles resulting from the KO are ranked by this probability and the higher probabilities represent the most promising KO interventions. Combinations of KOs are not feasible in interactive mode but they can be tested manually in the individual sample mode. Figure 3.3 shows a schema of the procedure.

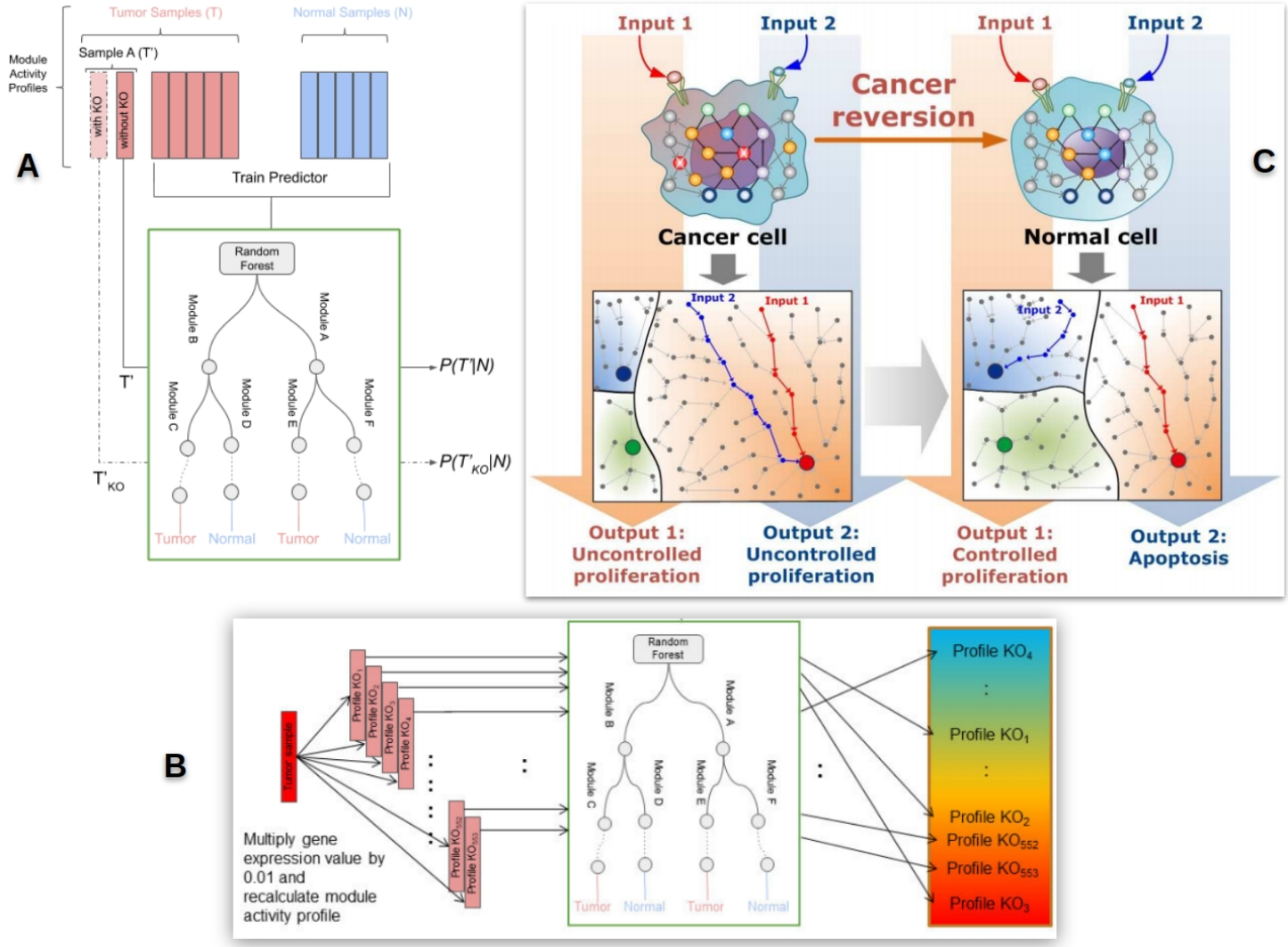


Figure 3.3: *Auto Knockout* functionality to find the optimal KO to revert a condition. A) Initially, a predictor is trained to distinguish between normal and tumour samples. B) Then, for the tumour sample problem, all the possible KOs are simulated by sequentially multiplying by 0.01 the expression values of any of the genes involved in all the modules. Then, the theoretical KO profiles are calculated for each simulated KO sample and the predictor is used to assign a probability to belong to a normal or a tumour class. The rank of more likely belonging to the normal class is a rank of the potential of transformation of a tumour sample into a normal sample. C) Illustration of how mechanism-based modelling can reveal the dynamical aspect of cancer reversion taken from Cho K-H et al., 2017 [92].

3.2.1.6 Web server development

Metabolizer is a web-based application that implements the above-described functionalities. Metabolizer web client has been developed in HTML5 with web components while the server component is written in R programming language. The program recodes gene expression data (either from microarray or from RNA-seq) into estimates of enzymatic activities along the sequence of reactions t at transform simple into complex metabolites or vice-versa. Metabolizer can be used for several purposes that include: (1) estimation of differential metabolic activity by comparing two

conditions, (2) construction of class predictors for further classification of new samples using metabolic activities as multigenic biomarkers; (3) search of therapeutic targets by predicting the ultimate impact of KOs on the final metabolite production activity of the modules, and (4) automatic detection of the optimal KO that makes the metabolic profile of an initial condition as close as possible to a final condition (e.g., the KO that reverts a disease to the normal status).

In addition to human metabolism, Metabolizer includes the metabolism of 5 more model species, namely mouse, rat, zebrafish, drosophila and worm, taken from the KEGG repository too. The input for Metabolizer consists of files of normalized gene expression values (in TSV format) along with an accompanying text file containing the experimental design. A tutorial explains in detail the required format. The results produced include a graphical output that represents the metabolic modules analyzed in which the sequences of enzymatic reactions that transform simple into complex metabolites are highlighted. In this way, disruptions or aimage5.pngctivations in the metabolite transformation chain can be easily visualized providing a straightforward interpretation of its real impact on the ultimate metabolite production activity. A convenient graphical interface, based on the CellMaps [93] libraries, provides an interactive view of the metabolic modules with configurable colour-coded representation of the metabolic modules and their components. In this interface, gene activity and module activities are simultaneously represented providing a visual, intuitive indication on relevant changes in the activity of the genes and their final impact in the activity of the modules (see Figure 3.4). Tables of the results that listing the modules with a significant change in the activity are provided, along with the statistics and the corresponding p-values. Metabolizer web server can be found at http://metabolizer.babelomics.org and the code is given at https://github.com/babelomics/metabolizer

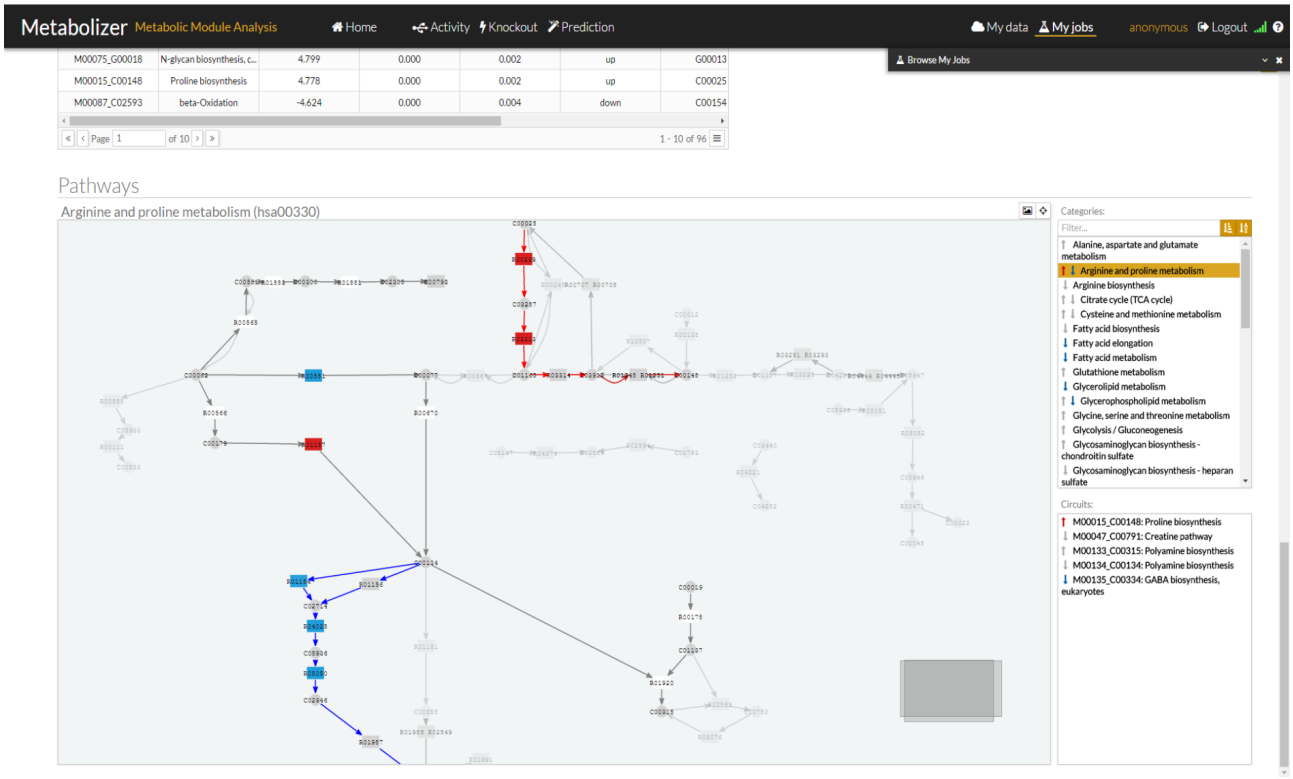


Figure 3.4: Metabolizer graphic interface with a representation of the modules. On the right side there is a list of KEGG pathways with arrows up or down in case they contain modules with up or down activations, respectively. When the arrow is grey, the change in activity is not significant. Red up arrows indicate a significant increase in activity and blue down arrow a significant decrease of activity in the module. Below the pathway list, there is another list with the modules within the pathway with the same code for arrows.

3.2.2 Evaluating the predictive power of Metabolizer

3.2.2.1 Samples and data processing

RNA-seq counts of breast invasive carcinoma (BRCA), liver hepatocellular carcinoma (LIHC), kidney renal clear cell carcinoma (KIRC), and prostate adenocarcinoma (PRAD) cancer types (a total 2256 cancer samples and 285 healthy reference tissue samples) were downloaded. The breakdown of sample sizes are given in Table 3.1). Unwanted technical variation, batch effect, was removed by COMBAT method [61] and the trimmed mean of M-values normalization method (TMM) [62] was used for gene expression normalization. Normalized gene expression values were log-transformed and re-scaled between 0 and 1. Breast Invasive Carcinoma subtype classifications were available through the cBIOportal (https://www.cbioportal.org/study/clinicalData?id=brca_tcga_pan_can_atlas_2018). Cell survival measurements after gene knockdown were taken from the Project Achilles 2.20.2 (https://portals.broadinstitute.org/achilles/datasets/15/download) [94]. Escherichia coli (E. coli) gene expression data in eight different combinations of carbon sources, nitrogen sources and electron acceptor conditions [44] was taken from

http://systemsbiology.ucsd.edu/In_Silico_Organisms/E_coli/E_coli_expression2. Metabolomics data for BRCA and KIRC were downloaded from the supplementary files of Terunuma et al. 2014, [95] and Hakimi et al. 2016 [96], respectively.

Cancer type	Abbreviation	Tumour samples	Normal samples
Breast invasive carcinoma	BRCA	1057	113
Kidney renal clear cell carcinoma	KIRC	526	72
Liver hepatocellular carcinoma	LIHC	294	48
Prostate adenocarcinoma	PRAD	379	52
Total		2256	285

Table 3.1: TCGA samples used in this study.

3.2.2.2 Sensitivity and specificity of models of metabolic module activity

To differentiate cancer from healthy samples, categorical models were built using module activities as features and RF as a supervised learning algorithm [89]. These categorical models were used to evaluate the sensitivity and specificity, the predictive power of modules for class membership prediction. Since the number of features is not too high (there are only 95 modules), feature selection was not considered necessary here. Specifically, we repeated 50 times the five-fold cross-validation on the dataset: two groups, one composed of normal samples and another, with the same size, composed of tumour samples randomly sampled were constructed. Four fifth parts were used to train a RF [89] predictor and the remaining fifth part was used to test the predictor with all the module activities (see Figure 3.6). Since the real labels of the fifth part are known, the correct and wrong assignations per class were used to calculate the area under the ROC curve (AUC).

3.2.2.3 Comparison of Metabolizer to other methods

Different approaches for the detection of different aspects of metabolic module activity have been proposed. In order to compare the accuracy of Metabolizer in detecting metabolic module activity, we have used a version of GSEA based on logistic regression [63] as implemented in the mdgsa Bioconductor package (http://bioconductor.org/packages/release/bioc/html/mdgsa.html) and a popular PT-based algorithm SPIA [97], as implemented in the SPIA Bioconductor package (http://bioconductor.org/packages/release/bioc/html/SPIA.html). For these methods, gene sets were defined using the genes within the metabolic modules. Additionally, the SPIA method also requires the topology of the modules. In order to adapt the modules to the pathway format needed for the SPIA function the relations between metabolites on a module are considered as activations. GSEA detects only differential activity while SPIA and Metabolizer also detect whether this different

activity implicates activation or deactivation (up/down regulation). Four cancers (Table 3.1) were used for the comparison. The sensitivity of the method was measured as to the number of modules detected as differentially active by comparing the four cancers in Table 3.1 with respect to their corresponding healthy tissues. The specificity was measured as the number of differentially active modules (false positives) found by each method in a comparison involving individuals of the same class.

In addition, we utilized a well-known version of CBM method [98], as implemented in the IMAT tool [99] using the human metabolic network Recon 2 V2.02 [100], to compare its performance against to Metabolizer as well. The IMAT tool maximizes the number of highly expressed reactions that are active and the number of lowly expressed reactions that are inactive. The reaction activity is inferred from the binarization of the corresponding gene expression values following a Boolean logic from gene-protein-reaction (GPR) rules within the context of metabolic networks [87]. CPLEX (V12.6.2) solver was used for solving Mixed Integer Linear Programming problems. Optimum solutions provide flux values of reactions and these flux values were used to classify reactions as active and inactive. All parameters were set as in the original article [98]. The binary results of reactions (active/inactive) were used to train a classifier. Since this CBM method is based on a pathway definition (Recon 2) [100] which is different from the KEGG metabolic modules used here [84], we use a different benchmarking framework in which reaction values are used as predictor features [101]. Given that classifiers based either on Module activities or on CBM reaction activities were able to distinguish cancer and normal tissues with almost 100% accuracy. For that reason, we challenged them with a more complex classification problem: distinguishing between cancer subtypes in the case of breast cancer. The BRCA dataset (Table 3.1) contains PAM50-defined [102] subtypes; Basal-like, HER2-enriched, Luminal A, and Luminal B of Breast Invasive Carcinoma [103]. The performance of a RF [89] based classifier trained using module activities by Metabolizer and reaction activities obtained by CBM were compared by five-fold cross validation, using gene expression-based classification as a gold standard. It is worth noting that only one gene belonging to the metabolic modules, PHGDH, was in the list of PAM50 genes used to define the BRCA subtypes.

3.2.2.4 Validation of KO predictions and case uses

3.2.2.4.1 An example of automatic optimal KO

To illustrate the potential of the Auto-KO option, we have used this tool to find KOs that would make a KIRC sample as similar as possible to a normal kidney sample in terms of metabolism. We used a balanced dataset that is composed of all 72 normal kidney samples available and 72 KIRC samples randomly sampled among all the available tumour samples and

used the Auto-KO option. Then, a class predictor is built that will be used to decide to what extent the tumor sample, after the KO, could be identified as likely as a normal sample. This approach used for each tumour sample that were in the list of 72 tumour samples. Then the results were given as the mean of these tumour samples. Most of the KOs do not have an effect that significantly reverts the metabolic tumor status towards that of a normal kidney, in a way that increases the probability of being recognized as normal by the predictor. Then, KOs were clustered based on their effect similarities (change in probability). The gene cluster which has the highest effect on the likelihood of being normal was used to create double KO pairs and to test their synergistic effects. Auto-KO option of the Metabolizer does not generate compute double KOs automatically, but users can apply multiple KOs manually.

3.2.2.4.2 Experimental validation in a cancer model of gastric adenocarcinoma of an optimal KO prediction in gastric cancer patients

Finally, as an additional validation, we used the optimal KO option of Metabolizer in a different cancer type, gastric cancer patients (STAD). Table 3.2 shows the predictions. The gene causing the strongest effect, DPYS, was found as essential in the catalogue of cancer dependencies [94]. The second predicted gene, UPB1, encodes an enzyme (β-ureidopropionase) that catalyzes the last step in the pyrimidine degradation pathway, required for epithelial-mesenchymal transition [104]. Using a cancer model of gastric adenocarcinoma (AGS cell line) we carried out a cell proliferation experiment upon depletion of UPB1 gene expression. All the following experimental validation work was done by the collaborators and in their wet-lab (Pujana's Lab, IDIBELL, Barcelona). The shRNAs targeting UPB1 were purchased from the MISSION (Sigma Aldrich) library, catalogue SHCLNG-NM_016327. Lentivirus production and transduction was performed following standard protocols and cell cultures were selected with puromycin for 72 h prior cell seeding for evaluation of proliferation/viability by methylthiazol tetrazolium (MTT)-based assays (Sigma-Aldrich). The data corresponds to sextuplicates and was replicated in different assays. UPB1 expression was detected with the Human Protein Atlas HPA000728 antibody (Sigma-Aldrich) and gene expression measured with primers 5′-TCGACCTAAACCTCTGCCAG-3′ and 5′-TAAGCCTGCCACACTTGCTA-3′, using PPP1CA as control.

Gene symbol	Entrez ID	*p(normal)* after KO	*p(normal)* before KO	Change in probability
DPYS	1807	0.468	0.33	0.138
UPB1	51733	0.468	0.33	0.138
GART	2618	0.416	0.33	0.086
ATIC	471	0.416	0.33	0.086
PAICS	10606	0.416	0.33	0.086
SLC27A5	10998	0.392	0.33	0.062
BAAT	570	0.392	0.33	0.062
HSD17B12	51144	0.368	0.33	0.038
TECR	9524	0.368	0.33	0.038
ALDH5A1	7915	0.354	0.33	0.024
ABAT	18	0.354	0.33	0.024

Table 3.2: Probabilities of STAD metabolic profiles being identified as normal cell metabolic profile after the KO of the gene.

3.2.2.4.3 Applying the method in other model organisms

E. coli can grow on different carbon sources using different metabolic routes [105]. A large-scale compendium of gene expression data of different E. coli growth in a variety of conditions is available [106]. Table 3.3 summarizes the eight conditions studied. We have used the gene expression values to derive the module activities for all conditions, looking for specific scenarios in which the activation of certain modules was expected. Of particular interest is the behaviour of Glycolysis and TCA cycle modules that should be affected by some of the conditions reported in the study that involve aerobic and anaerobic conditions with different carbon sources [106], providing thus insights into energy metabolism linked to cellular respiration. Figure 3.5 illustrates the knowledge given in the biology textbook, the routes used by E.coli under different forms of respiration.

Carbon Sources	Nitrogen Sources	Electron Acceptor (Respiration) Conditions	Conditions	# Samples
D-galactonate	NH4	Aerobic	Condition 1	2
Glucose	NH4	Aerobic	Condition 2	9
Glucose	NH4	Anaerobic	Condition 3	10
Glucose	Nitrate	Anaerobic	Condition 4	3
Glycerol	NH4	Aerobic	Condition 5	5
Lactate	NH4	Aerobic	Condition 6	6
L-galactonate	NH4	Aerobic	Condition 7	3
Thymidine	NH4	Aerobic	Condition 8	3
Total				**41**

Table 3.3: Details of Escherichia coli growth conditions and the number of samples for each condition.

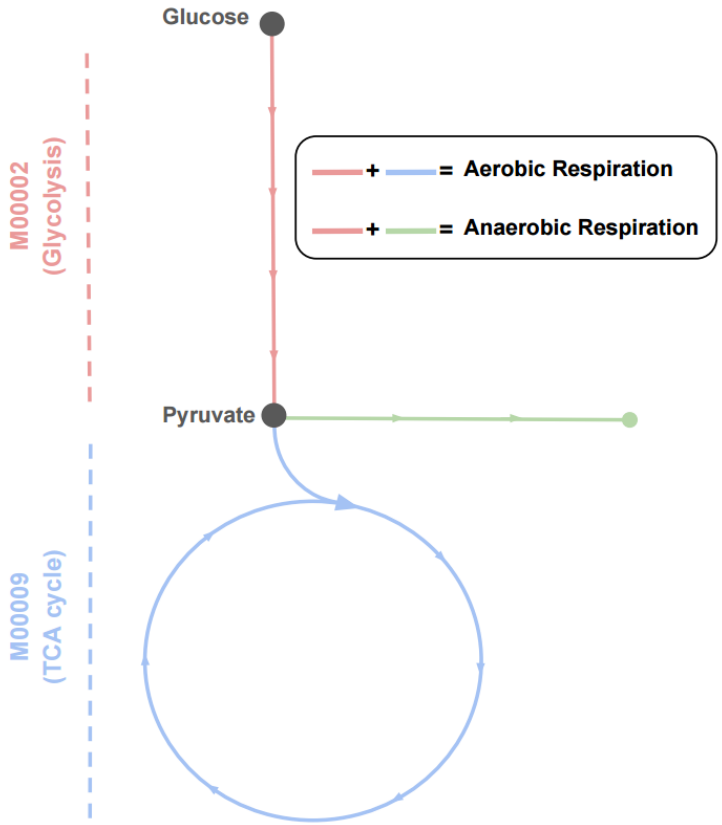


Figure 3.5: Metabolic routes used by E. Coli during the aerobic and anaerobic respiration. Glycolysis (red route, KEGG module Id: M00002) is used under both conditions. While aerobic respiration continues with oxidative phosphorylation through TCA cycle (blue route, KEGG module Id: M00009), anaerobic respiration continues with fermentation process for the production of lactic acid (animals and bacteria) and ethanol (yeast).

3.2.2.4.4 Concordance between module activity and concentration of final metabolite

To further assess the validity of the predictions, metabolomic data from breast and kidney cancer were then analyzed. We used the balance between the initial and final metabolite fold changes (ratio of the arithmetic mean values) as an indication of activation (relative increase in the final metabolite with respect to the initial one) or inactivation (relative decrease in the final metabolite with respect to the initial one).

3.3 Results and Discussion

Metabolizer web tool offers several functions. Under this chapter, all these functions were benchmarked, and the results were validated based on literature and wet-lab experiments. Below, we discussed all the results in the order given; differentially module activity, knockout simulations, and constructing predictors.

To evaluate the predictive power of Metabolizer, categorical models were built using

module activities as features and RF as a supervised learning algorithm. The models were evaluated using five-fold cross-validation using datasets with real cancer and control labels and also randomized labels. As it is shown in Table 3.4, the predictive power of the models which built using module activities as features is extremely high in four cancer types. The AUC in real class comparisons can be compared to the poor AUC values in artificial classes obtained by random permutation of cancer and control labels. This strongly suggests that module activities account for real biological features that change between cancers and normal tissues.

	BRCA	**BRCA random**	**LIHC**	**LIHC random**	**KIRC**	**KIRC random**	**PRAD**	**PRAD random**
Mean	1.000	0.495	1.000	0.525	0.999	0.477	0.998	0.491
Standard deviation	0	0.190	0	0.251	0.002	0.216	0.006	0.208
Median	1.000	0.5025	1.000	0.533	1.000	0.464	1.000	0.469
Median absolute deviation	0	0.205	0	0.312	0	0.243	0	0.228

Table 3.4: AUC values obtained for tumour types in Table 3.1, with the corresponding AUC values obtained when artificial classes are obtained by randomizing sample label.

In order to compare the accuracy of Metabolizer in detecting metabolic module activity, we set two different comparison scenarios based on the existing tools and their output. We first compared the capability for detecting differentially activated modules when cancer is compared to the corresponding unaffected tissue in four distinct cancer types: BRCA, LIHC, KIRC, and PRAD (Table 3.1). Second, in order to calculate the false positive rates, we compared 1000 times two artificial sample sets by random sampling of normal tissues maintaining the proportions of the real comparison. That is 102 vs. 11 for BRCA, 41 vs. 7 in LIHC, 63 vs. 9 in KIRC and 46 vs. 6 in PRAD. The same procedure was repeated using cancer samples. In this case, the proportions were 995 vs. 102 in BRCA, 253 vs. 41 in LIHC, 463 vs. 63 in KIRC and 333 vs. 46 in PRAD. For this contrast, we used the conventional approach based on unstructured gene sets, the GSEA [63], and an approach that takes into account the relationships among genes within gene sets, the SPIA [97]. Table 3.5 shows the number of modules found as differentially activated in the different cancers by the different methods. Metabolizer outperforms both the sensitivity and specificity of GSEA and SPIA. GSEA founds between 5 and 14 modules, depending on cancer, with averages ranging from 2 to 7 false positive (FP) modules. SPIA increases the specificity at the exchange of reducing the sensitivity, with a very low detection rate. Metabolizer increases by almost one order of magnitude both sensitivity and specificity (Table 3.5). In general, the results found by the methods were consistent across them, taking into account their different sensitivities. As expected, modules controlling the biosynthesis of nucleotide precursors [107] and Acetyl-CoA [108, 109] were found across cancers by GSEA and Metabolizer. However, several well-known metabolic

activities associated to cancer development and progression, such as increased production of l-Proline [110] and succinate, 9 or related to metastasis, such as fumarate, 4-aminobutanoate (GABA biosynthesis) [111] or N-acylsphingosine (Ceramide biosynthesis), [112] were found only by the more sensitive Metabolizer method.

Method	BRCA		LIHC		KIRC		PRAD	
	Found	FP	Found	FP	Found	FP	Found	FP
GSEA	8	3.5/1.8	5	2.8/6.0	14	7.4/3.4	5	2.7/2.3
SPIA	2	0.07/0.05	1	0.3/0.1	2	0.1/0.07	1	1.1/0.1
Metabolizer	81	0.008/0.06	77	0.04/0.04	77	0.03/0.03	73	0.05/0.05

Table 3.5: Number of modules found as differentially activated in the cancers listed in Table 3.1 by the different methods GSEA, SPIA, and Metabolizer. The number of false positives (FP) was calculated by comparing 1000 times two artificial sample sets by random sampling of normal tissues maintaining the proportions of the real comparison. That is 102 vs. 11 for BRCA, 41 vs. 7 in LIHC, 63 vs. 9 in KIRC and 46 vs. 6 in PRAD. The same procedure was repeated using cancer samples. In this case the proportions were 995 vs. 102 in BRCA, 253 vs. 41 in LIHC, 463 vs. 63 in KIRC and 333 vs. 46 in PRAD. The second column for each cancer type shows the average number of FPs obtained with normal samples/the same figure obtained from cancer samples

Since CBM analysis is based on a different type of pathway (Recon 2), the comparison cannot be carried out in the previous benchmarking framework that uses metabolic modules defined within KEGG pathways. Instead, we carried out a comparison of classification performances using a previously proposed benchmarking framework based on the use of reaction activities estimated by CBM as features for classification [101]. Given that cancer versus normal tissue was a quite naive classification problem for which both CBM and Metabolizer resulted in almost 100% classification accuracy. Thus, we used a more challenging classification problem: BRCA subtype prediction. Classification performances were carried out using a RF-based predictor with five-fold cross-validation. Since BRCA subtypes have been defined using the expression of 50 genes with the PAM50 classifier,43 the classification obtained using the expression of all genes is expected to provide an upper limit of classification performance. Figure 3.6 shows how module activities obtained with Metabolizer outperform CBM-based reaction activities in classifying all the BRCA subtypes. It is worth noting that there is no common list of features between gene-based and module based predictors. Only one gene belonging to the metabolic modules, PHGDH, was in the list of PAM50 genes used to define the BRCA subtypes.

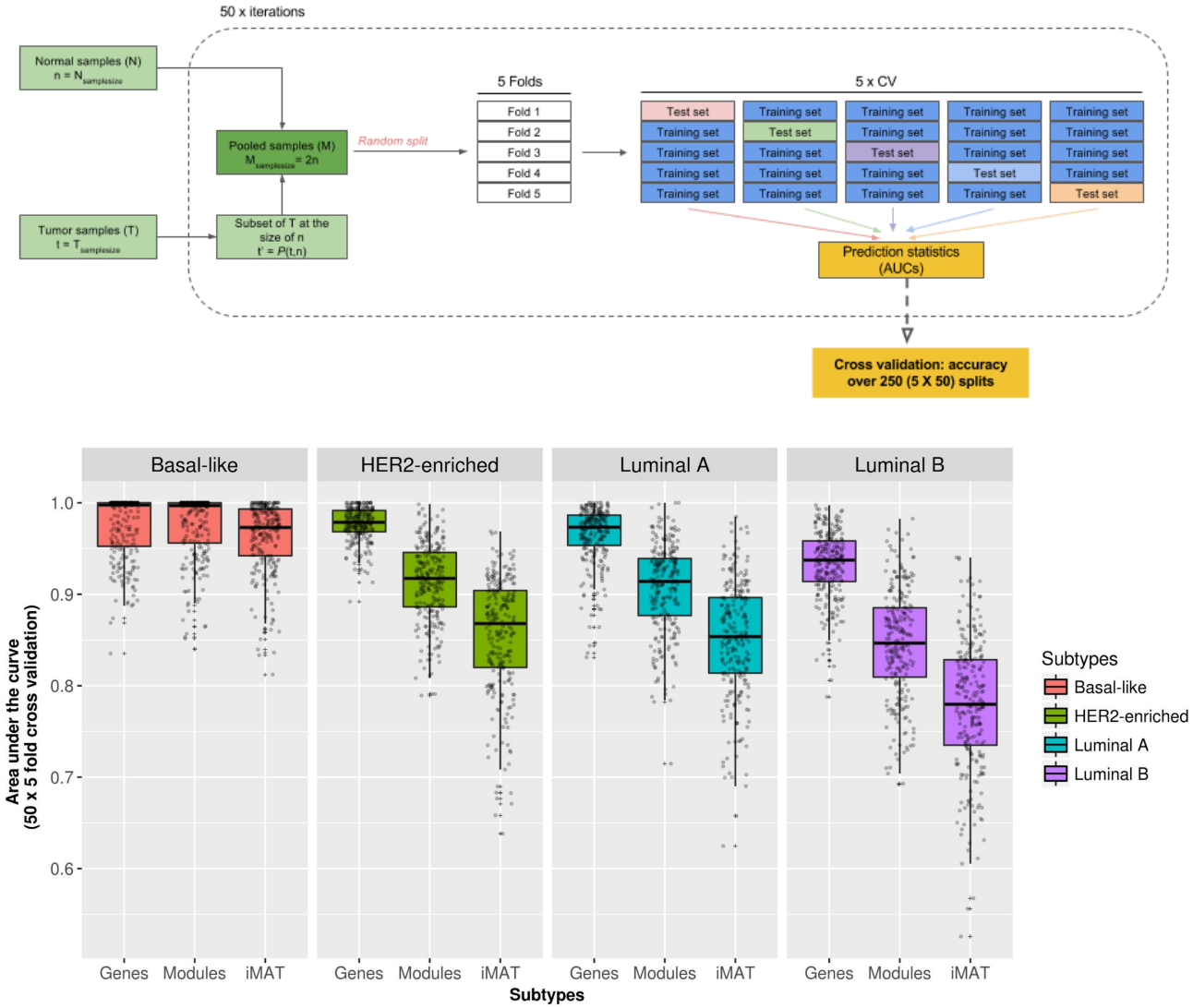


Figure 3.6: BRCA subtype classification performances obtained using module activities inferred with Metabolizer and CBM-based reaction activities by iMAT. BRCA subtypes are defined on the bases of PAM50 gene activities and therefore, gene expression is taken as the gold standard classification performance. Upper panel of this figure illustrates the 50 x 5 cross-validation applied to obtain classification performances.

Knockout option allows studying single gene KOs and the effect of drugs with known targets. When enough samples for at least two different conditions provided, Auto-KO option allows us to integrate omics data on top of biological pathways and apply machine learning algorithms to discover patient-specific potential drug targets automatically. In the given KIRC example, in a few cases the result of the KO changes the metabolic status of the tumor in a way that is identified as normal in approximately 25%. Figure 3.7 shows the KOs which have a similar strength of condition reverting effect. The details of the KO results for the genes in group G_1.5 are given in Table 3.6. This table lists the genes in which a KO produces changes in the metabolic profile of the tumor cell that make it more similar to the metabolic profile exhibited by a normal kidney cell. Some of these optimal KO predictions were known as cancer-related genes. For

example, HSD17B12, is a known cancer antigen [113], EBP is a long-known cancer estrogen receptor [114] or DHCR24 is a gene whose over-expression is related to bad prognostic in several cancers [115], which explain the potential predicted impact that their KOs have in the cancer metabolic profile. However, beyond the knowledge derived from the literature, other experimental evidence, such as the recent release of a large-scale map of cancer dependency [94], can be used to validate predictions made on the simulated KOs that would potentially reduce the cancer phenotype of cells and make them resemble normal cells.

Gene symbol	Entrez ID	p(normal) after KO	p(normal) before KO	Change in probability
HSD17B12	51144	0.348	0.92	0.256
TECR	9524	0.348	0.92	0.256
SC5D	6309	0.328	0.92	0.236
EBP	10682	0.328	0.92	0.236
DHCR24	1718	0.328	0.92	0.236
LSS	4047	0.328	0.92	0.236
TM7SF2	7108	0.328	0.92	0.236
NSDHL	50814	0.328	0.92	0.236
CYP51A1	1595	0.328	0.92	0.236
HSD17B7	51478	0.328	0.92	0.236
DHCR7	1717	0.328	0.92	0.236

Table 3.6: Probabilities of KIRC metabolic profiles being identified as normal cell metabolic profile after the KO of the gene.

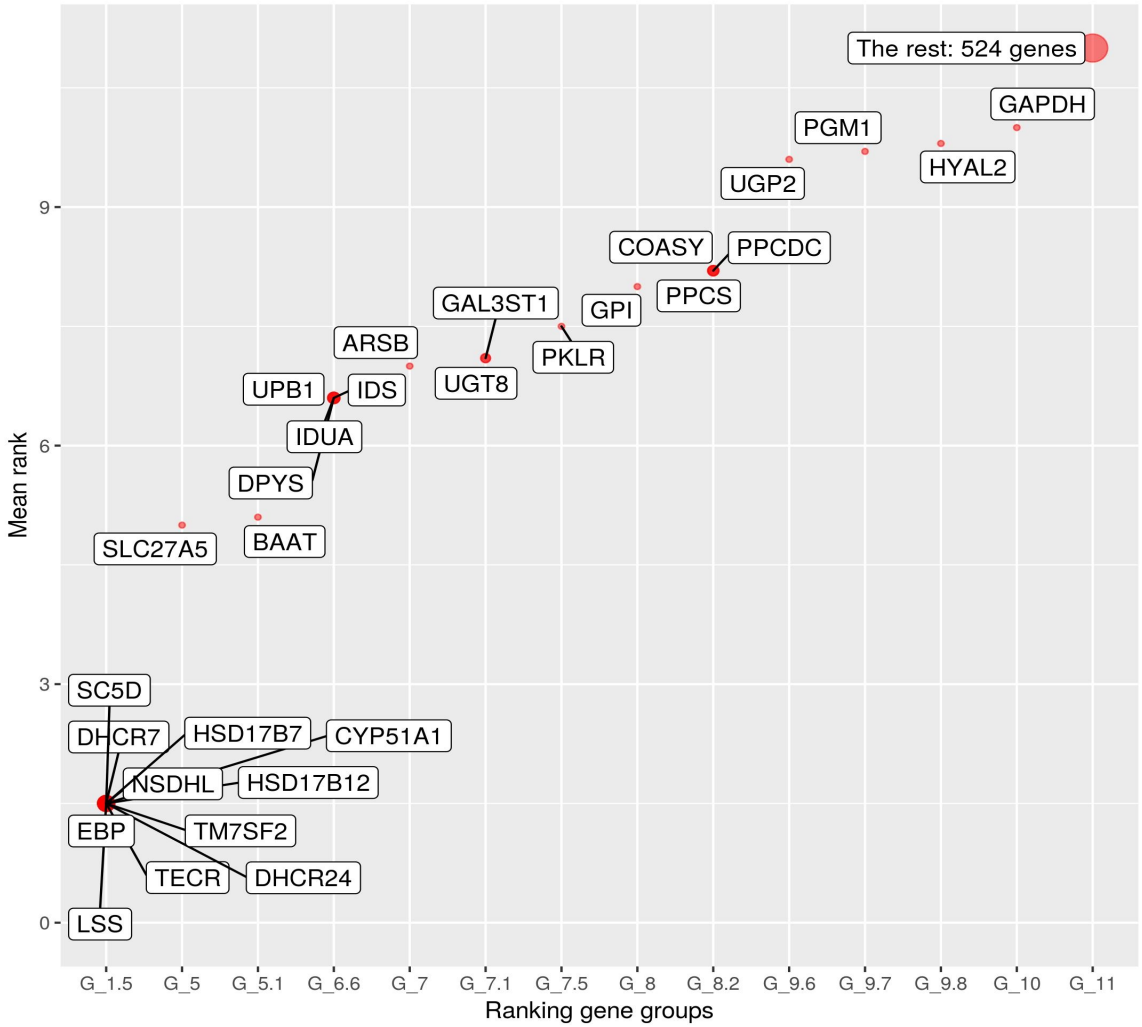


Figure 3.7: KOs which have similar strength of condition reverting effect in KIRC. Genes were ranked based on their effect, 1 is the best performing genes. Then the mean of ranks for 72 tumour samples was calculated (y-axis). Based on the mean ranks the genes were clustered (x-axis). The genes which had the highest impact on condition reverting are given in cluster G_1.5.

The expectation is that inhibitions of optimal KO genes should result in the reduction of the proliferative capability of the corresponding cell lines that could be interpreted as a reversion of cancer phenotype towards a normal cell (or at less, a non-proliferative cell). In spite of the fact that cancer outcome is a much more complex phenotype than the proliferation of a cell line, when genes in Table 3.6 are inhibited in the cancer dependency experiment [94] a reduction in the proliferation was observed for ten out of the eleven predicted optimal KOs (HSD17B12, TECR, SC5D, EBP, DHCR24, LSS, NSDHL, CYP51A1, HSD17B7, DHCR7) (see Figure 3.8). Moreover, in some cases, we were able to detect an increase in patient survival in patients with low expression of some of the optimal KO proteins in Table 3.6. Thus, according to Protein Atlas [116], low expression of TECR protein is significantly associated to better patient survival in urothelial cancer (see https://www.proteinatlas.org/ENSG00000099797-TECR/pathology/tissue/urothelial+cancer), and the same is observed in DHCR24 in endometrial cancer (https://www.proteinatlas.org/ENSG00000116133-DHCR24/pathology/tissue/endometrial+cancer),

LSS in urothelial cancer (https://www.proteinatlas.org/ENSG00000160285-LSS/pathology/tissue/urothelial+cancer), CYP51A1 in cervical cancer (https://www.proteinatlas.org/ENSG00000001630-CYP51A1/pathology/tissue/cervical+cancer), and HSD17B7 in renal cancer (https://www.proteinatlas.org/ENSG00000132196-HSD17B7/pathology/tissue/renal+cancer). However, Protein Atlas results that support Metabolizer predictions must be taken with some caution given that they implicitly make the assumption that lower cancer cell survival would be equivalent to higher patient survival.

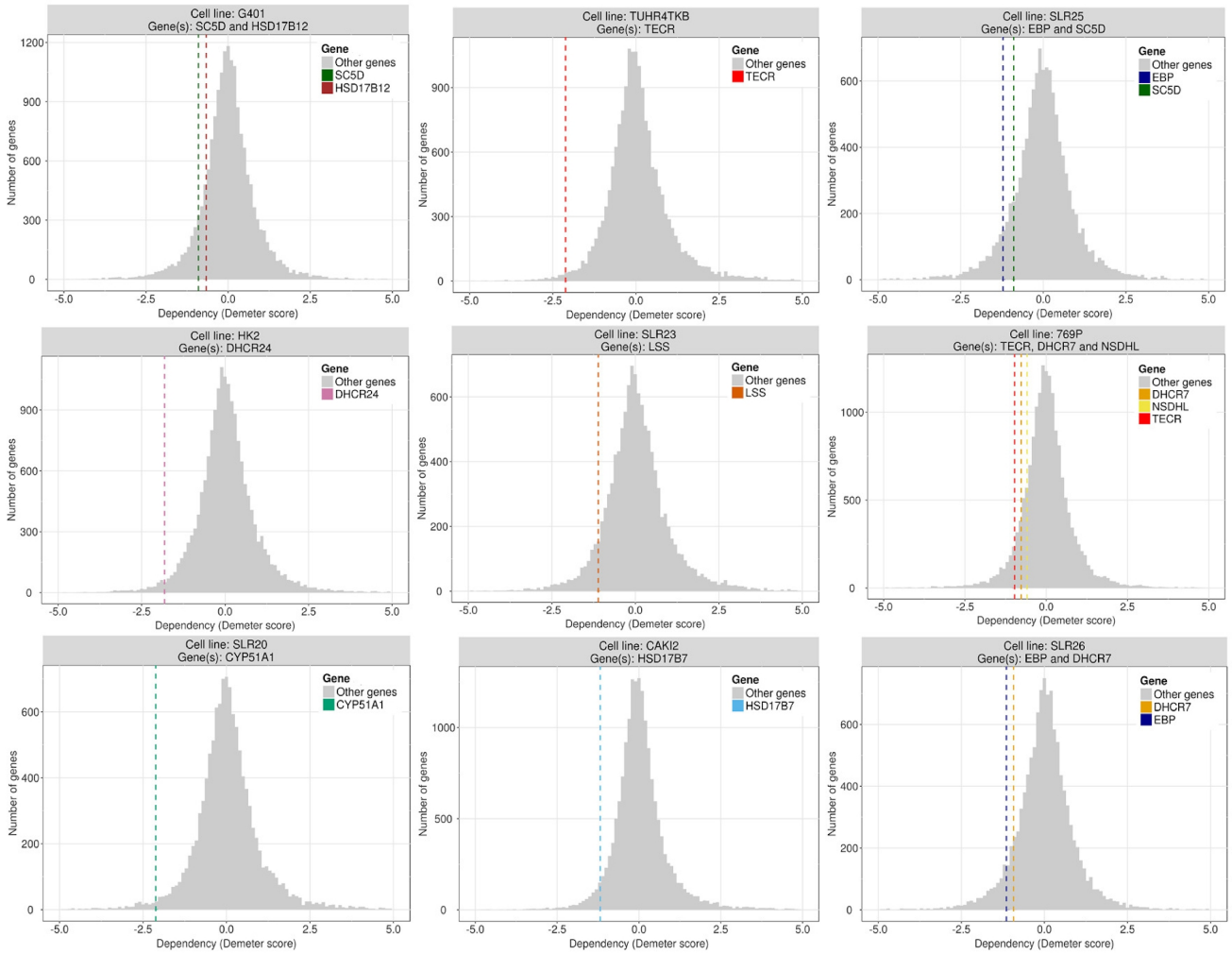


Figure 3.8: Essentiality (Demeter score) of genes predicted as optimal KOs with respect to the background distribution of essentiality values. Values below 0 indicate lower proliferation. From left to right and top to bottom: HSD17B12 and SC5Din cell line G401 (KIDNEY); TECR in cell line TUHR4TKB (KIDNEY) (this gene shows the same results in KMRC1 cell line of KIDNEY, data not shown); SC5D and EBP in SLR25 cell line (KIDNEY) (SC5D shows the same result in G401 cell line of SOFT_TISSUE, data not shown); DHCR24 in cell line HK2 (KIDNEY); LSS in cell line SLR23 (KIDNEY); NSDHL, DHCR7, and TECR in 769P cell line (KIDNEY); CYP51A1 in cell line SKRC20 (KIDNEY) (also less proliferative in SLR20 KIDNEY cell line, data not shown); HSD17B7 in cell line CAKI2 (KIDNEY); DHCR7 and EBP in cell line SLR26 (KIDNEY)

Drug repurposing has emerged as an alternative approach for rapid identification of effective therapeutics to complex diseases. Synergistic drug combinations using approved drugs identified from drug repurposing screens is a useful option which may overcome the problem of weak activity of individual drugs. Double KOs simulation is a computational strategy for predicting synergistic KO and drug pairs and biomarkers. Although a 25% increase in the probability of resembling a normal cell might look a small value, it is not expectable that a cancer cell becomes a normal cell with a unique intervention. However, if we assay combinations of double KOs between genes in Table 3.6, we observe a dramatic increase in the “normal” character of the cancer cell, as graphically depicted in Figure 3.9 Most of the combination have a similar effect that the single KOs alone (peak around 0.25). However, some combinations produce synergistic changes with a dramatic effect in the metabolic profile of KIRC cells that make them more similar to normal cellsthan to cancer cells (peak around a 0.5 of change in the probability).

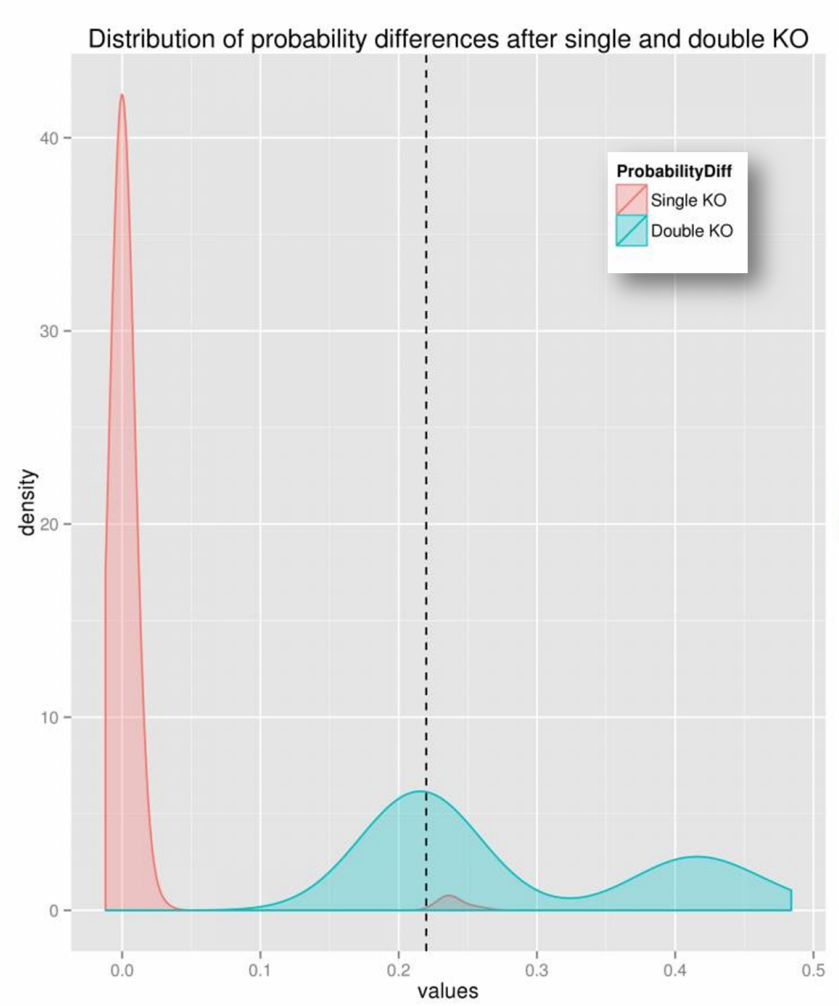


Figure 3.9: Distribution of the difference of probabilities that the predictor identifies a sample as a normal cell after and before the KOs of the corresponding genes (red distribution) or pair of genes (blue distribution).

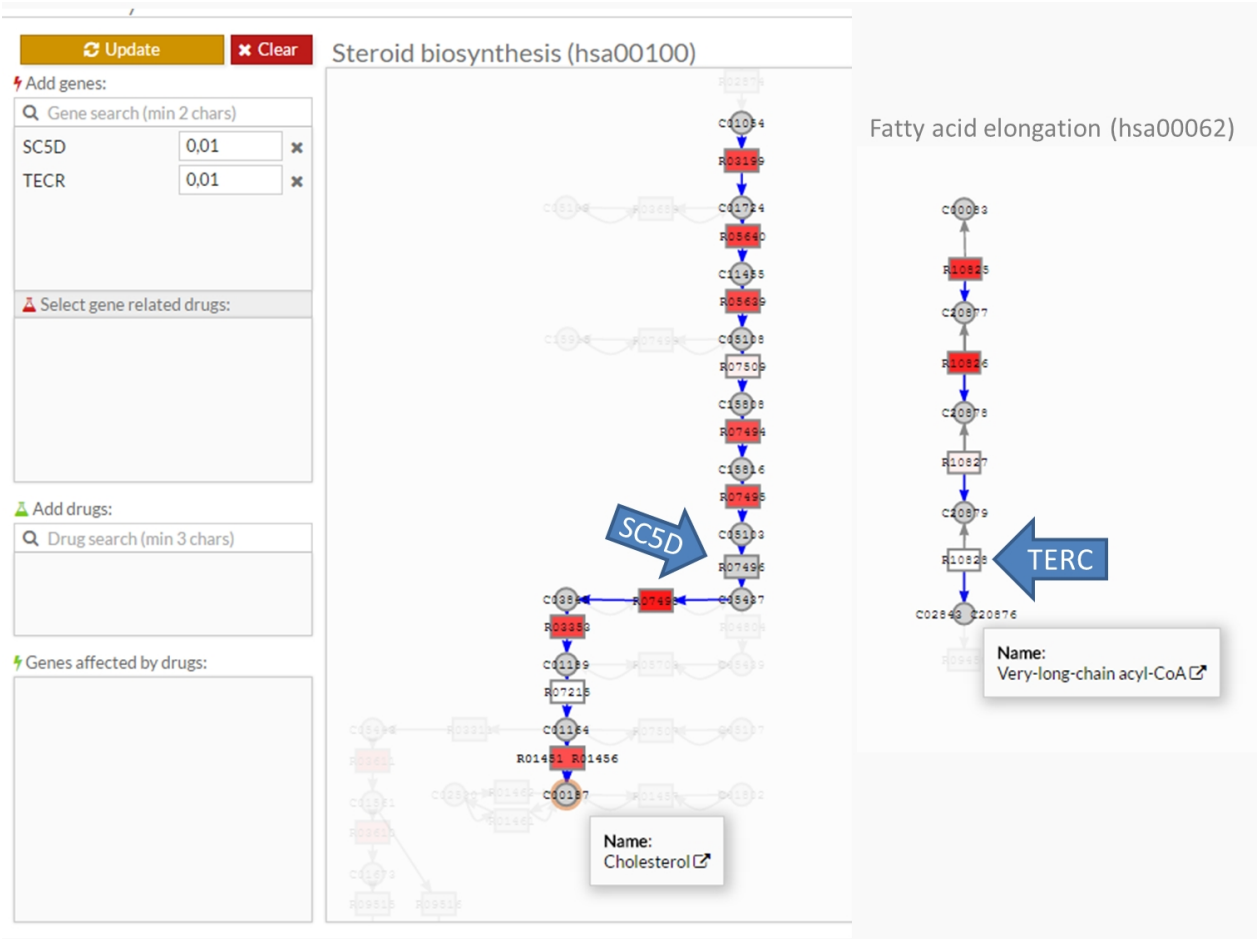


Figure 3.10: Representation of the modules corresponding to the Steroid biosynthesis (hsa00100) and Fatty acid elongation (hsa00062). Arrows point to the KO genes (SD5S and TERC). KOs were made by substituting the actual gene expression values by a low non-zero value of 0.01 (as can be seen in the Add genes box of the Metabolizer program. Genes in red indicate that they were active. At the end of both modules the resulting metabolite can be seen.

As an example, Figure 3.10 shows the double KO in genes SC5D and TERC, which affects to the Fatty acid elongation module and the Steroid biosynthesis module, the last one involved in the production of cholesterol. Actually, it is long known that tumour membranes are rich in cholesterol [117], suggesting that cholesterol utilization by cancer cells is an important feature of carcinogenesis and, probably, metastasis [118].

As an additional validation, we used the optimal KO option of Metabolizer in a different cancer type, gastric cancer patients (STAD). Table 3.2 shows the predictions. The gene causing the strongest effect, DPYS, was found as essential in the catalogue of cancer dependencies.56 The second predicted gene, UPB1, encodes an enzyme (β-ureidopropionase) that catalyzes the last step in the pyrimidine degradation pathway, required for epithelial-mesenchymal transition [118].

shRNA-mediated experimental KO conditions set, however, because of some technical issues, only UPB1 experiments were successfully completed. As anticipated by our prediction, three different short hairpin shRNA sequences directed to UPB1 caused a significant decrease in

cell proliferation (see Figure 3.11). This result constitutes an independent validation that reinforces the prediction made by the model proposed. Additionally, the inhibition of the rest of genes caused a remarkable reduction in the proliferation in the cancer dependency experiment, being in all the cases within the 10% most affected genes [94].

Finally, two more validations were done based on the request of the peer reviewers during the review process of the article of Metabolizer. Most of the metabolic modelling applications have been extensively dealing with bacterial organisms. We have used a classical bacterial model organism: Escherichia coli to demonstrate the validity and usefulness of the models of metabolic modules. Of particular interest is the behaviour of Glycolysis and TCA cycle modules that should be affected by some of the conditions reported in the study that involve aerobic and anaerobic conditions with different carbon sources [106], providing thus insights into energy metabolism linked to cellular respiration. E. coli samples which grown either aerobically or anaerobically with glucose present high activity of glycolysis (Conditions 2, 3 and 4 in Figure 3.12 A). Conversely, TCA cycle was almost inactive for anaerobic conditions, while it showed activity in aerobic conditions (conditions 3 and 4 in Figure 3.12 B). Under the anaerobic conditions, pyruvate cannot be converted into acetyl-CoA and therefore it does not enter the TCA cycle but rather continues with the fermentation process. When only the reaction activities corresponding to the enzymes contained in the modules (estimated using the IMAT tool [36]) are considered, it is difficult to detect similar condition-specific increases or decreases of activities (see Figure 3.12). This is probably due to the fact that, while module activities are describing whole biological processes, reaction activities are elementary pieces of such processes, shared by different modules. Therefore, it may happen that some reactions can be active within inactive modules because they are part of other active modules. For the same reason, similar behaviour is observed for gene expression activities (Figure 3.12). Consequently, module activities seem to be a better descriptor of the metabolic processes of the cell.

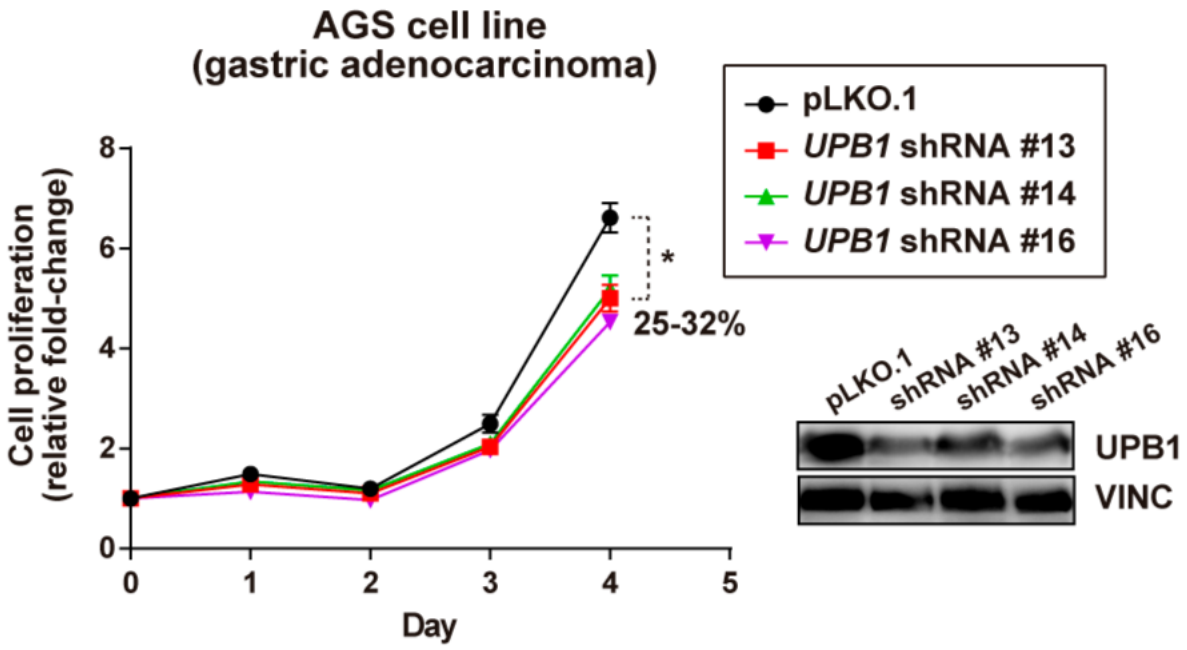


Figure 3.11: Experimental validation of an optimal KO prediction in gastric cancer patients. Relative cell proliferation of line AGS (stomach gastric adenocarcinoma) upon UPB1 expression depletion by three different MISSION shRNAs or transduced with control vector pLKO.1. The asterisk indicates significant differences (Mann–Whitney test p-values < 0.01). The percentage of reduction of cell proliferation is also shown. The prediction of UPB1 essentiality made by Metabolizer was confirmed by a relatively more sensitive behaviour.

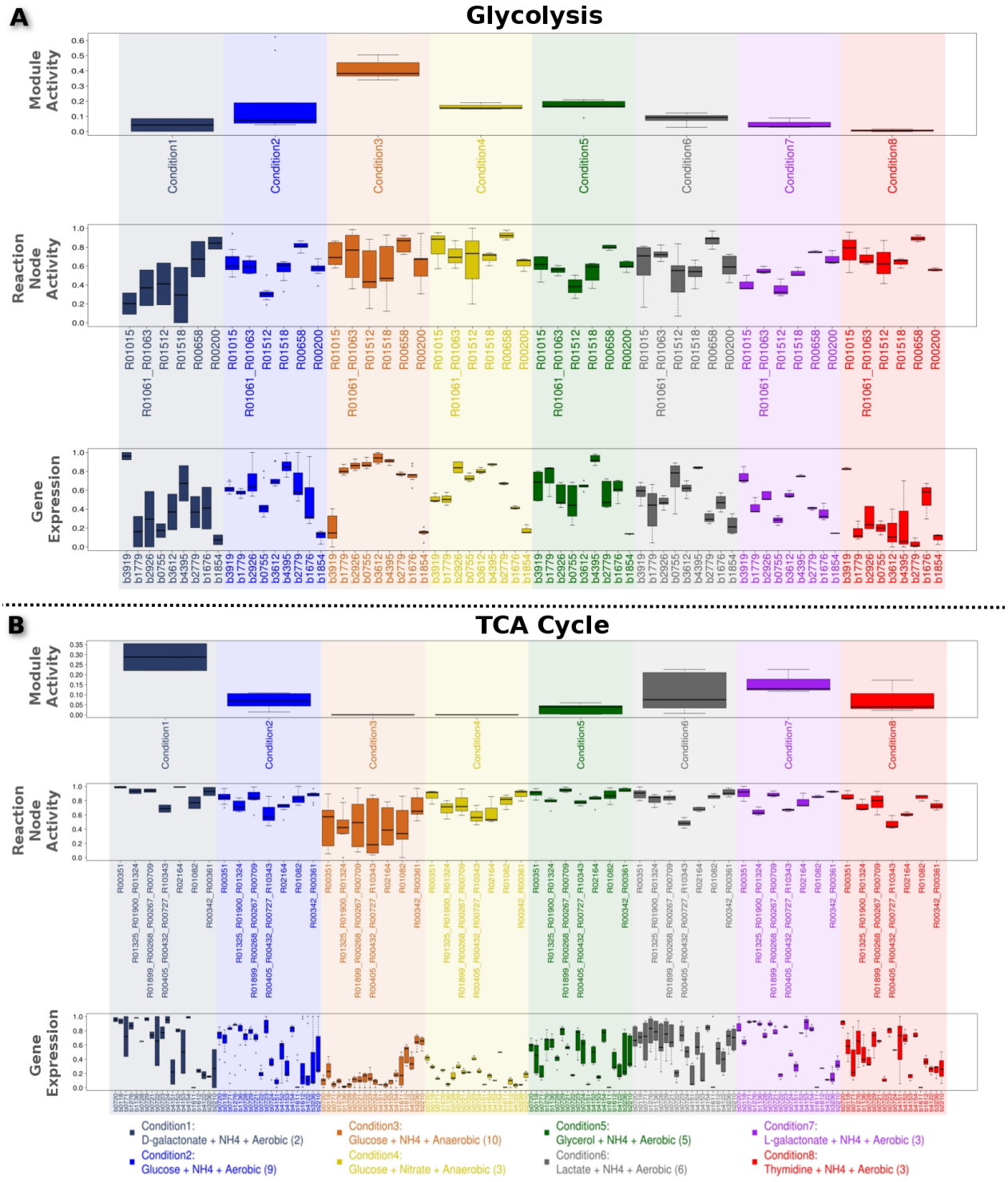


Figure 3.12: Two examples of module activity in E. coli growing under different conditions. Colours are showing different conditions. From navy blue to red; condition 1,2,3,4,5,6,7,8 (details in Table 3.3). Module activity, reaction node activity and gene expression values are given for two modules, M00002 and M00009. A module can be composed of several reactions and genes, due to this reason, from up to down the number of boxplots are getting more per condition. Y-axis shows the level of module activity, reaction node activity gene expression values. A) E. coli samples which grown either aerobically or anaerobically with glucose present high activity of glycolysis / M00002 (especially condition 3, but also 2 and 4). B) TCA-cycle / M00009 is almost inactive for anaerobic conditions, while shows activity in aerobic conditions (conditions 3 and 4).

		Initial Metabolite				Final Metabolite					
Module	Cancer type	Metabolite	Normal Median (SE)	Cancer Median (SE)	Fold Change	Metabolite	Normal Median (SE)	Cancer Median (SE)	Fold Change	Model prediction	Validation
M00035_1	*BRCA*	*C00073*	0.79 (0.08)	1.47 (0.1)	**1.9**	*C02291*	0.11 (0.03)	0.87 (0.87)	**7.8**	Up	**Correct**
	KIRC		1.49 (0.06)	0.68 (0.04)	**0.5**		NA	NA	NA	Down	NA
M00100	*BRCA*	*C00319*	0.08 (0.12)	0.92 (1.76)	**11.1**	*C00346*	0.4 (0.18)	6.85 (2.08)	**17.3**	Up	**Correct**
	KIRC		0.57 (0.09)	1.28 (0.29)	**2.2**		1.06 (0.03)	0.9 (0.05)	**0.8**	Down	**Correct**
M00135	*BRCA*	*C00134*	0.09 (0.06)	1.14 (0.9)	**12.9**	*C00334*	0.27 (0.02)	0.33 (0.11)	**1.2**	Down	**Correct**
	KIRC		0.55 (0.11)	2.11 (0.63)	**3.6**		2.19 (0.41)	0.65 (0.05)	**0.3**	Down	**Correct**

Table 3.7: Fold changes of metabolites from metabolomics data in BRCA and KIRC, and fold changes of predicted module activities.

Modules present the chain of biochemical reactions that transform of simple to complex metabolites or vice versa. The final metabolite of a module can be produced by different modules and also can be consumed by different biochemical reactions. Due to these reasons, the module activity can not be inferred as a direct estimate of the concentration of the final metabolite. However, we may expect to observe some correlation between assay based metabolite measurements and module activity predictions if the proposed model is accurate.

With the premise that both the initial and final metabolites of each module were measured in these studies, we found two datasets where three modules could be evaluated. For these settings, as it is shown in Table 3.7, all of our five predictions proved to be correct. Therefore, the predictions from our study are generally transferable metabolic activity levels.

Module	Main Category	Description/Name	Start molecule	End molecule
M00001	CH-LPD	Glycolysis (Embden-Meyerhof pathway), glucose => pyruvate	C00267	C00022
M00002	CH-LPD	Glycolysis, core module involving three-carbon compounds	C00111	C00022
M00003	CH-LPD	Gluconeogenesis, oxaloacetate => fructose-6P	C00036	C05345
M00004	CH-LPD	Pentose phosphate pathway (Pentose phosphate cycle)	C01172	Cycle: Pentose phosphate pathway
M00006	CH-LPD	Pentose phosphate pathway, oxidative phase, glucose 6P => ribulose 5P	C01172	C00199
M00007	CH-LPD	Pentose phosphate pathway, non-oxidative phase, fructose 6P => ribose 5P	C05345+C00118	C00117
M00009	CH-LPD	Citrate cycle (TCA cycle, Krebs cycle)	C00024+C00036	Cycle: Citrate cycle
M00010	CH-LPD	Citrate cycle, first carbon oxidation, oxaloacetate => 2-oxoglutarate	C00036+C00024	C00026
M00011	CH-LPD	Citrate cycle, second carbon oxidation, 2-oxoglutarate => oxaloacetate	C00026+C15972	C00036
M00013	CH-LPD	Malonate semialdehyde pathway, propanoyl-CoA => acetyl-CoA	C00100	C00024
M00014	CH-LPD	Glucuronate pathway (uronate pathway)	C00029	C00231
M00015	NUC-AA	Proline biosynthesis, glutamate => proline	C00025	C00148
M00020	NUC-AA	Serine biosynthesis, glycerate-3P => serine	C00197	C00065
M00027	NUC-AA	GABA (gamma-Aminobutyrate) shunt	C00025	C00042
M00029_1	NUC-AA	Urea cycle	C00014+C00049	C00122
M00029_2	NUC-AA	Urea cycle	C00014+C00049	C00086

M00032	NUC-AA	Lysine degradation, lysine => saccharopine => acetoacetyl-CoA	C00047	C00332
M00034_1	NUC-AA	Methionine salvage pathway	C00073	C00147
M00035_1	NUC-AA	Methionine degradation	C00073+C00065	C02291
M00036	NUC-AA	Leucine degradation, leucine => acetoacetate + acetyl-CoA	C00123	C00164
M00037	NUC-AA	Melatonin biosynthesis, tryptophan => serotonin => melatonin	C00078	C01598
M00042	NUC-AA	Catecholamine biosynthesis, tyrosine => dopamine => noradrenaline => adrenaline	C00082	C00788
M00043	NUC-AA	Thyroid hormone biosynthesis, tyrosine => triiodothyronine/thyroxine	C00082	C02465
M00044_1	NUC-AA	Tyrosine degradation, tyrosine => homogentisate	C00082	C00122
M00046_1	NUC-AA	Pyrimidine degradation, uracil => beta-alanine, thymine => 3-aminoisobutanoate	C00106	C00099
M00046_2	NUC-AA	Pyrimidine degradation, uracil => beta-alanine, thymine => 3-aminoisobutanoate	C00178	C05145
M00047	NUC-AA	Creatine pathway	C00062	C00791
M00048	NUC-AA	Inosine monophosphate biosynthesis, PRPP + glutamine => IM	C00119+C00064	C00130
M00049	NUC-AA	Adenine ribonucleotide biosynthesis, IMP => ADP,ATP	C00130	C00002
M00050	NUC-AA	Guanine ribonucleotide biosynthesis IMP => GDP,GTP	C00130	C00044
M00051_1	NUC-AA	Uridine monophosphate biosynthesis, glutamine (+ PRPP) => UMP	C00064+C00119	C00105
M00052	NUC-AA	Pyrimidine ribonucleotide biosynthesis, UMP => UDP/UTP,CDP/CTP	C00105	C00112
M00055	CH-LPD	N-glycan precursor biosynthesis	C00110	G00008
M00056_1	CH-LPD	O-glycan biosynthesis, mucin type core	C02189+G10611	G00025
M00056_2	CH-LPD	O-glycan biosynthesis, mucin type core	C02189+G10611	G00029
M00056_3	CH-LPD	O-glycan biosynthesis, mucin type core	C02189+G10611	G00031
M00057	CH-LPD	Glycosaminoglycan biosynthesis, linkage tetrasaccharide	C02189	G00157
M00058	CH-LPD	Glycosaminoglycan biosynthesis, chondroitin sulfate backbone	G00157	G00160
M00059	CH-LPD	Glycosaminoglycan biosynthesis, heparan sulfate backbone	G00157	G00164
M00065	CH-LPD	GPI-anchor biosynthesis, core oligosaccharide	C01194	G13044
M00066	CH-LPD	Lactosylceramide biosynthesis	C00195	C01290
M00067_1	CH-LPD	Cerebroside and sulfatide biosynthesis	C00195	C06125
M00067_2	CH-LPD	Cerebroside and sulfatide biosynthesis	C03201	C20825
M00068	CH-LPD	Glycosphingolipid biosynthesis, globo-series, LacCer => Gb4Cer	G00092	G00094
M00069	CH-LPD	Glycosphingolipid biosynthesis, ganglio series, LacCer => GT3	G00092	G00118
M00070	CH-LPD	Glycosphingolipid biosynthesis, lacto-series, LacCer => Lc4Cer	G00092	G00037
M00071	CH-LPD	Glycosphingolipid biosynthesis, neolacto-series, LacCer => nLc4Cer	G00092	G00050
M00073	CH-LPD	N-glycan precursor trimming	G00009	G00012
M00075_1	CH-LPD	N-glycan biosynthesis, complex type	G00013	G00019
M00075_2	CH-LPD	N-glycan biosynthesis, complex type	G00013	G00022
M00075_3	CH-LPD	N-glycan biosynthesis, complex type	G00013	G00018
M00076	CH-LPD	Dermatan sulfate degradation	C00426	G00872
M00077	CH-LPD	Chondroitin sulfate degradation	G12336	G00872
M00078	CH-LPD	Heparan sulfate degradation	C00925	G02632
M00079	CH-LPD	Keratan sulfate degradation	C00573	G01391
M00082	CH-LPD	Fatty acid biosynthesis, initiation	C00024	C05744
M00083	CH-LPD	Fatty acid biosynthesis, elongation	C03939	C05745
M00085	CH-LPD	Fatty acid biosynthesis, elongation, mitochondria	C00024	C00040
M00087	CH-LPD	beta-Oxidation	C00154	C02593
M00089	CH-LPD	Triacylglycerol biosynthesis	C00093	C00422
M00090	CH-LPD	Phosphatidylcholine (PC) biosynthesis, choline => PC	C00114	C00157

M00091	CH-LPD	Phosphatidylcholine (PC) biosynthesis, PE => PC	C00350	C00157
M00092	CH-LPD	Phosphatidylethanolamine (PE) biosynthesis, ethanolamine => PE	C00189	C00350
M00094_1	CH-LPD	Ceramide biosynthesis	C00154+C00065	C00195
M00095	CH-LPD	C5 isoprenoid biosynthesis, mevalonate pathway	C00024	C00235
M00098	CH-LPD	Acylglycerol degradation	C00422	C00116
M00099	CH-LPD	Sphingosine biosynthesis	C00154	C00319
M00100	CH-LPD	Sphingosine degradation	C00319	C00346
M00101	CH-LPD	Cholesterol biosynthesis, squalene 2,3-epoxide => cholesterol	C01054	C00187
M00103	CH-LPD	Cholecalciferol biosynthesis	C01164	C01673
M00104_1	CH-LPD	Bile acid biosynthesis, cholesterol => cholate	C00187	C00695
M00104_2	CH-LPD	Bile acid biosynthesis, cholesterol => cholate	C00187	C02528
M00106_1	CH-LPD	Conjugated bile acid biosynthesis, cholate => taurocholate/glycocholate	C00695	C05122
M00106_2	CH-LPD	Conjugated bile acid biosynthesis, cholate => taurocholate/glycocholate	C00695	C01921
M00107	CH-LPD	Steroid hormone biosynthesis, cholesterol => prognenolone => progesterone	C00187	C00410
M00108	CH-LPD	C21-Steroid hormone biosynthesis, progesterone => corticosterone/aldosterone	C00410	C01780
M00109	CH-LPD	C21-Steroid hormone biosynthesis, progesterone => cortisol/cortisone	C00410	C00762
M00110	CH-LPD	C19/C18-Steroid hormone biosynthesis, pregnenolone => androstenedione => estrone	C01953	C00468
M00118	NUC-AA	Glutathione biosynthesis, glutamate => glutathione	C00025	C00051
M00120	NUC-AA	Coenzyme A biosynthesis, pantothenate => CoA	C00864	C00010
M00128	NUC-AA	Ubiquinone biosynthesis, eukaryotes, 4-hydroxybenzoate => ubiquinone	C00156	C00399
M00130	CH-LPD	Inositol phosphate metabolism, PI=> PIP2 => Ins(1,4,5)P3 => Ins(1,3,4,5)P4	C01194	C01272
M00131	CH-LPD	Inositol phosphate metabolism, Ins(1,3,4,5)P4 => Ins(1,3,4)P3 => myo-inositol	C01272	C00137
M00132	CH-LPD	Inositol phosphate metabolism, Ins(1,3,4)P3 => phytate	C01243	C01204
M00133_1	NUC-AA	Polyamine biosynthesis, arginine => agmatine => putrescine => spermidine	C00062+C00019	C00315
M00134	NUC-AA	Polyamine biosynthesis, arginine => ornithine => putrescine	C00062	C00134
M00135	NUC-AA	GABA biosynthesis, eukaryotes, putrescine => GABA	C00134	C00334
M00141	NUC-AA	C1-unit interconversion, eukaryotes	C00101	Cycle: C1-unit interconversion
M00338_1	NUC-AA	Cysteine biosynthesis, homocysteine + serine => cysteine	C00065+C00155	C00097
M00367	NUC-AA	C10-C20 isoprenoid biosynthesis, non-plant eukaryotes	C00129	C00353
M00415	NUC-AA	Fatty acid biosynthesis, elongation, endoplasmic reticulum	C00083	C20876
M00549	CH-LPD	Nucleotide sugar biosynthesis, glucose => UDP-glucose	C00267	C00029
M00554	CH-LPD	Nucleotide sugar biosynthesis, galactose => UDP-galactose	C00124	C00052
M00632_1	CH-LPD	Galactose degradation, Leloir pathway, galactose => alpha-D-glucose-1P	C00124	C00103
M00741	CH-LPD	Propanoyl-CoA metabolism, propanoyl-CoA => succinyl-CoA	C00100	C00091

CH-LPD: Carbohydrate and lipid metabolism; **NUC-AA**: Nucleotide and amino acid metabolism

Table 3.8: Metabolic modules used in this study.

Chapter 4

A pan-cancer metabolic landscape based on gene expression integration into pathway modules

Chapter 4 is adapted from the following publication: “**Cubuk C**, Hidalgo MR, Amadoz A, et al. (2018). Gene expression integration into pathway modules reveals a pan-cancer metabolic landscape. Cancer Research, 78(21), 6059-6072. DOI: 110.1158/0008-5472.CAN-17-2705”. This chapter can be also considered as a case study which efforts to accomplish modelling the metabolism of cancer using mathematical models given in the previous chapters.

Chapter 4
THE PAN-CANCER METABOLIC LANDSCAPE

4.1 Overview and objectives

Metabolic reprogramming plays an important role in cancer development and progression and is a well-established neoplastic hallmark. Cancer cells need to adapt their metabolism to survive and proliferate under the metabolically compromised conditions provided by the tumour microenvironment. It is well known that the common proliferative phenotype of cancer cells require intensive support for the biosynthesis of cellular components and generation of energy, which overall are accomplished by reprogramming of metabolism. Warburg effect, enhanced aerobic glycolysis, is one of the well-known reprogramming observation done almost one century ago. Additionally, alterations in the synthesis of nucleotides, amino acids and lipids [119], mutations in metabolic genes and accumulations of key metabolites [120] have been reported. These observations, along with the discovery of the therapeutic potential of metabolic targets in cancer [121], has sparked a growing interest in cancer metabolism [122, 123]. Recent studies show that genes involved in metabolic pathways display a remarkable heterogeneity across various cancer types [124], which suggests that personalized therapies are likely to be successful if the context of the intervention is accurately depicted. In this context, synthetic lethality, defined as combined molecular perturbations with a drastic effect on cell viability, but with no individual effect, offers a promising range of potential therapeutic interventions based on cancer metabolic dependencies [125]. Recent studies have demonstrated that complex phenotypes or outcomes such as patient survival [37, 57] and drug activity [38] are better predicted by the inferred activity of pathways, than by the activity of their constituent genes and/or proteins. Indeed, mechanistic models of signal propagation have been successfully applied to predict complex phenotypes using estimates of signalling pathway activities inferred from gene expression data [38, 57], chapter 2 of this thesis]. In addition, such models provide important information about disease mechanisms and mode of action (MoA) of drugs [57]. This approach successfully extended to metabolism in the context of metabolic modules (Chapter 3). Here we generalize the application of this approach to describe the metabolic profiles and dependencies across 14 cancer types. This chapter reveals common and specific metabolic modules that influence patient survival and also identify metabolic dependencies based on targeted molecular predictions that point to novel therapeutic interventions.

4.2 Materials and methods

4.2.1 Data resources and processing

RNA-seq counts for a total of 9428 samples, 8319 corresponding to cancer and 649 to healthy reference tissue, belonging to 25 cancer types, (see Table 4.1), as well as their subtype stratification, were downloaded from the International Cancer Genome Consortium (ICGC) repository (https://dcc.icgc.org/releases/release_20/Projects). The trimmed mean of M-values (TMM) normalization method [62] was used for gene expression normalization. Expression data on responses to drugs were taken from GSE25066, GSE50948, GSE5462 datasets downloaded from GEO. Probes mapping in more than one gene were discarded. The median value of the probes mapping on a gene was used as the expression value for this gene. Microarray data normalization and background correction were done using the RMA method implemented in the affy Bioconductor package (https://bioconductor.org/packages/release/bioc/html/affy.html). Normalized expression datasets, both microarray and RNA-seq, were log-transformed and truncation by quantile 0.99 was applied. The COMBAT method [61] was used for batch effect correction. Finally, data were re-scaled between 0 and 1. Somatic mutation data, in MAF format as the output of MuTect2 variant aggregation and masking workflow, were taken from the CDG cancer portal; https://portal.gdc.cancer.gov/files/995c0111-d90b-4140-bee7-3845436c3b42 and https://portal.gdc.cancer.gov/files/da904cd3-79d7-4ae3-b6c0-e7127998b3e6 for BRCA and GBM, respectively. Cell line expression values and cell line survival data were taken from CCLE and Achilles projects. A total of 212 cell lines were used in this study (Supplementary Table S2, Cubuk et al. Cancer Research., 2018 [85]). Gene expression data were taken from the Cancer Cell Line Encyclopedia (https://portals.broadinstitute.org/ccle/) and preprocessed as given above. Cell survival measurements after gene KD were taken from the Project Achilles 2.4.3 (https://portals.broadinstitute.org/achilles/datasets/5/download) [126]. Survival validation data were taken from the new release 2.20.2 of the project Achilles (https://portals.broadinstitute.org/achilles/datasets/15/download) [94]. Clinical data were available through the cBIOportal (http://www.cbioportal.org/) [127]. These data included individual survival information that was used for survival analysis.

Cancer type	Abbr.	Tumor	Normal	Metastasis	Unknown	Alive	Deceased	Analysis type
Bladder Urothelial Carcinoma	**BLCA**	301	17	-	-	161	134	Diff/Cor/KM/Cox/Nor
Breast Invasive Carcinoma	**BRCA**	1057	113	7	-	900	146	Diff/Cor/KM/Cox/Nor
Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma	**CESC**	259	3	2	-	195	65	KM/Cox/Nor
Chronic Lymphocytic Leukemia	**CLLE**	311	-	-	-	-	-	Nor
Colon Adenocarcinoma	**COAD**	451	41	1	1	335	93	Diff/Cor/KM/Cox/Nor
Glioblastoma Multiforme	**GBM**	153	-	-	13	32	125	KM/Cox/Nor
Head and Neck Squamous Cell Carcinoma	**HNSC**	480	42	2	-	271	210	Diff/Cor/KM/Cox/Nor
Kidney Renal Clear Cell Carcinoma	**KIRC**	526	72	-	-	345	173	Diff/Cor/KM/Cox/Nor
Kidney Renal Papillary Cell Carcinoma	**KIRP**	222	32	-	-	189	32	Diff/Cor/KM/Cox/Nor
Acute Myeloid Leukemia	**LAML**	173	-	-	-	59	114	KM/Cox/Nor
Brain Lower Grade Glioma	**LGG**	439	-	-	14	340	109	KM/Cox/Nor
Liver Hepatocellular Carcinoma	**LIHC**	294	48	-	3	184	112	Diff/Cor/KM/Cox/Nor
Lung Adenocarcinoma	**LUAD**	486	55	-	2	305	173	Diff/Cor/KM/Cox/Nor
Lung Squamous Cell Carcinoma	**LUSC**	428	45	-	-	239	188	Diff/Cor/KM/Cox/Nor
Malignant Lymphoma	**MALY**	97	-	-	7	-	-	Nor
Ovarian Serous Cystadenocarcinoma	**OV**	342	-	-	34	107	160	KM/Cox/Nor
Pancreatic Adenocarcinoma	**PAAD**	142	3	-	-	64	77	KM/Cox/Nor
Pancreatic Cancer	**PACA**	83	-	-	10	-	-	Nor
Pancreatic Endocrine Neoplasms	**PAEN**	32	-	-	1	-	-	Nor
Prostate Adenocarcinoma	**PRAD**	379	52	-	-	367	7	Diff/Cor/KM/Cox/Nor
Rectum Adenocarcinoma	**READ**	153	9	-	1	126	26	Diff/Cor/KM/Cox/Nor
Skin Cutaneous melanoma	**SKCM**	80	1	353	-	54	28	KM/Cox/Nor
Gastric Adenocarcinoma	**STAD**	415	35	-	-	251	163	Diff/Cor/KM/Cox/Nor
Thyroid Papillary Carcinoma	**THCA**	500	58	8	-	488	16	Diff/Cor/KM/Cox/Nor
Uterine Corpus Endometrial Carcinoma	**UCEC**	516	23	-	1	421	88	Diff/Cor/KM/Cox/Nor
Total number of individuals	***9428***	***8319***	***649***	***373***	***87***	*5433*	*2239*	

Table 4.1: Cancer types used in this study and specific type of analysis in which the cancer was used. **Diff**: Differential module activity analysis, **Cor**: Cooperation of metabolic modules, **KM**: Kaplan-Meier, **Cox**: Cox multiple regression analysis, **Nor**: Batch effect correction and normalization.

4.2.2 Differential module activity estimation

Activity values for the modules were calculated using Metabolizer web tool, http://metabolizer.babelomics.org. Details of the method for module activity and the web tool are given in Chapter 3. The Wilcoxon test is used to assess the significance of the observed changes in module activity when samples of two conditions are compared. Since many modules were simultaneously tested, the popular FDR method [88] was used to correct for multiple testing effects.

4.2.3 Survival analysis

Kaplan-Meier (K-M) curves were used to relate module activity to patient survival in different cancers. The value of the activity estimated for each module in each individual was used to assess its relationship with individual patient survival. Calculations were carried out using survdiff function in the survival package of R (https://cran.r-project.org/web/packages/survival/). Cox regression analysis [128] was used to relate combined module activity to survival in the different cancers. Calculations were carried out using the coxph function in the survival of R package (https://cran.r-project.org/web/packages/survival/). A stepwise algorithm implemented in the step function of the stats R package (https://stat.ethz.ch/R-manual/R-devel/library/stats/html/step.html), was used to add or remove modules according to the significance of their contributions to explain survival in the multiple regression model. The step function uses the Akaike Information Criterion (AIC) to select the best model by iteratively adding and removing variables. Finally, the method yields a list of modules whose combination is significantly related to survival. Adjustment for multiple testing was made by the FDR method [88]. For patient stratification, based on predicted module activity, high and low module activity groups were defined using the 80% and 20% percentiles, respectively.

4.2.4 Module essentiality

4.2.4.1 Simulation of the effect of gene knockdowns on module activity

Given a set of gene expression values (wt expression), the activity of the modules was estimated as described above (wt activity). Then, the knocked down gene(s) expression value(s) were set to 0.001 (KD expression) and the activity of the modules was recalculated again (KD activity). The log-fold-change in module activities was then calculated from the comparison of KD and wt module activity profiles as;

Log-fold-change = log(KD module activity) - log(wt module activity)

4.2.4.2 Relationship between module activity and cell survival

To estimate module activity essentiality, cell lines were grouped by cancer type. For each cancer type, the impact of gene KDs on the activity of the modules was calculated as described above. Then, a Spearman correlation coefficient between log-fold-change values and cell survival, as described in the Project Achilles was calculated. Lower Achilles scores indicate higher mortality and, consequently, the essentiality of the KD gene. Positive correlations indicate essentiality in

module activity (the less activity the lower the Achilles index) in this particular cancer type.

4.2.5 Validation of the essentiality predictions

4.2.5.1 Independent dataset validation

Data on cancer dependencies, which include estimates of cell viabilities after gene KD, from the Project Achilles 2.20.2 was used to check the validity of the predictions made with the Project Achilles 2.4.3. It was expected that the inhibition of an onco-module would reduce the viability of cancer cells. Conversely, the inhibition of a tumour suppressor module should result in greater cell survival. In order to detect these increases or decreases, the Project Achilles 2.20.2 cell viability scores observed in the cell line in which an effect of KD on cell survival was predicted were compared with the scores reported for the other cell lines (background score). Increases or decreases in the mean values were taken as evidence of predicted effects on cell viability.

4.2.5.2 Experimental validation

The shRNAs targeting UPB1 were purchased from the MISSION library (Sigma Aldrich), catalogue SHCLNG-NM_016327. Lentivirus was produced and transduced following standard protocols and cell cultures were selected with puromycin for 72 hours before cell seeding for evaluation of proliferation/viability by methylthiazol tetrazolium (MTT)-based assays (Sigma-Aldrich). The data corresponds to sextuplicates and were replicated in different assays. UPB1 expression was detected with the Human Protein Atlas HPA000728 antibody (Sigma-Aldrich) and gene expression measured with primers 5´-TCGACCTAAACCTCTGCCAG-3’ and 5’-TAAGCCTGCCACACTTGCTA-3’, using PPP1CA as control.

4.3 Results

4.3.1 Data pre-processing

RNA-seq counts for 25 cancer types, totalling 9428 samples (Table 4.1) were downloaded from The International Cancer Genome Consortium (ICGC) repository. Principal component analysis (PCA) was used to detect possible batch effects. The results are shown by plotting samples with respect to disease status (Figure 4.1 A and B), sequencing centre (Figure 4.1 C and D) and project (Figure 4.1 E and F). An appreciable technical batch effect due to the sequencing centre was found (Figure 4.1 C) and this was corrected by application of the COMBAT [61] method (Figure 4.1 D). Samples were normalized and preprocessed as explained in Methods.

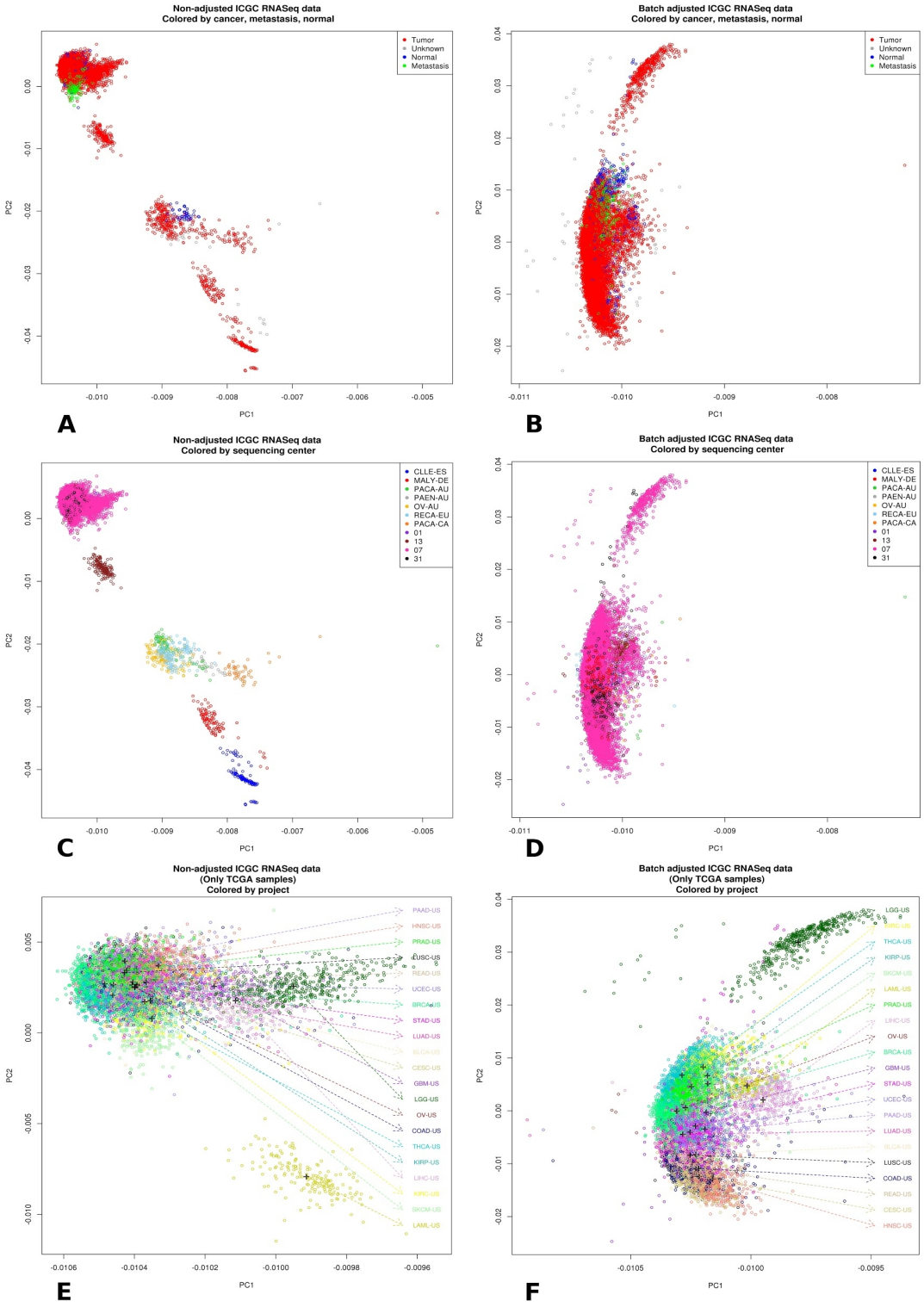


Figure 4.1: PCA plots for detecting batch effects.

4.3.2 Pan-cancer metabolic activity profiles

For this differential module activity analysis we used 14 cancer types in which at least a 5% of healthy reference samples were available (totalling 6299 cancer samples and 687 healthy samples). For each cancer type, the activity of the modules was calculated for all tumours and for

all healthy tissue samples as described in Methods. Briefly, gene expression profiles were converted into metabolic module activity profiles by applying formula (Figure 3.2) that takes into account the chain of metabolic reactions required to complete the transformation of simple into complex metabolites in each module. Next, the Wilcoxon test was used to assess differences between conditions. Figure 4.2 shows the significant activations and deactivations of modules in tumours with respect to the corresponding healthy tissue.

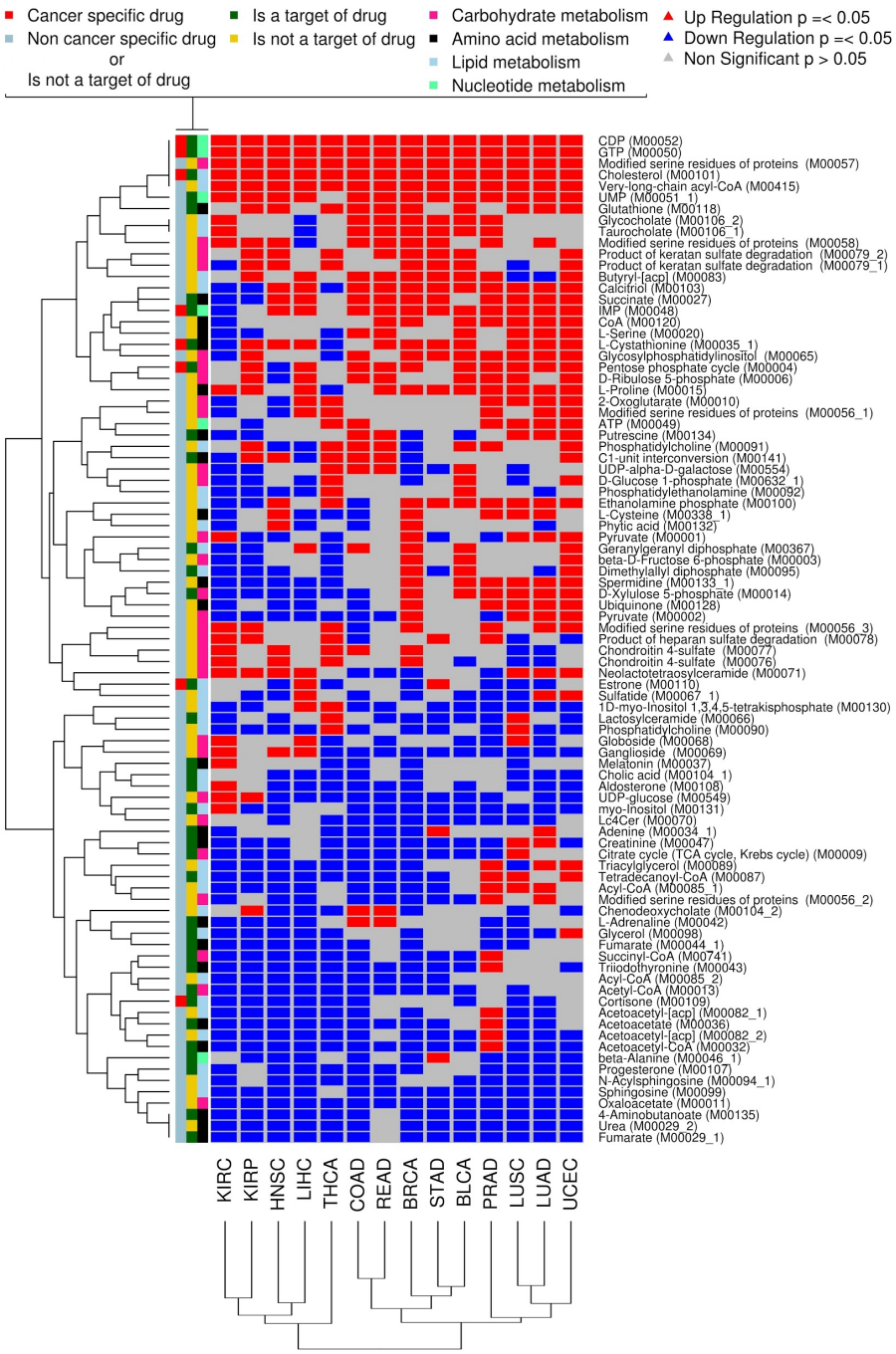


Figure 4.2: Heatmap with the significant (FDR-adjusted P < 0.05) changes in module activity when the 14 cancers analyzed were compared with the corresponding tissue of origin. Activity upregulation is represented in red and downregulation in blue. The left-most column represents modules in which one or several gene products are targets of cancer drugs; the second column represents modules in which one or several gene products are targets of other types of drugs; the third column represents the general metabolic categories: carbohydrate (CH), amino acid (AA), lipid (LP), or nucleotide (NT).

4.3.3 Metabolic modules may be altered by oncogenic mutations

Many cancer drivers are known to promote metabolic reprogramming in cancer. To test our predictions in this context, we analyzed the impact of relatively frequent oncogenic mutations in well-established drivers linked to metabolic reprogramming; AKT1 and PIK3CA in BRCA, and IDH1 in GBM (Table 4.2). Most of the variants were causing missense and in-frame shift sequence changes without clear evidence of pathogenicity or loss of function effect. The intronic, intergenic, non-coding exon, non-coding transcript and synonymous variants were excluded from the analysis (Supplementary Tables 8B, Cubuk et al. Cancer Research., 2018 [85]). The module activity of the groups below were compared;

Breast invasive carcinoma (BRCA):

- AKT1 mutated samples vs. AKT1 not mutated samples
- PIK3CA mutated samples vs. PIK3CA not mutated samples

Glioblastoma multiforme (GBM):

- IDH1 mutated samples vs. IDH1 not mutated samples

Consistent with a major role linked to metabolism, mutations in PIK3CA caused significant changes in the predicted values of many metabolic modules, with coherent changes in seven of the modules also being significantly altered by AKT1 mutations (Supplementary Tables 8C, Cubuk et al. Cancer Research., 2018 [85]). Among the altered modules in PIK3CA mutants, some of the findings were consistent with current knowledge. The largest predicted metabolic activation in PIK3CA mutants is found to be the M00027 module of GABA shunt (end metabolite succinate), which is consistent with data of reprogrammed glutamine metabolism in this setting (Supplementary Tables 8C, Cubuk et al. Cancer Research., 2018 [85] and [129]). In contrast, the activity of the M00034_1 module of methionine salvage is predicted to be higher in PIK3CA wild-type tumours, but interestingly this pathway becomes activated as a mechanism of resistance to PI3K inhibitors (Supplementary Tables 8C, Cubuk et al. Cancer Research., 2018 [85] and [130]). The IDH1 mutations in GBM caused fewer module alterations, probably in part because only seven mutated samples were included in the analysis. Nonetheless, the largest impact upon IDH1 mutations is predicted to be activation of proline biosynthesis (M00015), which links to metabolites downstream of IDH1 activity. Another predicted effect was the downregulation of glycosphingolipid biosynthesis (M00071), which is consistent with the major demand of citrate towards the substrate of the reaction catalyzed by IDH1. In turn, the major activation corresponds to components downstream of its activity that is related to proline biosynthesis (M00015) (Supplementary Tables 8C, Cubuk et al. Cancer Research., 2018 [85]). Thus, the predictions from this study may also support the identification of specific, cancer driver-linked, metabolic reprogramming and/or vulnerabilities.

Cancer Type	Gene	Samples with mutation	Samples without mutation
BRCA	*AKT1*	28	1020
	PIK3CA	307	741
GBM	*IDH1*	7	141

Table 4.2: Number of samples in each group that were used to test the impact of mutations over metabolic module activities.

4.3.4 Cooperation between metabolic modules

Metabolic modules do not function in isolation, but rather display highly correlated (positive or negative) patterns of activity that influence cancer development and/or progression [124]. However, how these correlations vary from normal tissue to cancer is poorly understood. Our results document a variable proportion of modules, ranging from 5.3% (in LIHC) to 26.9% (in READ), that are significantly positively correlated in normal tissue but not in the corresponding tumour. This proportion is smaller for negative correlations, ranging from 1.1% (in LUSC) to 10.6% (in BLCA). 10-35% of the activity of metabolic modules is uncoupled when normal and cancer metabolic activities are compared (Figure 4.3). Figure 4.4 represents in detail the modules whose activities are correlated in normal and/or cancer tissue and those in which the significance or direction of the correlation change.

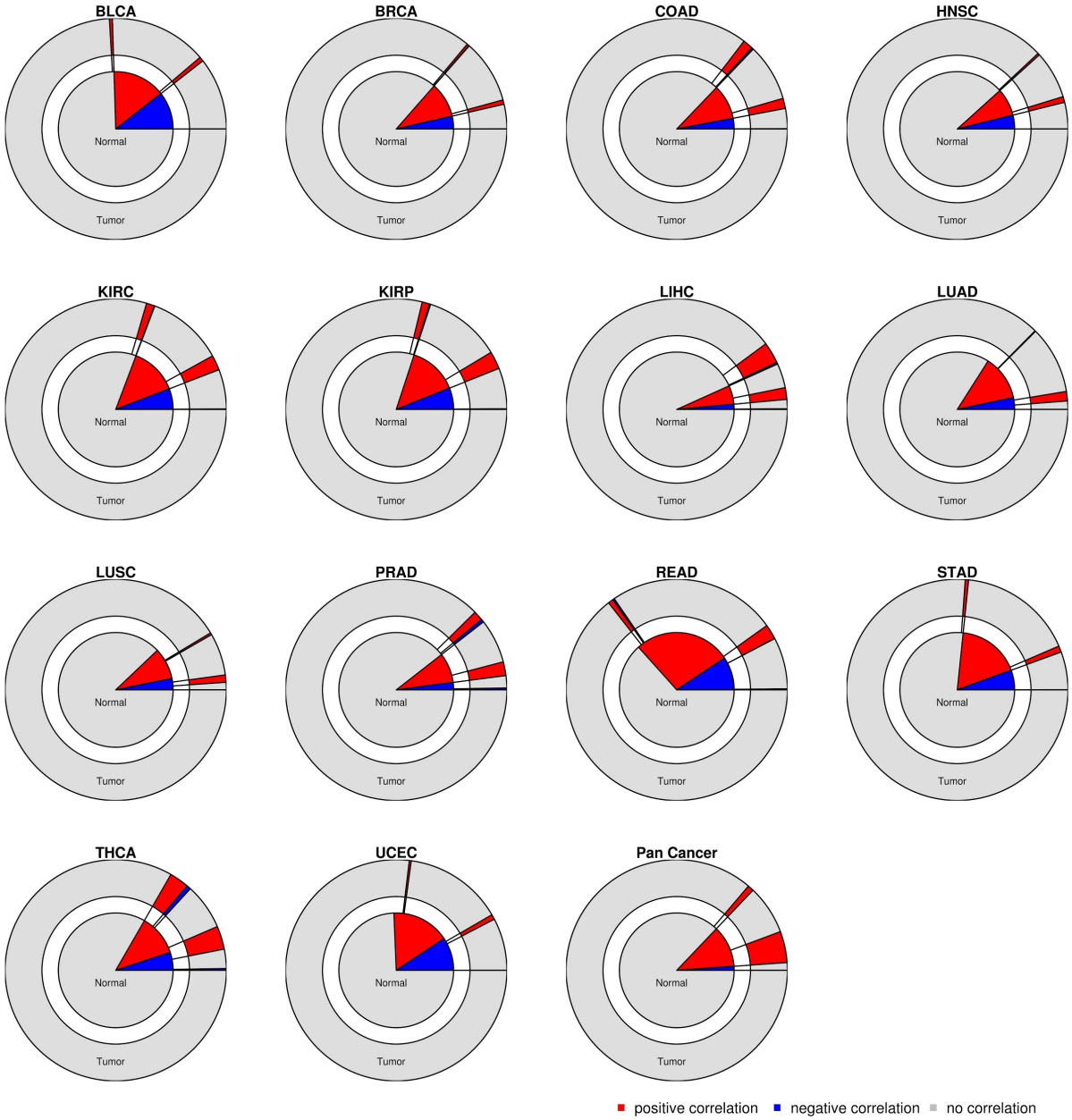


Figure 4.3: Cooperation between metabolic modules. Changes in correlations of module activities from the normal tissue (the inner circle) to the corresponding cancer type (outer circle). The proportion of positive correlations in the activity of the modules is represented in red, while the proportion of negative correlations is represented in blue.

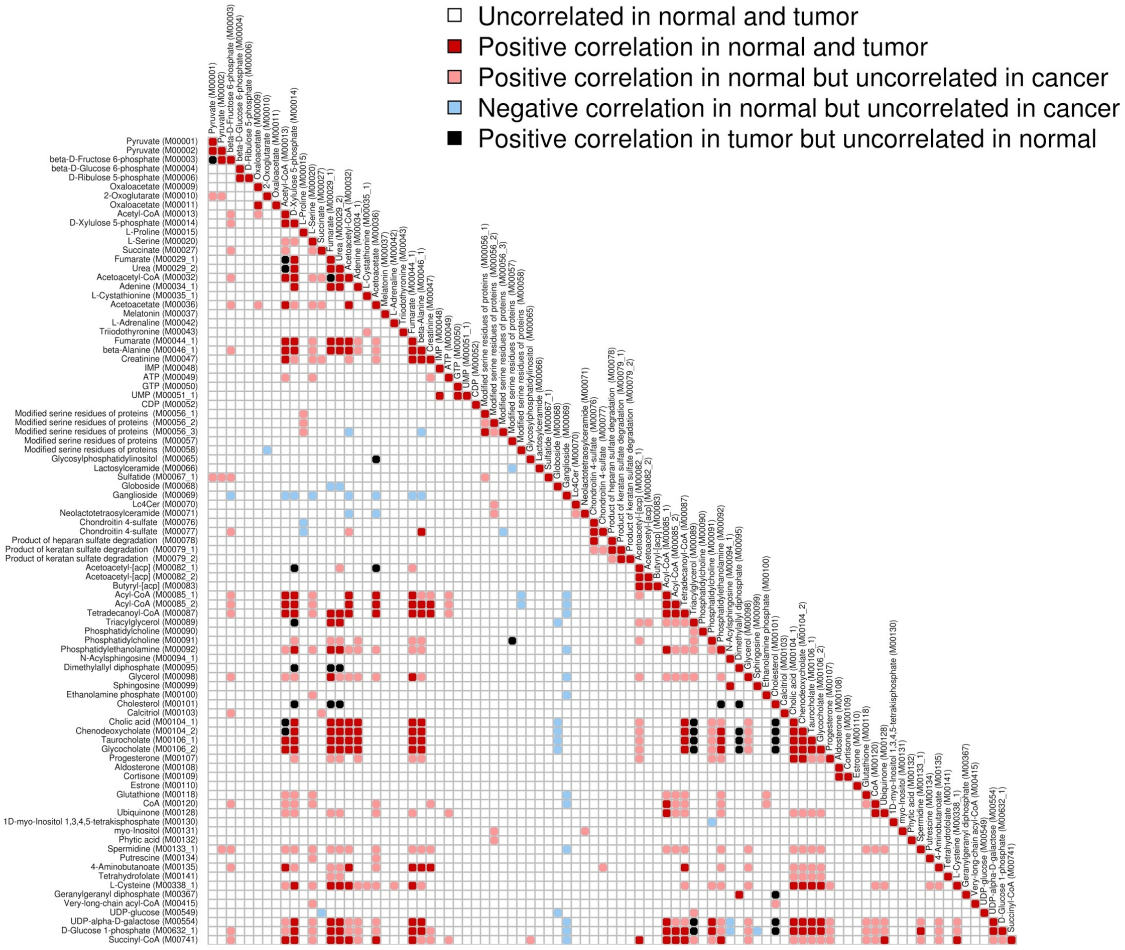


Figure 4.4: Detailed description of changes in pairwise module correlation. Colour code is as follows: pairs of modules with correlated activity both in normal tissues and in cancer in dark red, pairs of modules with activity correlated in normal tissues that lost the correlation in cancer in pale red, modules with negative correlation in their activities that lost the correlation in cancers in pale blue and modules uncorrelated in the normal tissue that appear correlated in cancer in black.

4.3.5 Modules associated with cancer outcome

Modules differentially activated between cancer and the corresponding normal tissue may highlight metabolic processes that are required for cancer development and/or progression. The availability of patient survival data in 21 cancer types (see Table 4.1) allows the identification of modules in which changes in activity are significantly associated with the progression of each cancer type. Supplementary Table 9 (Cubuk et al. Cancer Research., 2018 [85]) portrays the modules whose change in activity is significantly associated with poorer patient survival in at least one cancer type. Since the number of deceased patients and, in general, data on mortality follow-ups is limited in the ICGC repository, significant results were obtained for only a few modules. In particular, kidney (KIRC), liver (LIHC) and glioma (LGG) cancer types featured a remarkable number of modules influencing cancer outcome. Moreover, following from the observation of

correlated modules, the impact on survival may be further determined by combinations of their metabolic activities. Thus, we applied Cox multiple regression analysis [128] to find the combination of module activities that best accounted for patient survival. Supplementary Table 10 (Cubuk et al. Cancer Research., 2018 [85]) shows the combinations of module activities significantly related to survival in various cancer types. Previous results have shown that predicted activities of single or combined metabolic modules are associated with differences in cancer outcome, further emphasizing their fundamental role in cancer progression. In addition, we observed that the magnitude of their effect on survival was greater in some instances than for any of the individual activities of the genes that comprise a given module, which provides additional evidence that modules are real entities that contribute as whole units to cell functioning (Table 4.3).

Module	Module status	Module adj. p-value	Cancer type	Gene	Gene adj. p-value	Gene status
Malonate semialdehyde pathway	DOWN	1.80e-10	KIRC	EHHADH	1.1e-08	DOWN
Pyrimidine ribonucleotide biosynthesis	UP	3.50e-08	KIRC	NME1-NME2	1.1e-07	UP
C21-Steroid hormone biosynthesis	DOWN	3.50e-05	LGG	CYP17A1	4.7e-05	DOWN
Chondroitin sulfate degradation	DOWN	5.60e-05	KIRC	HYAL1	2.4e-04	DOWN
Citrate cycle, second carbon oxidation	DOWN	6.60e-05	KIRC	SDHD	1.0e-04	DOWN
Inositol phosphate metabolism	UP	1.40e-04	LGG	IPPK	0.02	DOWN
Pyrimidine degradation	DOWN	1.50e-04	KIRC	UPB1	4.1e-03	DOWN
Glycosphingolipid biosynthesis, neolacto-series	UP	2.30e-04	LUAD	B4GALT4	0.11	UP
Pentose phosphate pathway, oxidative phase	UP	4.80e-04	LGG	PGLS	0.76	UP
Glycosphingolipid biosynthesis, neolacto-series	UP	5.60e-04	KIRC	B4GALT2	2.7e-03	UP
Cerebroside and sulfatide biosynthesis	DOWN	6.50e-04	LGG	GAL3ST1	3.4e-03	DOWN
Glycolysis. core module involving three-carbon compounds	UP	2.20e-03	LIHC	TPI1	0.01	UP
Conjugated bile acid biosynthesis	DOWN	2.20e-03	LIHC	SLC27A5	9.6e-03	DOWN
Inositol phosphate	DOWN	2.20e-03	LIHC	INPP4B	9.6e-03	DOWN
Pyrimidine ribonucleotide biosynthesis	DOWN	2.50e-03	OV	NME3	0.33	DOWN
C19/C18-Steroid hormone biosynthesis	UP	3.60e-03	LIHC	CYP19A1	9.6e-03	UP
Inosine monophosphate biosynthesis	UP	3.60e-03	LIHC	ATIC	0.02	UP
Pentose phosphate pathway (Pentose phosphate cycle)	UP	4.70e-03	LGG	TKTL2	5.0e-03	DOWN
Inositol phosphate	DOWN	5.60e-03	BLCA	INPP1	0.03	DOWN
Heparan sulfate degradation	UP	5.90e-03	GBM	GNS	0.2	DOWN
Cholecalciferol biosynthesis	UP	6.30e-03	KIRC	CYP2R1	0.02	UP
Triacylglycerol biosynthesis	DOWN	6.40e-03	LIHC	PLPP1	9.6e-03	DOWN
GABA (gamma-Aminobutyrate) shunt	DOWN	7.70e-03	LGG	GAD1	0.02	DOWN
Ubiquinone biosynthesis, eukaryotes	DOWN	8.10e-03	LIHC	COQ7	0.03	DOWN
Cysteine biosynthesis	DOWN	8.30e-03	LIHC	CTH	0.11	DOWN
Propanoyl-CoA	DOWN	8.30e-03	LIHC	MCEE	9.6e-03	DOWN
Bile acid biosynthesis	DOWN	8.30e-03	LIHC	CYP27A1	9.6e-03	DOWN
Bile acid biosynthesis	DOWN	8.90e-03	LIHC	SLC27A5	9.6e-03	DOWN
Citrate cycle. first carbon oxidation	DOWN	0.01	LGG	CS	0.33	UP

Serine biosynthesis	UP	0.02	LIHC	PSPH	0.06	UP
Melatonin biosynthesis	DOWN	0.02	LGG	DDC	0.02	DOWN
Pentose phosphate pathway (Pentose phosphate cycle)	UP	0.02	LIHC	TKT	0.03	UP
Citrate cycle, second carbon oxidation	UP	0.02	LGG	SDHA	0.02	DOWN
Glycosphingolipid biosynthesis, globo-series	UP	0.02	LUAD	A4GALT	0.13	UP
Steroid hormone biosynthesis	UP	0.02	KIRC	CYP11A1	0.05	UP
Pyrimidine degradation	DOWN	0.02	LUAD	DPYS	0.07	DOWN
Pyrimidine ribonucleotide biosynthesis	UP	0.02	LIHC	AK9	0.03	DOWN
Methionine degradation	UP	0.03	UCEC	MAT2A	0.03	UP
Catecholamine biosynthesis	UP	0.03	UCEC	DDC	0.03	UP
Cysteine biosynthesis	UP	0.03	UCEC	CBS	0.09	UP
Pentose phosphate pathway (Pentose phosphate cycle)	UP	0.03	LUAD	RPE	0.07	UP
Citrate cycle, second carbon oxidation	DOWN	0.03	KIRP	FH	0.04	NA
beta-Oxidation	DOWN	0.03	LIHC	EHHADH	0.04	DOWN
Phosphatidylcholine (PC) biosynthesis	DOWN	0.03	LIHC	CHKA	0.12	UP
Guanine ribonucleotide biosynthesis	UP	0.03	KIRP	IMPDH2	0.55	UP
Catecholamine biosynthesis	DOWN	0.04	KIRP	PNMT	0.5	DOWN
GABA biosynthesis. eukaryotes	DOWN	0.04	UCEC	ALDH3A2	0.13	DOWN
Citrate cycle, second carbon oxidation	UP	0.04	UCEC	SDHC	0.16	UP
Glycolysis (Embden-Meyerhof pathway)	UP	0.04	LUAD	PGK2	0.05	UP
Propanoyl-CoA	DOWN	0.04	KIRP	PCCA	0.04	DOWN
O-glycan biosynthesis, mucin type core	DOWN	0.04	UCEC	GALNT2	0.09	UP
O-glycan biosynthesis, mucin type core	DOWN	0.04	UCEC	GALNT2	0.09	UP
Keratan sulfate degradation	UP	0.04	LIHC	GALNS	0.1	UP
Glycosphingolipid biosynthesis, globo-series	UP	0.04	KIRP	A4GALT	0.23	UP
Phosphatidylcholine (PC) biosynthesis	DOWN	0.04	UCEC	PEMT	0.05	DOWN
Glycolysis (Embden-Meyerhof pathway)	UP	0.05	HNSC	PKLR	0.2	DOWN

Table 4.3: Modules showing the strongest association with survival than any of their gene components.

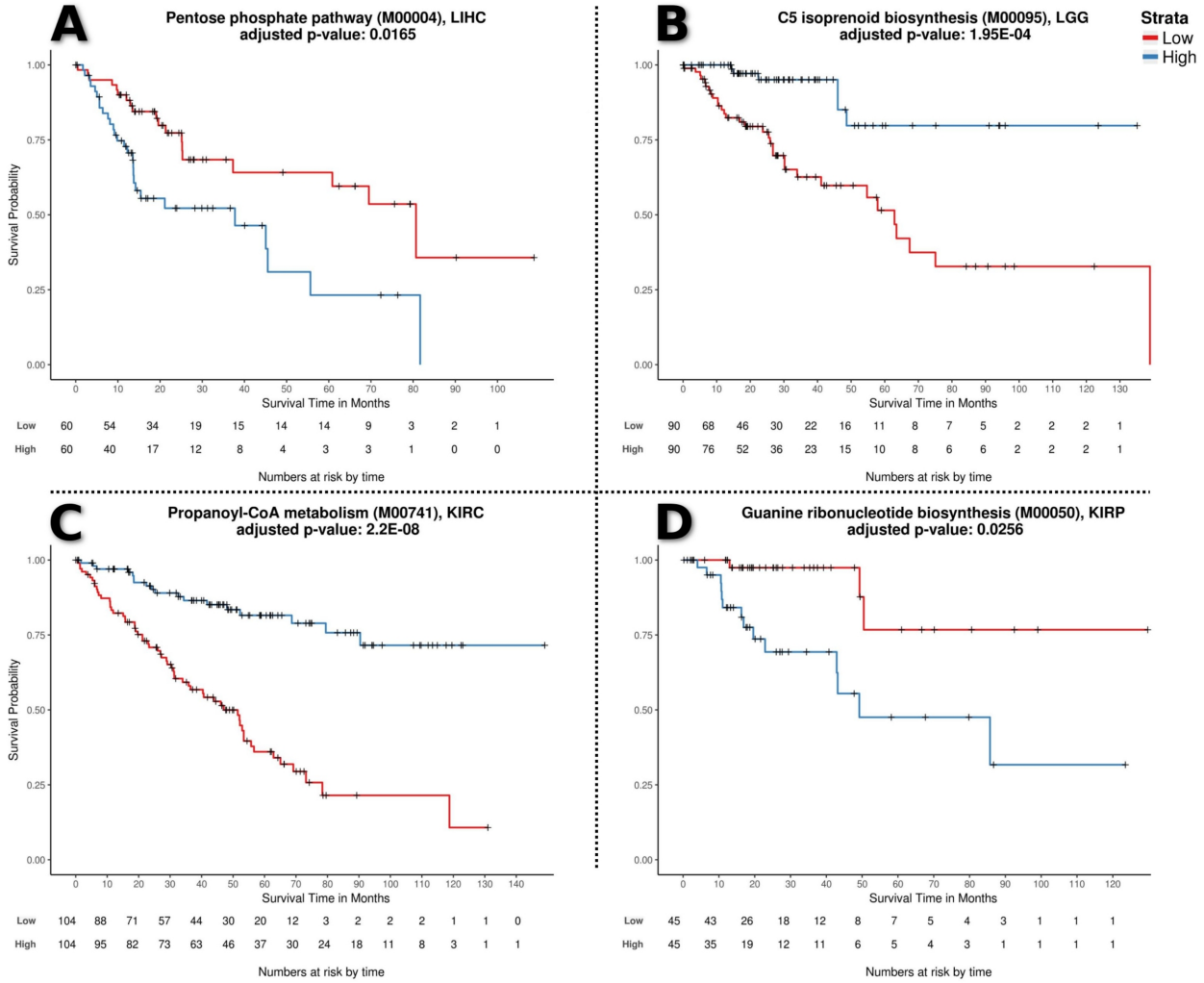


Figure 4.7: K-M plots showing the relationship between module activity and patient survival in different cancer types. High and low module activity groups were defined by patients in the 80% and 20% percentiles of module activity, respectively. The x-axis shows time in months and the number of patients at risk in the high activity and low activity groups. A) Pentose phosphate pathway in LICH. B) C5 isoprenoid biosynthesis in LGG. C) Propanoyl-CoA metabolism in KIRC. D) Guanine ribonucleotide biosynthesis in KIRP.

4.3.6 Essentiality and module activity

The availability of basal gene expression data from 212 cell lines of the Cancer Cell Line Encyclopedia, along with the release of the results of Project Achilles, which assessed the consequences of individual silencing of thousands of genes across many cancer cell lines, allows the influence of predicted metabolic module activities on cancer robustness to be evaluated. The effect of every gene expression KD on the activity of the corresponding module was calculated as the log-fold-change between the estimated activity using cell line gene expression values and the activity estimated by assigning a very low expression value (see Methods) to the KD gene. Subsequently, the correlations of the log-fold-change values with the Achilles score, which accounts for cell viability, were calculated. Given that different cancer types display specific

patterns of differential module activations, essentiality in modules is also expected to be specific to particular cancer types. Therefore, cell lines were grouped by cancer type to obtain the correlations between module activity and cell viability. Considering only significant correlations (FDR adjusted p-value < 0.05) with a correlation coefficient > 0.5 (or < -0.5) obtained from at least eight data points (cell lines x KD genes), a total of 20 modules in 12 cancer types showed significant positive (Table 4.4 and Figure 4.5A) or negative correlations (Table 4.4 and Figure 4.5B).

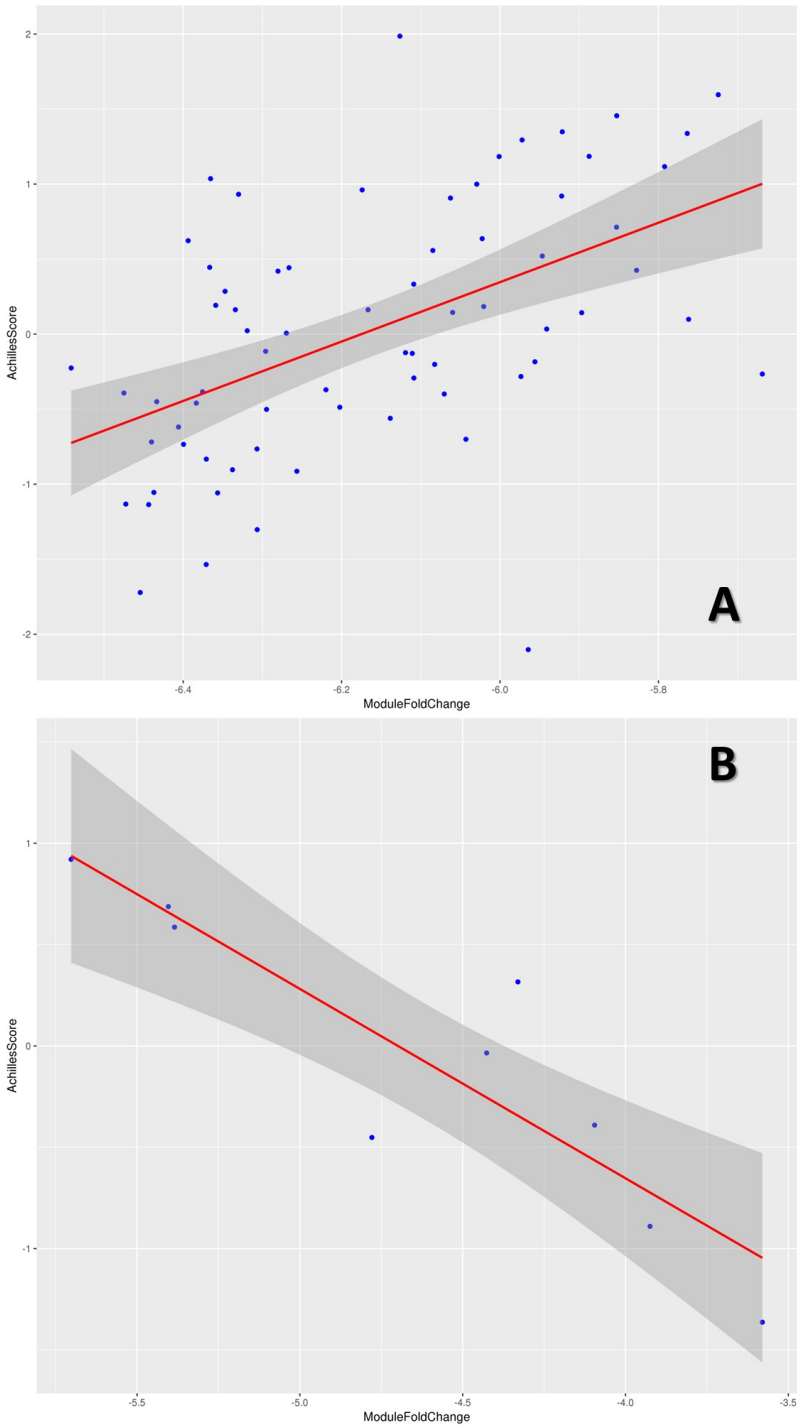


Figure 4.5: Correlation between module activity and cell survival. Correlation between increase in module activity, expressed as log-fold change (x-axis) and cell survival (y-axis) corresponding to gene knockdowns in A) heparan sulfate degradation module and B) bile acid biosynthesis module.

Module	KD Genes	Predicted KD	r	p-value	Tissue	Cell Lines	Module End Metabolite
Bile acid biosynthesis	*AKR1D1 CYP8B1 SLC27A5*	*CYP27A1- AMACR-ACOX2* CYP7A1 HSD17B4* SCP2-HSD3B7* ACOT8**	-0.883	0.003	Urinary tract	3	C00695: Cholic acid
Dermatan sulfate degradation	*IDS ARSB HYAL1*	*IDUA- HYAL4- SPAM1 HYAL3* HYAL2**	-0.812	0	Bone	6	G00872: Chondroitin 4-sulfate
C10-C20 isoprenoid biosynthesis	*IDI1 FDPS GGPS1*	*IDI2**	-0.692	0.016	Stomach	4	C00353: Geranylgeranyl diphosphate
Chondroitin sulfate degradation	*ARSB HYAL1*	*HYAL4- SPAM1 HYAL3* HYAL2**	-0.662	0.019	Bone	6	G00872: Chondroitin 4-sulfate
Inosine monophosphate biosynthesis	*ATIC ADSL PAICS PFAS*	*PPAT GART*	-0.622	0.035	Prostate	3	C00130: IMP
Serine biosynthesis	*PSAT1*	*PHGDH+ PSPH**	-0.61	0.03	Breast	13	C00065: L-Serine
Leucine degradation	*DLD BCKDHA IVD BCAT1*	*BCKDHB HMGCL HMGCLL1* AUH MCCC1 MCCC2* DBT* BCAT2*	-0.601	0.043	Urinary tract	3	C00164: Acetoacetate
beta-Oxidation	*ACAA1 HADHB EHHADH ECHS1*	*ACAA2* HADH HADHA*	-0.58	0	Esophagus	10	C02593: Tetradecanoyl-CoA
Nucleotide sugar biosynthesis	*PGM1 HK2 HK3 UGP2*	*PGM2 HK1* HKDC1**	-0.552	0.002	Skin	7	C00029: UDP-glucose
Pentose phosphate pathway (Pentose phosphate cycle)	*RPE PGD PGLS*	*GPI* TKT* TKTL1* TKTL2* RPIA* RPEL1* G6PD* TALDO1*	-0.541	0	Breast	13	C01172: beta-D-Glucose 6-phosphate
Sphingosine degradation	*SPHK1 SGPL1*	*SPHK2**	-0.532	0.017	Esophagus	10	C00346: Ethanolamine phosphate
Ceramide biosynthesis	*CERS5 DEGS2 DEGS1 SPTLC1 SPTLC2*	*CERS1* CERS3 CERS6* CERS2 CERS4* SPTLC3* KDSR**	-0.523	0.045	Prostate	3	C00195: N-Acylsphingosine
Melatonin biosynthesis	*AANAT*	*ASMT* DDC* TPH2* TPH1*	-0.515	0.044	Pancreas	16	C01598: Melatonin
Inositol phosphate metabolism	*ITPK1 IPMK*	*IPPK*	-0.505	0.025	Kidney	10	C01204: Phytic acid
Glycosphingolipid biosynthesis, ganglio series	*ST8SIA1 ST3GAL5*		-0.5	0	Haematopoietic	27	G00118: Ganglioside (GT3)
C10-C20 isoprenoid biosynthesis	*IDI1 FDPS GGPS1*	*IDI2**	0.5	0.022	Skin	7	C00353: Geranylgeranyl diphosphate
Heparan sulfate degradation	*IDS GNS*	*SGSH HPSE2+ IDUA HGSNAT- NAGLU* GUSB**	0.554	0	CNS	35	G02632: glycan
Pyrimidine degradation	*DPYD DPYS*	*UPB1**	0.574	0.035	Skin	7	C00099: beta-Alanine
Pyrimidine degradation	*DPYD DPYS*	*UPB1**	0.574	0.035	Skin	7	C05145: 3-Aminoisobutyric acid
Conjugated bile acid biosynthesis	*SLC27A5 BAAT*		0.6	0.006	Kidney	10	C05122: Taurocholate
Conjugated bile acid biosynthesis	*SLC27A5 BAAT*		0.6	0.006	Kidney	10	C01921: Glycocholate
Methionine salvage pathway	*ADI1 MRI1 SRM AMD1 MAT2B MAT1A*	*APIP+ MTAP* MAT2A* ENOPH1**	0.618	0.032	Soft tissue	2	C00147: Adenine

Polyamine biosynthesis	*SRM AMD1*	*AZIN2* AGMAT**	0.639	0	Haematopoietic	27	C00315: Spermidine
Nucleotide sugar biosynthesis	*PGM1 HK2 HK3 UGP2*	*PGM2* HK1 HKDC1**	0.641	0.025	Urinary tract	3	C00029: UDP-glucose
Inosine monophosphate biosynthesis	*ATIC ADSL PAICS PFAS*	*PPAT GART*	0.677	0	Bone	6	C00130: IMP
Dermatan sulfate degradation	*IDS ARSB HYAL1*	*IDUAI* HYAL4-SPAM1* HYAL3 HYAL2**	0.692	0	Esophagus	10	G00872: Chondroitin 4-sulfate
Pyrimidine degradation	*DPYD DPYS*	*UPB1*	0.762	0.037	Stomach	4	C00099: beta-Alanine
Pyrimidine degradation	*DPYD DPYS*	*UPB1*	0.762	0.037	Stomach	4	C05145: 3-Aminoisobutyric acid

Table 4.4: Essential modules. The first column contains the name of the module; the second column the genes knocked down in Project Achilles; the third column lists the other genes in the module, whose inhibition is predicted to cause inhibition of the corresponding module and therefore the same effect as the genes in the second column (*, confirmed effect; +, inconclusive effect; -, no information available for them); the fourth column states the correlation coefficient (r), whose positive and negative values respectively indicate an oncomodule and a tumor suppressor module; the fifth column the p-value; the sixth column the cancer tissue from which the cell lines were derived; the seventh column lists the number of different cell lines derived from each tissue and the last column the final metabolite of the module.

4.3.7 Validation of the gene essentiality predictions

We used a recently published study on cancer dependencies [94, 126] that provides extra data on cell survival after massive gene KD. The comparison of cell survival in the cancer types predicted with respect to survival in cancers validated 48 of the 77 predictions (62%), along with three less conclusive validations, which would result in a 66% validation rate, covering 24 of the 28 modules predicted to affect cell viability (see Table 4.4 and Supplementary Table S13, Cubuk et al. Cancer Research., 2018 [85]). This is an excellent proportion of validations, especially if we consider that the method used for validation can fail to detect real KD effects when the KD also markedly affects background survival. Actually, independent experiments can confirm inconclusive validations of predicted inhibitions of essential modules. An interesting example is a small molecule, CBR 5884, which inhibits PHGDH causing selective toxicity in breast cancer cell lines by inhibiting serine biosynthesis [131], as predicted (see Table 4.4 and Supplementary Table S13, Cubuk et al. Cancer Research., 2018 [85]). Finally, to further validate of our predictions (Table 4.4), the impairment of cell proliferation upon depletion of UPB1 gene expression was assessed in two models of gastric cancer (AGS and MKN45 cell lines). This gene encodes an enzyme (βureidopropionase) that catalyzes the final step in the pyrimidine degradation pathway, which in turn is required for epithelial-mesenchymal transition [104]. Thus, two short hairpin shRNA sequences directed to UPB1 caused a significant decrease in proliferation of the two gastric cancer cell lines (AGS and MKN45), as predicted by the model. Conversely, the inhibition in a colon adenocarcinoma cell line (SW480), predicted as non-essential by our model did not result in a

significant difference in growing (Figure 4.6), providing a negative control validation. Although additional experiments may be warranted to confirm cancer vulnerability or resistance based on predicted metabolic activities, these results can be considered independent validations that reinforce the predictions made by the model proposed (Table 4.4 and Supplementary Table S13, Cubuk et al. Cancer Research., 2018 [85]).

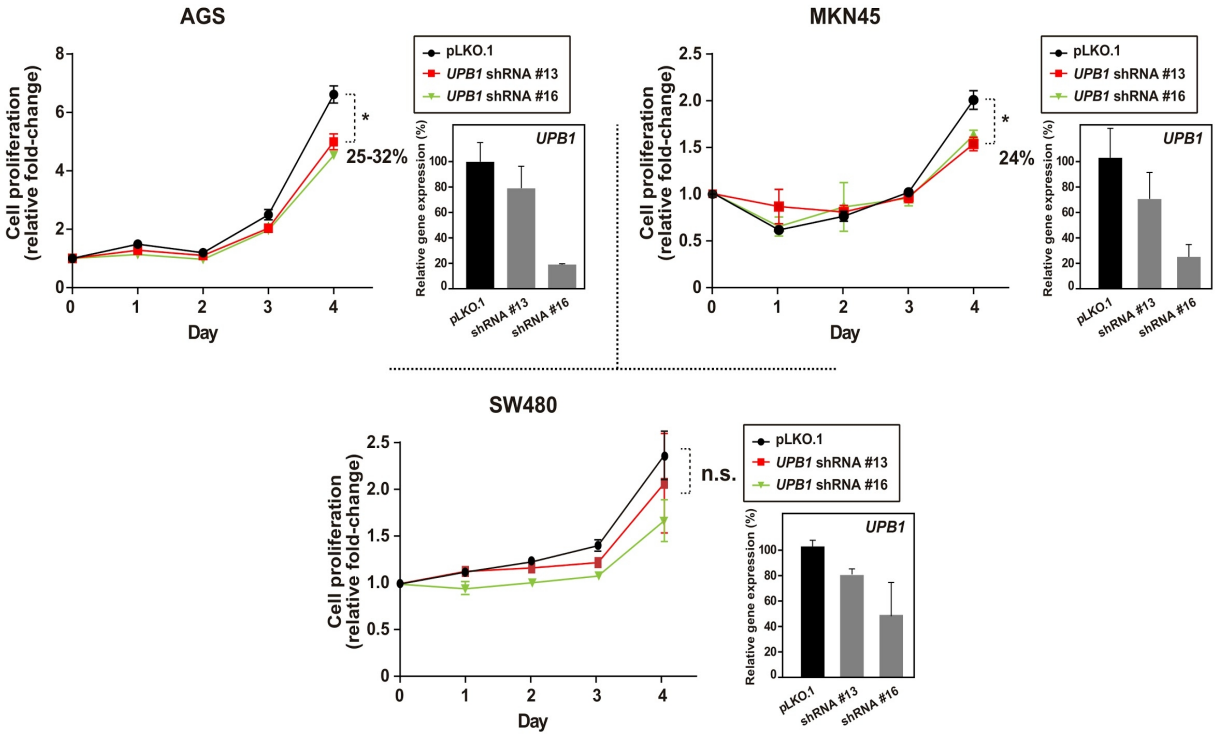


Figure 4.6: Graph showing relative cell proliferation upon UPB1 expression depletion (two different MISSION shRNAs were used as detailed in the inset) or transduction with control vector pLKO.1. The asterisk indicates significant differences (Mann–Whitney test $p < 0.01$) and the range of reduction (%) of cell proliferation is also shown. The prediction of UPB1 essentiality made by the model in lines AGS and MKN45 (stomach gastric adenocarcinoma) was confirmed by relatively more sensitive behavior, while UPB1 does not seem to be relatively sensitive in SW480 (colon adenocarcinoma), as predicted by the model as well.

4.3.8 Therapeutic targeting of metabolic modules

Onco-modules are effective candidates for treating cancer (individually or in combinations), but interventions that activate some tumour suppressor metabolic modules may also offer useful therapeutic strategies. Supplementary Table S14, Cubuk et al. Cancer Research., 2018 [85] lists 137 potential interventions with known drugs that are likely to affect cancer cell viability according to the predictions of the model proposed here.

4.4 Discussion

Although the role of metabolism in cancer has long been known [132, 133], the results presented here provide a more detailed, mechanistic view documenting the relevance of specific metabolic module activities in cancer. This study is based on the integration of gene expression data into metabolic pathway modules and, therefore, may be limited by the lack of correlation between gene and protein expression, and with metabolic activities. However, evaluation of RPPA data [134] and gene expression indicates that gene expression measures are generally a valuable proxy for protein expression and activities (Cubuk et al. Cancer Research., 2018 [85]).

As expected, the production of nucleotides and their precursors (CDP and GTP) shows recurrent significant activation in all cancer types when compared with the corresponding reference tissues (Figure 4.2). Other pervasively activated modules include well-known cancer metabolic dependencies, like cholesterol biosynthesis (M00101), which is consistent with its essential role in cell membranes and as a precursor of steroid hormones [135], and proline biosynthesis (M00015), which is essential in many carcinogenesis settings [110]. In addition, the predicted overexpression of L-cystathionine and L-cysteine across many cancer types may reflect a defect in S-adenosyl-L-methionine, which in turn is consistent with common DNA hypomethylation in cancer cells [136]. On the other hand, this study reveals metabolites whose production is significantly reduced in several cancers types. For example, the well-known Warburg effect, that is, the preference of cancer cells for anaerobic over aerobic metabolism is apparent in modules such as Citrate cycle, second carbon oxidation (M00011). It is also known that many human tumours do not express ASS1 [137], one of the key enzymes of the Urea cycle (M00029) module, which is systematically downregulated in almost all cancer types. Specific observations also support the relevance of the predicted metabolic activities. Examples of cancer metabolic specificities are: upregulation of leucine (M00036) and catecholamine metabolism (M00042) in prostate [138] and colorectal [139] cancer, respectively, and downregulation of glycosaminoglycan (M00058) and polyamine biosynthesis (M00134) in the liver [140] and breast [141] cancer, respectively. In turn, this study highlights less explored metabolic associations, such as downregulation of the pentose phosphate cycle (M00004) in head and neck cancer, or accumulation of cysteine (M00338_1) in several cancer types, which may indicate a link to altered metabolism of reactive oxygen species. Collectively, the results of this study depict biologically relevant metabolic profiles throughout human cancer and provide many novel hypotheses about metabolic alterations in the disease. Metabolic modules are also relevant for establishing the molecular basis that differentiates between cancer subtypes. Supplementary Table S7, (Cubuk et al. Cancer Research., 2018 [85]) provides a detailed survey of differential and common metabolic module activities when cancer subtypes are compared. Although a detailed description of the findings is beyond the scope of this manuscript it is worthwhile highlighting some observations, such as the significant specific

reduction of the activity of the module C21-Steroid hormone biosynthesis, progesterone (M00109) in basal-like breast cancer subtype (the only non-hormone dependent form of the disease) [142]. Experts in specific cancer types can use Supplementary Table S7, (Cubuk et al. Cancer Research., 2018 [85]) to identify relevant subtype-specific module activities that can be exploited for therapeutic purposes. Some of the modules that display different behaviors in cancer are expected to have a direct effect on patient survival. In spite of the limited patient survival data in the ICGC repository, Supplementary Table S9, (Cubuk et al. Cancer Research., 2018 [85]) demonstrates that a remarkable number of modules are associated with poorer patient survival. Specifically, a high level of activity of the pentose phosphate module was found to be significantly associated with poor survival in five cancer types (see Supplementary Table S9 (Cubuk et al. Cancer Research., 2018 [85]) and K-M plots in Figure 4.7A). This observation is consistent with the role of this module in the biosynthesis of nucleotides and NADPH, which is known to play a key role in facilitating cancer cells to cope with anabolic demands and to fight oxidative stress [143]. The analysis of metabolic modules reveals their role as ultimate mechanistic entities whose activity is related to cancer cell fate. For example, the expression of EHHADH has recently been associated with poor prognosis of KIRC [144], but the corresponding module, Malonate semialdehyde pathway (M00013) better predicts outcome (see Supplementary Table S11, Cubuk et al. Cancer Research., 2018 [85]). In fact, out of the 69 metabolic modules associated with differences in survival, a total of 27 (40%) modules (see Supplementary Table S11, Cubuk et al. Cancer Research., 2018 [85]) showed a stronger effect (based on hazard ratio estimations) than any of their corresponding genes. Other modules are also significantly related to survival in other cancer types as LGG (Figure 4.7B), KIRC (Figure 4.7C) or KIRP (Figure 4.7D). Moreover, in the same way, that genes are co-regulated in higher-level entities (metabolic modules), the activations and deactivations of metabolic modules are not an independent process and, in fact, proper cell functionality seems to require a high degree of module activity coordination. Figure 4.3 illustrates how only a few core processes originally correlated in the normal tissues maintain the correlation in all cancer types. An example of this concordance is the positive coordination between fumarate, succinyl-CoA, and urea, which indicates the expected link between the citric acid and urea cycles (Figure 4.4). Unexpectedly, some modules uncorrelated in normal tissue emerge as being coupled in tumours (see Figure 4.4). Thus, according to cancer metabolic demands, bile acids (e.g. cholic acid, M00104_1) is positively correlated with cholesterol (M00101) and triacylglycerol (M00089). In turn, the negative correlation of the previous metabolites with a glycosphingolipid (globoside, M00068), which is linked to differentiation and antigenicity [145], is lost in cancer. Similarly, cholesterol is positively correlated with nucleotide sugar biosynthesis (M00554 and M00632_1), but another glycosphingolipid (ganglioside, M00069) is negatively correlated with this process only in normal tissue. Collectively, these results further highlight the complexity of metabolic reprogramming in cancer. Available data on survival of cancer cell lines after extensive KD (Supplementary Table S2, Cubuk et al. Cancer Research., 2018 [85]) allowed the model to be

used to relate module activity to cell survival in cell lines. Positive correlations between module activity and cell viability (see Table 4.4) indicate that the corresponding module may play an essential role in the corresponding cancer type. Therefore, they can be classified as onco-modules. Such constitutively active modules include common cancer dependencies, like nucleotide sugar biosynthesis, necessary for cell proliferation, and heparan sulfate degradation, necessary for extracellular matrix biosynthesis and thereby, cancer progression and invasion [140]. Conversely, tumour suppressor modules showed negative correlations with cell viability possibly indicating constraints in cancer development and/or progression. These modules include bile acid biosynthesis (M00104), which produces metabolites known to induce apoptosis and inhibit cancer cell proliferation [146] (Table 4.4 and Figure 4.5B). In addition, the study identifies modules with contrary effects depending on the tissue of origin, which probably indicate specific cancer dependencies. For example, inosine monophosphate biosynthesis is positively (bone) or negatively (prostate) correlated depending on whether there is also reduced or enhanced oxidative phosphorylation, respectively [147]. The detection of onco-modules and tumour suppressor modules was used to suggest previously unidentified potentially actionable genes (Table 4.4) because the model proposed predicted an effect of their KD on the activity of the corresponding modules. Recently published extra data on cell survival after massive gene KD [94] was used to validate the predictions made, confirming these for 62% of the genes (48 of the 77 predictions) included in 86% of the modules (24 of the 28), which constitutes a high rate of validation. Given the level of accuracy of the predictions of the model of metabolic module activities, the obvious subsequent step was to predict the effect of drugs, with known targets within modules, in order to shed light on their mechanisms of action (MoA). Actually, components of some metabolic modules are targeted by well-known clinical drugs, such as gemcitabine, which is approved for the treatment of several cancer types (see Supplementary Table S14, Cubuk et al. Cancer Research., 2018 [85] and specifically DB00441 entry in DrugBank). This drug is a nucleoside analog that impairs DNA synthesis by specifically inhibiting the production process of GTP, CDP, and their precursor metabolites. In addition, consistent with recent findings for different cancer types [110, 148], targeting proline (M00015), and less frequently serine (M00020) metabolism, may be efficient strategies for cancer treatment. Additional observations may extend the applications of cancer drugs. The predicted activation of isoprenoid biosynthesis (M00095) in breast cancer is consistent with a potentially protective role of simvastatin in the progression of this cancer type [149]. Following from this observation, predicted metabolic activities support similar applications in the bladder and endometrial cancer [150]. Furthermore, the use of pamidronate, which is currently applied to target bone metastasis in breast cancer and multiple myeloma, and targets isoprenoid biosynthesis (M00367) module, might also be applied to bladder and endometrial cancer [151]. It is worth pointing out that other bisphosphonates show some benefit in these settings and in colorectal cancer [152], which was also predicted in this study. In addition, targeting accumulation of L-cystathionine (M00035_1) by azacitidine, which causes global DNA hypomethylation, may be

useful in at least 10 cancer types. The study also supports drug repurposing, like the potential use of an approved drug for rheumatoid arthritis, leflunomide (which targets UMP biosynthesis) to treat several cancer types [153]. Therefore, this study describes cancer metabolic dependencies that highlight novel therapeutic opportunities either by using current drugs or compounds, or by developing targeted approaches against essential gene products. It is worth noting that a total of 16 commonly mutated genes from the COSMIC database [154] were present in 11 modules. Although it is likely that some of the samples used in this study contained any of these mutations, the information about the mutational status of the genes in the modules provided in the ICGC repository was scarce and so we could not include this information in the model. However, if this information were available, two scenarios could be considered by the model used here: i) activating mutation (e.g. a translocation to another constitutive promoter), which will be detected in the gene expression level itself, and ii) loss-of-function mutation, which can be simulated in the model by setting the gene expression value to 0 (an expressed non-functional gene is mechanistically equivalent to a non-expressed gene) [43, 155]. Although Project Achilles has yielded abundant data, its results are far from exhaustive and, consequently, those obtained here can be considered an underestimate of the actual total number of modules that are essential in cancer.

Chapter 5

Conclusions

Chapter 5
CONCLUSIONS

5.1 Conclusions

In this thesis, Cubuk et al. developed computational systems biology approaches for modelling functional modules of cell signalling and metabolism. Using the bioinformatics framework given in this thesis, we are able to understand intra-cellular behaviours that lie behind the disease initiation and progression. Moreover, in light of this understanding, it is being feasible to integrate omics data on top of biological pathways and apply machine learning algorithms to discover patient-specific potential drug targets. The applications of genomics-guided therapeutics, and changing the paradigm for drug development through cell modelling are becoming an emerging need [156, 157]. Thus, the research done in this thesis is important and timely, due to the extreme relevance for systems biology and even more for systems medicine.

The conclusions of this thesis are summarised and organised below according to the goals originally defined in objectives;

5.1.1 A model of mechanistic pathway activity

MPA methods provide an innovative, biologically inspired alternative for the interpretation of transcriptomic experiments. They can be considered as the next-generation pathway analysis methods. MPA methods constitute an evolution of pathway analysis methods in which pathways are decomposed into elementary subpathways or circuits that potentially account for cell outcomes that can help to explain mechanistic features of phenotypes (disease mechanism, drug MoA, etc.). Here, we developed a new MPA method and benchmarked its sensitivity and specificity with the existing methods. From this comparison we concluded that, although most of the methods were highly specific, they presented remarkable differences in terms of sensitivity. From their relative performances, it can be concluded that a biologically realistic definition of the circuits like receptor-to-effector circuits within the pathways analyzed is a major determinant of the success of the method. However, the scoring approach, which accounts for the activity of the circuit, must also be representative of the biological activity of the cell. Thus, the propagation algorithm used by the method proposed, HiPathia, seems to be the most efficient solution, followed by scores based on differential gene expression, implemented in subSPIA, DEGraph and TAPPA. On the other hand, many MPA methods simply cannot handle loops and artificially disconnect them or even remove them from the calculations. However, our iterative method does not violate the topology and uses it with all given features. In any case, MPA methods have demonstrated to be more sensitive than the conventional functional analysis (ORA or GSEA) and represent a promising alternative for the

interpretation of genomic measurements. I believe that the demand on the MPA methods is and will be increasing because of the enhancing importance of systems medicine which is fundamental to face the challenges of diagnosis and treatment of complex diseases [158]. Since the MPA methods analyze cellular mechanisms rather than their components, they can be used as an alternative to compensate for the ineffective usage of single-gene biomarkers in the near future. MPA can help to discover actionable mechanistic biomarkers.

5.1.2 Metabolizer web tool for differential metabolic activity analysis and discovery of therapeutic targets using summarized metabolic pathway models

Metabolic module activities obtained under the proposed modelling method outperform other methods used to infer metabolic activity, such as GSEA [64], SPIA [97], or CBM [98] (as implemented in IMAT tool [99]). And, furthermore, we have validated most of the predictions made by the method in an independent dataset. These results show that metabolic modules can be considered a relevant type of functional module in cancer and probably also in other diseases related to metabolism. The program Metabolizer allows researchers to easily estimate metabolic module activities from gene expression measurements and use them for different purposes. Thus, the comparison between two conditions can throw light on the subjacent molecular mechanisms that make them different. In this way, disease mechanisms or drug mechanisms of action can easily be interpreted within the context of metabolism. Such comparisons can also be used to derive multigenic predictors with a mechanistic meaning, that have demonstrated to be useful to predict complex traits [38].

Diagnostic strategies are rapidly changing in cancer and other diseases because of the availability of increasingly affordable genomic analysis [159]. Therapies that specifically target genetic alterations are probing to be safer and more effective than traditional chemotherapies when used in the adequate patient population [160]. Perhaps, one of the most relevant aspects of modelling is that models allow predicting the effect of simulated gene expression profiles over the activity of metabolic modules, opening the door to anticipate the effect of the intervention on genes. In this respect, Metabolizer constitutes an extremely useful tool for finding putative actionable targets for a specific condition [68]. This is very relevant in the context of personalized medicine and can help in finding individualized therapeutic interventions for patients [67]. In fact, recent reports indicate that genes involved in metabolic pathways show a remarkable heterogeneity across different cancer patients [124]. This suggests that personalized therapies might likely be successful providing the context of the interventions can be properly explored and understood with a tool such as Metabolizer. For example, synthetic lethality, defined as genetic mutations or gene expression alterations with little or null individual effect on cell viability but that results in cell death when

combined, offers a promising range of potential therapeutic interventions [125] that can only be properly exploited in a framework such as the one provided by Metabolizer.

Therefore, Metabolizer can be considered an innovative tool that enables the use of standard measurements of gene expression in the context of the complexity of the metabolic network, with a direct application in the clinic as well as in research in animal models.

5.1.3 A pan-cancer metabolic landscape based on gene expression integration into pathway modules

Changes in the metabolic processes play a key role in cancer development and progression and this phenomenon is a recognized cancer hallmark (7). However, metabolic maps are complex and understanding the global implications in cancer of changes in the activity of processes or components is challenging. Recently, the metabolic map has been decomposed into modules, which consist of sequential reactions representing a summary of fundamental metabolic processes (20). Here we have explored the usefulness of modules to understand cancer metabolic profiles and their relation with cancer outcome and treatment. A simple model is used to predict module activity from the expression levels of its gene components. In a pan-cancer analysis, we demonstrate that the activity of certain modules changes significantly between cancers and the corresponding tissues of origin. We also report changes in the correlated activity of modules. The activity of several modules is significantly associated with cancer prognosis and, moreover, these associations are stronger for the module than for any of their constituent genes. This finding strongly supports the notion that the effect on the phenotype arises from the coordinated activity of the genes in the module. Therefore, essentiality at the gene level would be a consequence of the impact of the activity of the corresponding gene product on the activity of the module. The associations with the outcome and cell viability allow us to coin the concepts of tumour suppressor metabolic modules and onco-modules. The associations found between metabolic module activities and patient survival confirms that metabolic modules can be realistically modelled within the proposed framework. Finally, using this modelling framework, we propose potential therapeutic targets to inhibit metabolic reprogramming in cancer.

Certainly, the metabolic modules used in this modelling framework describe only a limited (although representative) portion of the whole known map of human metabolism. Therefore, the model presented here provides mechanistic insights into cell metabolic activities that are significantly linked to complex phenotypes, such as cancer prognosis, but probably has limitations in the accurate prediction of the fate of specific metabolites or phenotypes not affected by the metabolites resulting from the 95 metabolic modules used in the model. More comprehensive models that encompass larger portions of the metabolism will, no doubt, increase the reliability of

the predictions. We anticipate that the data and models produced will play an increasingly important role in personalized treatment (54).

Scientific Contributions

PUBLICATIONS:

- **Cubuk C**, Can FE, Peña-Chilet, M and Dopazo, J (2020) Mechanistic Models of Signaling Pathways Reveal the Drug Action Mechanisms behind Gender-Specific Gene Expression for Cancer Treatments. Cells 9, 1579.

- Garrett A, Callaway A, Durkie M, **Cubuk C**, Alikian M, Burghel G, Robinson R, Izatt L, Talukdar S, Side L, Cranston T, Palmer-Smith S, Baralle D, Berry I, Drummond J, Wallace A, Norbury G, Eccles D, Ellard S, Lalloo F, Evans D, Woodward E, Tischkowitz M, Hanson H and Turnbull C (2020) Cancer Variant Interpretation Group UK (CanVIG-UK): an exemplar national subspecialty multidisciplinary network. Journal of Medical Genetics, pp.jmedgenet-2019-106759.

- Petkevicius K, Virtue S, Bidault G, Jenkins B, **Cubuk C**, Morgantini C, Aouadi M, Dopazo J, Serlie M, Koulman A and Vidal-Puig A (2019) Accelerated phosphatidylcholine turnover in macrophages promotes adipose tissue inflammation in obesity. eLife 8, e47990.

- Menden MP, Wang D, Mason MJ, AstraZeneca-Sanger Drug Combination DREAM Consortium (**Cubuk C**) et al. (2019) Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. Nature Communications 10, 2674.

- **Cubuk C** (MSc thesis; written in Turkish and abstract in English) Discovery of novel miRNAs and their variant analysis using bioinformatics approaches: Citrus model organism. (https://bit.ly/2t6KkDw or https://bit.ly/2ZtPTbd)

- **Cubuk C**, Hidalgo M, Amadoz A, Rian K, Salavert F, Pujana M, Mateo F, Herranz C, Carbonell-Caballero J & Dopazo J (2019). Differential metabolic activity and discovery of therapeutic targets using summarized metabolic pathway models. npj Systems Biology and Applications, 5(1). (http://metabolizer.babelomics.org/ or https://github.com/babelomics/metabolizer)

- **Cubuk C**, Hidalgo M, Amadoz A, Pujana M, Mateo F, Herranz C, Carbonell-Caballero J & Dopazo J (2018) Gene Expression Integration into Pathway Modules Reveals a Pan-Cancer Metabolic Landscape. Cancer Research, 78(21), 6059-6072.

- Fourati S, Talla A, Mahmoudian M, The Respiratory Viral DREAM Challenge Consortium (**Cubuk C**) et al. (2018) A crowdsourced analysis to identify ab initio molecular signatures predictive of susceptibility to viral infection. Nature Communications 9, 4418.

- Amadoz A, Hidalgo M, **Cubuk C**, Carbonell-Caballero J & Dopazo J (2018) A comparison of mechanistic signaling pathway activity analysis methods. Briefings in Bioinformatics 20(5), 1655-1668.

- Ferreira PG, Muñoz-Aguirre M, Reverter F, Godinho C, Sousa A, Amadoz A, Sodaei R, Hidalgo M, Pervouchine D, Carbonell-Caballero J, Nurtdinov R, Breschi A, Oliveira P, **Cubuk C**, Aguet F, Oliveira C, Dopazo J, Sammeth M, Ardlie KG & Guigo R. (2018) The effects of death and post-mortem cold ischemia on human tissue transcriptomes. Nature Communications 9, 490.

- Hidalgo M, Amadoz A, **Cubuk C**, Carbonell-Caballero J & Dopazo J (2018) Models of cell signaling uncover molecular mechanisms of high-risk neuroblastoma and predict disease outcome. Biology Direct 13, 16.

- Carbonell-Caballero J, Amadoz A, Alonso R, Hidalgo M, **Cubuk C**, Conesa D, López-Quílez A & Dopazo J (2017) Reference genome assessment from a population scale perspective: an accurate profile of variability and
noise. Bioinformatics, btx482.

- Jiao Y, Hidalgo M, **Cubuk C**, Amadoz A, Carbonell-Caballero J, Vert JP & Dopazo J (2017) Signaling pathway activities improve prognosis for breast cancer. Technical Report, BioRxiv, 132357.

- Hidalgo M, **Cubuk C**, Amadoz A, Salavert F, Carbonell-Caballero J & Dopazo J (2017) High throughput estimation of functional cell activities reveals disease mechanisms and predicts relevant clinical outcomes. Oncotarget 8, 5160-5178. (http://hipathia.babelomics.org/ or
https://bioconductor.org/packages/release/bioc/html/hipathia.html)

- Salavert F, Hidalgo M, Amadoz A, **Cubuk C**, Medina I, Crespo D, Carbonell-Caballero J & Dopazo J (2016) Actionable pathways: interactive discovery of therapeutic targets using signaling pathway models. Nucleic Acids Research 44, W212-W216. (http://pathact.babelomics.org/)

- Sanghez V, **Cubuk C**, Sebastián-Leon P, Carobbio S, Dopazo J, Vidal-Puig A & Bartolomucci A (2016) Chronic subordination stress selectively downregulates the insulin signaling pathway in liver and skeletal muscle but not in adipose tissue of male mice. Stress 19, 214-224.

- Razzoli M, Frontini A, Gurney A, Mondini E, **Cubuk C**, Katz L, Cero C, Bolan P, Dopazo J, Vidal-Puig A, Cinti S & Bartolomucci A (2016) Stress-induced activation of brown adipose tissue prevents obesity in conditions of low adaptive thermogenesis. Molecular Metabolism 5, 19-33.

- Alonso R, Salavert F, Garcia-Garcia F, Carbonell-Caballero J, Bleda M, Garcia-Alonso L, Sanchis-Juan A, Perez-Gil D, Marin-Garcia P, Sanchez R, **Cubuk C**, Hidalgo M, Amadoz A, Hernansaiz-Ballesteros R, Alemán A, Tarraga J, Montaner D, Medina I & Dopazo J (2015) Babelomics 5.0: functional interpretation for new generations of genomic data. Nucleic Acids Research 43, W117-W121. (http://babelomics.org/)

ORAL PRESENTATIONS:

- **Cubuk C**. Big Data & Cancer Genomics Workshop 3: personalized treatments and drug discovery. Oncothon, Granada, Spain, November 12-14, 2018

- Cubuk C, Salavert F, Hidalgo M, Amadoz A, Carbonell-Caballero J & Dopazo J. Metabolizer: a web tool for analysis of modular architecture of metabolic pathways using transcriptomic data. Pitch Challenge Presentation, IV Meeting of PhD Students at UPV, Valencia, Spain, June 01, 2017.

- **Cubuk C**, Hidalgo M, Carbonell-Caballero J & Dopazo J. Signalling circuit activities as mechanism-based features to predict mode of action of chemicals. CAMDA 2015 Conference at ISMB/ECCB, Dublin, July 10-11, 2015.

- Hidalgo M, **Cubuk C**, Carbonell-Caballero J & Dopazo J. Functional hallmarks in clear cell renal cell carcinoma grade and stage progression revealed by changes in signalling circuit activities. CAMDA 2015 Conference at ISMB/ECCB, Dublin, July 10-11, 2015.

POSTERS:

- **Cubuk C**, Loucera C, Hidalgo M & Dopazo J. The metabolite abundances in coherence with the activities of signaling pathway circuits. 4th Disease Maps Community Meeting (DMCM2019), Sevilla, Spain, October 2-4, 2019.

- **Cubuk C**, Hidalgo M, Loucera C, Rian K, Pena-Chilet M, Falco M, Nepomuceno-Chamorro I, Milina-Abril H & Dopazo J. Interpreting genomic profiles with mechanistic models of pathways. XIV Symposium On Bioinformatics (JBI2018), Granada, Spain, November 12-14, 2018.

- **Cubuk C**, Hidalgo M, Amadoz A, Carbonell-Caballero J & Dopazo J. Gene expression integration into pathway modules reveals a pan-cancer metabolic landscape. XIV Symposium On Bioinformatics (JBI2018), Granada, Spain, November 12-14, 2018.

- **Cubuk C**, Hidalgo M, Carbonell-Caballero J & Dopazo J. Identification of key metabolic patterns of cancer using RNA-Seq data. 4th Conference on Constraint-Based Reconstruction and Analysis (COBRA 2015),Heidelberg, Germany, September 16-18, 2015, and XIII Symposium on Bioinformatics (JBI2016), UPV, Valencia, Spain, May 10-13, 2016.

- **Cubuk C** and & Dopazo J. Constraining metabolic model with gene expression data to understand functional differences. MLPM Summer School, Institut Curie in Paris, France, September 11-19, 2014.

CHALLENGES:

- **Cubuk C**, Hidalgo M, Loucera C, Rian K, Pena-Chilet M, Garrido-Rodriguez M, Falco M, Gundogdu P & Dopazo J. Dream Challenge: Single Cell Signaling in Breast Cancer Challenge, 2019.

- **Cubuk C**, Hidalgo M, Amadoz A, Carbonell-Caballero J & Dopazo J. DREAM Challenge: AstraZeneca-Sanger Drug Combination Prediction, 2017. (Published: doi.org/10.1038/s41467-019-09799-2)

- Amadoz A, Hidalgo M, **Cubuk C**, Carbonell-Caballero J & Dopazo J. DREAM Challenge: Discovering Dynamic Molecular Signatures in Response to Virus Exposure, 2016 (Published: doi.org/10.1038/s41467-018-06735-8).

- **Cubuk C**, Hidalgo M, Carbonell-Caballero J & Dopazo J. FDA SEQC Challenges (CAMDA 2015): SEQC Rat TGx - rat liver response to chemicals, 2015.

PARTICIPATION IN CONFERENCES, MEETINGS, WORKSHOPS:

- 02-04 October 2019, 4th Disease Maps Community Meeting (DMCM2019), Sevilla, Spain. https://disease-maps.org/DMCM2019

- 12-14 November 2018, Oncothon and JBI2018, Granada, Spain. http://oncothon.ptsgranada.com/ and http://jbi2018.ugr.es/

- 01 June 2017, IV Meeting of PhD Students at UPV, Valencia, Spain.
https://www.upv.es/contenidos/ENCDOC/indexi.html

- 21 October 2016, Alfried Krupp-Symposium, Munich, Germany.
http://mlpm.eu/summer-school/alfried-krupp-symposium/

- 20-21 October 2016, MLPM ITN Final Meeting, Munich, Germany.
http://mlpm.eu/summer-school/final-itn-meeting/

- 21-24 May 2016, Europe Genetics Conference, Barcelona, Spain.
https://www.eshg.org/home2016.0.html

- 19-20 May 2016, MLPM Closing Conference, (ESHG Symposium 2016), Barcelona, Spain.
http://mlpm.eu/summer-school/4th-annual-meeting/

- 10-12 May 2016, XIII Symposium on Bioinformatics, Valencia, Spain.
http://www.jbi2016.org/

- 14-18 March 2016, Team Working Event: The 2nd ITN March Retreat, Valencia, Spain.
I was the organizer. http://www.mlpm.eu/blog/team-working-event-the-2nd-itn-march-retreat/

- 21-25 September 2015, 3rd MLPM Summer School, Manchester, United Kingdom.
http://mlpm.eu/summer-school/summer-school-2015/

- 16-18 September 2015, 4th Conference on Constraint-Based Reconstruction and Analysis (COBRA 2015). Heidelberg, Germany. http://www.aiche.org/sbe/conferences/conference-on-constraint-based-reconstruction-and-analysis-cobra/2015

- 10-14 July 2015, ISMB/ECCB 2015 - 23rd Annual International Conference on Intelligent Systems for Molecular Biology and the 14th European Conference on Computational Biology, Dublin, Ireland. http://www.iscb.org/ismbeccb2015

- 8-10 April 2015, 6th Annual IMPPC Conference, Molecular Targets for Predictive and Personalized Medicine of Cancer, Barcelona, Spain.
http://www.imppc.org/media/upload/pdf/6th_annual_conference_booklet_2015_editora_3_54_1.pdf

- 2-4 March 2015, Team Working Event: 1st MLPM Mini-Hackathon in Basel, Basel, Switzerland.
http://www.mlpm.eu/blog/1st-mlpm-mini-hackathon-in-basel/

- 11-19 September 2014, 2nd MLPM Summer School, Paris, France.
http://mlpm.eu/summer-school/summer-school-2014/

- 12-14 May 2014. The Systems Biology Modelling Cycle (supported by BioPreDyn). EBI, Cambridge, United Kingdom. http://www.ebi.ac.uk/training/course/BioPreDyn2014

INTERNATIONAL PROJECTS INVOLVED:

- Nov. 2013 - Nov. 2017, Machine Learning for Personalized Medicine (MLPM-ITN), 7th Framework Programme, EU. (http://mlpm.eu/)

INTERNSHIPS AND SECONDMENTS

- **July - October 2016**, Secondment of MLPM Project, Pharmatics Limited. Edinburgh, Scotland, UK.

The projects which I worked on;

▪ Mendelian randomization using summary-level data.

▪ Adapting linear discriminant analysis (LDA) and prediction analysis for microarrays (PAM) to summary level data.

▪ An automated system for scientific literature search in PubMed.

- **February - May 2016**, Secondment of MLPM Project, Machine Learning & Computational Biology Lab., Basel, ETH Zürich, Switzerland.

The projects which I worked on;

▪ Assessment of gene essentiality using metabolic module activities.

▪ Correlation of metabolic module activities (before/after insilico knockout) with Achilles gene essentiality scores

▪ Predicting gene essentiality from metabolic module activities using machine learning methods (Random Forest, Support Vector Machines and Lasso).

References

1. *Medicinisch-chemische Untersuchungen - Google Books Available at: https://books.google.co.uk/books?id=YJRTAAAAcAAJ&pg=PA456&redir_esc=y#v=onepage&q&f=false [Accessed February 27, 2020].*
2. *Frederick Sanger - Biographical - NobelPrize.org Available at: https://www.nobelprize.org/prizes/chemistry/1980/sanger/biographical/ [Accessed February 27, 2020].*
3. *What Is Systems Biology · Institute for Systems Biology Available at: https://isbscience.org/about/what-is-systems-biology/ [Accessed February 27, 2020].*
4. *Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 27, 29–34.*
5. *Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R., et al. (2020). The reactome pathway knowledgebase. Nucleic Acids Res. 48, D498–D503.*
6. *Kuperstein, I., Bonnet, E., Nguyen, H. A., Cohen, D., Viara, E., Grieco, L., Fourquet, S., Calzone, L., Russo, C., Kondratova, M., et al. (2015). Atlas of Cancer Signalling Network: a systems biology resource for integrative analysis of cancer data with Google Maps. Oncogenesis 4, e160.*
7. *Fazekas, D., Koltai, M., Türei, D., Módos, D., Pálfy, M., Dúl, Z., Zsákai, L., Szalay-Bekő, M., Lenti, K., Farkas, I. J., et al. (2013). SignaLink 2 - a signaling pathway resource with multi-layered regulatory networks. BMC Syst. Biol. 7, 7.*
8. *Kelder, T., van Iersel, M. P., Hanspers, K., Kutmon, M., Conklin, B. R., Evelo, C. T., and Pico, A. R. (2012). WikiPathways: building research communities on biological pathways. Nucleic Acids Res. 40, D1301-7.*
9. *Türei, D., Korcsmáros, T., and Saez-Rodriguez, J. (2016). OmniPath: guidelines and gateway for literature-curated signaling pathway resources. Nat. Methods 13, 966–967.*
10. *Cerami, E. G., Gross, B. E., Demir, E., Rodchenkov, I., Babur, O., Anwar, N., Schultz, N., Bader, G. D., and Sander, C. (2011). Pathway Commons, a web resource for biological pathway data. Nucleic Acids Res. 39, D685-90.*
11. *Belinky, F., Nativ, N., Stelzer, G., Zimmerman, S., Iny Stein, T., Safran, M., and Lancet, D. (2015). PathCards: multi-source consolidation of human biological pathways. Database (Oxford) 2015.*

12. *Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25, 25–29.*

13. *Mootha, V. K., Lindgren, C. M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., et al. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat. Genet. 34, 267–273.*

14. *Zyla, J., Marczyk, M., Weiner, J., and Polanska, J. (2017). Ranking metrics in gene set enrichment analysis: do they matter? BMC Bioinformatics 18, 256.*

15. *Draghici, S., Khatri, P., Tarca, A. L., Amin, K., Done, A., Voichita, C., Georgescu, C., and Romero, R. (2007). A systems biology approach for pathway level analysis. Genome Res. 17, 1537–1545.*

16. *Sebastián-León, P., Carbonell, J., Salavert, F., Sanchez, R., Medina, I., and Dopazo, J. (2013). Inferring the functional effect of gene expression changes in signaling pathways. Nucleic Acids Res. 41, W213-7.*

17. *Hao, T., Wu, D., Zhao, L., Wang, Q., Wang, E., and Sun, J. (2018). The Genome-Scale Integrated Networks in Microorganisms. Front. Microbiol. 9, 296.*

18. *King, Z. A., Lu, J., Dräger, A., Miller, P., Federowicz, S., Lerman, J. A., Ebrahim, A., Palsson, B. O., and Lewis, N. E. (2016). BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. Nucleic Acids Res. 44, D515-22.*

19. *Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H.-Y., Cohoon, M., de Crécy-Lagard, V., Diaz, N., Disz, T., Edwards, R., et al. (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res. 33, 5691–5702.*

20. *Pornputtapong, N., Nookaew, I., and Nielsen, J. (2015). Human metabolic atlas: an online resource for human metabolism. Database (Oxford) 2015, bav068.*

21. *Chelliah, V., Laibe, C., and Le Novère, N. (2013). BioModels Database: a repository of mathematical models of biological processes. Methods Mol. Biol. 1021, 189–199.*

22. *Brunk, E., Sahoo, S., Zielinski, D. C., Altunkaya, A., Dräger, A., Mih, N., Gatto, F., Nilsson, A., Preciat Gonzalez, G. A., Aurich, M. K., et al. (2018). Recon3D enables a three-dimensional view of gene variation in human metabolism. Nat. Biotechnol. 36, 272–281.*

23. *Lewis, N. E., Nagarajan, H., and Palsson, B. O. (2012). Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. Nat. Rev. Microbiol. 10, 291–305.*

24. *Orth, J. D., Thiele, I., and Palsson, B. Ø. (2010). What is flux balance analysis? Nat. Biotechnol. 28, 245–248.*

25. *Opdam, S., Richelle, A., Kellman, B., Li, S., Zielinski, D. C., and Lewis, N. E. (2017). A Systematic Evaluation of Methods for Tailoring Genome-Scale Metabolic Models. Cell Syst. 4, 318-329.e6.*

26. *Thiele, I., and Palsson, B. Ø. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat. Protoc. 5, 93–121.*

27. *Hyduke, D. R., and Palsson, B. Ø. (2010). Towards genome-scale signalling network reconstructions. Nat. Rev. Genet. 11, 297–307.*

28. *Emmert-Streib, F., and Dehmer, M. (2018). Inference of Genome-Scale Gene Regulatory Networks: Are There Differences in Biological and Clinical Validations? MAKE 1, 138–148.*

29. *Oughtred, R., Stark, C., Breitkreutz, B.-J., Rust, J., Boucher, L., Chang, C., Kolas, N., O'Donnell, L., Leung, G., McAdam, R., et al. (2019). The BioGRID interaction database: 2019 update. Nucleic Acids Res. 47, D529–D541.*

30. *Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., Sacco, F., Palma, A., Nardozza, A. P., Santonico, E., et al. (2012). MINT, the molecular interaction database: 2012 update. Nucleic Acids Res. 40, D857-61.*

31. *Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N. H., Chavali, G., Chen, C., del-Toro, N., et al. (2014). The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res. 42, D358-63.*

32. *Siebenhaller, M., Nielsen, S. S., McGee, F., Balaur, I., Auffray, C., and Mazein, A. (2018). Human-like layout algorithms for signalling hypergraphs: outlining requirements. Brief. Bioinformatics.*

33. *Minguez, P., Götz, S., Montaner, D., Al-Shahrour, F., and Dopazo, J. (2009). SNOW, a web-based tool for the statistical analysis of protein-protein interaction networks. Nucleic Acids Res. 37, W109-14.*

34. *García-Alonso, L., Alonso, R., Vidal, E., Amadoz, A., de María, A., Minguez, P., Medina, I., and Dopazo, J. (2012). Discovering the hidden sub-network component in a ranked list of genes or proteins derived from genomic experiments. Nucleic Acids Res. 40, e158.*

35. *Ibáñez, M., Carbonell-Caballero, J., Such, E., García-Alonso, L., Liquori, A., López-Pavía, M., Llop, M., Alonso, C., Barragán, E., Gómez-Seguí, I., et al. (2018). The modular network structure of the mutational landscape of Acute Myeloid Leukemia. PLoS ONE 13, e0202926.*

36. *Oti, M., and Brunner, H. G. (2007). The modular nature of genetic diseases. Clin. Genet. 71,*

1–11.

37. *Fey, D., Halasz, M., Dreidax, D., Kennedy, S. P., Hastings, J. F., Rauch, N., Munoz, A. G., Pilkington, R., Fischer, M., Westermann, F., et al. (2015). Signaling pathway models as biomarkers: Patient-specific simulations of JNK activity predict the survival of neuroblastoma patients. Sci. Signal. 8, ra130.*

38. *Amadoz, A., Sebastian-Leon, P., Vidal, E., Salavert, F., and Dopazo, J. (2015). Using activation status of signaling pathways as mechanism-based biomarkers to predict drug sensitivity. Sci. Rep. 5, 18494.*

39. *Jaakkola, M. K., and Elo, L. L. (2016). Empirical comparison of structure-based pathway methods. Brief. Bioinformatics 17, 336–345.*

40. *Jacob, L., Neuvial, P., and Dudoit, S. (2012). More power via graph-structured tests for differential expression of gene networks. Ann. Appl. Stat. 6, 561–600.*

41. *Li, X., Shen, L., Shang, X., and Liu, W. (2015). Subpathway Analysis based on Signaling-Pathway Impact Analysis of Signaling Pathway. PLoS ONE 10, e0132813.*

42. *Martini, P., Sales, G., Massa, M. S., Chiogna, M., and Romualdi, C. (2013). Along signal paths: an empirical gene set approach exploiting pathway topology. Nucleic Acids Res. 41, e19.*

43. *Salavert, F., Hidago, M. R., Amadoz, A., Çubuk, C., Medina, I., Crespo, D., Carbonell-Caballero, J., and Dopazo, J. (2016). Actionable pathways: interactive discovery of therapeutic targets using signaling pathway models. Nucleic Acids Res. 44, W212-6.*

44. *Bader, G. D., Cary, M. P., and Sander, C. (2006). Pathguide: a pathway resource list. Nucleic Acids Res. 34, D504-6.*

45. *Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res. 42, D199-205.*

46. *Sebastian-Leon, P., Vidal, E., Minguez, P., Conesa, A., Tarazona, S., Amadoz, A., Armero, C., Salavert, F., Vidal-Puig, A., Montaner, D., et al. (2014). Understanding disease mechanisms with models of signaling pathway activities. BMC Syst. Biol. 8, 121.*

47. *Li, J., Rix, U., Fang, B., Bai, Y., Edwards, A., Colinge, J., Bennett, K. L., Gao, J., Song, L., Eschrich, S., et al. (2010). A chemical and phosphoproteomic characterization of dasatinib action in lung cancer. Nat. Chem. Biol. 6, 291–299.*

48. *Mitsos, A., Melas, I. N., Siminelakis, P., Chairakaki, A. D., Saez-Rodriguez, J., and Alexopoulos, L. G. (2009). Identifying drug effects via pathway alterations using an integer linear programming optimization formulation on phosphoproteomic data. PLoS Comput. Biol.*

5, e1000591.

49. *Efroni, S., Schaefer, C. F., and Buetow, K. H. (2007). Identification of key processes underlying cancer phenotypes using biologic pathway analysis. PLoS ONE 2, e425.*

50. *Kschischang, F. R., Frey, B. J., and Loeliger, H. A. (2001). Factor graphs and the sum-product algorithm. IEEE Trans. Inform. Theory 47, 498–519.*

51. *Massa, M. S., Chiogna, M., and Romualdi, C. (2010). Gene set analysis exploiting the topology of a pathway. BMC Syst. Biol. 4, 121.*

52. *Gao, S., and Wang, X. (2007). TAPPA: topological analysis of pathway phenotype association. Bioinformatics 23, 3100–3102.*

53. *Hung, J.-H., Whitfield, T. W., Yang, T.-H., Hu, Z., Weng, Z., and DeLisi, C. (2010). Identification of functional modules that correlate with phenotypic difference: the influence of network topology. Genome Biol. 11, R23.*

54. *UniProt Consortium (2015). UniProt: a hub for protein information. Nucleic Acids Res. 43, D204-12.*

55. *Hidalgo, M. R., Amadoz, A., Çubuk, C., Carbonell-Caballero, J., and Dopazo, J. (2018). Models of cell signaling uncover molecular mechanisms of high-risk neuroblastoma and predict disease outcome. Biol. Direct 13, 16.*

56. *Esteban-Medina, M., Peña-Chilet, M., Loucera, C., and Dopazo, J. (2019). Exploring the druggable space around the Fanconi anemia pathway using machine learning and mechanistic models. BMC Bioinformatics 20, 370.*

57. *Hidalgo, M. R., Cubuk, C., Amadoz, A., Salavert, F., Carbonell-Caballero, J., and Dopazo, J. (2017). High throughput estimation of functional cell activities reveals disease mechanisms and predicts relevant clinical outcomes. Oncotarget 8, 5160–5178.*

58. *Jiao, Y., Hidalgo, M. R., Cubuk, C., Amadoz, A., Carbonell-Caballero, J., Vert, J.-P., and Dopazo, J. (2017). Signaling pathway activities improve prognosis for breast cancer. BioRxiv.*

59. *Ferreira, P. G., Muñoz-Aguirre, M., Reverter, F., Sá Godinho, C. P., Sousa, A., Amadoz, A., Sodaei, R., Hidalgo, M. R., Pervouchine, D., Carbonell-Caballero, J., et al. (2018). The effects of death and post-mortem cold ischemia on human tissue transcriptomes. Nat. Commun. 9, 490.*

60. *Petkevicius, K., Virtue, S., Bidault, G., Jenkins, B., Çubuk, C., Morgantini, C., Aouadi, M., Dopazo, J., Serlie, M., Koulman, A., et al. (2019). Accelerated phosphatidylcholine turnover in macrophages promotes adipose tissue inflammation in obesity. BioRxiv.*

61. *Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray*

expression data using empirical Bayes methods. Biostatistics 8, 118–127.

62. *Robinson, M. D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 11, R25.*

63. *Montaner, D., and Dopazo, J. (2010). Multidimensional gene set analysis of genomic data. PLoS ONE 5, e10348.*

64. *Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 102, 15545–15550.*

65. *Al-Shahrour, F., Díaz-Uriarte, R., and Dopazo, J. (2005). Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information. Bioinformatics 21, 2988–2993.*

66. *Goeman, J. J., van de Geer, S. A., de Kort, F., and van Houwelingen, H. C. (2004). A global test for groups of genes: testing association with a clinical outcome. Bioinformatics 20, 93–99.*

67. *Fryburg, D. A., Song, D. H., Laifenfeld, D., and de Graaf, D. (2014). Systems diagnostics: anticipating the next generation of diagnostic tests based on mechanistic insight into disease. Drug Discov. Today 19, 108–112.*

68. *Dopazo, J. (2014). Genomics and transcriptomics in drug discovery. Drug Discov. Today 19, 126–132.*

69. *Koumakis, L., Kanterakis, A., Kartsaki, E., Chatzimina, M., Zervakis, M., Tsiknakis, M., Vassou, D., Kafetzopoulos, D., Marias, K., Moustakis, V., et al. (2016). MinePath: Mining for Phenotype Differential Sub-paths in Molecular Pathways. PLoS Comput. Biol. 12, e1005187.*

70. *Haynes, W. A., Higdon, R., Stanberry, L., Collins, D., and Kolker, E. (2013). Differential expression analysis for pathways. PLoS Comput. Biol. 9, e1002967.*

71. *Ibrahim, M. A.-H., Jassim, S., Cawthorne, M. A., and Langlands, K. (2012). A topology-based score for pathway enrichment. J. Comput. Biol. 19, 563–573.*

72. *Pavlova, N. N., and Thompson, C. B. (2016). The emerging hallmarks of cancer metabolism. Cell Metab. 23, 27–47.*

73. *Kaelin, W. G., and McKnight, S. L. (2013). Influence of metabolism on epigenetics and disease. Cell 153, 56–69.*

74. *Al-Shahrour, F., Díaz-Uriarte, R., and Dopazo, J. (2004). FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics 20, 578–580.*

75. *Khatri, P., Sirota, M., and Butte, A. J. (2012). Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput. Biol. 8, e1002375.*

76. *Iorio, F., Bosotti, R., Scacheri, E., Belcastro, V., Mithbaokar, P., Ferriero, R., Murino, L., Tagliaferri, R., Brunetti-Pierri, N., Isacchi, A., et al. (2010). Discovery of drug mode of action and drug repositioning from transcriptional responses. Proc Natl Acad Sci USA 107, 14621–14626.*

77. *Corsello, S. M., Bittker, J. A., Liu, Z., Gould, J., McCarren, P., Hirschman, J. E., Johnston, S. E., Vrcic, A., Wong, B., Khan, M., et al. (2017). The Drug Repurposing Hub: a next-generation drug library and information resource. Nat. Med. 23, 405–408.*

78. *Amadoz, A., Hidalgo, M. R., Çubuk, C., Carbonell-Caballero, J., and Dopazo, J. (2019). A comparison of mechanistic signaling pathway activity analysis methods. Brief. Bioinformatics 20, 1655–1668.*

79. *Bordbar, A., Monk, J. M., King, Z. A., and Palsson, B. O. (2014). Constraint-based models predict metabolic and associated cellular functions. Nat. Rev. Genet. 15, 107–120.*

80. *Folger, O., Jerby, L., Frezza, C., Gottlieb, E., Ruppin, E., and Shlomi, T. (2011). Predicting selective drug targets in cancer through metabolic networks. Mol. Syst. Biol. 7, 501.*

81. *Agren, R., Bordel, S., Mardinoglu, A., Pornputtapong, N., Nookaew, I., and Nielsen, J. (2012). Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT. PLoS Comput. Biol. 8, e1002518.*

82. *Fisher, J., and Henzinger, T. A. (2007). Executable cell biology. Nat. Biotechnol. 25, 1239–1249.*

83. *Heirendt, L., Arreckx, S., Pfau, T., Mendoza, S. N., Richelle, A., Heinken, A., Haraldsdóttir, H. S., Wachowiak, J., Keating, S. M., Vlasov, V., et al. (2019). Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. Nat. Protoc. 14, 639–702.*

84. *Muto, A., Kotera, M., Tokimatsu, T., Nakagawa, Z., Goto, S., and Kanehisa, M. (2013). Modular architecture of metabolic pathways revealed by conserved sequences of reactions. J. Chem. Inf. Model. 53, 613–622.*

85. *Cubuk, C., Hidalgo, M. R., Amadoz, A., Pujana, M. A., Mateo, F., Herranz, C., Carbonell-Caballero, J., and Dopazo, J. (2018). Gene Expression Integration into Pathway Modules Reveals a Pan-Cancer Metabolic Landscape. Cancer Res. 78, 6059–6072.*

86. *Çubuk, C., Hidalgo, M. R., Amadoz, A., Rian, K., Salavert, F., Pujana, M. A., Mateo, F., Herranz, C., Carbonell-Caballero, J., and Dopazo, J. (2019). Differential metabolic activity and discovery of therapeutic targets using summarized metabolic pathway models. npj Syst.*

Biol. Appl. 5, 7.

87. *Jensen, P. A., Lutz, K. A., and Papin, J. A. (2011). TIGER: Toolbox for integrating genome-scale metabolic models, expression data, and transcriptional regulatory networks. BMC Syst. Biol. 5, 147.*

88. *Yoav, B., and Yosef, H. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J.R. Statist 57, 289–300.*

89. *Breiman, L. (2001). Random Forests. Springer Science and Business Media LLC.*

90. *Vapnik, V. N. (1998). Statistical Learning Theory 1st ed. (New York: Wiley-interscience).*

91. *Wishart, D. S. (2008). DrugBank and its relevance to pharmacogenomics. Pharmacogenomics 9, 1155–1162.*

92. *Cho, K.-H., Lee, S., Kim, D., Shin, D., Joo, J. I., and Park, S.-M. (2017). Cancer reversion, a renewed challenge in systems biology. Current Opinion in Systems Biology 2, 49–58.*

93. *Salavert, F., García-Alonso, L., Sánchez, R., Alonso, R., Bleda, M., Medina, I., and Dopazo, J. (2016). Web-based network analysis and visualization using CellMaps. Bioinformatics 32, 3041–3043.*

94. *Tsherniak, A., Vazquez, F., Montgomery, P. G., Weir, B. A., Kryukov, G., Cowley, G. S., Gill, S., Harrington, W. F., Pantel, S., Krill-Burger, J. M., et al. (2017). Defining a cancer dependency map. Cell 170, 564-576.e16.*

95. *Terunuma, A., Putluri, N., Mishra, P., Mathé, E. A., Dorsey, T. H., Yi, M., Wallace, T. A., Issaq, H. J., Zhou, M., Killian, J. K., et al. (2014). MYC-driven accumulation of 2-hydroxyglutarate is associated with breast cancer prognosis. J. Clin. Invest. 124, 398–412.*

96. *Hakimi, A. A., Reznik, E., Lee, C.-H., Creighton, C. J., Brannon, A. R., Luna, A., Aksoy, B. A., Liu, E. M., Shen, R., Lee, W., et al. (2016). An integrated metabolic atlas of clear cell renal cell carcinoma. Cancer Cell 29, 104–116.*

97. *Tarca, A. L., Draghici, S., Khatri, P., Hassan, S. S., Mittal, P., Kim, J.-S., Kim, C. J., Kusanovic, J. P., and Romero, R. (2009). A novel signaling pathway impact analysis. Bioinformatics 25, 75–82.*

98. *Shlomi, T., Cabili, M. N., Herrgård, M. J., Palsson, B. Ø., and Ruppin, E. (2008). Network-based prediction of human tissue-specific metabolism. Nat. Biotechnol. 26, 1003–1010.*

99. *Zur, H., Ruppin, E., and Shlomi, T. (2010). iMAT: an integrative metabolic analysis tool. Bioinformatics 26, 3140–3142.*

100. *Thiele, I., Swainston, N., Fleming, R. M. T., Hoppe, A., Sahoo, S., Aurich, M. K., Haraldsdottir, H., Mo, M. L., Rolfsson, O., Stobbe, M. D., et al. (2013). A community-driven*

global reconstruction of human metabolism. Nat. Biotechnol. 31, 419–425.

101. Auslander, N., Wagner, A., Oberhardt, M., and Ruppin, E. (2016). Data-Driven Metabolic Pathway Compositions Enhance Cancer Survival Prediction. PLoS Comput. Biol. 12, e1005125.

102. Chia, S. K., Bramwell, V. H., Tu, D., Shepherd, L. E., Jiang, S., Vickery, T., Mardis, E., Leung, S., Ung, K., Pritchard, K. I., et al. (2012). A 50-gene intrinsic subtype classifier for prognosis and prediction of benefit from adjuvant tamoxifen. Clin. Cancer Res. 18, 4465–4472.

103. Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. Nature 490, 61–70.

104. Shaul, Y. D., Freinkman, E., Comb, W. C., Cantor, J. R., Tam, W. L., Thiru, P., Kim, D., Kanarek, N., Pacold, M. E., Chen, W. W., et al. (2014). Dihydropyrimidine accumulation is required for the epithelial-mesenchymal transition. Cell 158, 1094–1109.

105. Fong, S. S., Nanchen, A., Palsson, B. O., and Sauer, U. (2006). Latent pathway activation and increased pathway capacity enable Escherichia coli adaptation to loss of key metabolic enzymes. J. Biol. Chem. 281, 8024–8033.

106. Lewis, N. E., Cho, B.-K., Knight, E. M., and Palsson, B. O. (2009). Gene expression profiling and the use of genome-scale in silico models of Escherichia coli for analysis: providing context for content. J. Bacteriol. 191, 3437–3444.

107. Jordheim, L. P., Durantel, D., Zoulim, F., and Dumontet, C. (2013). Advances in the development of nucleoside and nucleotide analogues for cancer and viral diseases. Nat. Rev. Drug Discov. 12, 447–464.

108. Schug, Z. T., Vande Voorde, J., and Gottlieb, E. (2016). The metabolic fate of acetate in cancer. Nat. Rev. Cancer 16, 708–717.

109. Kamphorst, J. J., Chung, M. K., Fan, J., and Rabinowitz, J. D. (2014). Quantitative analysis of acetyl-CoA production in hypoxic cancer cells reveals substantial contribution from acetate. Cancer Metab. 2, 23.

110. Sahu, N., Dela Cruz, D., Gao, M., Sandoval, W., Haverty, P. M., Liu, J., Stephan, J.-P., Haley, B., Classon, M., Hatzivassiliou, G., et al. (2016). Proline Starvation Induces Unresolved ER Stress and Hinders mTORC1-Dependent Tumorigenesis. Cell Metab. 24, 753–761.

111. Neman, J., Termini, J., Wilczynski, S., Vaidehi, N., Choy, C., Kowolik, C. M., Li, H., Hambrecht, A. C., Roberts, E., and Jandial, R. (2014). Human breast cancer metastases to the brain display GABAergic properties in the neural niche. Proc Natl Acad Sci USA 111,

984–989.

112. *Beloribi-Djefaflia, S., Vasseur, S., and Guillaumond, F. (2016). Lipid metabolic reprogramming in cancer cells. Oncogenesis 5, e189.*

113. *Visus, C., Ito, D., Dhir, R., Szczepanski, M. J., Chang, Y. J., Latimer, J. J., Grant, S. G., and DeLeo, A. B. (2011). Identification of Hydroxysteroid (17β) dehydrogenase type 12 (HSD17B12) as a CD8+ T-cell-defined human tumor antigen of human carcinomas. Cancer Immunol. Immunother. 60, 919–929.*

114. *McGuire, W. L. (1973). Estrogen receptors in human breast cancer. J. Clin. Invest. 52, 73–77.*

115. *Di Stasi, D., Vallacchi, V., Campi, V., Ranzani, T., Daniotti, M., Chiodini, E., Fiorentini, S., Greeve, I., Prinetti, A., Rivoltini, L., et al. (2005). DHCR24 gene expression is upregulated in melanoma metastases and associated to resistance to oxidative stress-induced apoptosis. Int. J. Cancer 115, 224–230.*

116. *Uhlen, M., Zhang, C., Lee, S., Sjöstedt, E., Fagerberg, L., Bidkhori, G., Benfeitas, R., Arif, M., Liu, Z., Edfors, F., et al. (2017). A pathology atlas of the human cancer transcriptome. Science 357.*

117. *Elegbede, J. A., Elson, C. E., Qureshi, A., Dennis, W. H., and Yatvin, M. B. (1986). Increasing the thermosensitivity of a mammary tumor (CA755) through dietary modification. Eur. J. Cancer Clin. Oncol. 22, 607–615.*

118. *Antalis, C. J., Uchida, A., Buhman, K. K., and Siddiqui, R. A. (2011). Migration of MDA-MB-231 breast cancer cells depends on the availability of exogenous lipids and cholesterol esterification. Clin. Exp. Metastasis 28, 733–741.*

119. *Vander Heiden, M. G. (2011). Targeting cancer metabolism: a therapeutic window opens. Nat. Rev. Drug Discov. 10, 671–684.*

120. *Dang, L., White, D. W., Gross, S., Bennett, B. D., Bittinger, M. A., Driggers, E. M., Fantin, V. R., Jang, H. G., Jin, S., Keenan, M. C., et al. (2009). Cancer-associated IDH1 mutations produce 2-hydroxyglutarate. Nature 462, 739–744.*

121. *Anastasiou, D., Yu, Y., Israelsen, W. J., Jiang, J.-K., Boxer, M. B., Hong, B. S., Tempel, W., Dimov, S., Shen, M., Jha, A., et al. (2012). Pyruvate kinase M2 activators promote tetramer formation and suppress tumorigenesis. Nat. Chem. Biol. 8, 839–847.*

122. *Hsu, P. P., and Sabatini, D. M. (2008). Cancer cell metabolism: Warburg and beyond. Cell 134, 703–707.*

123. *Deberardinis, R. J., Sayed, N., Ditsworth, D., and Thompson, C. B. (2008). Brick by brick: metabolism and tumor cell growth. Curr. Opin. Genet. Dev. 18, 54–61.*

124. *Hu, J., Locasale, J. W., Bielas, J. H., O'Sullivan, J., Sheahan, K., Cantley, L. C., Vander Heiden, M. G., and Vitkup, D. (2013). Heterogeneity of tumor-induced gene expression changes in the human metabolic network. Nat. Biotechnol. 31, 522–529.*

125. *Pfister, S. X., Markkanen, E., Jiang, Y., Sarkar, S., Woodcock, M., Orlando, G., Mavrommati, I., Pai, C.-C., Zalmas, L.-P., Drobnitzky, N., et al. (2015). Inhibiting WEE1 Selectively Kills Histone H3K36me3-Deficient Cancers by dNTP Starvation. Cancer Cell 28, 557–568.*

126. *Cowley, G. S., Weir, B. A., Vazquez, F., Tamayo, P., Scott, J. A., Rusin, S., East-Seletsky, A., Ali, L. D., Gerath, W. F., Pantel, S. E., et al. (2014). Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies. Sci. Data 1, 140035.*

127. *Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci. Signal. 6, pl1.*

128. *D.R., C. (1972). Regression Models and Life-Tables. J R Stat Soc Series B Stat Methodol 34, 187–220.*

129. *Hao, Y., Samuels, Y., Li, Q., Krokowski, D., Guan, B.-J., Wang, C., Jin, Z., Dong, B., Cao, B., Feng, X., et al. (2016). Oncogenic PIK3CA mutations reprogram glutamine metabolism in colorectal cancer. Nat. Commun. 7, 11971.*

130. *Makinoshima, H., Umemura, S., Suzuki, A., Nakanishi, H., Maruyama, A., Udagawa, H., Mimaki, S., Matsumoto, S., Niho, S., Ishii, G., et al. (2018). Metabolic Determinants of Sensitivity to Phosphatidylinositol 3-Kinase Pathway Inhibitor in Small-Cell Lung Carcinoma. Cancer Res. 78, 2179–2190.*

131. *Mullarky, E., Lucki, N. C., Beheshti Zavareh, R., Anglin, J. L., Gomes, A. P., Nicolay, B. N., Wong, J. C. Y., Christen, S., Takahashi, H., Singh, P. K., et al. (2016). Identification of a small molecule inhibitor of 3-phosphoglycerate dehydrogenase to target serine biosynthesis in cancers. Proc Natl Acad Sci USA 113, 1778–1783.*

132. *Warburg, O. (1925). The metabolism of carcinoma cells. J. Cancer Res. 9, 148–163.*

133. *Carracedo, A., Cantley, L. C., and Pandolfi, P. P. (2013). Cancer metabolism: fatty acid oxidation in the limelight. Nat. Rev. Cancer 13, 227–232.*

134. *Akbani, R., Ng, P. K. S., Werner, H. M. J., Shahmoradgoli, M., Zhang, F., Ju, Z., Liu, W., Yang, J.-Y., Yoshihara, K., Li, J., et al. (2014). A pan-cancer proteomic perspective on The Cancer Genome Atlas. Nat. Commun. 5, 3887.*

135. *Kuzu, O. F., Noory, M. A., and Robertson, G. P. (2016). The role of cholesterol in cancer. Cancer Res. 76, 2063–2070.*

136. *Ehrlich, M. (2009). DNA hypomethylation in cancer cells. Epigenomics 1, 239–259.*

137. *Dillon, B. J., Prieto, V. G., Curley, S. A., Ensor, C. M., Holtsberg, F. W., Bomalaski, J. S., and Clark, M. A. (2004). Incidence and distribution of argininosuccinate synthetase deficiency in human cancers: a method for identifying cancers sensitive to arginine deprivation. Cancer 100, 826–833.*

138. *Wang, Q., Tiffen, J., Bailey, C. G., Lehman, M. L., Ritchie, W., Fazli, L., Metierre, C., Feng, Y. J., Li, E., Gleave, M., et al. (2013). Targeting amino acid transport in metastatic castration-resistant prostate cancer: effects on cell cycle, cell growth, and tumor development. J Natl Cancer Inst 105, 1463–1473.*

139. *Coelho, M., Moz, M., Correia, G., Teixeira, A., Medeiros, R., and Ribeiro, L. (2015). Antiproliferative effects of β-blockers on human colorectal cancer cells. Oncol. Rep. 33, 2513–2520.*

140. *Sasisekharan, R., Shriver, Z., Venkataraman, G., and Narayanasami, U. (2002). Roles of heparan-sulphate glycosaminoglycans in cancer. Nat. Rev. Cancer 2, 521–528.*

141. *Huang, Y., Keen, J. C., Pledgie, A., Marton, L. J., Zhu, T., Sukumar, S., Park, B. H., Blair, B., Brenner, K., Casero, R. A., et al. (2006). Polyamine analogues down-regulate estrogen receptor alpha expression in human breast cancer cells. J. Biol. Chem. 281, 19055–19063.*

142. *Fadare, O., and Tavassoli, F. A. (2008). Clinical and pathologic aspects of basal-like breast cancers. Nat. Clin. Pract. Oncol. 5, 149–159.*

143. *Patra, K. C., and Hay, N. (2014). The pentose phosphate pathway and cancer. Trends Biochem. Sci. 39, 347–354.*

144. *Dimitrieva, S., Schlapbach, R., and Rehrauer, H. (2016). Prognostic value of cross-omics screening for kidney clear cell renal cancer survival. Biol. Direct 11, 68.*

145. *Hakomori, S. (1984). Glycosphingolipids as differentiation-dependent, tumor-associated markers and as regulators of cell proliferation. Trends Biochem. Sci. 9, 453–459.*

146. *Martinez, J. D., Stratagoules, E. D., LaRue, J. M., Powell, A. A., Gause, P. R., Craven, M. T., Payne, C. M., Powell, M. B., Gerner, E. W., and Earnest, D. L. (1998). Different bile acids exhibit distinct biological effects: the tumor promoter deoxycholic acid induces apoptosis and the chemopreventive agent ursodeoxycholic acid inhibits cell proliferation. Nutr. Cancer 31, 111–118.*

147. *Newman, A. C., and Maddocks, O. D. K. (2017). One-carbon metabolism in cancer. Br. J. Cancer 116, 1499–1504.*

148. *Amelio, I., Cutruzzolá, F., Antonov, A., Agostini, M., and Melino, G. (2014). Serine and glycine metabolism in cancer. Trends Biochem. Sci. 39, 191–198.*

149. *Ahern, T. P., Lash, T. L., Damkier, P., Christiansen, P. M., and Cronin-Fenton, D. P. (2014). Statins and breast cancer prognosis: evidence and opportunities. Lancet Oncol. 15, e461-8.*

150. *Wang, G., Cao, R., Wang, Y., Qian, G., Dan, H. C., Jiang, W., Ju, L., Wu, M., Xiao, Y., and Wang, X. (2016). Simvastatin induces cell cycle arrest and inhibits proliferation of bladder cancer cells via PPARγ signalling pathway. Sci. Rep. 6, 35783.*

151. *Coleman, R. E. (2000). Management of bone metastases. Oncologist 5, 463–470.*

152. *Rennert, G., Pinchev, M., Rennert, H. S., and Gruber, S. B. (2011). Use of bisphosphonates and reduced risk of colorectal cancer. J. Clin. Oncol. 29, 1146–1150.*

153. *Mathur, D., Stratikopoulos, E., Ozturk, S., Steinbach, N., Pegno, S., Schoenfeld, S., Yong, R., Murty, V. V., Asara, J. M., Cantley, L. C., et al. (2017). PTEN regulates glutamine flux to pyrimidine synthesis and sensitivity to dihydroorotate dehydrogenase inhibition. Cancer Discov. 7, 380–390.*

154. *Forbes, S. A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C., Ward, S., et al. (2015). COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. 43, D805-11.*

155. *Hernansaiz-Ballesteros, R. D., Salavert, F., Sebastián-León, P., Alemán, A., Medina, I., and Dopazo, J. (2015). Assessing the impact of mutations found in next generation sequencing data over human signaling pathways. Nucleic Acids Res. 43, W270-5.*

156. *Personalized Medicine | Drug Discovery Platform | Cellworks Pipeline Available at: https://cellworks.life/technology/pipeline [Accessed March 11, 2020].*

157. *Science | Achilles Therapeutics Available at: https://achillestx.com/science/ [Accessed March 11, 2020].*

158. *Gustafsson, M., Nestor, C. E., Zhang, H., Barabási, A.-L., Baranzini, S., Brunak, S., Chung, K. F., Federoff, H. J., Gavin, A.-C., Meehan, R. R., et al. (2014). Modules, networks and systems medicine for understanding disease and aiding diagnosis. Genome Med. 6, 82.*

159. *Jones, S., Anagnostou, V., Lytle, K., Parpart-Li, S., Nesselbush, M., Riley, D. R., Shukla, M., Chesnick, B., Kadan, M., Papp, E., et al. (2015). Personalized genomic analyses for cancer mutation discovery and interpretation. Sci. Transl. Med. 7, 283ra53.*

160. *Stegmeier, F., Warmuth, M., Sellers, W. R., and Dorsch, M. (2010). Targeted cancer therapies in the twenty-first century: lessons from imatinib. Clin. Pharmacol. Ther. 87, 543–552.*