UNIVERSIDADE D
COIMBRA

Tânia Luísa Barbosa Barata

Metabolic Modeling of Cancer Stem Cells and Interplay Between Cancer Metabolism and Epigenetics

Tese no âmbito do Doutoramento em Biociências ramo de especialização em Biologia Celular e Molecular orientada pelo Doutor Ricardo Neves Pires das Neves, Doutora Paula Cristina Veríssimo Pires e Professor Doutor Miguel Francisco Pereira Almeida Rocha e apresentada ao Departamento de Ciências da Vida da Faculdade de Ciências e Tecnologia da Universidade de Coimbra.

Dezembro de 2022

Faculdade de Ciências e Tecnologia da Universidade de Coimbra
Departamento de Ciências da Vida

Metabolic Modeling of Cancer Stem Cells and Interplay Between Cancer Metabolism and Epigenetics

Tânia Luísa Barbosa Barata

Tese no âmbito do Doutoramento em Biociências ramo de especialização em Biologia Celular e Molecular orientada pelo Doutor Ricardo Neves Pires das Neves, Doutora Paula Cristina Veríssimo Pires e Professor Doutor Miguel Francisco Pereira Almeida Rocha e apresentada ao Departamento de Ciências da Vida da Faculdade de Ciências e Tecnologia da Universidade de Coimbra.

Dezembro de 2022

1 2 9 0

UNIVERSIDADE D
COIMBRA

Para o meu avô

Funding

This thesis was supported by:
FCT – Fundação para a Ciência e a Tecnologia under projects *UIDB/04539/2020* and *UIDP/04539/2020* (Strategic Plan CIBB). We also acknowledge the PhD studentship with reference *SFRH/BD/123028/2016* funded by *FCT*.

Agradecimentos

Antes de mais, gostaria de agradecer aos meus pais e ao meu irmão. Obrigada pelo apoio e por estarem sempre disponíveis para ouvir as minhas peripécias. Sinto uma enorme gratidão pela paciência, carinho e sacrifícios que fizeram para que eu pudesse sonhar em ingressar num doutoramento.

De seguida, gostaria de agradecer aos meus orientadores. Ao Dr. Miguel Rocha pela disponibilidade para me orientar na área de Bioinformática, apesar dos inúmeros orientandos que tinha à data. Agradeço o facto de ter arranjado tempo na sua agenda atarefada para me ajudar sempre que necessitava e pela perspetiva realista que transmitia quando as coisas corriam menos bem. Ao Dr. Ricardo Neves quero agradecer a disponibilidade pelo acompanhamento, e pelas críticas construtivas na parte biológica da tese, bem como pelo entusiasmo que transmite nas discussões científicas.

Gostaria ainda de agradecer aos meus orientadores "oficiosos", os colegas do grupo. Agradeço especialmente ao Vítor Vieira, Rúben Rodrigues, Vítor Pereira e Sophia Santos por se mostrarem sempre disponíveis para me dar dicas e auxiliar quando encontrava dificuldades técnicas. Obrigada a todos os restantes colegas pelo ambiente relaxado, simpático e de entre-ajuda. Foi divertido trabalhar num grupo que funciona como um.

Por fim quero agradecer à Raquel Cunha, Tiago Sousa e Alexandra Ferreira pelo constante encorajamento e pela amizade ao longo deste período.

Abstract

Cancer is a disease with a high mortality rate whose incidence has risen in the past years. Cancer Stem Cells (CSCs) are known to contribute to cancer aggressiveness, metastasis, chemo/radio-therapy resistance, and tumor recurrence. Furthermore, recent studies have emphasized the importance of metabolic reprogramming of CSCs for the maintenance and progression of the cancer phenotype through the fulfillment of the energetic requirements and the supply of substrates fundamental for fast-cell growth. Therefore, it is of paramount importance to develop therapeutic strategies tailored to target the metabolism of CSCs.

Among the factors that can contribute to cancer onset and progression is the imbalance of epigenetic regulatory mechanisms like DNA methylation, which can promote aberrant gene expression profiles without affecting the DNA sequence. Those mechanisms can influence the transcription of genes encoding signaling and regulatory proteins, but also metabolic enzymes. Hence, the disruption of epigenetic regulation may induce metabolic shifts that contribute to the acquisition of cancerous phenotypes. Likewise, since some metabolites are substrates and cofactors of epigenetic regulators, their availability can impair epigenetic mechanisms, in such a way that metabolic shifts may feed cancer progression and onset through epigenetic deregulation. Given the interplay between metabolism, cancer, and epigenetics, recent research has been developed on the frontiers between those biological processes in an attempt to identify new disease mechanisms and potential therapeutic targets.

Genome-Scale Metabolic Models (GSMMs) are mathematical representations of a network of all metabolic reactions in a cell and genes encoding enzymes catalyzing those reactions that can be used to simulate *in silico* the metabolic state of cells. GSMMs have been very useful in the past to predict metabolic phenotypes of different organisms and cell types, including cancer cells. Nevertheless, GSMMs of CSCs or of differentiated Cancer Cells (CCs) embodying the interaction between metabolism and DNA methylation, which is an epigenetic

mechanism, have not been developed so far.

In the first study presented in this thesis, computational GSMMs were built for CSCs and CCs of ten different tissues using reconstruction algorithms. The best reconstruction strategy was selected and implemented to obtain the models. Models were gapfilled to be able to simulate growth and perform metabolic tasks, and then, they were validated through the comparison of simulated essential genes and lethal genes identified from gene knockout experiments. Flux simulations were used to predict metabolic phenotypes, identify potential therapeutic targets, and spot already-known Transcription Factors (TFs), miRNAs, and antimetabolites that could be used as part of drug repurposing strategies against cancer. Furthermore, results were in accordance with experimental evidence, provided insights into new metabolic mechanisms for already known agents, and allowed for the identification of potential new targets and compounds that could be interesting for further *in vitro* and *in vivo* validation.

In the second study of this thesis, a generic GSMM of a human cell integrating DNA methylation or demethylation reactions collected from literature and databases was first obtained, and then, the best of different reconstruction strategies was identified and applied to the generic model to create GSMMs for 31 human cancer cell lines. Genome Scale Metabolic Models enhanced with Enzymatic Constraints using Kinetic and Omics data (GECKOs) were subsequently built based on those GSMMs to improve the accuracy of the simulated reaction fluxes without the need to pre-define uptake or secretion rates for any metabolite. Furthermore, cell-line specific DNA methylation levels were included in the models in the shape of coefficients of DNA composition reaction in an effort to depict the influence of metabolism over global DNA methylation in different cancer cell lines. Flux simulations demonstrated the ability of these models to provide simulated fluxes of exchange reactions similar to the equivalent experimentally measured uptake/secretion rates and to make good functional predictions. These models might be useful in the future identification of potential new therapeutic targets against cancer.

Keywords: Cancer Stem Cells (CSCs), Genome-Scale Metabolic Models (GSMMs), differentiated Cancer Cells (CCs), DNA methylation, Genome Scale Metabolic Models enhanced with Enzymatic Constraints using Kinetic and Omics data (GECKOs).

Resumo

O cancro é uma doença com elevada taxa de mortalidade cuja incidência tem aumentado nos últimos anos. As Células Cancerígenas Estaminais (CSCs) são conhecidas por contribuírem para a agressividade, metastização, resistência à quimio/radioterapia, e recorrência do cancro. Além disso, estudos recentes têm enfatizado a importância da reprogramação metabólica das CSCs para a manutenção e progressão do fenótipo cancerígeno através da satisfação de necessidades energéticas e o fornecimento de metabolitos fundamentais para o rápido crescimento celular. Consequentemente, é indispensável desenvolver estratégias terapêuticas específicas dirigidas ao metabolismo das CSCs.

Entre os fatores que podem contribuir para o aparecimento e progressão do cancro está o desequilíbrio de mecanismos regulatórios epigenéticos como a metilação do DNA, que pode promover perfis de expressão génica aberrantes sem afetar a sequência do DNA. Esses mecanismos podem influenciar a transcrição de genes que codificam proteínas de sinalização e proteínas regulatórias, mas também enzimas. Logo, a perturbação de regulação epigenética pode induzir alterações metabólicas que contribuem para a aquisição de fenótipos cancerígenos. Do mesmo modo, visto que alguns metabolitos são substratos e cofatores de reguladores epigenéticos, a sua disponibilidade pode afetar mecanismos epigenéticos de tal forma que as alterações metabólicas podem contribuir para o aparecimento e a progressão do cancro através de desregulação epigenética. Dada a interação entre metabolismo, cancro e epigenética, investigação recente tem sido desenvolvida na fronteira entre esses processos biológicos numa tentativa de identificar novos mecanismos de doença e potenciais alvos terapêuticos.

Modelos metabólicos à escala genómica (GSMMs) são representações matemáticas da rede composta por todas as reações metabólicas de uma célula e dos genes que codificam enzimas que catalisam essas reações, que podem ser usados para simular *in silico* o estado metabólico das células. GSMMs têm sido muito úteis na previsão de fenótipos metabólicos de diferentes organismos e

tipos celulares, incluindo células cancerígenas. No entanto, GSMMs de CSCs ou de células cancerígenas diferenciadas (CCs) materializando a interação entre metabolismo e metilação de DNA, o qual é um mecanismo epigenético, não foram desenvolvidos até ao momento.

No primeiro estudo apresentado nesta tese, GSMMs foram construídos para CSCs e CCs de dez diferentes tecidos, usando algoritmos de reconstrução. A melhor estratégia de reconstrução foi selecionada e implementada para obter os modelos. Os modelos foram "gapfilled" para ser possível simular crescimento e para que pudessem realizar "tasks" metabólicas, e de seguida foram validados através da comparação de genes essenciais simulados e genes letais identificados com experiências de "knockout" génico. Simulações de fluxo foram usadas para prever fenótipos metabólicos, identificar potenciais alvos terapêuticos e detetar factores de transcrição (TFs), miRNAs e antimetabolitos já conhecidos que possam ser usados como estratégia de reutilização de medicamentos contra o cancro. Além disso, os resultados estavam de acordo com a evidência experimental, forneceram uma perceção de novos mecanismos metabólicos para agentes já conhecidos, e permitiram a identificação de novos alvos potenciais e compostos que possam ser interessantes para posterior validação *in vitro* e *in vivo*.

No segundo estudo desta tese, um GSMM genérico de uma célula humana, integrando reações de metilação e demetilação do DNA recolhidas da literatura e bases de dados foi primeiramente obtido, e depois, a melhor de diferentes estratégias de reconstrução foi identificada e aplicada ao modelo genérico para criar GSMMs de 31 linhas celulares de cancro humano. Modelos metabólicos à escala genómica melhorados com limitações enzimáticas usando dados cinéticos e ómicos (GECKOs) foram posteriormente construídos a partir de GSMMs para melhorar a precisão da simulação de fluxos sem a necessidade de pré-definir taxas de captação ou produção de qualquer metabolito. Além disso, níveis de metilação do DNA específicos para cada linha celular foram incluídos nos modelos sob a forma de coeficientes na reação de composição do DNA com o propósito de retratar a influência do metabolismo sobre a metilação global do DNA in diferentes linhas celulares de cancro. As simulações de fluxo demonstraram a capacidade destes modelos de simular fluxos de reações de troca similares às equivalentes taxas de captação ou secreção e de fazer boas previsões funcionais. Estes modelos poderão ser utilizados na identificação futura de novos potenciais alvos terapêuticos contra o cancro.

Palavras Chave: células cancerígenas estaminais (CSCs), modelos metabólicos à escala genómica (GSMMs), células cancerígenas diferenciadas (CCs), metilação do DNA, modelos metabólicos à escala genómica melhorados com limitações enzimáticas usando dados cinéticos e ómicos (GECKOs).

Contents

List of Figures

List of Tables

Abbreviations

α-KG alpha-Ketoglutarate

k_{cat} turnover number

2-HG 2-Hydroxyglutarate

3PG 3-Phosphoglycerate

5hmC 5-hydroxymethyl cytosine

5mC 5-methyl cytosine

ACC Acetyl-CoA Carboxylase

acetyl-CoA acetyl-Coenzyme A

ACL ATP Citrate Lyase

ACSS2 Acetyl-CoA Synthetase 2

ADRP Adipose Differentiation Related Protein

AGC Aspartate-Glutamate Carrier

AhR Aryl hydrocarbon Receptor

ALDH Aldehyde Dehydrogenase

ALT Alanine aminotransferase

AML Acute Myeloid Leukemia

AMPK AMP-activated Protein Kinase

ASCT2 Alanine, Serine, Cysteine Transporter 2

AST Aspartate aminotransferase

ATF4 Activating Transcription Factor 4

ATGL Adipose Tissue Triacylglycerol Lipase

ATP Adenosine Triphosphate

CAD Carbamoyl-phosphate synthetase 2, Aspartate transcarbamylase and Dihydroorotase

CeCaFDB Central Carbon Metabolic Flux Database

CGI CpG islands

CIC Citrate carrier

CoA Co-enzyme A

CRR Core Reaction-Required

CSC Cancer Stem Cell

DDBJ DNA Data Bank of Japan

DHAP Dihydroxyacetone phosphate

DHFR Dihydrofolate reductase

DNMT DNA methyltransferase

dTMP deoxythymidine monophosphate

dUMP deoxyuridine monophosphate

E4P Erythrose-4-Phosphate

EC Enzyme Commission

EDC Expression Data-Compatible

EFM Elementary Flux Mode

EHMN Edinburgh Human Metabolic Network

EMT Epithelial-to-Mesenchymal Transition

ENA European Nucleotide Archive

ETC Electron Transport Chain

F1,6BP Fructose-1,6-bisphosphate

F2,6BP Fructose-2,6-biphosphate

F6P Fructose-6-phosphate

FABP Fatty Acid Binding Protein

FACS Fluoresce Activated Cell Sorting

FAD Flavin Adenine Dinucleotide

$FADH_2$ reduced Flavin Adenine Dinucleotide

FAO Fatty Acid Oxidation

FASN Fatty Acid Synthetase

FBA Flux Balance Analysis

FBAwMC FBA with Molecular Crowding

FBS Fetal Bovine Serum

FH Fumarate Hydratase

FVA Flux Variability Analysis

G3P Glyceraldehyde-3-Phosphate

G6P Glucose-6-Phosphate

G6PD Glucose-6-Phosphate Dehydrogenase

GC Guanine-cytosine

GDC Genomic Data Commons

GDH Glutamate dehydrogenase

GECKO Genome Scale Metabolic Models enhanced with Enzymatic Constraints using Kinetic and Omics data

GEO Gene Expression Omnibus

GIM_3E Gene Inactivation Moderated by Metabolism, Metabolomics and Expression

GIMME Gene Inactivation Moderated by Metabolism and Expression

GIMMEp Gene Inactivity Moderated by Metabolism and Expression by proteome

GlcNAc N-Acetylglucosamine

GLS Glutaminase

GLUT Glucose Transporter

GO Gene Ontology

GPR Gene-Protein-Reaction

GRM3 Metabotropic glutamate receptor 3

GSH Glutathione

GSK3β Glycogen Synthase Kinase-3 beta

GSMM Genome-Scale Metabolic Model

GSSG Glutathione disulfide

GTEx Genotype-Tissue Expression

GTPA Gene-Transcript-Protein-Association

HAT Histone acetyltransferase

HBP Hexosamine Biosynthetic Pathway

HCC Hepatocellular carcinoma

hCYS Homocysteine

HDAC Histone deacetylase

HDM Histone demethylase

HIF Hypoxia-Inducible Factor

HK Hexokinase

HMDB Human Metabolome Database

HMR Human Metabolic Reaction

HMT Histone methyltransferase

HPA Human Protein Atlas

HPM Human Proteome Map

HPRD Human Protein Reference Database

HSL Hormone-Sensitive Lipase

IDH Isocitrate dehydrogenase

IDO1 Indoleamine 2,3-dioxygenase 1

IL-4 Interleukin 4

iMAT integrated Metabolic Analysis Tool

INIT Integrative Network Inference for Tissues

JHDM Jumonji-domain-containing histone demethylase

LAT1 Lactate Transporter 1

LD Lipid Droplet

LDH Lactate dehydrogenase

LDLR LDL Receptor

LKB1 Liver Kinase B1

LP Linear Programming

LSD Lysine-Specific Demethylase

MBA Model Building Algorithm

mCADRE metabolic Context-specificity Assessed by Deterministic Reaction Evaluation

MCT Monocarboxylic Acid Transporter

ME Metabolism and gene Expression

Met methionine

meTHF 5,10-methylene-THF

MFA Metabolic Flux Analysis

MILP Mixed-Integer Linear Programming

MMP Matrix metalloproteinase

MOMA Minimization Of Metabolic Adjustment

MPA Metabolic Pathway Analysis

MPC Mitochondrial Pyruvate Carrier

MS Methionine Synthetase

MTD Metabolic Task-Derived

mTHF methyl-THF

MTHFD Methylene-THF Dehydrogenase

mTOR mammalian Target Of Rapamycin

mTORC1 mTOR Complex 1

MUFA Monounsaturated fatty acid

NAD^+ oxidized Nicotinamide Adenine Dinucleotide

NADH reduced Nicotinamide Adenine Dinucleotide

NADPH reduced Nicotinamide Adenine Dinucleotide Phosphate

NAFLD Non-Alcoholic Fat Liver Disease

NAM Nicotinamide

NF-kB Nuclear Factor Kappa light chain enhancer of activated B cells

NGS Next-Generation Sequencing

NH_4^+ Ammonia

NMR Nuclear Magnetic Resonance

NRF2 Nuclear factor erythroid-derived 2

OAA Oxaloacetate

OFR Objective Function-Required

OGC oxoglutarate(KG)-glutamate carrier

OGT GlcNAc transferase

OXPHOS Oxidative Phosphorylation

PDAC Pancreatic ductal adenocarcinoma

PDH Pyruvate Dehydrogenase

PDK Pyruvate Dehydrogenase Kinase

pFBA parsimonious enzyme usage FBA

PFK Phosphofructokinase-1

PFK2 Phosphofructokinase-2

PGC1α Peroxisome proliferator-activator 1 alpha

PGM Phosphoglycerate mutase

PHGDH Phosphoglycerate dehydrogenase

PI3K Phospho-Inositol 3 Kinase

PK Pyruvate Kinase

PKM2 Pyruvate Kinase M2

PPP Pentose-Phosphate Pathway

PRPS2 Ribose-phosphate pyrophosphokinase 2

PSAT1 Phosphoserine aminotransferase

PSPH Phosphoserine phosphatase

PTM Post-Translational Modification

R5P Ribose-5-Phosphate

Rb Retinoblastoma

ROOM Regulatory On/Off Minimization

ROS Reactive Oxygen Species

RTK Receptor Tyrosine Kinase

Ru5P Ribulose-5-Phosphate

SAH S-Adenosyl-Homocysteine

SAM S-Adenosyl-Methionine

SC normal Stem Cell

SCD Stearoyl-CoA Desaturase

SDH Succinate dehydrogenase

SHMT Serine hydroxymethyltransferase

SSP Serine Synthesis Pathway

stMFA stoichiometric MFA

TAG Triacylglyceride

TC Transport Classification

TCA Tricarboxylic acid

TDO2 Tryptophan 2,3-dioxygenase

TET Ten-eleven translocation

THF Tetrahydrofolate

TIGAR TP53-Inducible Glycolysis and Apoptosis Regulator

tINIT task-driven INIT

TKT Transketolase

Treg Regulatory T-cell

Tsc2 Tuberous sclerosis complex protein 2

TYMS Thymidylate synthase

VEGF Vascular Endothelial Growth Factor

VMH Virtual Metabolic Human

WBM Whole-Body-Metabolic

Xu5P Xylulose-5-Phosphate

Thesis Outline

Cancer is a disease of high incidence and mortality resulting from the deregulation of a wide range of molecular networks, and culminating in abnormal cell proliferation and evasion. One of those biological networks is metabolism. In fact, metabolic reprogramming is considered one of the hallmarks of cancer on account of its effect on cell survival and growth. In addition, many metabolites act as substrates and cofactors of epigenetic enzymes, while epigenetic modifications affect the expression of different genes, including those coding for metabolic enzymes. Therefore, emphasis has been given to the interplay between epigenetic regulation and metabolism in cancer.

Concurrently, there has been an increase in awareness of the importance of studying Cancer Stem Cells (CSCs), as they are believed to be the "origin" of cancer development, due to their stem cell-like properties, which also drive cancer aggressiveness, metastasis, recurrence, and evasion to conventional treatments.

In the past decades, a technical push has been put into place to create *in silico* metabolic models of distinct organisms and cell types at genome-scale, as a result of the rise in availability of *omics* data deposited in public databases, as well as the evolution in the mathematical formulation of metabolic models.

Even though the ability of cancer cells to proliferate fostered the reconstruction of metabolic models for those cells with cell growth as the main metabolic objective, few studies have attempted to build models specifically for CSCs, or that could emphasize the metabolism and epigenetic cross-talk in cancer.

The main motivation for the present thesis was to fill that knowledge gap, by creating tissue-specific *in silico* genome-scale metabolic models that could be used to: **i)** predict metabolic phenotypes specific to CSCs in comparison with their differentiated counterparts; **ii)** suggest potential therapeutic targets as well as known biological compounds that could be used in drug repurposing methodologies against cancer; **iii)** offer a mechanistic interpretation to the influence of metabolism in global DNA methylation in cancer.

To meet that goal a general introduction was made to the theme in chapter 1, followed by two studies presented in chapters 2 and 3, respectively, and a final conclusion summarizing the relevant findings of this thesis in chapter 4.

– The first study tries to clarify the specific objectives **i)** and **ii)** and has already been published: T. Barata, V. Vieira, R. Rodrigues, R. Pires das Neves, Rocha M., "Reconstruction of tissue-specific genome-scale metabolic models for human cancer stem cells", *Computers in Biology and Medicine*, vol. 142, pp.1-12, 2022 (DOI: 10.1016/J.COMPBIOMED.2021.105177).

– The second study tackles the objective **iii)**.

I declare that this thesis was written and organized by me, and I confirm that it has not been previously submitted, in whole or in part, to obtain another academic degree. I confirm that the work described was done by me and by the co-authors, in the case of joint publications.

The work presented in chapter 2 was conceptualized by me, Vítor Vieira, Rúben Rodrigues and Miguel Rocha. I performed all the analyzes and adapted some code previously created by co-authors in the sections 2.2.1 and 2.2.3. I wrote the first draft of the paper and incorporated later suggestions from other authors.

The work presented in chapter 3 was conceptualized by me, Vítor Pereira, Ricardo Neves and Miguel Rocha. I performed all the analyzes and the curation of the mass balance of reactions was done with the help of Sophia Santos. I wrote the first draft of the study and incorporated later suggestions from other authors.

Chapter 1

Introduction

1.1 Cancer

Just in 2020, around 19 million new cases of cancer have arisen and almost 10 million people died of this disease around the globe [1]. Cancer is a disease characterized by aberrant cell proliferation that can affect any type of cell in the body. Regardless of the large variability of cell types that may be affected, only a few cancer types are frequent [2]. In fact, only six cancer types (breast, lung, colon, prostate, stomach, liver, and cervical cancer) accounted for more than half of the cancer cases worldwide in 2020 [1].

The high incidence and mortality of this disease makes its study of paramount importance. This section focuses on traits of cancer and cancer stem cells, mainly metabolic reprogramming strategies applied by those cells to meet their need for long-term survival and proliferation.

1.1.1 Cancer hallmarks and origin

Several common characteristics to all types of cancer have been defined. The traditional cancer hallmarks comprise the ability to self-produce proliferative signals, block growth suppressors (such as TP53 and retinoblastoma-associated proteins), evade programmed cell death, induce angiogenesis, activate metastasis and enable replicative immortality, i.e. the ability to proliferate many times without entering in differentiation or apoptotic state. Two more hallmarks, not so extensively reported, are evasion of immunological destruction and reprogrammed cellular metabolism to support energetic requirements of neoplastic proliferation [3].

Despite its standard features, cancer exhibits wide complexity. Cancer di-

versity reflects not only in the variability of tissues affected, but also in its genetic heterogeneity among patients with the same type of cancer and cancer cells within the same tumor. Intra- and inter-tumor variability are a hurdle that conventional and even personalized therapeutic strategies cannot completely overcome. The main reasons are human-to-human genetic and epigenetic diversity, as well as subclone variety [4].

Many tumors present organized variability, where subclones are ordered in a hierarchical structure. *Clonal evolution theory* states that hierarchical subclone variability results from intra-tumor evolution through cumulative mutations and selection in different microenvironments. Conversely, the *Cancer Stem Cell hypothesis* claims that a small population of Cancer Stem Cells (CSCs), sharing the same pluripotency properties as normal Stem Cells (SCs), can generate all cells within the tumor in a hierarchical fashion, on account of their capacity to differentiate into distinct cell lineages [5].

1.1.2 Cancer stem cells

The study of CSCs has been influenced by research on SCs due to the similarity between these two. In addition to multipotency, CSCs, just like SCs, have been regarded as a rare population of quiescent cells that seldomly divide, but with the ability to self-renew [6]. CSCs have been described as cells capable of asymmetric division, giving rise to some daughter cells and other transient amplifying cells. While the first contribute to cancer maintenance, the latter rapidly divide, eventually differentiating into poorly tumorigenic bulk non-stem cells [7]. Nevertheless, both normal and cancer stem cells were shown to be abundant and actively divide in many tissues, like the epidermis and intestinal crypts. Furthermore, recent reports show that stem cell progeny does not necessarily have distinct fates [7].

Stem cells survive in a specific micro-environment, called stem cell niche. The niche comprises physical, paracrine, and even metabolic cues from neighbor supporting-stromal cells, extracellular matrix, nervous cells, and sometimes even endocrine signals transported by surrounding blood vessels that regulate gene expression and signaling pathways of stem cells, enabling the maintenance of their unique properties [8]. When there is no space within the stem cell niche for new daughter (cancer/normal) stem cells, they "fall off" outside the niche and start to differentiate. In that case, all the progeny shares the same fate, become differentiated cells, while no new daughter stem cells are created. This process is known as neutral competition. However, it is not irreversible. Tran-

sient amplifying and even fully differentiated daughter cells can re-enter the niche and dedifferentiate to replace dead (cancer/normal) stem cells. This plastic behavior makes CSCs difficult to eradicate because even when abolished with tailored therapies they can regenerate afterwards from differentiated cells [7].

A striking difference between cancer and normal stem cells is that the former become increasingly independent from the niche, as they can reproduce signals on their own, assuming a stem cell phenotype even outside of their original microenvironment. This CSC trait also halts differentiation inside the tumor. More CSCs than non-CSCs are produced, enabling a shallower subclone hierarchy in many tumor tissues, as opposed to a broad one usually observed in corresponding normal tissues [7].

Bulk cancer cells may detach from a primary tumor, enter the blood circulation, and eventually be transported to distant sites, but only CSCs have the tumorigenic properties needed to create a new tumor. Therefore, CSCs can colonize other organs and give origin to secondary tumors or metastases, reproducing all the subclone variability observed in primary tumors [6]. This is particularly relevant, as metastasis is accountable for 90% of cancer-related deaths [9]. Additionally, CSCs show resistance to traditional cancer treatments that target fast-diving cells, such as chemotherapy and radiotherapy, due to the potential to maintain a slow diving state in most cancers, to increase expression of drug efflux pumps and efficient DNA repair mechanisms [6]. That trait, together with the abovementioned CSC phenotype-plasticity, favors tumor reoccurrence, driving therapy failure.

CSCs secrete factors like Hypoxia-Inducible Factor 1 (HIF1) which in turn induces the release of pro-angiogenic factors that expand the blood vessel network within and around the tumor, allowing the increase in nutrient supply necessary for the fast growth of solid tumors. Simultaneously, CSCs can survive deep within tumors where hypoxia, which would be toxic to normal cells, prevails. That inward positioning protects CSCs from the effect of chemotherapeutic drugs, and at the same time, provides hypoxic conditions that shift CSCs gene expression towards CSC survival, self-renewal, and invasiveness [6, 10].

Immune evasion is another key property of CSCs. CSCs may escape the immune system surveillance by expressing factors that avoid their eradication by immune cells and therefore foster their survival outside primary tumors, in circulation, or at secondary tumor sites. For example, reports have shown that high levels of anti-apoptotic proteins found in CSCs protect them from NK and cytotoxic T cells [11].

Epithelial-to-Mesenchymal Transition (EMT), defined as a change from ep-

ithelial to mesenchymal phenotype and identified by the change of cell surface markers (like decrease in E-cadherin and increase of N-cadherin and vimentin) has also been associated with CSC phenotype [10]. By acquisition of mesenchymal properties, CSCs alter their cytoskeleton and lose cell-to-cell adhesion, gaining the capacity to migrate, invade surrounding tissues and metastasize [10, 11]. The link between CSCs and EMT has also been established with immunomodulatory effects, and cell plasticity [10]. However, there has been discussion on whether EMT is fundamental for CSC identity, as a return to an epithelial phenotype was described to be essential for metastatic growth. Nevertheless, some studies suggest that those contradictory observations result from environmental signals that make cells transit between a CSC and a cancer non-stem cell state [7].

Different methods are usually conjugated to identify CSCs, such as Fluoresce Activated Cell Sorting (FACS) based on stem and cancer cell markers, tumorsphere formation, Aldehyde Dehydrogenase (ALDH) enzyme activity [6], detection of overexpression of pathways associated with CSCs, and side population sorting, i.e. isolation of cells with higher ability to efflux dyes and drugs [6]. Nevertheless, the main criterion to identify CSCs is still nowadays the assessment of tumorigenicity, attained with the detection of new tumors after inoculation of a small number of cells in immune-deficient mice over successive passages, a process known as secondary transplantation [6, 7].

Overall, CSCs can be considered as highly plastic tumorigenic quiescent cells, prone to survive in adverse environments and with high resistance to traditional cancer treatments. Characteristics that turn them into the main entities promoting cancer aggressiveness, drug resistance, metastases, and cancer reoccurrence.

1.1.3 Metabolic reprogramming of cancer cells

One of the hallmarks of cancer is the ability to reprogram metabolism to fulfill the energetic and biosynthetic needs of highly proliferating cancer cells [12]. In 1930, Otto Warburg discovered that cancer cells rely more on glycolysis than on Oxidative Phosphorylation (OXPHOS) even when oxygen is abundantly available. This got known as aerobic glycolysis or Warburg effect [13–15]. Since then, breakthroughs in biochemistry and cellular biology fields fomented progress in research for the causes and consequences of metabolic reprogramming in cancer cells [16]. The current scientific knowledge on cancer cell metabolism can be organized into four tiers: nutrient-uptake modifications,

alteration of preferential intracellular metabolic pathways, effects of metabolism on tumor niche, and effects of metabolism on cancer epigenetics [16].

1.1.3.1 Nutrient-uptake modifications in cancer cells

In order to fulfill their goal of survival and growth, cancer cells need to have appropriate inputs of biosynthetic elements, like carbon and nitrogen. Since glucose and glutamine are the main sources of carbon and nitrogen for cancer cells, these cells must acquire cellular mechanisms to uphold a sufficient uptake of those nutrients from the extracellular environment [17].

In some cancers, mutations in Phospho-Inositol 3 Kinase (PI3K), its inhibitors (PTEN and INPP4B), or upstream Receptor Tyrosine Kinases (RTKs) leads to an increase of expression of the Glucose Transporter 1 (GLUT1) gene and movement of GLUT1 protein within the cell to the cytoplasmatic membrane, allowing the import of glucose [16] (Figure 1.1). Similarly, oncogenes like c-Myc, Kras, and Yap/Taz also overexpress GLUT1 in cancer cells, whereas Yap/Taz and loss-off-function mutations of TP53 tumor suppressor gene enhance GLUT3 expression (another glucose transporter) [17]. Another gene which transcription is increased upon aberrant activation of the PI3K/Akt pathway is the gene coding for Hexokinase 2 (HK2) [16]. Upon activation, HK2 binds to mitochondria and uses mitochondrial Adenosine Triphosphate (ATP) to catalyze the first step of glycolysis, the conversion of glucose to Glucose-6-Phosphate (G6P). In this way, glucose is quickly phosphorylated and remains trapped within the cell, without being able to move to the extracellular environment [13, 17] (Figure 1.1).

In order to improve the import of glutamine in cancer cells, the oncogene c-Myc and Yap/Taz signaling [17] promote the expression of glutamine transporter Alanine, Serine, Cysteine Transporter 2 (ASCT2), and of enzymes that facilitate glutamine uptake by converting it to glutamate (Glutaminase – GLS, Ribose-phosphate pyrophosphokinase 2 – PRPS2, and Carbamoyl-phosphate synthetase 2, Aspartate transcarbamylase and Dihydroorotase – CAD enzymes), which remains trapped in the cell (Figure 1.1). Furthermore, the outflux of glutamine is coupled to an influx of essential amino acids, such as leucine, through the Lactate Transporter 1 (LAT1). Therefore, tumors harboring c-Myc mutations indirectly contribute to the import of essential amino acids [16]. The tumor microenvironment also plays a role in the influx of glutamine. Receptor binding of Interleukin 4 (IL-4) inflammatory cytokine and uptake of extracellular lactate respectively promotes the expression and stabilizes Myc, indirectly inducing

ASCT2 expression [17]. Besides, loss-of-function mutations in tumor suppressor gene Retinoblastoma (Rb) may as well upregulate the intake of glutamine, as it releases E2F3 activator transcription factor from inhibition by Rb, allowing ASCT2 gene expression [16] (Figure 1.1).

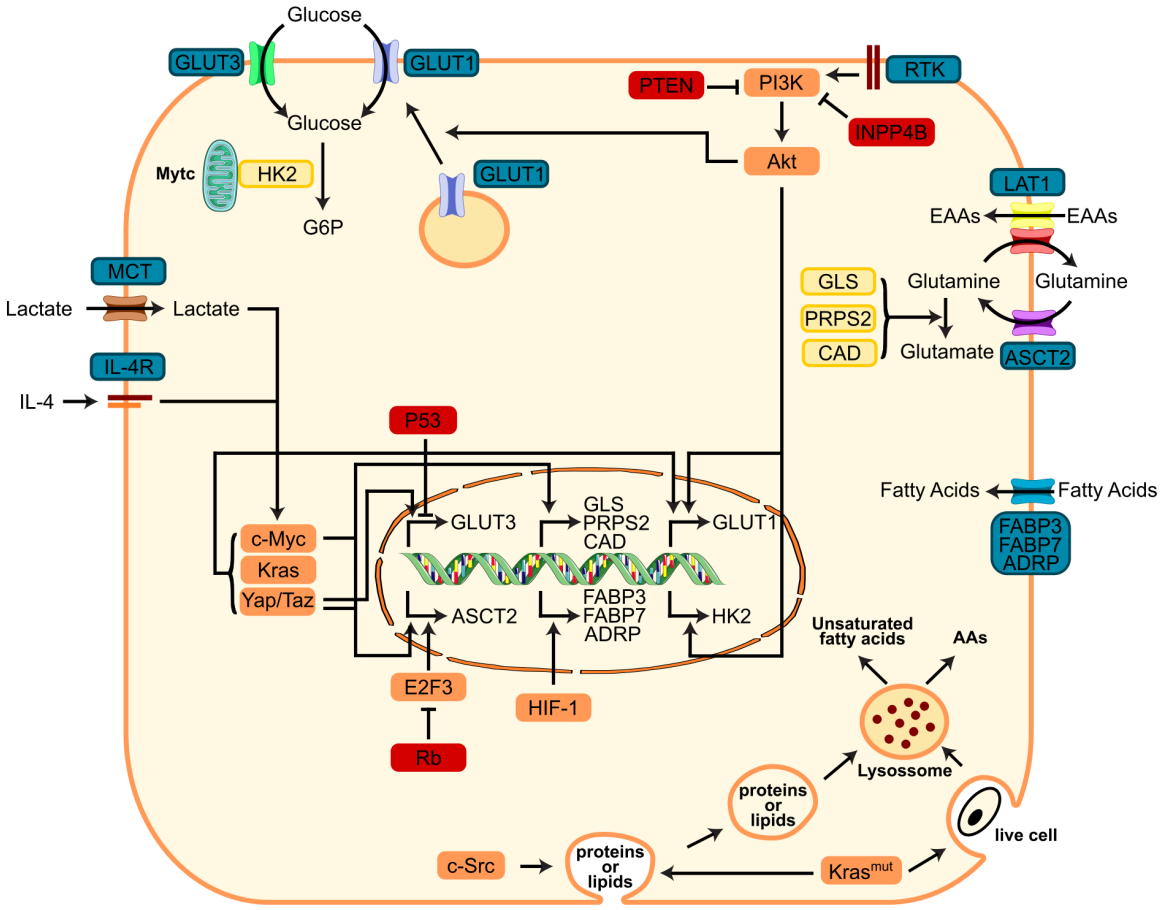

Figure 1.1: Nutrient-uptake strategies in cancer cells. Names of nutrient transporter proteins are in blue and metabolic enzymes in yellow. Proteins which genes suffer loss-of-function mutations with cancer are in red and remaining proteins in orange. AAs: Amino Acids; EAAs: Essential Amino Acids; IL-4R: IL-4 Receptor; MCT: Mono-Carboxylate Transporter; Mytc: mitochondria; P53: protein coded by TP53 gene. Redrawn figure based on Pavlova et al.(2016) [16], Park et al.(2020) [17], and `smart.servier.com`.

In addition, fatty acid uptake can as well be increased in cancer cells. For example, under hypoxia, cancer cells activate HIF1α which induces the expression of fatty acid receptors, like Fatty Acid Binding Protein 3 (FABP3), Fatty Acid Binding Protein 7 (FABP7), and Adipose Differentiation Related Protein (ADRP) [17] (Figure 1.1).

Although the increase in nutrients uptake is beneficial, it is not enough for cancer cells to meet their metabolic needs, specifically when the blood vessel formation does not keep up with the fast growth of tumor mass. In that case, cancer cells face nutrient shortage and hypoxia and are forced to use opportunistic strategies of nutrient acquisition. In some cancers, mutated Kras and c-Src genes promote an actin cytoskeleton remodeling that allows macropinocytosis, i.e. extracellular macromolecules (e.g proteins or lipids) are involved in vesicles and imported to the interior of cytoplasm where they are fused with lysosomes and eventually degraded. One such case is when cells are under hypoxic conditions and for this reason, cannot drive the oxygen-consuming process of unsaturated fatty acid synthesis. Instead, they import lipids by macropinocytosis. Furthermore, apoptotic bodies and even live cells can be exposed to phagocytosis or entosis by cancer cells, as opportunistic ways to obtain nutrients [16] (Figure 1.1).

1.1.3.2 Preferential intracellular metabolic pathways

Cancer cells not only modify their nutrient-uptake mechanisms but also change the metabolic pathways through which those nutrients are processed.

i) Glycolysis and intersecting pathways

In aerobiosis, normal cells use mostly aerobic rather than anaerobic respiration, since OXPHOS produces energy (ATP) from glucose more efficiently than glycolysis. Strikingly, one of the most described metabolic traits of cancer cells is the ability to undergo aerobic glycolysis (also known as the Warburg effect), i.e. to rely on glycolysis even upon high oxygen concentrations. Back in 1930, when this effect was uncovered, Otto Warburg suggested cancer cells were forced to rely on glycolysis due to mutations causing mitochondrial damage or malfunction. However, this theory was refuted later on by scientific evidence demonstrating that cancer mitochondria can carry out normal OXPHOS. Then, an alternative hypothesis emerged, implying that glycolysis allowed fast energy production, which was fundamental for fast-dividing cancer cells [13]. However, evidence now suggests glycolysis acts as a means to produce reducing equivalents and biosynthetic precursors of macromolecules for cancer cell proliferation more than a source for fast ATP production [13, 16].

As mentioned above, deregulation of signaling pathways in cancer cells leads to an increase of glucose transporters at the cell surface and HK2 activation, which foments aerobic glycolysis. Nonetheless, cancer cell signaling interferes

with other glycolytic enzymes. In glycolysis, Phosphofructokinase-1 (PFK) promotes the formation of Fructose-1,6-bisphosphate (F1,6BP) from Fructose-6-phosphate (F6P), whereas Phosphofructokinase-2 (PFK2) catalyzes the conversion of F6P to Fructose-2,6-biphosphate (F2,6BP), which functions as an activator of PFK [18] (Figure 1.2). In cancer, activation of Yap/Taz upregulates an isoform of PFK2 (the PFKFB3) and the increase in PI3K/Akt signaling activates PFK2 enzyme, while Myc and TP53 mutations enhance PFK activity. Furthermore, TP53 loss-of-function mutations reduce the expression of TP53-Inducible Glycolysis and Apoptosis Regulator (TIGAR). As TIGAR promotes F2,6BP degradation, TP53 mutations release glycolysis from the indirect inhibitory effect of TIGAR [17] (Figure 1.2).

The intense use of glycolysis in cancer cells fosters the metabolic flux through intersecting pathways, such as the Pentose-Phosphate Pathway (PPP) and the Hexosamine Biosynthetic Pathway (HBP).

Three glycolytic intermediaries can engage in the PPP: G6P, F6P, and Glyceraldehyde-3-Phosphate (G3P). In cancer, Myc mutations upregulate Glucose-6-Phosphate Dehydrogenase (G6PD) and Transketolase (TKT) enzymes, which respectively catalyze the conversion of G6P to 6-phosphogluconolactone, and of G3P to Xylulose-5-Phosphate (Xu5P) or of F6P to Xu5P [17, 19]. An intense PI3K/Akt signaling can as well increase the expression of G6PD and directly phosphorylate TKT, activating it. Furthermore, loss-of-function of TP53 prevents its binding to and inhibition of G6PD, consequently enhancing PPP activity [17] (Figure 1.2).

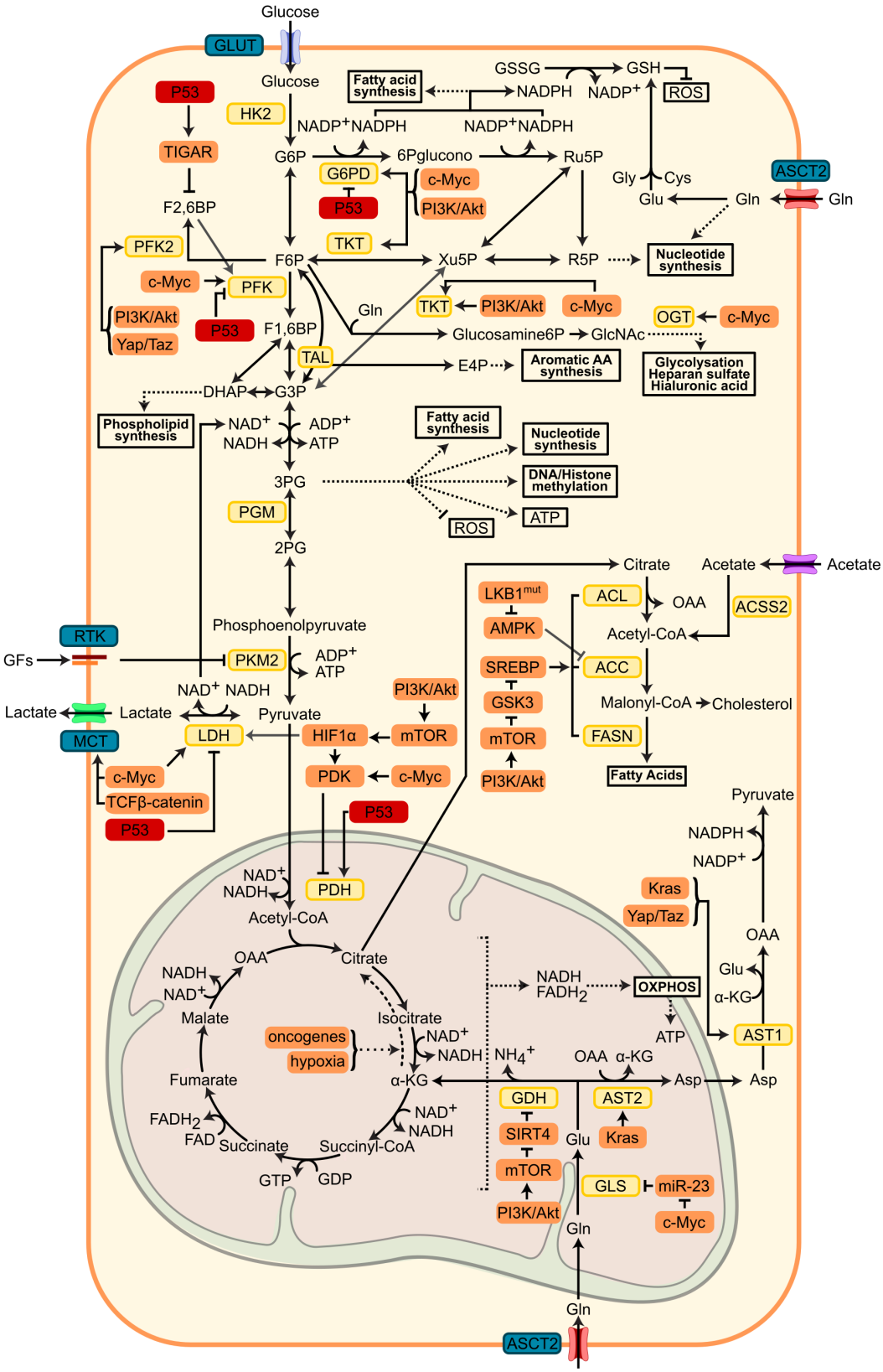

Figure 1.2: Preferential intracellular metabolic pathways in cancer cells. Names of nutrient transporter proteins are in blue and metabolic enzymes in yellow. Proteins which genes suffer loss-of-function mutations with cancer are in red and remaining proteins in orange.

2PG: 2-Phospho-Glycerate; 6PGlucono: 6-Phospho-Gluconolactone; ADP^+: Adenosine Di-Phosphate; Asp: Aspartate; Cys: Cysteine; FAD: Flavin Adenine Dinucleotide; GDP: Guanosine Di-Phosphate; Gln: Glutamine; Glu: Glutamate; Glucosamine-6P: Glucosamine-6-phosphate; Gly: Glycine; GTP: Guanosine Tri-Phosphate; TAL: transaldoase; P53: protein coded by TP53 gene. Redrawn figure based on Pavlova et al. (2016) [16] , Park et al. (2020) [17], Kim et al. (2012) [19] and `smart.servier.com`.

However, the reactions catalyzed by G6PD and TKT are reversible. Therefore, it is mainly the accumulation of glycolytic intermediates resulting from tight regulation of glycolytic enzymes (like HK2, PFK, Phosphoglycerate mutase - PGM, and Pyruvate Kinase M2 - PKM2) by cancer cells that determines carbon flux diversion into the PPP [13]. Ribulose-5-Phosphate (Ru5P), a precursor for Ribose-5-Phosphate (R5P) which in turn is used for synthesis of nucleotides, can be produced in the oxidative and non-oxidative phases of PPP, although only the oxidative phase generates reduced Nicotinamide Adenine Dinucleotide Phosphate (NADPH) [13]. The diversion of carbon flux from glycolysis to the PPP pathway is therefore not only essential for nucleotide formation, but also for fatty acid synthesis and production of Glutathione (GSH) from Glutathione disulfide (GSSG), as NADPH is consumed in the two last-mentioned processes. On the other hand, GSH acts as an antioxidant by reducing Reactive Oxygen Species (ROS). Therefore, an increase in NADPH production through PPP protects cancer cells from oxidative damage, preventing their apoptosis [17]. Furthermore, Erythrose-4-Phosphate (E4P), a precursor for the synthesis of aromatic amino acids, can also be produced in the PPP non-oxidative phase [20]. So, an increase of flux in PPP may as well favor amino acid synthesis in cancer cells (Figure 1.2).

The HBP pathway is initiated with the conversion of the glycolysis intermediary F6P and glutamine to glucosamine-6-phosphate, which is then transformed into N-Acetylglucosamine (GlcNAc) [16], the building block for glycosyl side chains of proteins and lipids [12]. Hence, the increase in flux through glycolysis and consequently through HBP in cancer contributes to protein-function deregulation by aberrant glycosylation, besides increasing the synthesis of heparan sulfate and hyaluronic acid. The first is a component essential for formation of new cancer-cell membranes while the last is an extracellular component that favors the tumor microenvironment [21]. Also, cancer cells harboring Myc mutations promote glycosylation by raising the activity of GlcNAc transferase (OGT), the enzyme catalyzing the transfer of GlcNAc to proteins, through chaperon mediated stabilization [16] (Figure 1.2).

Additionally, the boost in glycolysis indirectly enhances the formation of glycerol-3-phosphate through the accumulation of the glycolytic intermediate Dihydroxyacetone phosphate (DHAP). glycerol-3-phosphate is used in the biosynthesis of phospholipids, which in turn are fundamental to the establishment of new cell membranes [16] (Figure 1.2).

Another pathway that diverges from glycolysis is the Serine Synthesis Pathway (SSP). The first reaction of this pathway involves the transition of the glycolytic intermediate 3-Phosphoglycerate (3PG) to 3-phospho-oxypyruvate with concomitant production of reduced Nicotinamide Adenine Dinucleotide (NADH), and it is catalyzed by the enzyme Phosphoglycerate dehydrogenase (PHGDH) (Figure 1.3). Then, Phosphoserine aminotransferase (PSAT1) fosters the transfer of an amino group from glutamate to 3-phospho-oxypyruvate, producing 3-phosphoserine and alpha-Ketoglutarate (α-KG), which in turn is a Tricarboxylic acid (TCA) cycle intermediate for energy production and anabolic reactions [17]. At the end of the pathway, 3-phosphoserine loses a phosphate group and it is converted to serine by the activity of Phosphoserine phosphatase (PSPH) [22]. Serine can then give origin to another amino acid, glycine, while providing a carbon atom to the one-carbon cycle, through the activity of the enzyme Serine hydroxymethyltransferase (SHMT) (Figure 1.3).

The one-carbon cycle is composed of the folate and methionine (Met) cycles. Besides contributing to S-Adenosyl-Methionine (SAM) production and subsequent epigenetic regulation (histone and DNA methylation) by providing one carbon to the Met cycle, the folate cycle has other functions. In the folate cycle, different one-carbon-carrier-tetrahydrofolate species, like Tetrahydrofolate (THF), methyl-THF (mTHF), 5,10-methylene-THF (meTHF), and 10-formylTHF interconvert through a series of redox reactions, some of which lead to NADPH/NADH production. meTHF can be used by the enzyme Thymidylate synthase (TYMS) to convert deoxyuridine monophosphate (dUMP) into the nucleotide deoxythymidine monophosphate (dTMP). 10-formylTHF may be introduced into the structure of purines during nucleotide *de novo* synthesis, hydrolyzed to formate in a reaction that produces ATP, or completely oxidized to carbon dioxide in a reaction where NADPH is produced [23]. On the other hand, the Homocysteine (hCYS) from the Met cycle may be converted through the transsulfuration pathway to cysteine. Cysteine, along with glutamate and glycine allows the production of GSH [22] (Figure 1.3).

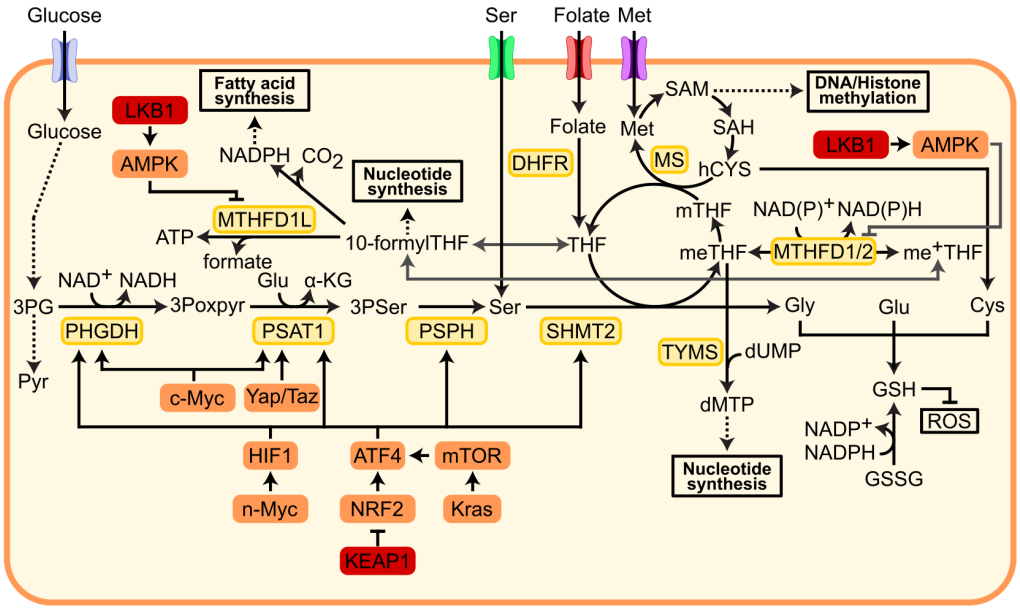

Figure 1.3: Serine Synthesis Pathway (SSP) and one-carbon metabolism in cancer cells. In the one-carbon cycle, folic acid (vitamin B9) is first reduced to Tetrahydrofolate (THF) by the activity of Dihydrofolate reductase (DHFR). The folate cycle then begins with the transfer of one-carbon unit from serine to THF, giving origin to 5,10-methylene-THF (meTHF) and glycine. Although to a smaller extent, glycine can as well serve as a one-carbon giver, through a reaction catalyzed by the glycine decarboxylase (not shown) [22, 23]. Thereafter, meTHF can be reduced to methyl-THF (mTHF), which in turn provides one carbon atom in the form of a methyl group to Homocysteine (hCYS) from the methionine (Met) cycle and regenerates THF, closing the folate cycle. In the Met cycle, when hCYS receives a methyl group from the folate cycle, generates Met, in a reaction catalyzed by Methionine Synthetase (MS). The ensuing adenylation of Met drives the production of S-Adenosyl-Methionine (SAM), which provides methyl groups for methylation of enzymes, histones, and DNA. Conversely, SAM is demethylated into S-Adenosyl-Homocysteine (SAH), which is then transformed back to hCYS, completing the Met cycle [22–24]. Names of nutrient transporter proteins are in blue and metabolic enzymes in yellow. Proteins which genes suffer loss-of-function mutations with cancer are in red and remaining proteins in orange. 3Pser: 3-phospho-serine 3Poxpyr: 3-phospho-oxypyruvate; Cys: Cysteine; Glu: Glutamate; Gly: Glycine; P53: protein coded by TP53 gene; Pyr: Pyruvate; Ser: Serine. Redrawn figure based on Ducker et al. (2017) [23], Rosenzweig et al. [22], Ryall et al. [24] and `smart.servier.com`.

Therefore, in cancer, the increase in glycolysis leads to the accumulation of 3PG, which upregulates the SSP and consequently enhances the one-carbon cycle. This in turn promotes nucleotide, ATP, and NADPH synthesis through the folate cycle; induces SAM production through the Met cycle; and fosters GSH

synthesis both through the Met cycle and through serine to glycine conversion. Consequently, these events allow the fulfillment of the cancer need for nucleic acid synthesis, energy, control of cellular redox status, lipid synthesis, signaling, and epigenetic remodeling [22] (Figures 1.2 and 1.3).

Besides the increase in glucose-derived carbon through the SSP, cancer cells also directly upregulate SSP and folate cycle enzymes. For example, mutant c-Myc upregulates the expression of PHGDH and PSAT1 genes in cancer cells, while intense Yap/Taz signaling upregulates PSAT1 [16, 17] (Figure 1.3). In normal cells, HIF1 serves as a redox regulator in hypoxia, since it promotes the expression of SSP and SHMT genes, leading to an increase in reducing equivalents that prevent ROS accumulation resulting from hypoxia. In cancer, mutant n-Myc induces a similar effect by further activating HIF1 [22].

Another redox regulator in normal cells that is upregulated in cancer is the Nuclear factor erythroid-derived 2 (NRF2). In normal cells, oxidative stress induces NRF2. In cancer, loss-of-function mutations in tumor suppressor KEAP1 release NRF2 from KEAP1 inhibition. Then, NRF2-induced upregulation of Activating Transcription Factor 4 (ATF4) unleashes transcription of SHMT2 and SSP genes [22]. In normal cells, mTOR works as a nutrient-sensor by also activating ATF4 upon low levels of serine, leading to an increase in serine pools, while in cancer, Kras oncogene stimulates the effects of ATF4 [22].

Another important regulator of the one-carbon cycle is AMP-activated Protein Kinase (AMPK). Upon low levels of ATP (and correspondingly high levels of AMP), AMPK is activated and downregulates the expression of Methylene-THF Dehydrogenase, which is an enzyme with multifunctional isoforms (MTHFD1/2/1L) involved in the interconversion of one-carbon-carrier-THF species in the folate cycle. By down-regulating the folate cycle when ATP levels are low, AMPK allows normal cells to stop ATP-consuming anabolic processes to help restore ATP levels. In cancer, the loss of function in tumor suppressor Liver Kinase B1 (LKB1), which is an AMPK activator has the opposite effect of inducing folate metabolism [22] (Figure 1.3).

One of the most relevant enzymes in promoting aerobic glycolysis is the Pyruvate Kinase (PK). In glycolysis, PK catalyzes the transformation of phosphoenolpyruvate into pyruvate [18]. Many cancers use the less active form of PK, the PKM2, besides inhibiting its activity through growth-factor induced signaling, leading to an accumulation of glycolytic intermediaries upstream of phosphoenolpyruvate (Figure 1.2). This contributes to the diversion of carbon flux to the anabolic pathways mentioned above [13, 16].

The end-product of glycolysis, pyruvate, is converted to lactate through the activity of Lactate dehydrogenase (LDH). Then, the lactate is secreted to the exterior of the cell through Monocarboxylic Acid Transporters (MCTs). Therefore, MCTs avoid the accumulation of lactate inside the cell, preventing the reverse lactate to pyruvate conversion and a destructive acidic environment inside the cell [18] (Figure 1.2).

Pyruvate decarboxylation is the irreversible conversion of pyruvate to acetyl-Coenzyme A (acetyl-CoA) catalyzed by the enzyme Pyruvate Dehydrogenase (PDH) in the mitochondrial matrix, and it is determinant for entry of carbon into the TCA cycle and subsequent OXPHOS. Pyruvate Dehydrogenase Kinase enzymes (PDKs) phosphorylate PDH, inhibiting its activity [18].

In cancer, Myc mutations overexpress both LDH, MCT1 [17] and PDK1 [16], TCFβ-catenin transcription induces MCT1 upregulation, and loss-of-function mutations in TP53 release LDH from inhibition and avoid activation of PDH [17] (Figure 1.2). Furthermore, under hypoxia, the stabilization of HIF1α by PI3K/Akt/mTOR signaling induces PDK1 [13] and LDH activity in cancer [17] (Figure 1.2). These events divert pyruvate from conversion to acetyl-CoA and subsequently avoid entry into TCA cycle [13, 17]. Furthermore, the pyruvate-to-lactate reaction produces oxidized Nicotinamide Adenine Dinucleotide (NAD^+), which can be used as a substrate for glycolytic reactions where NADH is produced, indirectly sustaining glycolysis [16].

The aerobic glycolysis where pyruvate is diverted to lactate production avoids an excess of carbon flux through the TCA cycle and associated OXPHOS, while at the same time allows NADPH production through the PPP. Therefore, aerobic glycolysis can be thought of as a mechanism implemented by cancer cells to avoid the production of ROS, as a disproportionate amount of ROS can lead to apoptosis. Nevertheless, cancer cells can still produce ROS due to tumor suppressor and oncogenic pathways that affect cellular enzymes, the tumor microenvironments (hypoxia or inflammation), and in some cases even mitochondrial dysfunction. In fact, moderate levels of ROS observed in cancer cells can be beneficial. ROS may contribute to tumorigenesis, cancer progression, and chemoresistance [25].

ii) Glutamine metabolism

As mentioned above, cancer cells increase uptake of glutamine because it is a source of carbon and nitrogen. In the cytoplasm, glutamine can be combined with glycolytic intermediate F6P and feed the HBP pathway, contributing to

glycosylation. Besides, glutamine provides nitrogen for pyrimidine and purine nucleotide synthesis and increases antioxidant potential by GSH generation after being converted to glutamate (Figure 1.2).

In mitochondria, the enzyme GLS catalyzes the conversion of glutamine to glutamate. Glutamate can then either be metabolized to α-KG and Ammonia (NH_4^+) through the activity of Glutamate dehydrogenase (GDH/GLUD1), or to non-essential amino acids and an α-ketoacid through the activity of transaminases/aminotransferases, such as the Aspartate aminotransferase (AST/GOT) or the Alanine aminotransferase (ALT) [16, 17].

In cancer cells, aberrant c-Myc activation restraints miR-23, releasing GLS translation from miR-23 inhibition and triggering its activity, while PI3K-Akt-mTOR enhanced signaling inhibits SIRT4 deacetylation which releases GDH from SIRT4 inhibition (Figure 1.2). High levels of Kras also induce the activity of both the mitochondrial and cytosolic ASTs (AST2/GOT2 and ASTT1/GOT1 respectively) in cancer. In mitochondria, AST2 catalyzes a reversible reaction in the direction where an amino group from glutamate is transferred to an α-ketoacid (which can be Oxaloacetate - OAA) to give origin to aspartate and a new α-ketoacid (which can be α-KG) [17, 26]. In turn, aspartate can be transported to the cytoplasm and be converted to OAA by the reverse of the abovementioned reaction, this time catalyzed by AST1. OAA can then be subsequently converted to pyruvate, in a reaction where NADPH is produced [17] (Figure 1.2). So, the abovementioned events in cancer cells, together with the increase in glutamine uptake, promote amino acid synthesis and generation of antioxidative potential.

Most importantly, an increase in glutamine metabolism allows anaplerosis, i.e. replenish of TCA cycle intermediates, through the production of α-KG. Usually, mitochondrial acetyl-CoA is produced from pyruvate decarboxylation or obtained from Fatty Acid Oxidation (FAO) [27]. In catabolic conditions, mitochondrial acetyl-CoA is routed to TCA cycle and OXPHOS, while in an anabolic state it can be transported to cytoplasm either directly through the carnitine shuttle or through the citrate-malate-pyruvate shuttle upon previous conversion to citrate [27]. In the cytoplasm, ATP Citrate Lyase (ACL) produces acetyl-CoA from the mitochondrial-imported citrate, which in turn may be used in protein acetylation or as a substrate for fatty acid synthesis [12].

However, due to aerobic glycolysis, cancer cells cannot obtain much acetyl-CoA through pyruvate decarboxylation. In cancer, oncogenes and hypoxia promote a context in which excess of α-KG is converted to mitochondrial citrate,

in a reaction named reductive carboxylation [16] (Figure 1.2). This reaction, which is the reverse of a TCA cycle reaction, increases the pool of citrate, allowing cancer to synthesize more fatty acids and consequently meet its need for high phospholipid membrane synthesis. Furthermore, this event enables cancer to feed the TCA cycle with intermediates that are used for NADH and reduced Flavin Adenine Dinucleotide ($FADH_2$) production, contributing to some energy production through OXPHOS [16].

iii) Fatty acid metabolism

The fatty acid synthesis starts with the conversion of cytosolic citrate to acetyl-CoA and OAA, catalyzed by ACL. Then, Acetyl-CoA Carboxylase (ACC) converts acetyl-CoA to malonyl-CoA, which is in turn set up into long fatty acid chains by Fatty Acid Synthetase (FASN) (Figure 1.2). Furthermore, malonyl-CoA can also be used for cholesterol synthesis [16]. In normal conditions, low lipid levels activate the SREBPs transcription factors by proteolysis and the protein product enters the nucleus, inducing the expression of genes coding for the abovementioned fatty acid enzymes [17]. In cancer, high PI3K/Akt and mTOR signaling restrain GSK3β, releasing nuclear SREBP from GSK3β inhibition, consequently enhancing fatty acid synthesis [13, 17]. In some other cancers, mutant LKB1 may inhibit LKB1-AMPK signaling and subsequently free ACC from inactivation by AMPK [17]. Some tumors have also been shown to overexpress Acetyl-CoA Synthetase 2 (ACSS2), an enzyme that catalyzes the transformation of imported extracellular acetate to acetyl-CoA, which contributes to fatty acid synthesis [16].

These events together with an increase of fatty acid uptake allow cancer cells to increase the pool of lipids to build new cellular membranes, alter membrane composition to include more saturated fatty acids, which are more resistant to oxidative damage [16], and use them as secondary messengers in cancer cell signaling [17].

1.1.3.3 Effects of metabolism on tumor niche

In the same way cell metabolism influences cancer cells, it also affects the cancer healthy-neighbor cells. The excess of lactic acid secreted by cancer cells hinders T-cell and dendritic cell activation, and migration of monocytes, while simultaneously attracting and promoting an immunosuppressive phenotype in macrophages [13, 16]. On the other hand, macrophages release cytokines and

growth factors that promote cancer growth, invasion, and metastasis [13].

Besides its role in suppressing antitumor immune-response, lactic acid may as well foster angiogenesis, which renders nutrients and oxygen for cancer growth. Specifically, it activates HIF1α, NF-kB, and PI3K in endothelial cells, and stimulates VEGF release in fibroblasts [16].

Diffusion of carbon dioxide and export of H^+ coupled to efflux of lactate acidifies the cancer extracellular environment, reducing pH and promoting the proteolytic activity of cathepsins and Matrix metalloproteinases (MMPs), which in turn degrade the extracellular matrix and boost tumor invasion. Additionally, stimulation of hyaluronic acid production in fibroblasts by lactate plays a role as well in cancer invasiveness [16].

Some tumors overexpress the enzymes Indoleamine 2,3-dioxygenase 1 (IDO1) and Tryptophan 2,3-dioxygenase (TDO2) that assist the conversion of tryptophan into kynurenine, leading to an extracellular shortage of tryptophan. As tryptophan is an essential amino acid, its depletion induces T-cell apoptosis. Furthermore, kynurenine enhances Treg cells, which suppress effector T-cells, and has an autoregulatory effect by inducing extracellular matrix degradation through interaction with Aryl hydrocarbon Receptor (AhR) in the cancer cells themselves [16].

1.1.3.4 Effects of metabolism on cancer epigenetics

Unlike genetic mutations, epigenetic alterations do not change the DNA sequence, but rather refer to the covalent modifications of DNA bases, histones, and protein complexes controlling nucleosome positioning, that modify DNA packing and chromatin accessibility to transcriptional machinery, consequently affecting gene expression [28].

Mutations on genes that code for DNA and histone regulatory enzymes, histones themselves, or nucleosome positioning complexes contribute to a gene expression shift in cancer [28]. Furthermore, since different metabolites also work as substrates or cofactors of epigenetic regulators, mutations on metabolic enzymes or proteins regulating them can as well change gene expression in cancer through epigenetic modification [24, 28].

The most described epigenetic modifiers are DNA and histone methyltransferases/demethylases and histone acetyltransferases/deacetylases.

DNA methyltransferases (DNMTs) are enzymes that catalyze the transfer of a methyl group from SAM to the 5th position of a cytosine residue in DNA, producing a 5-methyl cytosine (5mC) and the by-product S-Adenosyl-

Homocysteine (SAH) (Figure 1.4). Methylated cytosines are mostly found in regions of a high density of Guanine-cytosine (GC) located in the promoters of genes and called CpG islands (CGI). DNA methylation in those areas usually promotes chromatin condensation, mostly resulting in repression of gene expression. Additionally, methylation may attract proteins that bind histone-modifying enzymes, indirectly inhibiting transcription [28].

On the other hand, DNA demethylation (reverse reaction of DNA methylation) starts with hydroxylation of 5mC into 5-hydroxymethyl cytosine (5hmC), a step that is catalyzed by the Ten-eleven translocation (TET) family of dioxygenases [28, 29] (Figure 1.4). TET demethylases require α-KG and iron for their activity, overall promote chromatin relaxation and subsequent induce gene expression [24, 28].

Like DNMTs, Histone methyltransferases (HMTs) consume SAM and generate SAH, but instead of methylating DNA, they transfer the methyl group to a lysine or arginine amino acid in the N-tail of a histone protein [24, 28]. Histone methylation may result in either gene expression induction or repression depending on which amino acid residue is methylated, the location of the methyl group within the residue and the number of methyl groups transferred. It can also affect either gene promoters or enhancers (Figure 1.4). For example, deposition of three methyl groups in lysine 4 of histone H3 (H3K4me3) lead to promoter activation, and mono-methylation at same lysine residue (H3K4me1) marks active enhancers, while trimethylated H3 at lysine 27 (H3K27me3) or trimethylated H3 at lysine 9 (H3K9me3) affect enhancers and indirectly repress gene expression [28].

The removal of methylation marks from histones has the opposite effect on gene expression than of the methylation and it is promoted by Histone demethylases (HDMs). There are two families of HDMs. One is the Lysine-Specific Demethylase (LSD) family. Enzymes of this family act specifically on mono- and di-methylated H3K4 and H3K9 marks while using Flavin Adenine Dinucleotide (FAD) as a cofactor and reducing it to $FADH_2$. The other is the family of Jumonji-domain-containing histone demethylases (JHDMs) that requires α-KG and converts it to succinate [24] (Figure 1.4).

Histone acetyltransferases (HATs) catalyze the transfer of an acetyl group from acetyl-CoA to a lysine residue in histone tails and release Co-enzyme A (CoA), while Histone deacetylases (HDACs) remove the acetyl group from histones. There are two types of HDACs, those that depend on zinc ion and convert the acetyl group to acetate (HDACs I, II, and IV), and those, known as sirtuins (HDACs III/SIRT 1-7), which convert NAD^+ and the acetyl group to

Nicotinamide (NAM) and 2'-O-Acetyl-ADP-Ribose [24]. Overall, histone acetylation activates gene transcription and may occur either in gene promoters or enhancers (Figure 1.4). For example, acetylated H3 at lysine 27 (H3K27ac) induces gene expression though enhancer activation [28].

As mentioned before, accumulation of the glycolytic intermediate 3PG and upregulation of one-carbon cycle enzymes in cancer leads to an increase in SAM production. Since DNMTs utilize SAM, the rise in SAM levels induces site-specific gene promoter hypermethylation and consequent transcriptional repression of tumor suppressor and cell-cycle checkpoint genes, which may trigger cancer formation and growth [22, 28]. On the other hand, deposition of methyl groups in histones by the SAM-dependent HMTs can as well silence tumor suppressor genes through for example the H3K9me repressive histone mark [28], or instead enhance oncogene expression through activating histone marks, such as H3K4me3 [30].

Both the activity of the TET DNA-demethylases and JHDM family of HDMs are dependent on the levels of the TCA cycle intermediate α-KG, which serves as a co-substrate of those enzymes (Figure 1.4). In some cancers, the decrease in α-KG levels promoted by a local reduction of glutamine availability leads to histone hypermethylation and subsequent block of differentiation genes, which facilitates the cancerous phenotype [28]. Furthermore, TETs and JHDMs are inhibited by their reaction product succinate and by the consecutive metabolite of succinate in the TCA cycle, the fumarate. In some cancers, there is a loss-of-function mutation in Succinate dehydrogenase (SDH), the enzyme that converts succinate to fumarate, or in Fumarate Hydratase (FH), the enzyme catalyzing the reaction of fumarate to malate, which leads to the respective accumulation of succinate or fumarate [16]. The accumulation of those metabolites impairs the activity of TET demethylases and JHDMs and induces global DNA hypermethylation (Figure 1.4). Among other effects, this event reduces for instance the transcription of miR-200, which consequently promotes an EMT phenotype characteristic of some cancers [30]. Additionally, some cancers carry a neomorphic mutation in the gene coding for TCA cycle enzyme Isocitrate dehydrogenase (IDH) so that instead of producing α-KG from isocitrate as it should in normal cells, the mutant enzyme converts α-KG to 2-Hydroxyglutarate (2-HG). As 2-HG is structurally similar to α-KG, it acts as a competitive inhibitor of TET demethylases and JHDMs, producing similar effects to SDH and FH loss-of-function mutations [16] (Figure 1.4).

Interestingly, even errors in the metabolite-transport system can affect α-KG-regulated epigenetic modifiers. For example, defects in the enzyme that

transports pyruvate from the cytoplasm to mitochondria, the Mitochondrial Pyruvate Carrier (MPC), besides promoting the glycolytic switch, may also cause tumorigenesis through decrease in α-KG production and consequent histone hypermethylation [28]. Also, defects on the oxoglutarate(KG)-glutamate carrier (OGC), which transports α-KG out of mitochondria when working as part of the malate-aspartate shuttle, have been associated with hypermethylation phenotype in some cancers [28] (Figure 1.4).

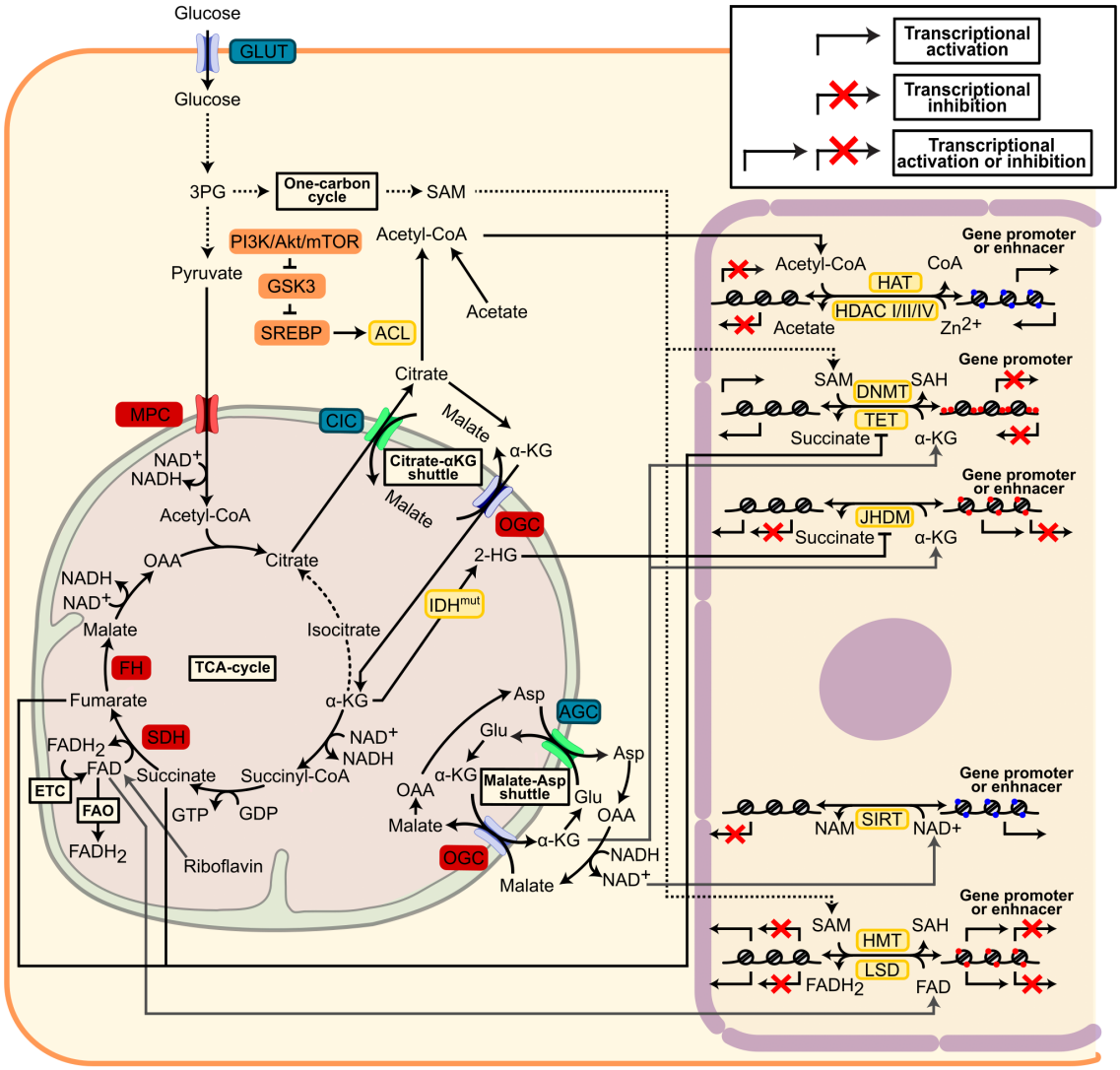

Figure 1.4: Influence of metabolism on epigenetic regulation of cancer cells. Names of proteins (enzymes or transporters) which suffer loss-of-function mutations with cancer are in red. Remaining transporter proteins and metabolic enzymes are in blue and in yellow, respectively. Other signaling proteins are in orange. Asp: Aspartate. Redrawn figure based on Saggese et al. (2020) [28] , Ryall et al. (2015) [24], Menon et al. (2020) [30], Pavlova et al. (2016) [16], and `smart.servier.com`.

The other family of HDMs named LSD requires FAD to work. FAD can be synthesized de novo from riboflavin (vitamin B2), reduced to $FADH_2$ upon acyl-CoA to 2-enoyl-CoA conversion during FAO or by SDH enzyme in TCA cycle, and obtained from $FADH_2$ oxidation in the Electron Transport Chain (ETC). Therefore, FAO, TCA cycle, ETC, and vitamin B2 availability can influence LSDs action (Figure 1.4). LSD1 activity has been associated with the promotion of differentiation through demethylation of activator histone mark H3K4me1 in enhancers of master regulator transcription factors (Oct4/Nanog/Sox2) in stem cells [24]. This suggests that reduction in LSD1 activity could play a role in cancer stem cell formation. However, reports show LSD1 can either have tumor suppressor or oncogenic function, depending on the context [31, 32].

Another important metabolite is acetyl-CoA, as histone acetylation relies on it. On account of the diversion of pyruvate to lactate production in cancer cells, the main source of acetyl-CoA is the mitochondrial citrate derived from glutamine. Nevertheless, acetyl-CoA can also be obtained from pyruvate decarboxylation, and even acetate [30] (Figure 1.4). As part of the citrate-α-KG shuttle, OGC's uptake of α-KG to mitochondria is associated with the export of isocitrate/citrate by the Citrate carrier (CIC). Isocitrate/citrate is then converted to acetyl-CoA in the cytoplasm, where is used by HATs. So sometimes, defects on OGC are responsible for cancerous phenotype through changes in acetyl-CoA levels [28] (Figure 1.4). Another mechanism through which HATs activity may be affected in cancer is the induction of expression of the ACL enzyme by the PI3K-Akt signaling, mentioned before.

On the other hand, sirtuins (HDACs) require NAD^+ to function. In the malate-aspartate shuttle, OGC-mediated import of malate is coupled to Aspartate-Glutamate Carrier (AGC) export of aspartate to cytoplasm. In turn, the cytoplasmatic aspartate is converted to OAA and subsequently to malate in a reaction where NAD^+ is produced (Figure 1.4). Therefore, OGC malfunctioning affects sirtuins activity, which hampers cell differentiation and induces cancer growth [28].

It is important to note that histones can suffer other less well-studied modifications besides methylation and acetylation, such as O-GlcNAcylation, formylation, propionylation, succinylation, that can affect chromatin remodeling [16, 30]. Furthermore, chromatin modifications induced by metabolic changes in cancer may as well affect the expression of metabolic enzymes, creating an interplay between epigenetics and metabolic regulation [16, 28].

1.1.4 Metabolic traits of cancer stem cells

Throughout different studies, CSCs have shown to hold a wide variety of metabolic traits and experimental limitations are one of the reasons for the observed phenotypic variability. There are few studies on CSC metabolism and the ones that exist apply different methods for CSCs isolation since even the definition of CSC is ambiguous [12]. Different labs isolate CSCs from fresh tumors based on distinct surface markers, some of which are not completely reliable or can be lost during sample preparation [33]. Also, many studies use CSCs retrieved from approved cell lines grown as 3D spheroids, and although survival and ability to grow in an anchorage-independent way are traits of stem cells, just 1% of spheroids have shown to be true *bonafide* CSCs. Furthermore, cell lines often do not capture the phenotypic variability of CSCs since they result from clonal amplification throughout many cell passages [33].

On the other hand, the metabolic phenotype of CSCs frequently changes with the tissue microenvironment. For example, glioma stem cells mainly rely on OXPHOS but use glycolysis when oxidative metabolism is inhibited. Metastatic breast CSCs with tropism to the liver and not to the lung or bone, undergo aerobic glycolysis due to the gluconeogenic nature of the liver tissue [7]. Leukemic CSCs adjacent to gonadal adipose tissue express high levels of fatty acid transporters and can oxidize fatty acids provided from lipolysis of the adipose tissue [34]. As most studies are *in vitro*, the abovementioned effects of the niche over CSCs are usually not taken into account in experimental settings [12]. Also, many studies use cells cultured in high levels of glucose and oxygen, which favors glycolysis as an energy production pathway and therefore may not recapitulate *in vivo* metabolic state of CSCs [7].

Although no common metabolic traits can be identified across CSCs of all cancer types and tumor micro-environments, CSCs always show distinct metabolic characteristics in comparison with the corresponding differentiated cancer cells. Those metabolic features are discussed next.

i) Glycolysis

Overall, most normal SCs, like for example the hematopoietic stem cells, remain in hypoxic niches where they use glycolysis instead of OXPHOS. Although OXPHOS is a more efficient mode of energy acquisition, it promotes ROS accumulation, which can be especially detrimental for cells like the SCs that must maintain an undamaged self-renewal capacity throughout the entire life of an organism [7, 12]. Furthermore, during cell reprogramming of differentiated cells, which rely mainly on OXPHOS, to induced pluripotent stem cells, there is an

increase in expression of glycolytic genes that precedes that of pluripotent markers. This event suggests the OXPHOS-to-glycolytic switch is the cause rather than the consequence of stemness acquisition [12].

Likewise, most CSCs depend more on glycolysis, produce more lactate, and show enhanced ATP levels and reduced mitochondrial respiration in comparison with differentiated cancer non-stem cells [12, 18]. Tumors with CSCs that were shown to have the glycolytic phenotype are osteosarcoma, glioblastoma, lung, breast, ovarian, hepatocellular, nasopharyngeal, and colon cancers [18, 33]. The increase in glycolysis in those CSCs is explained by a rise in gene expression of glucose transporters, the glycolytic enzymes like HK2, PFK2, and PKM2, and also LDH [12]. As mentioned before when contrasting cancer cells with normal cells, the enhanced glycolytic activity observed in CSCs in comparison with differentiated cancer cells is advantageous for fast ATP, NADPH, amino acid, and nucleotide production. Indeed, another enzyme usually overexpressed in CSCs is the G6PD, which further contributes to nucleotide synthesis [12]. Furthermore, an increase in glycolytic PFK enzyme allows activation of Yap-TED and subsequent transcriptional activation of genes involved in the CSC features of cell migration and EMT [17]. Also, as mentioned before the efflux of lactate acidifies the cancer micro-environment, activating MMPs transcription through NF-kB signaling. In turn, MMPs degrade the extracellular matrix, promoting tissue invasion, which is fundamental for metastasis formation [17].

It has been proven that inhibition of glycolysis or glucose uptake can induce CSC death [18]. However, some studies emphasize glycolysis is not *per se* responsible for stemness, but instead it is an alternative source of energy to which CSCs turn to when trying to reduce cytotoxic ROS levels, through reduction of OXPHOS [33]. In glioblastoma xenografts, for example, the underexpression of SDH subunit B (SDHB) causes mitochondrial malfunction and consequent ROS production. ROS activates HIF, promoting the expression of glycolytic genes, which in turn enhances glycolysis and reduces the synthesis of OXPHOS-induced ROS [12].

ii) Oxidative phosphorylation

While most normal and cancer stem cells rely on glycolysis for energy production, sometimes they may depend on other metabolic pathways, such as OXPHOS. A classic example is the normal muscle stem cells/satellite cells. These cells present high OXPHOS activity because they lay in aerobic niches close to blood vessels. In this case, the electron carriers needed for OXPHOS

are provided by FAO. Furthermore, when those cells differentiate to more committed states they undergo a metabolic shift towards glycolysis concomitant with an epigenetic reprogramming event [7, 12]. Another example of stem cell dependence on OXPHOS is that of normal $Lrg5^+$ stem cells in the intestinal crypt. These highly proliferative stem cells use lactate-derived pyruvate to feed the TCA cycle and OXPHOS, while lactate is supplied by the surrounding glycolytic epithelial cells named *Paneth* cells. Interestingly, the high ROS levels stemming from OXPHOS activity in intestinal stem cells are not harmful but instead induce cell differentiation through a signaling event [7, 12].

Likewise, some studies show CSCs prefer to use OXPHOS instead of glycolysis, consume less glucose, synthesize less lactate and produce a higher amount of ATP than differentiated cancer cells [18]. CSCs with OXPHOS phenotype have been found in Acute Myeloid Leukemia (AML), glioblastoma, melanoma, pancreatic, ovarian, breast, lung, and papillary thyroid cancer [33].

The main advantage of opting for OXPHOS is to obtain energy upon nutrient shortage conditions since it is a more efficient method of energy production. In fact, $CD44^+CD117^+$ ovarian CSCs and $CD133^+$ Pancreatic ductal adenocarcinoma (PDAC) CSCs are less sensitive to glucose and glutamine shortage than corresponding differentiated cancer cells [33]. Another possible explanation for the OXPHOS phenotype is the availability of nutrients in the tumor niche that directly feed the TCA cycle. One such situation occurs when cancer-associated fibroblasts and differentiated cancer cells secrete lactate which is then used by CSCs for TCA cycle anaplerosis, a process known as reverse Warburg effect [18,33]. Another case is the secretion of alanine by pancreatic stellate cells which can be converted to pyruvate by ALT in pancreatic CSCs, feeding a TCA cycle derived-OXPHOS in CSCs [33].

As mentioned before, moderate levels of ROS promote tumorigenesis and cancer progression. This might be the reason why ROS production through the OXPHOS in CSCs is not detrimental. Furthermore, the increase in mitochondrial OXPHOS has been associated with chemotherapeutic resistance [18], while mitochondrial biogenesis has been identified as a driver of CSC survival and stemness [33]. Although the specific mechanisms by which these processes occur have not been described yet, studies identified some key players. One of those is the mitochondrial biogenesis regulator and transcription factor Peroxisome proliferator-activator 1 alpha (PGC1α) [33]. Nevertheless, overexpression of this transcription factor in CSCs has been associated with either increase in OXPHOS or an increase in glycolysis. Indeed, it is thought that the metabolic

plasticity provided by PGC1α activation is important for metastatic CSCs to adapt to external nutrient conditions of the site they are trying to colonize during secondary tumor formation [17].

iii) Glutamine

Glutamine metabolism is essential to control cell toxicity through the regulation of ROS levels. This is especially relevant for cells that experience ECM-detachment, like metastatic CSCs [17]. In lung and pancreatic cancer, glutamine is essential for CSC stemness through the production of GSH. The reduction in ROS levels resultant from the increase in GSH production prevents phosphorylation and subsequent degradation of β-catenin. Active β-catenin consequently promotes the expression of stem cell markers, like Sox2 [35]. As mentioned before, glutamine-derived glutamate can be converted to OAA and then to pyruvate while simultaneously producing NADPH, in a non-canonical pathway. In a study in PDAC, CSCs exhibited that non-canonical behavior to maintain redox balance and avoid high ROS levels, preventing cell death. Furthermore, glutamine deprivation not only decreased self-renewal and expression of stemness genes but also enhanced CSC sensitivity to radiotherapy [36]. Glutamine has as well been implicated as a player in resistance to chemotherapy in colorectal cancer. Metformin negatively affects CSCs through different mechanisms, one of them is the activation of AMPK, which phosphorylates Tuberous sclerosis complex protein 2 (Tsc2). On the other hand, Tsc2 inhibits mTORC1 leading to a decrease in protein synthesis and cell growth [37]. In fact, metformin-resistant colon CSCs, unlike sensitive cells, do not activate AMPK, show higher expression of glutamine transporter ASCT2, and lose their resistance when in glutamine-depleted medium [38].

Apart from induction of survival and drug resistance, glutamine is essential for cell invasion. Upregulation of the GLS enzyme boosts the conversion of glutamine to glutamate, which upon secretion works as a ligand for Metabotropic glutamate receptor 3 (GRM3) receptor, eventually inducing the movement of a MMP to the cell surface and therefore promoting tissue invasion [17].

iv) Lipid metabolism

As mentioned before, an increase in fatty acid synthesis through the accumulation of glycolytic intermediates is essential for cancer development, as it contributes to the formation of new cellular membranes and lipids that act as secondary messengers in cell signaling. That is especially true for CSCs. In

comparison with differentiated cancer cells, CSCs of some tumors (like breast, melanoma, and glioblastoma) show high levels of expression of SREBP, the abovementioned master regulator transcription factor for fatty acid synthesis [33]. In gliomas, FASN expression is increased and its inhibition decreases stem cell marker expression, as well as proliferation and migration of CSCs [12]. Furthermore, overall high levels of FASN, ACL, and ACC in CSCs have been associated with stemness, metastasis, and tumor recurrence [39].

Also, in support of the role of lipids in CSC signaling, some sphingolipids and eicosanoids were shown to activate self-renewal or survival signaling pathways, like Notch, Akt, or NF-kB, inducing CSC proliferation and tumorigenicity in breast, bladder and ovarian cancers [33].

De novo synthesized fatty acids not only contribute to membrane remodeling and cell signaling but also energy storage in the form of Triacylglycerides (TAGs) inside Lipid Droplets (LDs), in the cytoplasm. LDs are spherical organelles made of retinyl and cholesteryl esters, together with TAGs. Besides fatty acid storage, LDs protect lipids from harmful ROS-mediated peroxidation [40]. They have been associated with cancer aggressiveness, support tumorigenicity of CSCs, and may be utilized as CSC markers [33,39]. In glycolytic CSCs, most pyruvate is routed to lactic acid synthesis. So, those CSCs primarily use glutamine to indirectly produce acetyl-CoA-derived fatty acids first through conversion of glutamate to α-KG and subsequently to citrate.

Nevertheless, CSCs may obtain the substrate for fatty acid synthesis, acetyl-CoA directly from acetate, or directly get fatty acids from CSC niche [40]. In agreement with this, CSCs were observed to exhibit high levels of the fatty acid transporter $CD36^+$ in different cancer types [12]. Metabolism of CSCs is plastic and therefore when these cells are under glucose-deprived conditions or when glycolysis is inhibited because of a chemotherapeutic drug, they use the TAGs stored in LDs instead of glucose to obtain energy. LDs undergo one of two processes: lipophagy, mentioned as an LD autophagy event; or lipolysis, described as the conversion of TAGs to fatty acid through the activity of lipases like Adipose Tissue Triacylglycerol Lipase (ATGL) or Hormone-Sensitive Lipase (HSL) [40]. Free fatty acids interact with acetyl-CoA to form acyl-CoA which in turn associates with carnitine. Carnitine carries the acyl moiety to mitochondria and transfers it to another CoA molecule producing mitochondrial acyl-CoA, while carnitine moves back from mitochondria to the cytoplasm. From that point on, mitochondrial acyl-CoA feeds FAO, producing reducing equivalents that are later used for ATP production through OXPHOS [40]. In this way,

CSCs can survive glucose shortage and escape chemotherapeutic treatments targeting glycolysis. In fact, besides fatty acid synthesis and fatty acid uptake, high FAO activity has been extensively reported as a characteristic of CSCs [33].

Aside from fatty acid metabolism, the lipidic composition of the cell membrane of CSCs is as well peculiar and can facilitate CSC tracing. CSCs have more Monounsaturated fatty acids (MUFAs) than their corresponding differentiated cancer cells, which provides higher fluidity to cellular membranes and consequently contributes to an increase in metastatic potential. In that way, elevated MUFAs expression can be utilized as a CSC marker [39]. One of the main players in CSC phenotype, the NF-kB, regulates the expression and activation of lipid desaturases like the enzyme Stearoyl-CoA Desaturase (SCD). This leads to an increase in unsaturated lipids, mainly MUFAs, which in turn activate one of the most significant pathways in cancer and normal stem cells, Wnt/β-catenin signaling pathway [33,39]. Furthermore, SCD activation has been shown to stabilize Yap/Taz promoting stemness and chemotherapy resistance in lung cancer [39].

High levels of other lipids, such as the cell membrane-component cholesterol, have as well demonstrated to be vital to self-renewal and tumor formation [33]. In CSCs, cholesterol uptake is raised through the intense use of LDLRs, the receptors for low-density lipoproteins which carry cholesterol [33, 39], and via cholesterol synthesis driven by an increase in mevalonate pathway [33]. The importance of cholesterol for CSC phenotype is possibly due to its contribution to lipid rafts [39]. Lipid rafts are cholesterol-enriched micro-areas in the cell membrane that regulate cell adhesion and cell signaling through the signaling proteins they contain. Since membrane lipid composition affects protein recruitment and interaction [33], change in lipid raft composition contributes to loss of integrin-mediated cell adhesion and extracellular matrix degradation, promoting cancer invasiveness [41].

v) Other metabolic features

As mentioned before in the epigenetics section, mutations in IDH enzymes may lead to DNA hyper-methylation through the reduction of activity of TET demethylases via 2-HG accumulation. These mutations are sometimes associated with the CSC phenotype. In leukemia, for example, the mutation promotes CSC self-renewal and impairs differentiation [12].

The metabolism of other amino acids besides glutamine, like serine and glycine, have as well been associated with CSC traits. For instance, in colorectal

cancer, CSCs have high levels of enzymes that transport and catabolize lysine which in turn activates Wnt/β-catenin signaling, contributing to self-renewal and metastasis [12,33]. Other metabolic traits of CSCs include enhanced purine synthesis mediated through Myc activation and overexpression of purine synthesis enzymes [12,33], a rise in the use of PPP (to decrease ROS levels through NADPH) and ketone bodies, and a high hyaluronic acid production through a HIF-mediated induction of glycolysis and consequent increase in the use of HBP pathway [33].

Overall, CSCs activate metabolic pathways that allow them to obtain energy and anabolic substrates to grow in the specific microenvironmental niches where they locate while favoring pathways and the activation of transcription factors, like the previously mentioned NRF2 [17, 33], that increase NADPH and GSH levels, protecting CSCs from high ROS levels usually faced during metastasis formation or radio/chemotherapy.

1.2 Metabolic modeling

Metabolic models are *in silico* mathematical representations of metabolic reactions and associated metabolites. A sub-class of these models, named Genome-Scale Metabolic Models (GSMMs), has the purpose to cover all metabolic reactions inside a cell with the intent to analyze cellular metabolic capabilities, simulate metabolic phenotypes, and/or optimize metabolite production. As GSMMs are a collection of mass-balanced biochemical reactions, it is theoretically possible to determine the rate of any reaction given that a sufficient number of known parameters are known, such as the enzyme catalytic activity and metabolite concentration.

Nevertheless, these parameters are difficult to measure experimentally, and although their values are reported in the literature for some enzymes/reactions, they often cannot be found for all reactions involved in a model at the genome scale. In order to overcome this limitation, constraints based on biologically reasonable assumptions were imposed on these models, allowing to build what is known as constraint-based models.

This section reviews the fundamentals of constraint-based GSMMs, mentions the historical contributions to build human GSMMs, and describes algorithms frequently used to perform simulation and analysis on those models. It also addresses omics-data integration methods for reconstruction of context-specific models capable of mimicking the metabolic behavior of specific cell-

types. Furthermore, this section presents some studies of context-specific modeling in human cells, particularly focused in cancer cells and finishes with the basis for enhancing models with enzymatic constraints.

1.2.1 Genome-Scale Metabolic Models (GSMMs)

The traditionally reductionist research methodology applied in Biology consists either in fully describing each component (e.g. the structure of a gene, enzyme, miRNA) of a system (i.e. cell), or removing each component at a time to understand its function (e.g. to evaluate whether it impacts growth, survival, or differentiation). Besides, the way a component interacts with others is typically assessed by cutting the connections to other components one at a time or in combinations [42]. This strategy allows to build visual maps that enable scientists to grasp some simple processes, but as the complexity increases, it becomes increasingly more difficult for a human, who can keep track of only a limited number of variables, to understand the system just by visual inspection of such maps. Furthermore, cellular phenotypes may not necessarily be due to the presence of a component or set of components, but instead the combinations of tuned levels of those components [42]. Hence, there is a need to represent those systems in a quantifiable and unified language which is amenable for a computer to analyze [42], to make accurate phenotypic predictions and suggest viable non-intuitive solutions for biological problems.

The advent of whole-genome sequencing and subsequent refinement of the technology to a high-throughput scale, where it became affordable to sequence genomes of organisms in a short time [43], the increase in the availability of other omics datasets (like proteomics or metabolomics) [44,45] in publicly available databases, together with the evolution of computational power [42,44] and mathematical modeling [43] fostered the development of Systems Biology. This relatively new field of research, instead of just focusing on the description and qualitative interactions of components of a biological system, applies *in silico* modeling to quantitatively study the interplay between those components and how it affects biological function [42, 45].

A similar transition from a reductionist to a holistic view of biological systems was observed with metabolic modeling. Following the sequencing of the first genome in 1995, of the *Haemophilus influenzae* [46], scientists were able to build the first metabolic model at the genome-scale for the same organism, four years later [47]. Since then, several GSMMs have been created to model the metabolism of different organisms across all five taxonomic kingdoms [48].

Relevant examples are *Escherichia coli* (a reference gram-negative bacteria), *Mycobacterium tuberculosis* (a pathogenic gram-positive bacteria), *Methanosarcina acetivorans* (an archea able to synthesize methane), *Saccharomyces cerevisiae* (a yeast), and Arabidopsis thaliana (a reference organism in plant research) [48].

As mentioned before, GSMMs comprise all known mass-balanced metabolic reactions of an organism, but an extra layer of information can be added on top, the Gene-Protein-Reaction (GPR) rules. In other words, the metabolic reaction network can be expanded with the associations of each metabolic reaction to the corresponding enzyme or enzymes that catalyze them and to the genes that encode those enzymes in that organism [48]. This is useful to test the effect of specific mutations such us gene knockout, and gene expression up-regulation or down-regulation in the metabolic network. Advantages are the possibility to predict essential genes or to suggest more realistic strain design methods (than those solely based on manipulation of reactions). Also, GPR associations allow the integration with transcriptomics data in the construction of tissue-specific models [49].

There are four stages to build a GSMM:

i) Create a draft reconstruction.

This stage starts with structural and functional annotation of the organism's genome. The start and end of each gene are annotated in the genome sequence (structural annotation) and the sequence of those genes that code for proteins is converted to amino acid sequence (to account for genetic code redundancy) [50]. The amino acid sequence is then compared with protein sequences of other organisms deposited in databases. Using alignment tools, it is possible to identify similar sequences of phylogenetic close organisms that are already annotated and, therefore, assume that the 'query' gene has the same function as the 'matched' gene (functional annotation). Genes encoding metabolic enzymes and transport proteins can be spotted using the appropriate Gene Ontology (GO) annotation categories and respectively assigned to an Enzyme Commission (EC) or Transport Classification (TC) number. EC and TC numbers can then be used to map the genes to the corresponding metabolic and transport reactions, by resorting to databases of biochemical reactions, such as KEGG or BRENDA [51].

ii) Manual refinement.

This is a time-consuming stage involving the manual curation of the reconstructed draft, where each reaction is evaluated based on literature or experimental evidence supporting its existence in the specific organism [51]. During

this stage of the reconstruction, special care should be given to some aspects, such as: reactions should be mass and charge-balanced; the standard Gibbs free energy of products and reagents is used to determine reaction directionality; non-catalyzed and intracellular transport reactions must be included; information on biomass composition, determined *in vitro* or estimated from genomic data, should be converted into a pseudo-reaction representing cell growth, called biomass reaction; and the amount of ATP needed for cell growth and cell maintenance should be respectively included in the biomass and in a specific ATP maintenance reaction [51].

iii) Conversion to a mathematical model.

This is an automated step where the refined metabolic reconstruction is converted into a mathematical format [51].

iv) Network evaluation.

In this final stage, dead-end metabolites (internal metabolites that are either only produced or consumed) are identified together with the directly or indirectly associated blocked reactions (reactions with no flux in any condition). Then, by overlapping the model's network at those points (dead-ends and blocked reactions) with maps of the generic metabolism is possible to identify reactions that may fill the gaps in the model network. This is known as gap filling and should be backed up by organism-specific literature evidence. Then, the model is evaluated. The criteria used in the model evaluation are the ability to secret certain by-products at a specific yield, present known metabolic inabilities (e.g. be auxotrophic for specific nutrients), ability to produce all biomass precursors present in the biomass equation, and simulate a growth rate similar to the one measured in the lab. Another strategy frequently applied to validate the model is to identify genes essential for growth in the model and compare with experimental gene essentiality datasets [51].

Stages **ii)** to **iv)** in the reconstruction process are iterated several times until the performance of the model is satisfactory. So, the reconstruction can be thought of as an iterative process that takes months to years.

GSMMs have been used to model reference organisms of interest for scientific research and industry, as they assist in microbial strain design and optimization for the production of industrially useful bio-compounds [48], like amino acids [52] and biofuels [53]. They have also been applied to identify drug targets in pathogens [54], and to study organisms' interplay, such as cross-feeding events in microbial communities [55], the human gut microbiome [56], or host-pathogen

interactions [57].

Another possible use for these models is the prediction of isozymes and promiscuous enzymes. The detection of genes that encode enzymes essential for cell survival (when the gene is knocked out leads to cell death) in the models, but which are shown to be non-essential *in vivo*, signal that the model (and hence the scientific knowledge) is missing another gene encoding for an enzyme (i.e. isozyme) that catalyzes the same essential reaction. By identifying genes with similar sequences to those (using BLAST) it is possible to select potential isozymes that are then experimentally validated [48]. On the other hand, BLAST can be used to identify other potential functions (i.e. reactions) of a metabolic gene for which the main function (i.e. associated reaction in the model) is already known. It is then easy to identify genes/enzymes in the model with the potential to have a side function that is also the main function of another different gene/enzyme (i.e. it is possible to spot potential promiscuous genes/enzymes). Before validating the promiscuous capacity of such genes/enzymes in the lab, the list of potential promiscuous entities can be narrowed down by testing *in silico* whether one can maintain the metabolic function when the other is knockout [48]. When representing the human metabolism, GSMMs may as well explain diseased mechanisms and suggest drug targets for therapeutical interventions [58]. Next, the efforts of the scientific community to build human GSMMs are reviewed.

1.2.2 Human Genome-Scale Metabolic Models

In 2004, the project aiming to fully sequence the human genome reached its end [59], and only three years later the first generic GSMM for humans, Recon1, was reconstructed [60]. In the same year, another generic reconstruction emerged, the Edinburgh Human Metabolic Network (EHMN) [61]. Since then, these models underwent successive updates and gave origin to new models. In 2009, a context-specific model incorporating the information of Recon1 and EHMN was built specifically for hepatocytes, the HepatoNet1 [62], while the Human Metabolic Reaction (HMR) [63], a generic model based on those same two models and with information on KEGG and HumanCyc was created in 2012. HMR was updated [64] two years later to the version HMR2 that condensed the information of the previous HMR model with new data on lipid metabolism [64]. Both HMR models have served as a starting point for the reconstruction of context-specific models of different tissues, like hepatocytes (to study non-alcoholic fat liver disease) [64], adipocytes (to understand obe-

sity) [65], myocytes (to model diabetes) [66], and cancer cells [63].

In 2013, an updated version of Recon1 which integrated information of EHMN and HepatoNet1 was created [67]. That model was called Recon 2 and underwent subsequent updates in the following years. Recon 2.1 had generic metabolites replaced by specific ones to ensure carbon balancing [68]. Recon 2.2 improved mass and charge balancing of all reactions and enhanced the representation of energy generation, achieving a model that correctly predicts ATP flux under different carbon sources [69]. Another evolution of Recon2 is Recon2M2, which was upgraded with Gene-Transcript-Protein-Associations (GTPAs) so that it could account for the isoforms resultant of metabolic genes subjected to alternative splicing [70].

The most recent model of the Recon series is Recon3D, a model released in 2018 that includes three-dimensional metabolite and protein structure and integrates data from HMR2 with Recon2, together with more reactions associated with drug-metabolism, transport, host-microbiome interactions, and absorption and metabolism of dietary compounds [71]. Note that visualization tools and databases that allow to inspect and query data of this kind of models have also been developed. Examples are the ReconMap [72] for visualization of information deposited in Recon2 and the Virtual Metabolic Human (VMH) [73] database that links the Recon3D metabolic model with human microbiome reconstructions and visual data maps, as well as knowledge on disease and nutrition. In 2020, two Whole-Body-Metabolic (WBM) models based on Recon3D were published: one for the male, called Harvey, and another for the female, named Harvetta [74]. Omics data and literature were used to reconstruct compartments depicting the organs of humans of both genres intertwined by other compartments which represented the body fluids [74].

In the same year, the most recent generic human GSMM was built, the Human1 [75]. It reconciles the information of HMR2 with Recon3D and iHsa [76] (an improved version of Recon2) to produce a unified GSMM of 13417 reactions, 10138 metabolites and 3625 genes [75]. In comparison with previous models, this reconstruction excluded duplicated or unnecessary reactions and metabolites, rebalanced and corrected reversibility of reactions, and updated the biomass reaction composition. Changes made during model development and future changes can be tracked by a version control open-source framework so that the community may engage in the model refinement. Furthermore, a web-portal, Metabolic Atlas (`www.metabolicatlas.org`) was made available to facilitate exploration and visualization of the data in the model [75].

1.2.3 Constraint-based models

The standard strategy to mathematically model biochemical reactions is to use kinetic models. To build kinetic (a.k.a. dynamic) models some assumptions must be made. One of them is that the compartments within the system under study (e.g. cytoplasm, organelles, or the extracellular compartment) must be small enough to assume there are no spatial gradients (i.e. values of physical parameters do not change with position within the compartment), and be large enough so that compounds are not crowded and hence stochastic events, like for example molecules reacting together due to a random clash [77], cannot happen.

The second law of thermodynamics is another physicochemical assumption that must be followed. It can be understood as the difference of Gibbs free energy between products and substrates (ΔG) that determines if a reaction is irreversible (when $\Delta G \neq 0$) and in that case in which direction occurs (forward if $\Delta G < 0$, reverse if $\Delta G > 0$) [51, 78]. Another premise is that there must be mass and charge balance inside the system which constrains the system stoichiometrically. In other words, the stoichiometric coefficients of metabolites in any reaction are such that the number of atoms of each element and charge in substrates is the same as in the products [78]. The mass balance applies not only to elements within the same reaction, but also to internal metabolites across reactions where those metabolites are consumed or produced. Therefore, in Figure 1.5, the variation in concentration of metabolite A with time can be represented by:

$$\frac{d[A]}{dt} = c_{A,v_1} v_1 + c_{A,v_2} v_2 \tag{1.1}$$

where $[A]$ is the concentration of A, t is time, v_1 and v_2 are respectively the rates of reactions that produce and consume A, while c_{A,v_1} and c_{A,v_2} are the corresponding stoichiometric coefficients. Those coefficients are positive when the metabolite is a product and negative when it is a substrate [79]. Hence, the above equation could be simplified to:

$$\frac{d[A]}{dt} = v_1 - v_2 \tag{1.2}$$

Note, however, that concentration depends on cell volume which can vary with cell growth, so an extra term, $\mu[A]$ (where μ is growth rate), is sometimes subtracted to the right side of the equations 1.1 and 1.2 to compensate for the increase in volume, but because $\mu[A]$ is much smaller than the other terms it is often omitted ($\mu[A] \approx 0$).

In kinetic modeling, v_1 and v_2 are replaced by functions dependent on $[A]$ that

contain kinetic constants [77, 79]. For example, when using Michaelis-Menten kinetics the v_{max}, the maximum rate a reaction can have, and k_M, the substrate concentration at which the reaction rate is half of v_{max}, i.e. the Michaelis-Menten constant, have to be known [79,80]. Such constants can be found in the literature for reactions of models of small dimensions. However, for large models such as GSMMs, kinetic parameters of many reactions are unknown [77, 81] and even if they were, the amount of time and computer power needed to solve a system of that many differential equations (each corresponding to a metabolite used by the cell) would be substantial [81]. Therefore, to model the metabolism at the genome scale, it is frequent to make another assumption: that the cell is in a quasi-steady state. The quasi-steady state ascertains that there is almost no variation in the concentration of metabolites over time. In other words, that the same amount of a metabolite produced in a reaction(s) is consumed by other reaction(s) in the system so that its concentration remains stable, i.e. the metabolite neither substantially accumulates nor depletes overtime. The quasi-steady state assumption can be made because, as long as the cell is not exposed to perturbations [81] (like change in concentration of an external substrate, pH, or temperature), the rate of metabolic reactions is much faster than that of other cell processes (transcriptional regulation, cell cycle) and environmental changes [82]. Therefore, it is a realistic assumption that is applied in specific situations and imposes constraints over the possible solutions (i.e. values of reaction rates) a system may present. Hence, the models where steady state is assumed are called constraint-based models.

Since there is no variation in concentration with time for internal metabolites when using constraint-based modeling, the above differential equation 1.2 is simplified to the linear equation $v_1 - v_2 = 0$, where the concentration of metabolite and the kinetic parameters disappear. As there is no need to determine kinetic parameters, unlike kinetic modeling, constraint-based modeling can be applied to GSMMs for which no kinetic parameters are known. In addition, the linear nature of the equations in a constraint-based model makes it much faster and easy to find a solution, as it requires less computer power. Nevertheless, it should be noted that kinetic models are much more precise than constraint-based models and perfectly amenable to represent specific pathways and even metabolic networks of small and medium scale [81].

To represent the complete system of linear equations of a constraint-based model the following equation is used:

$$S.v = 0 \tag{1.3}$$

where S is a matrix with the stoichiometric constraints applied to the system, and v is a vector of variables that stand for fluxes/rates of each reaction in the system. The matrix S has m rows and n columns, representing the metabolites and reactions, respectively. Each value in the matrix is the stoichiometric coefficient of a metabolite m in reaction n. Positive coefficients show the metabolite is produced and negative values that the metabolite is consumed [81, 83]. Note that external metabolites, the ones participating in exchange reactions (i.e. reactions that connect metabolites of the most external compartment with the outside of the system) are not represented as rows in the stoichiometric matrix S, since they are not bound to mass balance (Figure 1.5). This fact assures that the concentration of external metabolites (the ones moving in or out of the system) can change with time, as it happens *in vitro* with cell medium components and products released from the cell.

Furthermore, the fluxes of reactions in vector v are subjected to reversibility constraints. Those are mathematically represented by inequalities that define a lower and upper limit to the flux of each reaction. Reactions in the forward direction can carry no flux when the reaction does not occur or positive flux values if it occurs (i.e. flux v_n of reaction n is $0 \leq v_n \leq +\infty$), while reactions in the reverse direction can carry no flux or have negative flux values (i.e. $-\infty \leq v_n \leq 0$). Reversible reactions can assume any flux value (i.e. $-\infty \leq vn \leq +\infty$) [81].

In order to algebraically solve the system of equations representing the constraint-based model (i.e. to obtain values for reaction fluxes), the system has to be determined. In other words, the number of variables (fluxes) must be the same as the number of equations (internal metabolites) for a solution to be found. However, in most biological systems, there is often a higher number of reactions than internal metabolites, so these systems are said to be underdetermined [81, 84]. In that case, there is more than one possible solution for the system of equations, so the null space of S (i.e. any v that satisfies the equation 1.3) can be thought of as a convex polyhedron in a 3-dimensional space, the limits of which are imposed by mass balance and capacity constrains (lower and upper flux bounds), as well as the quasi-steady state assumption (Figure 1.6) [85].

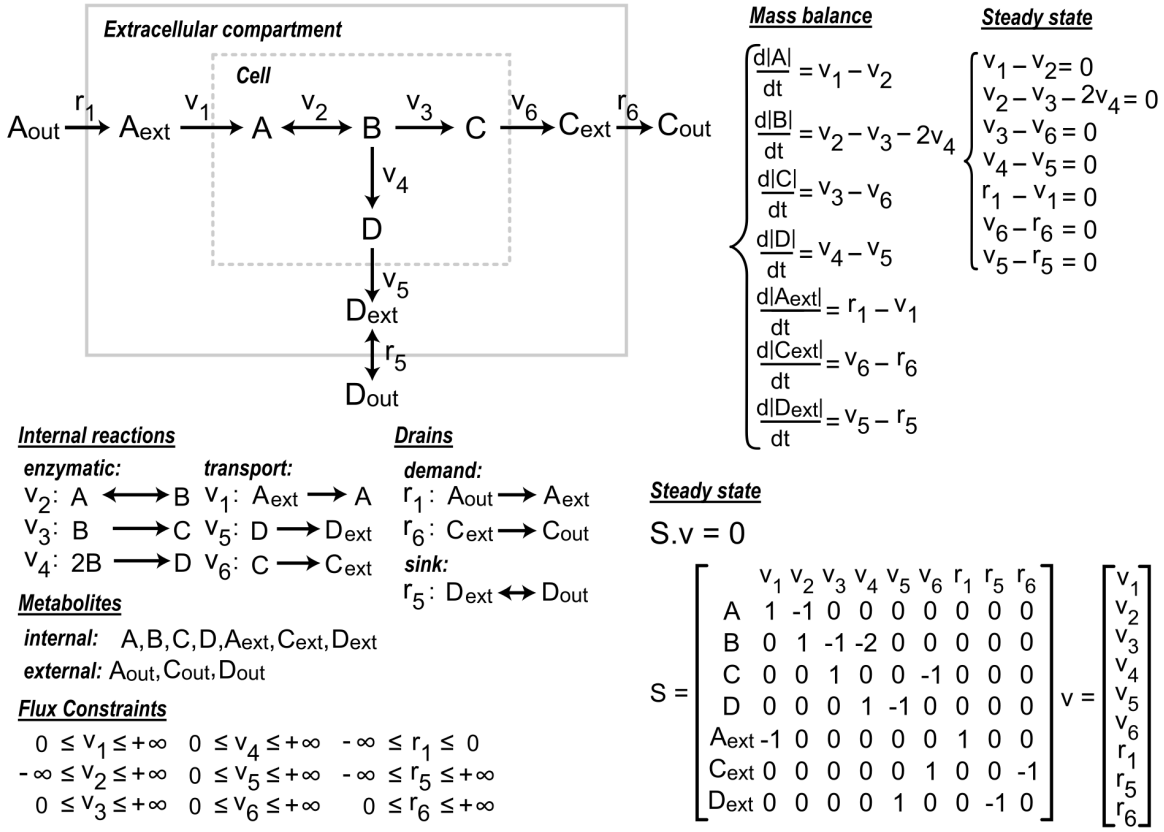

Figure 1.5: Toy metabolic model. **Left**: a metabolic system including a cellular and an extracellular compartment. Internal reactions can interconvert metabolites (enzymatic) or transport them across compartments (transport reactions). Drains are added to allow the entry or removal of a compound from the system. In this case, all drains are also exchange reactions, as they connect the most external compartment with the outside of the system. Up and lower bounds constraint fluxes capacities and define reactions direction. **Top center**: A system of differential equations based on the mass-balance assumption representing the system at left. **Top right**: the same equations but with steady state assumptions. **Bottom right**: mathematical representation of the constraint-based model based on the metabolic network at left. External metabolites are not represented in the stoichiometric matrix S, since they are not mass balanced.

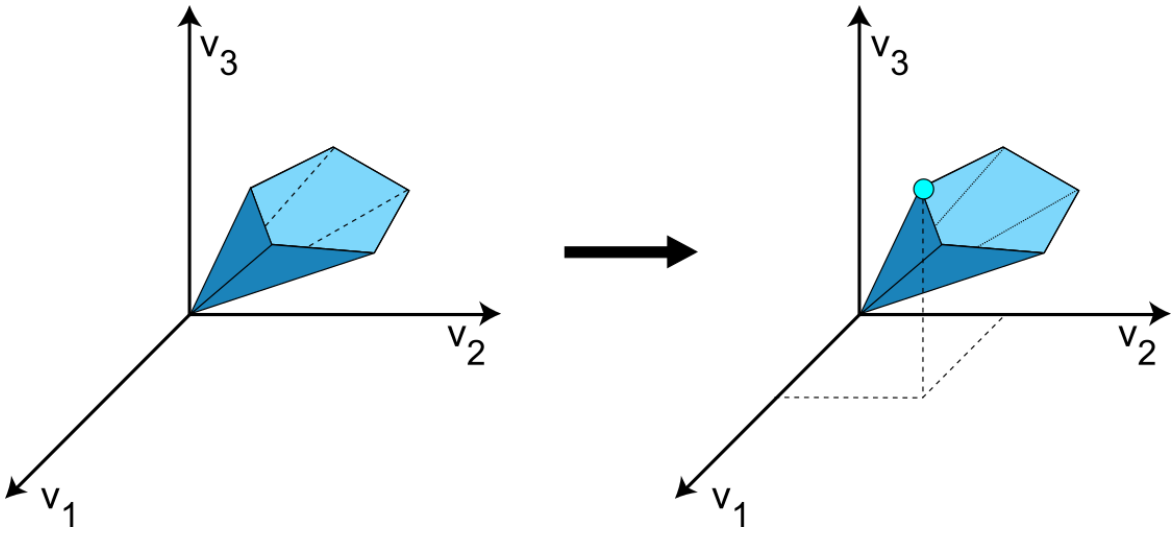

Figure 1.6: Representation of the solution space of a toy constrained-based model with three reactions, in which each coordinate represents a reaction flux. Left: The solution space is a pointed polyhedron due to stoichiometric and flux capacity constraints. Right: The blue point represents an optimal solution provided by FBA. Redrawn figure based on Orth et al. (2010) [85].

To convert an underdetermined system into a determined one, there is a need to first identify the degrees of freedom, that is, the number of variables that must be known to make the system solvable. In basic systems, where there are no linearly dependent rows or columns in the matrix S, this is simply the difference between the number of columns (reactions) and rows (metabolites), whereas when the matrix S has linear dependencies, the rank of the matrix must be considered in the calculation. Then, scientists may experimentally measure the same number of fluxes as the degrees of freedom to make the system determined. Usually, those are fluxes of exchange reactions, as they are easier to obtain. If that is not sufficient to make the system solvable, then internal fluxes may be further assessed through substrate labeling with ^{13}C followed by quantification of labeled carbon contained within the products, using mass spectrometry [84].

When the number of experimentally quantified fluxes is higher than the degrees of freedom, the system becomes over-determined (i.e. the number of equations/metabolites is superior to the number of variables/unknown fluxes). In such situations, a better approach than selecting just a few measured fluxes (in the same amount as degrees of freedom) and algebraically estimating the remaining unknown ones is to use a method named overdetermined Metabolic Flux Analysis (MFA). With this method, the values of all fluxes (known and unknown) are re-estimated by reducing differences between known experimen-

tally measured fluxes and *in silico* estimated fluxes of corresponding reactions. A typical approach is to do a linear least square regression, where the sum of squares of the errors associated with those differences is minimized [84]. Overdetermined MFA may work just with fluxes of exchange reactions (obtained through exometabolomics experiments), in which case is called stoichiometric MFA (stMFA), or use as well internal fluxes measured through ^{13}C tracing, which is known as ^{13}C-MFA [81].

1.2.4 Methods for phenotype simulation

Although it can be possible in some cases to transform an underdetermined system into a determined one by experimentally obtaining fluxes for some reactions and arithmetically estimating the others, such a procedure is typically unfeasible for large systems, as the degrees of freedom are often too high [84]. Furthermore, each time the model is needed to make a simulation under new environmental settings, a different experimental setup needs to be put into place to acquire the minimal number of fluxes needed to solve the system. Therefore, computational methods have been developed to estimate flux distributions under different environmental conditions, or in other words, to make phenotypic predictions.

One type of methods allows phenotypic simulations in under-determined constraint-based models by making reasonable biological assumptions and applying optimization strategies to choose a flux distribution within the solution space, i.e. to select one of the flux vectors that solves the equation 1.3, which may also be perceived as a point within the flux polyhedron shown in Figure 1.6. That class of methods turns the constraint-based numeric problem into an optimization problem, whose mathematical formulation, besides the abovementioned constraints, includes an objective function.

The most commonly applied method of this class is Flux Balance Analysis (FBA) [83]. FBA consists in maximization or minimization of a flux of a single reaction (has one objective) or a linear combination of fluxes of different reactions (multi-objective function) [83], where the mathematical definition of the objective function is:

$$max/min : Z = c^T.v \tag{1.4}$$

where Z is the objective function, v is the vector of all variables/reaction fluxes, and c is a row vector of weights given to each reaction. Whenever the weight of a reaction is different from zero, it means the function is trying to optimize (maximize or minimize depending on the signal) the flux of that reaction. The

magnitude of the weight determines how much each reaction contributes to the objective function [85]. For example, if the objective is to maximize a function in which the weight of reaction A is 1 while the weight of reaction B is 2, then the flux of B will be maximized twice as much as A.

Most of the time, the objective function assumed in FBA is one that tries to maximize growth (i.e. the flux of biomass reaction) [79]. This assumption, grounded in the idea of adaptative evolution [86], is proven experimentally, as cells try to maximize their proliferation (i.e. biomass) in the exponential growth phase, which is valid for both procaryotes and eukaryotes [79]. Besides, experimental data shows a good correlation with FBA predicted phenotypes in wild-type organisms, in relation to growth yield, flux values, and substrate uptake rates [87]. Nevertheless, alternative objectives may be used with FBA, such as to maximize ATP production, minimize the sum of squares of all fluxes (assuming there is maximum enzymatic efficiency during growth), minimize the number of active reactions [79], maximize the formation of a biotechnologically relevant compound [83], minimize nutrient uptake rates [86] or minimize redox potential (assuming cells reduce oxidizing reactions to conserve energy) [88].

One of the main disadvantages of FBA is that when several possible flux distributions fulfill the same optimal value of the objective function (i.e. there is more than one optimal solution) it does not have a criterion to select one among those [83]. Therefore, for situations where multiple optima can be observed, parsimonious enzyme usage FBA (pFBA) is an alternative that, even though it does not always guarantee a unique solution, can alleviate the problem. pFBA consists of two steps. First, an FBA analysis identifies the maximum objective value (e.g., maximum growth rate), which is then applied as a flux constraint (e.g., biomass reaction flux lower bound is set to that value), while a new linear optimization problem is run whose objective function is the minimization of the sum of the of all fluxes [83, 89]. This procedure narrows down the number of suitable optimal distributions by keeping overall flux to a minimum, while assuring the maximization of an objective (e.g. growth) [83] and it is based on the biologically reasonable assumption of reduced cellular enzyme usage [90].

FBA and pFBA make accurate phenotypic predictions in wild-type organisms, but the same does not apply to phenotype simulation with mutant organisms. Mutants created *in vitro* are not under the same evolutionary pressure as wild-type strains and did not have time to develop a reaction flux regulatory mechanism that optimizes a biological objective such as growth, i.e., mutants need time to fully adapt. Hence, phenotypic simulation methods for non-adapted mutants were developed based on the more reasonable premise that such mu-

tants are likely to display a sub-optimal flux distribution intermediate between the wild-type and mutant optimum i.e., the mutant metabolic phenotype is very close to the wild-type one [87]. Examples of such methods are Minimization Of Metabolic Adjustment (MOMA), which minimizes the sum of squared differences between wild-type and mutant fluxes of each relevant reaction in the wild-type distribution [87] and Regulatory On/Off Minimization (ROOM) [91], that minimizes the number of reactions of a mutant organism whose fluxes are significantly different from the corresponding ones in the wild-type.

Before implementing any of the aforementioned methods for phenotypic simulation in mutants, there is a need to define which reactions are affected by the mutations and to what degree. First, the set of mutated genes is defined and the relative expression values of each mutated gene in relation to a reference wild-type strain are obtained. If the relative expression value is above one, it means the gene is being over-expressed. In case it is below one, then the gene is under-expressed, and when the value is zero, the gene is knocked out [89].

Afterwards, the gene expression value is propagated into a reaction value. In those cases where there is a one-to-one correspondence of a gene to a protein, the reaction value assumes the relative expression value of the corresponding gene, whereas in many-to-one correspondence, the Boolean GPR rules of the model must be first translated into functions that can deal with numerical expression values [89]. If the reaction is catalyzed by a protein complex, a situation usually represented by an *AND* operator in GPR rules, the value associated with the reaction is the minimum of the values of genes coding the proteins of the complex. This is because, albeit the expression of all genes encoding all the enzyme sub-units is needed for the reaction to happen, the expression of the gene with the minimum transcriptional level acts as a bottleneck for the formation of the enzyme complex [89].

On the other hand, if the reaction can be catalyzed by more than one enzyme (i.e., there are isozymes), a case represented by an *OR* operator in GPR rules, then the value attributed to the reaction can be the average [89], sum or maximum of the expression levels of the genes, depending on one's choice [92]. The average or sum can be applied when there is the assumption that although the genes involved work independently, both isozymes catalyze the reaction when both genes are expressed [89], while the maximum function is applied upon the assumption that the isozyme encoded by the most expressed gene is the one that most influences the reaction [92].

Once the affected reactions have been identified and the degree of influence of the mutation on those have been assessed, that information is translated into

flux constraints over the reactions. Assuming v_i as reaction flux value of the reference (i.e., wild-type strain) and p as the calculated relative expression value for the reaction in the mutant: if $p > 1$, then the reaction will be overexpressed and the flux is constrained to be larger than v_i multiplied by p; if $p < 1$, then the reaction will be underexpressed and the flux is constrained to be lower than v_i multiplied by p. Depending on whether the flux value of the reference organism is positive or negative, this results in constrains to the upper or lower bounds of the fluxes (see Table 1.1) [89]. After adding the flux constraints, phenotypes can be simulated.

Table 1.1: Constraints applied to mutant reaction fluxes.

	Positive reference flux	**Negative reference flux**
Overexpression ($p > 1$)	$p.v_i < flux < UB^*$	$LB^* < flux < p.v_i$
Underexpression ($p < 1$)	$0 < flux < p.v_i$	$p.v_i < flux < 0$

* UB and LB are original upper and lower bounds, respectively, of the reference flux.

There are other phenotype prediction methods that, unlike the ones mentioned before, do not need to optimize for a biological objective and, therefore, are not biased. These are collectively known as Metabolic Pathway Analysis (MPA) methods, which try to represent all possible phenotypes that a metabolic network may have, depending solely on the aforementioned stoichiometric, thermodynamic, and steady-state constraints of constraint-based models, that is, without making any assumptions on a biological purpose [78].

Within the scope of MPA, Elementary Flux Modes (EFMs) analysis is one of the best-known methods [49]. EFMs are non-decomposable flux distributions, representing subnetworks within the metabolic network, that upon combination can describe all the feasible flux distributions that a model may theoretically present. Biochemically, EFMs may be thought of as subnetworks/paths that connect sets of metabolites in a network without unnecessary loops and therefore represent the basic metabolic entities of a system [81, 93], while in the visual representation of the constraint-based solution space described in Figure 1.6 they can be understood as the edges of flux polyhedron [49].

From a phenotypic prediction standpoint, EFMs are useful to identify organisms and the substrates that each of those requires to produce a metabolite of interest or to observe all the phenotypic consequences of deleting specific genes or reactions. However, the enumeration of all EFMs is computationally demanding, and it can be infeasible when applied to large-scale models. For example,

a medium-scale model can show hundreds of millions of EFMs, which enumeration can be time-consuming. In comparison, the abovementioned optimization algorithms are faster [79, 93].

1.2.5 Methods to analyze metabolic models

Albeit utilized in the prediction of specific metabolic flux distributions, GSMMs may as well be used to study the generic metabolic limits and properties of an organism. Flux Variability Analysis (FVA) is one of the methodologies which can meet that purpose. Whilst it does not identify all possible flux distributions (or the best one) that lead(s) to a specific metabolic phenotype, FVA uncovers the range of flux variability that is allowed within the metabolic constraints of the system [94].

Specifically, it applies Linear Programming (LP) to a pair of objective functions that are the minimization and subsequent maximization of the flux of each reaction, to identify the lower and upper flux bounds that the reaction may theoretically present. When alternate optima (several distinct flux distributions that achieve the same optimal objective value) exist, FVA can be applied to identify the minimum or maximum flux values admitted by each reaction in at least one of those alternative flux distributions. In that case, an FBA must first be performed to identify the optimal objective value (e.g., highest growth rate), which is set as an additional flux constraint to the model (e.g., the highest growth rate is set as lower bound of the biomass reaction), followed by maximization and minimization of the flux of each reaction [94].

FVA enables the detection of reactions that are essential to maintain an optimal phenotype (e.g. highest growth rate) regardless of the flux distribution that the system may use to achieve that optimality. Basically, a reaction where the admissible flux range (interval between maximum and minimum values) does not comprise a zero value in an FVA can be considered essential [95]. Conversely, reactions where both the maximum and minimum values are zero, are said to be blocked, as they are inactive in any flux distribution that follows the system constraints [96]. With FVA, it is also possible to spot potentially interesting reactions that show low flexibility (narrow flux range). Hence, FVA allows to identify reactions that characterize or oppose the phenotype of interest and even provides an overview of phenotype redundance, as few essential reactions suggest the organism may use alternative pathways to meet its phenotypic goal [94, 95].

FVA may not only be applied to optimal, but also sub-optimal conditions. For example, flux variability can be computed to a smaller percentage of maxi-

mum growth. This is useful, for instance, to detect reactions that are essential for survival in non-optimal growth conditions [97].

Note that the aforementioned EFMs can be used as well to analyze metabolic networks, namely to determine their robustness, through simple approaches such as comparing the number of EFMs before and after a reaction knockout [98].

1.2.6 Context-specific models

Multicellular organisms undergo cell differentiation during their development. Although all cells of an organism share the same genetic code, differentiated cells of distinct tissues execute different functions due to variation in gene transcription, mRNA translation levels, and protein Post-Translational Modifications (PTMs). In that sense, generic GSMMs, solely built on the genetic context of the organism, cannot accurately describe the specific metabolism of each cell type. Similarly, unicellular organisms show distinct gene/protein expression patterns depending on the available substrates and environmental conditions, which affects their metabolism. Hence, in both cases, there is a need to integrate condition/cell type-specific omics data into generic GSMMs to reduce the solution space and consequently attain more accurate phenotypic predictions [99].

As previously mentioned, genomics data serves as a foundation to build organism-specific GSMMs suitable to study wild-type or mutated unicellular organisms. On the other hand, cell type/condition-specific GSMMs, whilst built upon generic models, require the integration of other types of omics data: transcriptomics, proteomics, metabolomics, and/or fluxomics.

The main types of omics data are described in Table 1.2. From those, fluxomics is the easiest type to integrate with GSMMs, as it only requires to constraint flux bounds to the values measured. Also, it is the most accurate because, unlike other methods, it does not depend on the assumption of a direct relationship between transcript/enzyme levels and reaction flux. In fact, one of the techniques mentioned above to make predictions with overdetermined constrained models, the MFA, utilizes fluxomics data. However, this type of data is scarce and difficult to obtain for large models.

With respect to metabolomics, the data integration may be performed in either of two ways. One strategy concerns the detection of which metabolites are produced *in vivo* that are not obtained *in silico*, leading to the inclusion of reactions producing those metabolites in the final model, a process sometimes applied to close gaps in reconstructed models. Another way is to relax the

steady-state assumption to allow the empirical accumulation (at an empirical rate) for metabolites which production levels are backed up by metabolomics evidence. Drawbacks of metabolomics data are frequent insufficient coverage and low confidence in the metabolite identification process.

Proteomics is sometimes included in context-specific reconstruction because, unlike transcriptomics which only reflects the gene expression levels of a cell, it also accounts for the effect of mRNA translation over the metabolic flux. However, proteomics coverage is often limited. Hence, as advancements in high-throughput sequencing increased the coverage and speed while reducing the cost of RNA sequencing, the majority of omics-integration methods were developed for transcriptomics data [99–101].

Omics-integration methods can be classified depending on whether they integrate absolute or relative values, to depict the metabolism under a single condition or differences in metabolism between conditions, respectively. When dealing with gene expression, one might assume that the latter approach is better, as it overcomes the limitation of lack of proportionality between transcript and flux levels. However, that is not the case, since none of the two types of methods outperforms the other [100].

Optionally, integration methods may be grouped as discrete or continuous, depending on whether omics data are discretized (e.g. classified as high/moderate/low or on/off, accordingly with arbitrary thresholds) or not. Intuitively, it would seem better to not discretize, but there is no proof that methods using continuous data perform better than those that discretize values. Furthermore, discretization has advantages, such as robustness to data noise, and reduction of reliance on the proportionality assumption between fluxes and omics data. Also, when the discretization is binary (on/off) the integration with the logic-based GPR rules becomes easier than with continuous values [100].

These methods may as well be split into two groups, accordingly with their ability to directly produce flux distributions or context-specific models that can later be used by traditional phenotype prediction algorithms. Nonetheless, some methods may be included in both categories, since they return both a context-specific model and a metabolic flux distribution for the complete model compatible with condition-specific omics data [100].

Another classification system is based on the mathematical objective. In that regard, methods can be grouped into three different main types: Objective Function-Required (OFR), Expression Data-Compatible (EDC), and Core Reaction-Required (CRR) [101].

Table 1.2: Main types of omics data useful to build constrained-based models

Type	**Description**	**Deposited in**
Genomics	Full-genome sequences obtained with Next-Generation Sequencing (NGS) techniques	GenBank [102], European Nucleotide Archive (ENA) [103], DNA Data Bank of Japan (DDBJ) [104], Genomic Data Commons (GDC) [105]
Transcriptomics	mRNA levels quantified by microarrays or RNA-sequencing. It allows to know which splicing isoforms are transcribed and how much	Human Protein Atlas (HPA) [106], Gene Expression Omnibus (GEO) [107], RNA-seq Atlas [108], ArrayExpress [109], Genotype-Tissue Expression (GTEx) [110], GDC [105]
Proteomics	Protein levels quantified by mass-spectrometry or western blotting	HPA [106], Human Protein Reference Database (HPRD) [111], Human Proteome Map (HPM) [112]
Metabolomics	Metabolites present inside the cells or in the extracellular media (exometabolites) quantified by mass-spectrometry or Nuclear Magnetic Resonance (NMR)	MetaboLights [113], Human Metabolome Database (HMDB) [114], Metabolomics Workbench [115]
Fluxomics	Metabolite turnover rates (i.e. reaction rates) determined by the proportion of carbon atoms with ^{13}C in reaction products upon labeling of substrates with ^{13}C. Quantification of ^{13}C is achieved with mass-spectrometry	Central Carbon Metabolic Flux Database (CeCaFDB)) [116]

i) Objective Function-Required (OFR) methods

As the name suggests, Objective Function-Required (OFR) methods [101] utilize a required metabolic function. In other words, they apply FBA to find the objective value that optimizes a metabolic function representing a biological objective, such as growth. A fraction of the objective value (e.g., at least 90% of growth) is then imposed as a minimal flux constraint, while a penalty function denoting the differences between simulated fluxes and gene expression levels is minimized [99].

The founding method of this class is Gene Inactivation Moderated by Metabolism and Expression (GIMME) [117]. In that algorithm, the gene expression level is converted to the associated reaction expression level using the functions mentioned before that apply GPR rules to numerical values. Then, the difference between the reaction expression level and a user-defined threshold is determined. For the reactions where the expression level is lower than the threshold, that difference multiplied by the reaction flux is defined as a reaction penalty. The sum of those penalties represents the inconsistency score (i.e., penalty function), which is minimized. This formulation penalizes (gives lower flux values and eventually excludes) reactions associated with low expressed genes but carrying high fluxes. While the penalization step excludes reactions (to reduce the inconsistency between expression and flux values), the required metabolic function guarantees the inclusion of low expressed reactions essential for the model operability [100, 117]. GIMME outputs both a flux distribution and a contextualized GSMM and integrates absolute expression as continuous values [100].

GIMME gave origin to other methods, like Gene Inactivity Moderated by Metabolism and Expression by proteome (GIMMEp), which contains a similar formulation to GIMME but utilizes proteomics instead of transcriptomics [118]. Gene Inactivation Moderated by Metabolism, Metabolomics and Expression (GIM_3E) [119] is also a GIMME-derived algorithm. However, this method is substantially distinct. Unlike GIMME, which gives a penalty score just to the reactions with lower expression than the threshold, GIM_3E attributes a penalty to all reactions. Specifically, each gene gets a penalty that is the difference between the maximum verified gene expression and the expression level of the particular gene. That penalty is propagated to a reaction penalty (using GPR rules) and multiplied by the reaction flux. The penalty function is the minimization of the sum of each reaction penalty. GIM_3E also allows the integration of metabolomics data by defining a non-zero minimal flux value for reactions

producing metabolites identified in the metabolomics dataset and constrains reversible reactions to proceed in only one direction, creating a Mixed-Integer Linear Programming (MILP) formulation that is more computationally expensive than the LP one of GIMME and GIMMEp [99, 119].

The advantage of OFR methods is the fact that the optimization for a biological objective guarantees operability, which improves the prediction of growth rate in comparison to other types of methods [100]. Nevertheless, whilst the definition of an objective is straightforward for prokaryotes, that is not the case for eukaryotes, particularly for differentiated cells, that fulfill distinct (sometimes not well defined) purposes for the organism's survival. For example, the main 'goal' of a neuron is not necessarily to grow, but instead to contribute to the overall functionality of the organism [99,100]. Therefore, methods solely focused on solving the inconsistencies between flux predictions and gene expression levels, without requiring the definition of a biological objective, were developed. Those methods are collectively called EDC [101].

ii) Expression Data-Compatible (EDC) methods

Expression Data-Compatible (EDC) methods [101] are algorithms that maximize the matches between flux states (i.e., active/inactive) and omics data states (expressed/not expressed) [99]. One of those, the E-Flux algorithm [120], integrates absolute gene expression values in a continuous approach. Specifically, it constrains the upper bound of reaction fluxes with continuous values that depend on the normalized expression values of associated genes, followed by phenotype simulation. This is a simple method that utilizes an LP formulation [101] and only produces a flux distribution i.e., no context-specific model is output.

Other EDC methods, that apply a more complex MILP formulation, may assign to each reaction a binary variable that assumes a value of one or zero depending on whether a match between simulated flux and omics data states is assumed in a specific solution [99]. One of such methods is the integrated Metabolic Analysis Tool (iMAT). This algorithm deals with absolute data that have been discretized [100]. Hence, the first step in the iMAT pipeline is gene/protein expression discretization, followed by propagation of the gene/protein expression score to a reaction score. Then, reactions are grouped into highly or lowly expressed depending on their reaction scores. iMAT introduces flux constraints that define what is a match in each group of reactions. For highly expressed reactions (R_H), a match is set to happen when the reaction

is active, i.e. when it carries a flux that is higher than a positive threshold ε if the reaction is in the forward direction and lower than -ε if it is in the reverse direction. For lowly expressed (R_L) reactions, a match occurs when the reaction is inactive, i.e. the flux is between -ε and ε. On top of those constrains, iMAT maximizes the sum of the mentioned binary variables (each representing a reaction), that is, it tries to increase the number of matches [121].

Besides producing a context-specific model, iMAT outputs a flux distribution. However, sometimes several flux distributions may provide the same maximum value for the sum of matches. In such situations, iMAT applies an adapted FVA approach. Specifically, the similarity between simulated fluxes and expression scores is assessed for each reaction, firstly when the reaction is forced to be active and then when it is forced to be inactive. If the similarity is higher when the reaction is active, the reaction is included, otherwise, it is excluded. If the similarity is equal in both situations, then the reaction is set as undetermined [99].

Integrative Network Inference for Tissues (INIT) [63] is another algorithm of this class that also accepts gene expression as an input although it was originally designed to integrate proteomics data [100]. INIT deals with absolute values, similarly to iMAT. However, it can integrate them both in a discrete or continuous way, depending on whether the user decides to assign distinct arbitrary reaction values/weights to different ranges of expression or set the reaction weight as a function dependent on the expression level [99]. Nonetheless, in both cases, a negative weight is attributed to reactions associated with low gene/protein expression levels [63]. Unlike iMAT, INIT does not integrate omics data as constraints. Instead, it directly includes them in the objective function, as reaction weights [99]. INIT maximizes the sum of the products of the weight given (by the expression level) to each reaction and the binary variable representing the inclusion (when its value is one) or exclusion (when value is zero) of that reaction in the final reconstruction. Therefore, the algorithm excludes reactions (i.e., gives a zero value to the binary variable) when those reactions carry negative weights (i.e., are lowly expressed), while including reactions (i.e., gives value one to the binary variable) when reactions carry positive weights (i.e., are highly expressed) [63]. Another characteristic of INIT is the ability to integrate metabolomics data in a qualitative manner, since a minimal value for net accumulation is allowed for internal metabolites that are proven to be produced from metabolomics evidence, while assuming a steady-state for the remaining ones [63].

EDC methods can better model cells for which a required metabolic function is unknown [99] and, therefore, guarantee a higher similarity between flux distributions and expression data, in those situations, than OFR methods. Nevertheless, the lack of a required functional objective can sometimes hamper the functionality of models reconstructed with EDC algorithms (e.g. models may not allow cell growth).

This fact led to the rise of a new class of algorithms, Metabolic Task-Derived (MTD), that assures both the abovementioned similarity and operability, but unlike OFR algorithms, does not require the flux distribution to carry a minimal value for a specific metabolic function. Instead, it guarantees that models can perform (one or more) metabolic tasks. In other words, although the context-specific model must be able to perform a metabolic task given the appropriate conditions, not all allowed flux distributions of the model need to have flux through the reactions of the task(s) [101]. The tasks can be, for example, the production or consumption of a metabolite, or even the activation of a whole pathway, under specific conditions [99].

One of the most well-known MTD methods is the task-driven INIT (tINIT) [122]. tINIT has a similar formulation to the abovementioned INIT algorithm, the only difference is that the reconstructed models must be able to perform a set of user-defined metabolic tasks and that reactions are constrained to operate in only one direction, which introduces an additional binary variable [99].

iii) Core Reaction-Required (CRR) methods

The final class, named Core Reaction-Required (CRR) [101], comprises methods that only produce models (no flux distributions are obtained), and that are based on the categorization of reactions into core or non-core, for those of which there are or there are not, respectively, enough evidences that they should be included in the context-specific model. CRR algorithms try to achieve flux consistency, i.e. every reaction of the reconstructed context-specific model has to be active at least in one of the allowed flux distributions, which is equivalent to say that the model cannot have blocked reactions. More importantly, CRR methods assure that as many as possible core reactions are included in the final context-specific model. Although FVA could be applied to identify blocked reactions, CRR methods often apply alternative algorithms that are less computationally intensive to speed up the process [99].

Model Building Algorithm (MBA) [123] is a CRR algorithm in which the

set of core reactions is further split into two subgroups: C_H and C_M. C_H is the group of core reactions that have a high likelihood of being included in the context-specific model because they belong to well-known pathways that have been manually curated, whereas C_M is the subset of core reactions with a moderate likelihood of being part of the model as inclusion is solely supported by high-throughput transcript/proteomics data. In the first step, the algorithm randomly removes one non-core reaction from the generic model and evaluates its consistency. That is, checks whether any reaction of the core reactions becomes blocked due to the retrieval of the non-core reaction. If no reaction in the C_H and only a limited number of reactions in C_M become blocked, then the reaction is removed together with all the blocked reactions, otherwise, the model is kept the same. Note that the maximum number of reactions that are allowed to be blocked in C_M, for the non-core reaction to be removed, is the product of a user-defined threshold (ε) and the number of non-core reactions that become blocked by the removal. Therefore, the value assigned to the threshold can determine if the final model includes a high number of C_M reactions (when the ε is low) or if it is more parsimonious (when ε is high).

The aforementioned process is repeated for every non-core reaction in any arbitrary order. Since inactive reactions are removed in each step without reposition, the order of removal of non-core reactions affects the composition of the context-specific model. To overcome that limitation, MBA is repeated to create 1000 intermediary models and all reactions are ranked based on the frequency by which each reaction is present in those. Finally, a model is built where all reactions of C_H are included and each other reaction is added in the order corresponding to their rank. Each time a reaction is added, the final model is checked for consistency and the process ends when a consistent model is found. This results in a model including all C_H reactions, as many as possible C_M reactions, and a minimum set of non-core reactions necessary to fill the gaps [123].

Unlike MBA, metabolic Context-specificity Assessed by Deterministic Reaction Evaluation (mCADRE) [124] does not generate intermediary models, it directly defines a final model. Furthermore, the selection of core reactions is automated, while the core of high confidence reactions in MBA is based on manual curation. Therefore, mCADRE is a less time-consuming alternative to MBA [99]. In mCADRE, gene expression levels of many samples of the same condition (e.g., a specific cell type) are firstly discretized into expressed or not expressed. Then, a frequency of expression among the samples is computed for each gene. GPR rules are utilized to transpose those values into reaction expression-based evidence values that, together with a user-defined threshold,

are used to split reactions into two groups: core (above the threshold) or non-core (below the threshold). Non-core reactions are ranked accordingly with the frequency of expression, the network connectivity (i.e., the number of 'neighboring' reactions in the metabolic network), and a level of confidence based on the type of evidence that supports the inclusion of the reaction in the generic model (e.g., *in silico*, experimental proof, or evidence for a related organism). Afterwards, each non-core reaction is removed from the generic model in the inverse order of the mentioned ranking. Each time a reaction is removed, the model is checked for consistency. As a rule, a non-core reaction and the reactions that are blocked upon its removal are only discarded when the exclusion of the former does not affect the production of any key-metabolite (a metabolite which production is confirmed by metabolomics data) and no blocked reaction is part of the core. However, blockage of core reactions may be accepted as long as it does not affect a key metabolite if there is evidence that the reaction to be removed is never expressed in that condition (expression values are zero for all samples), that its removal is needed to allow flux through core reactions and the number of non-core reactions versus core reactions that are blocked obeys to a specific ratio [99, 124].

Both MBA and mCADRE utilize a MILP formulation, which is computationally intensive. On the other hand, a different CRR algorithm, known as FASTCORE [125] applies LP, and therefore is several orders of magnitude faster than MBA, producing models in seconds. Unlike MBA and mCADRE, FASTCORE does not establish a specific criterion to define a core. In fact, it provides the freedom to combine any type of omics data or bibliographic evidence to decide which reactions belong to the core. This increases the confidence in the definition of the core, as missing information in one data type can be complemented by the other [99, 125]. Once the core is defined, the algorithm tries to solve two LP problems. The first is to maximize the number of irreversible core reactions that have a flux value above a predefined small positive constant ε, i.e., increase the number of core reactions that are active (not blocked). Those active reactions (that carry flux) after the first optimization are then used as constraints for the second LP problem, where their fluxes are set as higher than ε. Those constraints assure that whatever is the solution for the second LP problem, it keeps those core reactions active. The second LP problem tries to reduce the number of active non-core reactions that are included in the context-specific model. It does that by minimizing the sum of absolute values of the non-core reaction fluxes (i.e. minimize the L1 norm of the flux vector of the non-core reactions). The non-core reactions that are active after the second optimization

and core reactions that are active in the solution for the first LP problem are included in the context-specific model. The process is then iterated using the remaining core and non-core reactions (i.e. those not included in the context model) until a consistent model that contains all core and a minimal number of non-core reactions is found [125]. To deal with reversible reactions the algorithm tests both forward and reversible directions, by changing the sign of the corresponding column in the stoichiometric matrix [99].

There are many other algorithms for omics data integration, besides the ones mentioned here. Furthermore, although the choice of the reconstruction method can significantly affect the structure of the context-specific model or flux distribution, none of those can be considered better than the other, as their performance varies depending on the specific case that is being modeled [100]. A procedure suggested to overcome that limitation is to integrate the same omics data sample with different algorithms and subsequently evaluate the accuracy of the corresponding reconstructed models with experimental data. The best-performing algorithm is then applied to reconstruct models for other samples obtained with the same methods and under the same experimental conditions, but for which no validation data is available [126]. The strategies often used to validate context-specific models and related flux distributions are comparisons of the activated metabolic pathways and metabolic tasks simulated *in silico* with what is reported in the literature for that context (e.g. cell type) [127], of flux values with experimentally obtained fluxomics data [126], or of simulated essential genes with lethal genes identified in 'survival' experiments using gene knockout screens (like CRISPR or RNAi screens) [92, 128].

1.2.7 Context-specific models of human cancer

Context-specific models have been built to reproduce *in silico* the metabolism of human cells like adipocytes [65], myocytes [66], hepatocytes [64], endothelial cells [129], or macrophages [130]. Furthermore, models have been reconstructed for cells of both healthy and diseased individuals. A variety of diseases have been modeled with this approach, including Non-Alcoholic Fat Liver Disease (NAFLD) [64], diabetes [66], obesity [65], sepsis [129], viral infections [130], and several types of cancers [62, 63]. Overall, these models allowed to identify metabolic traits, spot potential drug targets, and predict the metabolic response to drug treatments [48, 131, 132]. The most relevant context-specific metabolic models reconstructed until now to model human cells, with a particular focus on cancer models, are summarized in Table 1.3.

The main advantage of modeling cancer cells is that it is safer to assume their biological purpose is to increase biomass production, as those cells are known to have a great potential to proliferate [131]. This allows using phenotype prediction techniques like FBA and context-specific reconstruction algorithms based on a biological objective. Nevertheless, when trying to assert the metabolic differences between cancer cells and metabolic tissues there is a need to build models for reference tissues as well (i.e., for healthy cells of corresponding tissues), and it is difficult to define a biological objective for differentiated human cells, though the development of integration methods that utilize tissue-specific metabolic tasks has mitigated the problem [131].

Furthermore, the assumption of growth cannot apply to cancer cells that are in a quiescent state, which is the case of some CSCs. Another limitation for the accurate reconstruction of cancer models and tissue-specific models for healthy cells is the lack of knowledge about the composition of the biomass reaction in the specific context, which is distinct from the generalized one usually applied [126]. For cancer, this is particularly problematic, since they more easily adapt their metabolic uptake to survive in environments with scarce nutrients than normal cells [133]. The niche is also known to affect cancer cells and, therefore, the lack of knowledge on how the surrounding environment influences the metabolism can result in inaccurate predictions. Although it is easier to model cancer cell metabolism in *in vitro* settings, as the cell culture media formulations are widely available, the exact rate at which nutrients are consumed is usually unknown (unless explicitly measured), which hampers the definition of accurate flux bounds for external model reactions. Also, human cells culture media often have components like Fetal Bovine Serum (FBS), the composition of which is unknown, making it impossible to identify exactly which exchange reactions should be active in the model. This is especially relevant as it has been proven that constraints in exchange fluxes significantly improve the ability of cancer models for phenotype prediction [126].

Table 1.3: Studies using context specific metabolic models

Date	Description	Generic model	Method	Reference
2010	Model (HepatoNet1) predicted metabolic states of hepatocytes in different physiological conditions, including detoxification of ammonia	Recon1	Manual	[62]
2011	Application of enzyme capacity constraints to generic model predicted the Warburg effect, the preference for glutamine uptake of cancer cells, and metabolic phases observed during cancer progression	Recon1	Inclusion of enzyme capacity constrains	[134]
2012	Models for 69 types of healthy cells and 16 types of cancer cells allowed to identify cancer-specific metabolic features and potential targets	HMR	INIT	[63]
2013	Adipocyte model predicted metabolic differences between obese and lean subjects and suggested therapeutic targets to treat obesity	HMR	Manual	[65]
2014	Human hepatocyte model predicted markers of different stages of NAFLD and identified potential therapeutic targets for the most aggressive state	HMR2	Manual	[64]
2014	Comparative analysis of a model of Hepatocellular carcinoma (HCC) summarizing the data of different patients, 6 personalized models of HCC and models of 83 types of healthy cells predicted antimetabolites potentially effective against HCC	HMR2	tINIT	[122]
2019	Integration of single-cell RNA-seq data of patients with breast and lung cancer, together with bulk-metabolomics data into a generic model to create single-cell flux distributions. It allowed to characterize the metabolic heterogeneity within tumor and identify metabolic interactions between cancer cell populations	HMRcore with additional pathways and GPR rules	scFBA	[135]

1.2.8 GSMMs enhanced with Enzymatic Constraints using Kinetic and Omics data (GECKOs)

Even with constraints on exometabolite uptake rates, traditional constraint-based models may provide inaccurate flux distributions, because they are based on the assumption that reaction rates are solely limited by substrate availability, while it is known that enzyme catalytic activity and concentration also influence the reaction rate. In such a model, the reaction flux will increase infinitely as the substrate concentration increases [136]. However, in reality, the reaction velocity just increases with substrate concentration until the catalytic sites of all enzyme molecules are occupied with the substrate. Beyond that point, even if the substrate concentration increases, the velocity of the reaction does not increase, that is, the reaction rate is limited by the enzyme levels [80].

One of the first strategies developed to depict the enzymatic-induced metabolic flux restriction *in silico* was FBA with Molecular Crowding (FBAwMC), which consists of adding constraints on the total cell volume occupied by all metabolic enzymes. However, this approach does not allow direct integration of proteomics data. Other strategies, like Metabolism and gene Expression (ME) models, integrate metabolic models with the cellular machinery necessary to synthesize proteins, from gene transcription to protein translation and maturation [136]. Nevertheless, although these models have been shown to make accurate predictions, they are very complex and entail detailed parameters that are not usually available for most organisms, like those describing how proteins fold and maturate [137].

In 2017, a new approach that allowed direct integration of enzyme kinetics and proteomics data (i.e. enzyme abundance) was developed, known as the Genome Scale Metabolic Models enhanced with Enzymatic Constraints using Kinetic and Omics data (GECKO). GECKO models expand the S matrix (Figure 1.5) of constrained-based models (where columns and rows represent, respectively, reactions' stoichiometry and metabolites' mass balance) by adding new rows that depict enzymes and new columns that define enzyme usage reactions. Protein levels are introduced as upper bounds for each enzyme usage reaction, whereas kinetic information, in the form of the inverse of the turnover number (k_{cat}) of each enzyme-reaction pair is inserted as stoichiometric coefficients [137].

The mathematical formulation of GECKO models is based on the following:

$$v = k.[ES] \tag{1.5}$$

which is the rate law (the rate at which the enzyme forms the product) for the first-ordered reaction:

$$\mathrm{ES} \xrightarrow{\mathrm{k}} \mathrm{E} + \mathrm{P} \tag{1.6}$$

where v is the reaction rate and $[ES]$ is the concentration of the enzyme-substrate complex, while E and P are the free enzyme and product respectively. v is limited by a maximal reaction velocity. Therefore:

$$v_j \leq V_{max} \Leftrightarrow v_j \leq k_{cat}^{ij}.[E_i] \tag{1.7}$$

where $[E_i]$ is the concentration of the total enzyme E_i (free and substrate-bounded enzyme), v_j is the rate of the reaction R_j (in mmol.gDW^{-1}.h^{-1}), and k_{cat} is the turnover number (in h^{-1}) of enzyme E_i in reaction R_j. Note that it is possible to assume that $[ES] = [E_i]$ when the reaction is at its maximal velocity because at that rate the active sites of all enzyme molecules are occupied with the substrate. Although enzymes are not consumed in reactions, they are used for a short period to catalyze them. To represent that enzyme usage in the following generic reaction R_j:

$$R_j : \mathrm{A} + \mathrm{B} \longrightarrow \mathrm{C} + \mathrm{D} \tag{1.8}$$

it is possible to introduce the enzyme as a pseudo-metabolite which does not affect the mass balance of the reaction:

$$R_j : \mathrm{n_{ij}} \cdot \mathrm{E}_i + \mathrm{A} + \mathrm{B} \xrightarrow{\mathrm{v}_j} \mathrm{C} + \mathrm{D} \tag{1.9}$$

To keep the mass balance of the enzyme after introducing it in the reactions it catalyzes, an overall enzyme usage pseudo-reaction EU_i that supplies enzyme is introduced:

$$EU_i : \quad \xrightarrow{\mathrm{e}_i} \mathrm{E}_i \tag{1.10}$$

In addition, the flux of that reaction, e_i (in mmol/gDW), can be constrained by the enzyme's concentration:

$$0 \leq e_i \leq [E_i] \tag{1.11}$$

Note that the units of flux are not mmol.gDW^{-1}.h^{-1} in this case, because this is not a real reaction, but rather a pseudo-reaction built just to represent enzyme usage. Combining 1.9 with 1.10 and assuming the steady state, the mass balance for enzyme E_i is:

$$-n_{ij}.v_j + e_i = 0 \tag{1.12}$$

Joining equations 1.11 and 1.12, the following is obtained:

$$v_j = \frac{e_i}{n_{ij}} \leq \frac{1}{n_{ij}}.[E_i] \tag{1.13}$$

By comparing equations 1.13 and 1.7, it is possible to conclude that:

$$n_{ij} = \frac{1}{k_{cat}^{ij}} \tag{1.14}$$

This means that in GECKO models, the stoichiometric coefficient of an enzyme in a reaction catalyzed by that enzyme must be the inverse of the k_{cat} (in h). As the units of this stoichiometric coefficient are hour, when multiplying by the reaction flux, the metabolic flux (mmol.gDW^{-1}.h^{-1}) is converted/corrected to the units of the flux of the enzyme usage reaction (mmol/gDW) [137].

Note that the metabolic reactions must be irreversible for this formulation to be applied. So, all reversible reactions in traditional constrained-based models are split into two reactions in GECKO models. One in the forward and the other in the reverse direction, both with the same enzyme and possibly different k_{cat} values (as k_{cat} depends on the enzyme affinity for the substrate) [137].

Other specific relationships between enzymes and reactions were also considered:

i) when a reaction has *isozymes*, the reaction is split into as many reactions as available isozymes. Each resulting reaction is catalyzed by one isozyme with a corresponding specific k_{cat} value. Besides, an arm reaction is introduced to keep the original upper bound in place. The arm reaction produces a pseudo-metabolite from the substrates, which is then used as a substrate by the isozymes-catalyzed reactions. This way, the flux of all those isozyme-catalyzed reactions can be constrained at one point, the upper bound of the arm reaction, and the original upper bound can be kept. The mass balance of each of these isozymes corresponds to one row that is added to the original stoichiometric coefficients matrix, S.

ii) for reactions catalyzed by *promiscuous enzymes* no specific action (for e.g. to split) is needed. The same enzyme is used as pseudo-substrate for each reaction, possibly with distinct k_{cat} values (as the substrate is different). Hence, there will be more than one non-zero coefficient in the row representing the enzyme's mass-balance in matrix S. Moreover, only one enzyme usage pseudo-reaction exists per enzyme, so reactions of promiscuous enzymes share the same amount of available enzyme.

iii) for an *enzyme complex* associated reaction, all proteins that are part of the complex are used as the reaction pseudo-metabolites, which implies that there is more than one nonzero coefficient in the column representing the reaction's stoichiometry in matrix S. The k_{cat} value is the same for all proteins of the complex and the stoichiometric coefficient of each protein is the product of the inverse of the k_{cat} and protein's stoichiometric coefficient in the enzyme complex [137].

When proteomics data is not available, instead of constraining each enzyme usage reaction with the enzyme concentration, it is possible to limit the total amount of enzyme and let the GECKO model choose which amount of each protein to use. The steps to implement this approach are:

i) insert another pseudo-metabolite that represents all enzymes in the model, named E_{pool}

ii) introduce a usage pseudo-reaction for E_{pool}:

$$ER_{pool}: \quad \xrightarrow{\mathrm{e_{pool}}} \mathrm{E_{pool}} \tag{1.15}$$

where flux, e_{pool}, units are in g/gDW.

iii) limit that usage reaction by total protein content as:

$$e_{pool} \leq P_{total}.f.\sigma \tag{1.16}$$

where the P_{total} is the mass of protein per cell mass (in g/gDW), f is the mass fraction of enzymes that are accounted for in the model out of all proteins, and σ is a fitted parameter that represents the *in vivo* average saturation of all enzymes.

iv) replace enzyme usage pseudo-reactions by pseudo-reactions that draw from the enzyme pool towards each corresponding enzyme. A representation of one of those reactions is:

$$ER_i: \mathrm{MW}_i \cdot \mathrm{E_{pool}} \xrightarrow{\mathrm{e}_i} \mathrm{E}_i \tag{1.17}$$

Since the flux e_i of this reaction is in mmol/gDW and the flux of E_{pool} usage reaction, e_{pool}, is given in g/gDW, to use the same units when doing the mass balance, the stoichiometric coefficient of the enzyme pool (E_{pool}) has to be the molecular weight (MW_i) in g/mmol of the enzyme [137].

Note that, when doing the mass balance for E_{pool} (combining 1.15 and 1.17):

$$e_{pool} - \sum_{i}^{P} MW_i.e_i = 0 \tag{1.18}$$

By matching equations 1.16 and 1.18, it is possible to obtain:

$$\sum_{i}^{P} MW_i.e_i \leq \sigma.f.P_{total} \tag{1.19}$$

This is equivalent to the formulation of a previous approach based on FBAwMC that accounted for enzyme limitation when individual concentration was unknown [137].

GECKO models have been able to mimic different physiological situations. One of those is overflow metabolism, i.e. the switch from exclusively OXPHOS metabolism to a mixture of it with fermentation when the growth rate surpasses a certain value. At high growth rates, there is a large flux of substrate and the amount of enzyme needed to bind that substrate becomes the limiting factor, i.e. the protein mass concentration limit is reached. Although respiratory enzymes are more energetically efficient (each enzyme molecule produces more ATP from same amount of substrate) than glycolytic enzymes, they are heavier and therefore fewer molecules of the former than of the latter are sufficient to reach the cells' limit of protein mass. Therefore, at high growth rates, the cell favors the use of more mass-efficient glycolytic enzymes molecules that in sufficient quantity can overall produce more energy than a few of the heavier respiratory enzyme molecules. Such biological behavior can only be simulated in models that integrate protein levels, like GECKOs. Another situation where these models showed more accurate predictions than traditional GSMMs is when estimating maximum growth under different carbon sources [137], and they have been particularly useful when the real flux values for uptake of exometabolites are unknown or can not be estimated from medium composition [75].

GECKOs have been used to model different organisms [137–139] and, recently, human cells [75]. Their main disadvantage is the need for kinetic data, that may be difficult to obtain for certain organisms or environmental conditions.

Overall, the simulation accuracy of context-specific metabolic models for a complex organism like the human one is still very limited. However, there is hope that integration of signaling and gene regulatory networks with metabolic

models will improve the functionality of those models as it would account for the influence of important regulators that are excluded from traditional metabolic models [48,140,141]. Also, it is expected that the future development of sequencing techniques and fast metabolic model reconstruction strategies will promote the creation of patient-specific models useful for personalized medicine [48,132].

Chapter 2

Reconstruction of Tissue-Specific Genome-Scale Metabolic Models for Human Cancer Stem Cells

2.1 Introduction

Like healthy stem cells, Cancer Stem Cells (CSCs) can differentiate into different cell types and hold self-renewal ability (i.e. give origin to new daughter CSCs), although they may keep themselves in a quiescent state (where they do not divide) for a long time [142]. These stem cell properties allow CSCs to drive cancer progression, and metastasis [143], while simultaneously escaping conventional radio/chemotherapy treatments, which are aimed at actively dividing differentiated cancer cells [142]. Hence, besides promoting cancer aggressiveness, CSCs are responsible for tumor recurrence after treatment [143]. Furthermore, according to the CSC hypothesis, CSCs carry a tumorigenesis potential, and are thought to give rise to most cell types within a tumor [5]. Therefore, there is an undeniable need to develop therapeutic strategies that specifically target CSCs to eradicate cancer.

Mutations can lead to cancer (stem) cell formation and indirectly impact metabolism, but metabolism may as well promote both cancer and stem cell phenotypes [5, 144]. Cancer and stem cells undergo metabolic reprogramming that allows them to obtain the basic metabolites and energy that support anabolic processes required for cell growth, like the synthesis of nucleic acids, proteins, and cell membrane triacylglycerols [5, 144]. Additionally, metabolites also serve as substrates and co-factors of enzymes responsible for DNA and histone modifications, contributing to epigenetic regulation of events, such as proliferation, differentiation and cell survival, in both cancer and stem cells [145, 146].

With the development of next-generation sequencing technologies, more

omics data is becoming available, allowing for the construction of computational models at the genome scale [48]. Genome-Scale Metabolic Models (GSMMs) are mathematical representations of metabolic reactions' stoichiometry, which include an additional mapping between genes, proteins, and reactions [48]. Generic GSMMs represent all metabolic reactions that may happen inside any cell of an organism, allowing for the study of the interplay of those reactions from a quantitative point of view, enabling counterintuitive and insightful *in silico* predictions of cellular metabolic behavior [48, 49, 83]. The most up to date generic GSMM for human cells is *Human1* [75], which integrates knowledge from several previous models, addressing known issues they present, such as incorrect reaction reversibility and existence of unnecessary reactions. It is a comprehensive model that covers 13,417 reactions, 4,164 unique metabolites and 3,625 genes [75].

Since the genes expressed (or their expression levels) change depending on the type of tissue, cells of different tissues often show distinct metabolic profiles. Therefore, a generic model *per se*, comprising all potential metabolic capabilities of a human cell, often does not allow to make accurate and precise predictions [147]. To address this challenge, a generic GSMM can serve as a basis for the reconstruction of tissue-specific GSMMs, through the integration of cell-type specific omics data, usually transcriptomics, to model the metabolic characteristics of specific tissues [147]. Such context-specific models were built for a variety of human tissues in the past, like myocytes [66], adipocytes [65], hepatocytes [64], endothelial cells [129] or macrophages [130], both to model normal cells and cells affected by morbidities [63]. In fact, different tissue-specific models have been created for a range of human diseases, such as obesity [65], diabetes [66], sepsis [129], NAFLD [64], viral infections [130] and different types of cancers [62, 63]. Those models were fundamental to detect metabolic traits [62, 63], spot potential drug targets [148], and predict the metabolic response to drug treatments [132].

Although tissue-specific GSMMs have been created for differentiated/bulk Cancer Cells (CCs) [149, 150] and normal stem cells [151] in the past, to the best of our knowledge, there was just one study that has previously reported the reconstruction of GSMMs for CSCs and for only one tissue: the liver [152]. Furthermore, the tissue-specific models of that study were based on a previous generic model of human cells, the *Recon2*. In this study, we build GSMMs for human CSCs and differentiated CCs of ten different tissues, based on the most up-to-date *Human1* generic metabolic model [75]. Flux simulations with these models allowed us to identify metabolic traits specific of CSCs (in comparison

with their differentiated counterparts) that are common to most types of tissues. Additionally, further analyses enabled the detection of essential genes and metabolites, Transcription Factors (TFs), and miRNAs that potentially revert the CSC phenotype and could be good candidates for experimental validation.

2.2 Results

In this work, metabolic models were reconstructed for CSCs and CCs of ten different tissues (AML, glioblastoma, lung, prostate, liver, ovary, breast, kidney, head and neck, and pancreatic cancers), by integrating gene expression data from ten individual datasets with the generic genome-scale metabolic model for human cells, the *Human1* [75]. Each raw dataset (either RNA-seq or microarray data) contained information pertaining CSCs and the corresponding differentiated CCs of one tissue, and was processed and analyzed independently of others. Only datasets with more than one donor or cell line and from peer-reviewed studies were selected. The inclusion of studies followed the published criteria on CSC phenotype definition: mainly the presence of CSC surface markers (Table C.1), ability to form oncospheres, and tumorigenicity in mice (Table C.2 – column 'Criteria for study inclusion'). Raw transcriptomics data was normalized with pipelines appropriate to each data type (details in Materials and Methods), and the best parameters for gene expression integration with GSMMs were tested. A flow diagram of the overall methodology of the study is shown in Figure A.7. Details of the pre-processing steps are depicted in Figure A.8 and A.9.

2.2.1 Best strategies for transcriptomics data integration into CSC metabolic models

Before reconstructing tissue-specific metabolic models for CSCs, it was necessary to select the best strategies/parameters for transcriptomics data integration into the generic metabolic model. First, to reduce the weight that highly expressed genes have in comparison with lowly expressed genes, each dataset of normalized gene expression data has gone through min-max scaling. Such normalization was essential to observe the differences between experimental conditions in hierarchical clustering (Figure A.1). Afterwards, different thresholding approaches were assessed. A global threshold strategy implies that a unique threshold value is applied to all genes, and genes with expression above that threshold are regarded as active (i.e., *on*), while remaining genes are deemed

inactive (i.e., *off*). In contrast, when using a local threshold strategy, a threshold value is computed per gene. The global approach predicts as active fewer cell-type-specific genes than the local strategy, and, therefore, metabolic networks resulting from the local approach present higher tissue specificity [153]. On the other hand, the local approach may allow for highly expressed genes with ubiquitous expression across distinct cell types to be predicted as inactive in some samples, if the expression is slightly lower than the local threshold in that sample [153]. Hence, we chose to test one global threshold (*global* strategy), a combination of one global threshold and one local threshold (*local 1* strategy), and a combination of two global thresholds and a local threshold (*local 2* strategy). Variations of *local 1* and *local 2*, the *local 1B* and *local 2B* strategies, where genes defined as *on*/*off* by global thresholds were given higher/lower gene scores than those defined by the local threshold, were also assessed (details in Materials and Methods). This procedure is to account for situations where a gene A, although more expressed than a gene B, has a lower score because the expression value is closer to a global threshold than to a local threshold.

Apart from gene thresholds, there was an evaluation of two gene to reaction scores conversion strategies: the *min-max* and *min-sum* as well as the decision to use the threshold calculation either for all genes or just the ones coding for metabolic enzymes (details provided in Materials and Methods). Results show that no integration strategy for transcriptomics data is consistently better than the others across all datasets, when using the proportion of simulations where the average Euclidean distance was smaller than the distance observed in sample groups (Figure A.2, for details of the assessment metric see Materials and Methods). Also, no strategy creates more variability in the reaction scores than the others (Figure A.3).

We selected the three-parameter combinations resulting in the lowest values for the above-mentioned metric, in one cell line of an RNA-seq study (Huh7) and another cell line of a microarray study (HCC1937) where there is gene lethality information available. Models were reconstructed with FASTCORE and INIT algorithms. The Mathews Correlation Coeficient (MCC) score was calculated between simulated and experimentally verified lethal genes, for each combination of parameters and each reconstruction algorithm in both cell lines.

The strategies rendering the best MCC scores for the cell line of the microarray study and the cell line of the RNA-seq study are shown in Table 2.1. Although the best-observed MCC scores were relatively low ($\simeq 0.3$), they are similar to the values previously reported for reconstructed models of human cancer cell lines [75].

Table 2.1: Best strategies for transcriptomics data integration.

Study	***T.S.***	***L.G.***	***U.G.***	***L.T.***	***G.C.***	***R.A.***	***Genes***	***Score***
microarray	local2	10th	90th	10th	min-max	FASTCORE	metabolic	0.37
RNA-seq	local2	25th	90th	10th	min-sum	INIT	metabolic	0.38

T.S.: Threshold Strategy; ***L.G.***: Lower Global Threshold; ***U.G.***: Upper Global Threshold; ***L.T.***: Local Threshold; ***G.C.***: Gene to reaction Score Conversion strategy; ***R.A.***: Reconstruction Algorithm; ***Genes***: Type of genes used in gene threshold calculation; ***Score***: MCC score

2.2.2 Flux simulation predicts metabolic pathways with higher flux in CSCs than CCs

The best transcriptomics data integration strategies mentioned above were used to reconstruct models for each donor/cell line in each study/tissue. Successfully reconstructed models were able to perform 57 essential metabolic tasks and grow in Ham's medium. Only donors/cell lines with successfully reconstructed models in all cell types of a study (i.e., matched donors of CCs and CSCs of each dataset) were used for further downstream analyses. In terms of model composition, the total number of active reactions and the distribution of active reactions across metabolic subsystems/pathways varied depending on the tissue (Figure A.4-A,C). Also, the overall number of reactions, genes, and metabolites was similar across the two cell types (Figure A.4-B). A parsimonious enzyme usage FBA (pFBA) demonstrated that CSCs of five tissues have significantly more flux than CCs in reactions of the *Pentose phosphate pathway* (liver, pancreas, head and neck, ovary, kidney), *Pyrimidine metabolism* (liver, AML, lung, head and neck, kidney), and *Oxidative phosphorylation* (liver, glioblastoma, breast, lung, kidney) (Figure 2.1). CSCs of four tissues are enriched in *Purine metabolism* (liver, AML, lung, prostate), *Glycolysis/Gluconeogenesis* (liver, prostate, ovary, kidney), *Tricarboxylic Acid Cycle and glyoxylate/dicarboxylate metabolism* (liver, pancreas, glioblastoma, breast) and *Folate metabolism* (liver, pancreas, glioblastoma, lung) pathways, whereas CSCs of three tissues are enriched in *Valine, leucine and isoleucine metabolism* (liver, head and neck, kidney), *Alanine, aspartate and glutamate metabolism* (liver, pancreas, glioblastoma), and *Nucleotide metabolism* (pancreas, AML, lung) pathways.

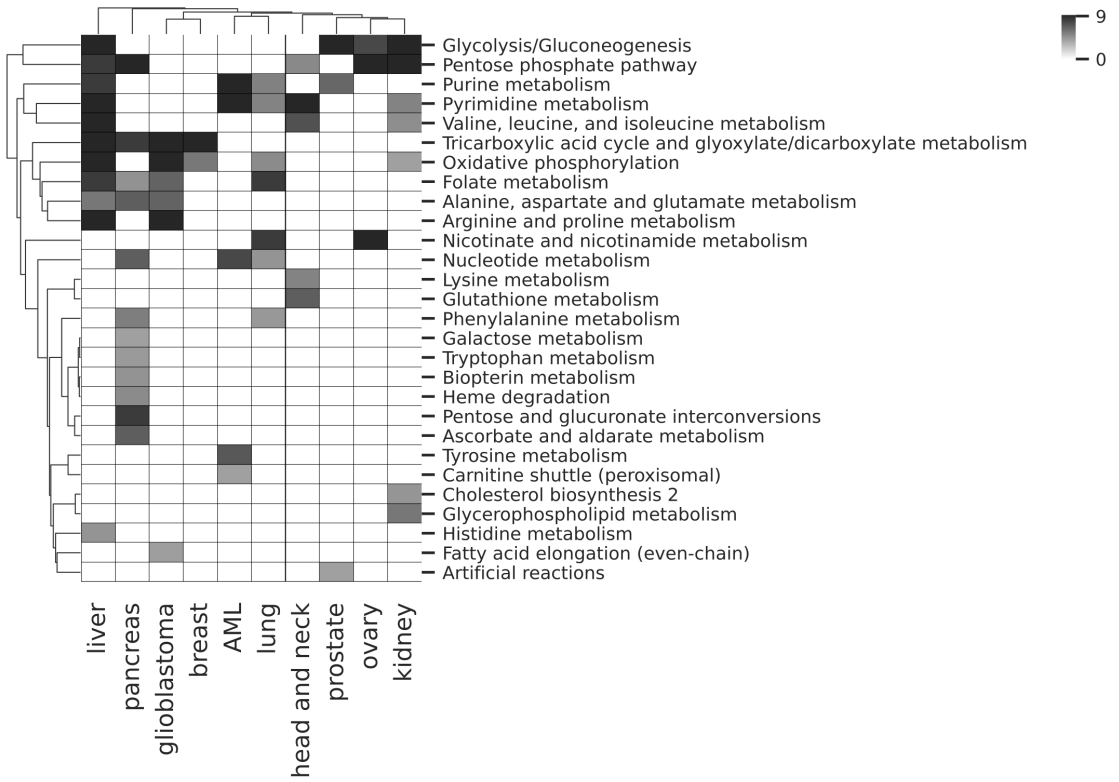

Figure 2.1: Metabolic pathways activated in models of different cell types. For each tissue, reactions were ranked in a list by the difference between CSCs and CCs median absolute flux values. A ranked reaction-set enrichment analysis test was applied to identify the reaction subsystems over-represented (p-value < 0.05) at the top (with more flux in CSCs than in CCs) of the list. Only over-represented subsystems have color and values are the $-\log(p$-value). Dendrograms show Euclidian distance with the average agglomeration method.

2.2.3 Prediction of essential genes, metabolites and antimetabolites

In the present work, we identified essential genes and metabolites by simulating their individual removal and subsequent selection of those that decrease the flux of the biomass reaction. The only gene predicted as essential in models of CSCs, but not in models of CCs of five tissues (pancreas, prostate, lung, liver, ovary) is CRAT (Figure 2.2-A), which is also predicted as essential for both models of CSCs and CCs in breast, AML, and glioblastoma (Figure A.5-A). The GSTM1 gene was predicted as essential in CSCs, but not in CCs, of four tissues (pancreas, glioblastoma, liver, head and neck), while ELOVL1 was predicted as essential in CSCs only of four tissues (pancreas, lung, liver, ovary) and in both CSCs and CCs of two tissues (breast, AML) (Figure 2.2-A and

A.5-A). Genes predicted as essential in both models of CSCs and CCs in all tissues are FECH and PPOX (Figure A.5-A).

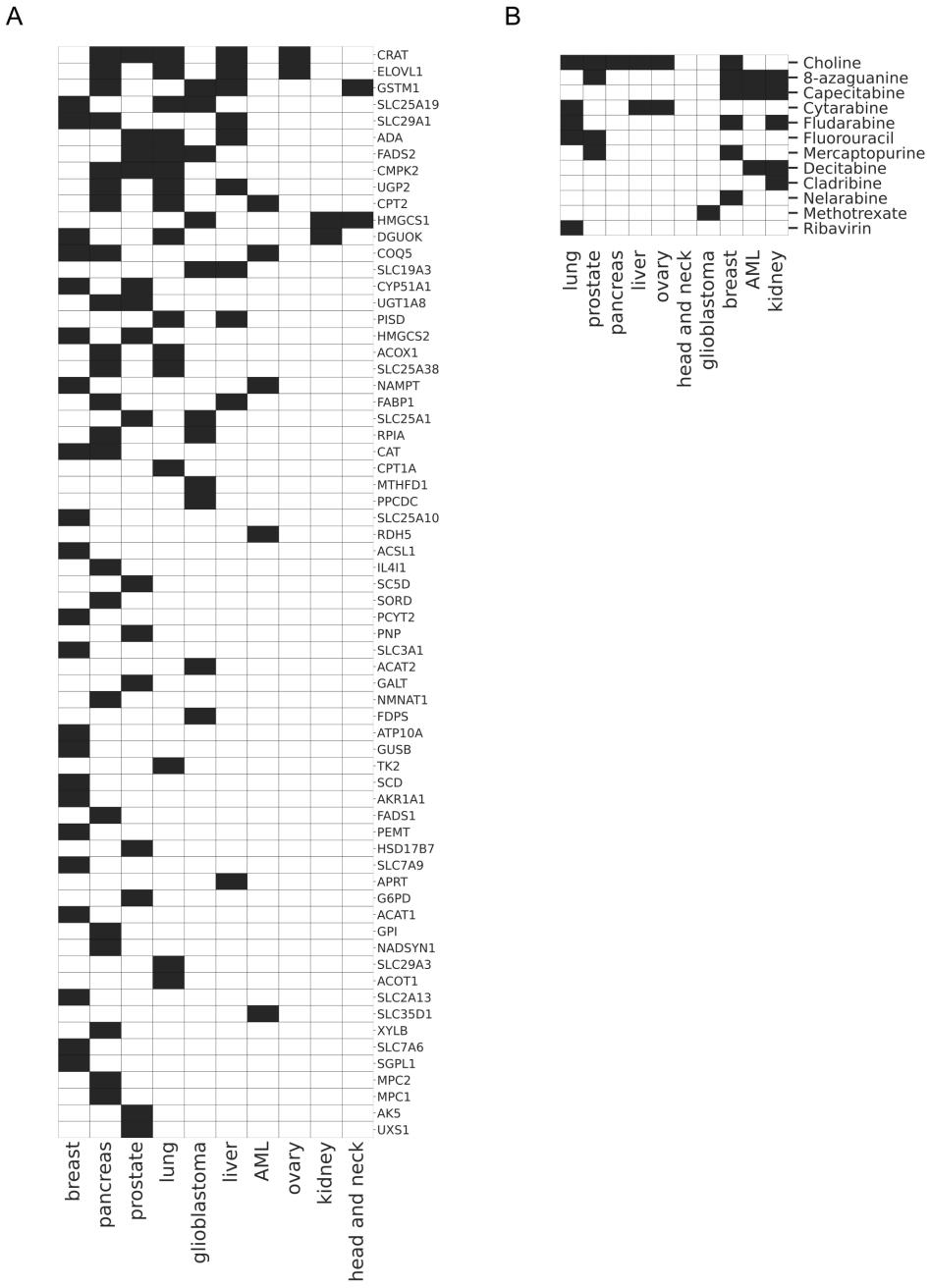

Figure 2.2: Essential genes and antimetabolites predicted only in CSCs. **A**: essential genes predicted for models of CSCs, but not for models of CCs. **B**: antimetabolites predicted to block the effect of the EMs that are specific for CSCs.

The Essential Metabolite (EM) specifically found for CSCs in most tissues was malonyl-carnitin, which was found as essential only for CSCs of four tissues

(liver, ovary, lung, and pancreas – see Table C.3), and for both CSCs and CCs of two tissues (AML and breast) (Table C.4). The metabolites that were essential for all tissues in both CSCs and CCs are nicotinamide D-ribonucleotide and nicotinamide (Table C.4). Both are part of nicotinamide metabolism, which is necessary for NAD synthesis [114], an EM of all cells.

An antimetabolite is a structural analog of a natural metabolite that can interact with the same targets as the metabolite, but that is not functional. It works as a competitive inhibitor of the metabolite, abrogating its effect. Here, we identified antimetabolites that might prevent the effect of EMs detected for CCs and/or CSCs. *Choline* is one of those suggested in this analysis because it is considered an antimetabolite by DrugBank (Figure 2.2-B). However, *choline* is also a metabolite in these metabolic models and it is an important metabolite for any cell, as it is necessary for the synthesis of cellular membrane glycerophospholipids [75]. Another suggested antimetabolite is *8-azaguanine* that is predicted to block the effect of EMs specific for CSCs in four tissues (prostate, breast, AML, kidney – see Figure 2.2-B) and of those common for both CSCs and CCs in four tissues (liver, prostate, breast, pancreas – see Figure A.5-B). Other anti-neoplastic antimetabolites suggested in this analysis are *capecitabine*, *fludarabine*, *cytarabine*, *mercaptopurine*, *decitabine*, *fluorouracil*, *ribavirin*, *cladribine*, *nelarabine* and *methotrexate* (see Figure 2.2-B).

2.2.4 Transcription factors and miRNAs that may potentially affect cell survival

To identify Transcription Factors (TFs) that may potentially cause CSC death, we first identified genes associated to reactions where the flux positively correlates with biomass in Flux Variability Analysis (FVA). Then, we found TFs that upon knockout or knockdown decrease the expression of those genes in databases. FLI1 is one of the suggested TFs that significantly targets genes associated with biomass only in CSCs (not CCs) of two tissues (head and neck, pancreas), although it also targets some genes in CSCs of other cancers (Figure 2.3-A). Results suggest HNF1A is a good target for a knockout in cancer of four tissues (ovary, liver, kidney, head and neck), although it only seems to significantly affect genes specific of CCs (not CSCs) (Figure A.6-A). Similarly, FOXA1 significantly interferes with genes correlated with biomass only in CCs of three tissues (liver, kidney, head and neck) (Figure A.6-A).

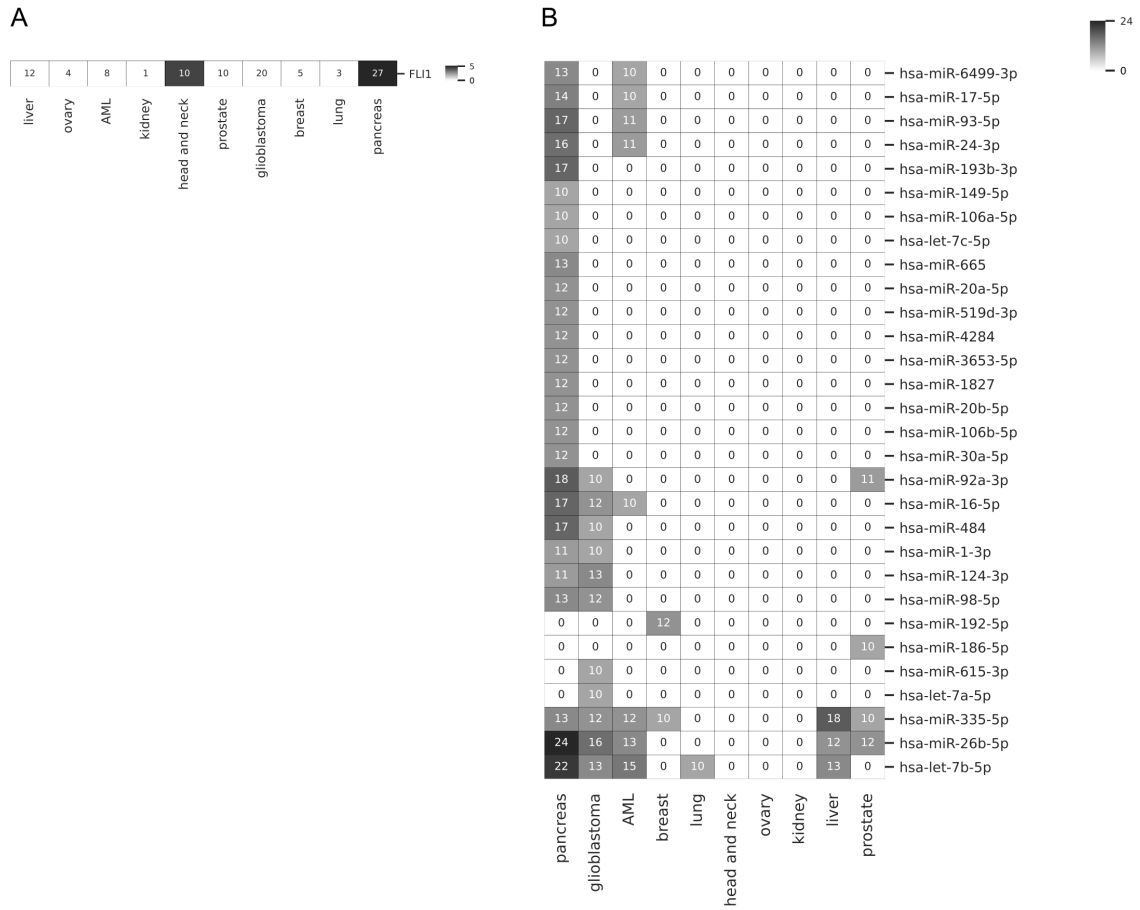

Figure 2.3: TFs and miRs with potential to target genes that are correlated with biomass only in CSCs. **A**: TFs that when knocked-out decrease expression of genes that are directly correlated with biomass in models of CSCs but not in models of CCs. Only TFs significantly targeting those genes have color (adjusted p-value < 0.05) and color intensity is $-\log$(adjusted p-value). Each number counts the genes associated with biomass targeted by the TF. **B**: miRs that target genes that are directly correlated with biomass in models of CSCs but not in models of CCs. miRs targeting the top 10 numbers of targets in each tissue (maybe more than 10 miRs per tissue if different miRs have the same number of targets) are shown. Color intensity and numbers reflect the number of genes associated with biomass targeted by the miR.

As most miRNAs (miRs) work as negative gene expression regulators, we overlapped genes directly correlated with biomass with targets in a miRNA-target gene database, to identify miRs that can potentially decrease the expression of those genes and, therefore, prevent cancer phenotypes. *hsa-miR-335-5p*, *hsa-miR-26b-5p*, and *hsa-let-7b-5p* target genes directly associated with biomass only in CSCs in six (pancreas, glioblastoma, AML, breast, liver, and prostate),

five (pancreas, glioblastoma, AML, liver, and prostate) and another five tissues (pancreas, glioblastoma, AML, lung, liver), respectively (Figure 2.3-B). However, none of these three miRs specifically targets CSCs, as they also target directly associated biomass genes of both CSCs and CCs in all tissues (Figure A.6-B).

2.3 Discussion

To the best of our knowledge, this is the first time that genome-scale metabolic models are reconstructed for both CSCs and CCs of ten different tissues. Such models were built before for liver CSCs, but were based on a generic metabolic model of human cells that is no longer up to date, the *HMR 2.0* [152]. Conversely, the metabolic models in the present work are built upon the most recent generic metabolic model for human cells, the *Human1* [75]. Moreover, different transcriptome data integration strategies and reconstruction algorithms were tested in this work before choosing the best one to apply to all models, which has advantages over arbitrarily choosing a gene threshold or a reconstruction algorithm.

The metabolic pathways more prevalent in reconstructed models of CSCs, in comparison with those of CCs, agree with reported experimental evidence. For example, intensive use of glycolysis was already described for some CSCs in comparison with CCs [154, 155], which is in line with a higher flux through the *Glycolysis/Gluconeogenesis* pathway observed in this study. It is thought that glycolytic CSCs have the advantage of obtaining energy under low oxygen levels, while simultaneously avoiding death through the reduction in oxidative phosphorylation-induced ROS-production, which is especially important for cells that divide frequently, like the case of CSCs. Nevertheless, *Oxidative phosphorylation*, as well as mitochondrial biogenesis [156], have been described as metabolic traits of CSCs of some tissues, unlike the differentiated counterparts that preferentially use glycolysis [157–161], which is also observed in this analysis. The preference of CSCs for this pathway has been attributed to the need to efficiently obtain energy in glucose-deprived microenvironments [159], or in niches rich in nutrients, that can feed the *Tricarboxylic acid cycle*, such as lactate [162] or alanine [163]. In line with this fact, *Tricarboxylic acid cycle and glyoxylate/carboxylate metabolism*, as well as *Alanine, aspartate and glutamate metabolism*, present more flux in models of CSCs of some tissues in this study. Moreover, a higher use of both glycolysis and oxidative phosphorylation for CSCs of the liver and kidney is observed. This simultaneous increased usage of both pathways

has been described before [161]. A possible explanation is that glycolysis can be fundamental for oxidative phosphorylation through pyruvate-promoted TCA cycle activation, and/or important to produce intermediary glycolytic metabolites that are substrates of the anabolic pathways that are essential for dividing cells [154, 155]. Interestingly, the observed increase in glutamate metabolism (*Alanine, aspartate and glutamate metabolism*) has also been experimentally verified in CSCs of some cancers. For example, glutamine-derived glutamate is a substrate for glutathione synthesis, which has shown to reduce Reactive Oxygen Species (ROS) levels in CSCs of lung and pancreas, preventing β-catenin degradation, and, therefore, indirectly maintaining the expression of stem cell markers [35]. In another study of pancreatic CSCs, glutamate caused a decrease in ROS levels and prevented cell death, through the production of NADPH concomitantly with its conversion to oxaloacetate, and then to pyruvate [36]. Furthermore, glutamate has been shown to indirectly induce movement of a matrix metalloproteinase to the cell surface, which in turn promotes matrix degradation and subsequent cell invasion [164], a trait often linked to CSCs [143].

The higher use of *Pentose phosphate pathway* observed in models of CSCs is also supported by literature [165]. The accepted explanation is that CSCs, mainly those that obtain energy from oxidative phosphorylation, produce more ROS, which can induce cell damage. Therefore, CSCs divert carbon flux from glycolysis to the *Pentose phosphate pathway* to increase production of NADPH and, consequently, down-regulate ROS levels [159]. A similar justification is often provided to explain the resistance of CSCs to chemotherapeutic treatments [166]. On the other hand, the use of the *Pentose phosphate pathway* by CSCs, allows them to increase fatty acid synthesis and nucleotide formation [154, 155], which are essential for their fast growth.

In line with this observation, models of CSCs obtained in this work show a higher flux in *Pyrimidine*, *Purine* and *Nucleotide metabolism* than in CCs, at least for some tissues. Another interesting pathway, that is significantly more represented in models of CSCs than in CCs (in four tissues), is *Folate metabolism*. The folate cycle is important for DNA methylation (through the donation of one carbon to the methionine cycle), glutathione formation (through NADPH production), lipid synthesis (due to NADPH generation), and nucleic acid synthesis, which are relevant processes in cancer [22]. Furthermore, previous studies state that folate can dedifferentiate glial cells to proliferative stem cells that express the pluripotent TFs Sox2 and Oct4, and that may be the reason for the rise in pediatric brain tumors following implementation of acid folic fortification in food in the U.S. [167].

It is also important to note that *Valine, leucine and isoleucine metabolism* has more flux in models of CSCs of three tissues, as these essential Branched-Chain Amino Acids (BCAAs) were reported to be important in several cancers. In those cancers, BCAAs can work as building blocks for protein synthesis, give origin to glutamate (by transfer of an amino group to αKG), which upon conversion to glutamine induces nucleotide synthesis, and can be oxidized to acetyl-CoA and succinyl-CoA that feed the TCA cycle contributing to energy production [168]. Furthermore, a BCAA metabolic enzyme, the branched-chain aminotransferase 1 (BCAT1) correlates with cancer aggressiveness [168], which is a characteristic often associated with CSCs [143].

This study predicted *CRAT* as an essential gene for both CSCs and CCs in two tissues, and as an essential gene specifically for CSCs in five tissues. *CRAT* is part of the carnitine-shuttle and codes for the enzyme carnitine O-acetyltransferase, which is primarily located in mitochondria and catalyzes the inclusion or removal of carnitine from acyl-CoA [169]. Since acetyl-CoA cannot be directly transferred from mitochondria to the cytoplasm, the CRAT-induced transfer of the acetyl-group to carnitine allows the resulting acetyl-carnitine to cross the mitochondrial membrane and transfer the acetyl-group to cytosolic CoA, indirectly promoting the movement of acetyl-CoA from the mitochondrial matrix to cytosol [169]. Cytosolic acetyl-CoA can then be used in fatty acid synthesis, which is essential for the formation of new cell membranes during cancer cell proliferation [169], be used in acetylation-induced deregulation of cytosolic protein function, an important contributing mechanism for the cancerous phenotype [170], or even be transferred again through the carnitine shuttle to the nucleus, where it contributes to histone and TF acetylation, fostering cancer cell growth through gene expression regulation [169, 170]. Furthermore, by decreasing the mitochondrial levels of acetyl-CoA, *CRAT* works as a buffer which prevents excessive mitochondrial protein acetylation, while releasing glycolysis pyruvate dehydrogenase from the acetyl-CoA-promoted inhibition, consequently unlocking glucose oxidation [169,171]. In cancer, the release of glucose oxidation from acetyl-CoA-induced block, together with the abovementioned promotion of fatty acid synthesis and the increase in Fatty Acid Oxidation (FAO), fostered by the need to replenish mitochondrial acetyl-CoA levels, ensues the contribution of high *CRAT* activity for cancer metabolic flexibility [171].

GSTM1 is a gene that codes for Glutathione S-Transferase Mu 1, an enzyme of the GST family. GSTs are known to conjugate cytotoxic compounds with glutathione, protecting against carcinogen-induced oxidative stress [172], which is known to cause gene mutations [173]. That is the reason suggested for

the association of GSTM1 null mutation (which abrogates the enzyme activity) with increased risk of cancer in some studies [172]. However, many other studies showed no relation between GSTM1 null phenotype and risk of different types of cancer [174–177]. Additionally, GSTs were also shown to protect cells from chemotherapy-induced oxidative stress, playing a role in anti-cancer drug resistance [178]. Here, GSTM1 is suggested as an essential gene for CSCs of four tissues, but it is not essential in corresponding differentiated CCs. As these models are exclusively metabolic, the reduction in biomass upon GSTM1 knock-out is probably related to reduced ability in preventing oxidation of amino-acids, since GSTM1 codes for proteins involved in seven reactions of *Phenylalanine, tyrosine and tryptophan biosynthesis* in *Human1*. ELOVL1 was predicted as essential in CSCs of four tissues and both CSCs and CCs of two tissues. This gene codes for an enzyme involved in the synthesis of very-long-chain fatty acids. High levels of ELOVL1 and increased fatty acid elongation have been observed in colorectal cancers [179, 180], and its silencing, together with other genes regulating lipid metabolism, affects the viability of breast cancer cells [181].

FECH and PPOX were predicted as essential in both models of CSCs and CCs in all tissues. FECH codes for ferrochelatase, which is the last enzyme of the heme biosynthetic pathway, that catalyzes the addition of a Fe^{2+} to protoporphyrin yielding a heme protein [182]. PPOX codes for protoporphyrinogen oxidase, which catalyzes the conversion of protoporphyrinogen IX to protoporphyrin, and, therefore, is also involved in heme production [183]. The essential role found for these genes in all tissues of both cell types suggests that they are important to the organism. In fact, heme is of vital importance due to its involvement in several biological processes, including oxygen transport, energy production and drug metabolism. The multifaceted nature of heme renders it as the best candidate molecule exploited/controlled by tumor cells to modulate their energetic metabolism, interact with the microenvironment and sustain proliferation and survival.

The Essential Metabolite (EM) identified in most tissues specifically for CSCs was malonyl-carnitine, which is part of the carnitine shuttle (more information in the Metabolic Atlas [75]), necessary for FAO and fatty acid synthesis. This result is in line with the intense use of fatty acid metabolism by CSCs to obtain and store energy [39]. The present analysis also suggested antimetabolites with the potential to block the effect of EMs in CSCs and/or CCs. The anticancer effect of most of these antimetabolites has already been validated, being described in Table 2.2. Interestingly, although it is known that *Methotrexate* does not efficiently cross the blood-brain barrier (BBB), its efficacy against glioblas-

toma is being studied alone or in combination with other compounds [184], and our analysis suggests it can be effective against glioblastoma CSCs. Furthermore, the present analysis hints that *Ribavirin* could be a potential antimetabolite to tackle lung CSCs. *Ribavirin* is a guanosine nucleoside analogue primarily used in treatment of Hepatitis C and other RNA virus, and its potential ability to treat AML is currently under study [114]. Nevertheless, to the best of our knowledge no connection between *ribavirin* and death of CSCs has been established before, although a recent study suggests it could be used to treat lung cancer [185].

Table 2.2: Antimetabolites identified in this analysis with proven anti-cancer effect.

Antimetabolite	***Role****
8-azaguanine	Small molecule with antineoplastic activity that stimulates cell differentiation and as a purine analogue that competes with guanidine for incorporation into tRNA.
Capecitabine	Drug enzymatically converted to fluorouracil, which in turn inhibits DNA synthesis. It is used in the treatment of metastatic and colorectal cancers.
Fludarabine	Purine analogue used in the treatment of leukemia.
Cytarabine	Pyrimidine analogue used in the treatment of leukemia.
Mercaptopurine	Analog of the purine bases adenine and hypoxanthine.
Decitabine	Pyrimidine nucleoside analogue used to treat Myelodysplastic syndromes.
Fluorouracil	Pyrimidine analogue used to treat diverse solid tumors, such as colon, rectal, breast, gastric, pancreatic, ovarian, bladder, and liver cancer.
Cladribine	Purine nucleoside analogue, primarily utilized to treat hairy cell leukemia.
Nelarabine	Purine nucleoside analogue used to treat T cell lymphoblastic leukemia or lymphoma.
Methotrexate	Folate derivative which affects enzymes of nucleotide synthesis, like dihydrofolate reductase, and it is used to treat leukemias and solid tumors. It has been suggested to treat glioblastoma.

* Anti-cancer role is supported by the antimetabolite description in DrugBank [114] and PubChem [186].

FLI1 was identified in this work as a TF that upon knockout could potentially lead to CSC death in two tissues: head and neck, and pancreas. FLI1 is highly expressed in different cancers [187–189], and the chromosomal translocation of the gene coding this TF produces a mutant that activates a genetic program essential for tumor maintenance in Ewing's sarcoma [190]. Additionally, inhibition of that mutant protein has shown to be effective in targeting CSCs of Ewing's sarcoma [191], while its expression seems to be associated with other aggressive tumors [192]. The FLI1 effect has been attributed to cell cycle regulation in different cancers [193], and to the overexpression of 3-phosphoglycerate dehydrogenase, an enzyme of the serine synthesis metabolic pathway, in Ew-

ing's sarcoma [194]. Furthermore, FLI1 transcriptional activity is increased by acetylation [195], a process that is regulated by the availability of the metabolite acetyl-CoA [27]. These facts, together with our results, emphasize the need to study potential metabolic mechanisms either triggered by or regulating FLI1 in cancers other than Ewing's sarcoma.

Our results also suggest that HNF1A and FOXA1 are potential targets to knockout in cancer of at least three tissues. However, reported literature for these TFs either validates their role as oncogenes [196–198] or as tumor suppressors [199, 200]. Therefore, their role appears to be dependent on the cellular context and may not be good candidates for experimental testing for that reason.

Three miRs were identified as potentially useful in reverting cancer phenotype (both CSCs and CCs) in at least five tissues: *hsa-miR-335-5p*, *hsa-miR-26b-5p* and *hsa-let-7b-5p*. Down-regulation of *hsa-miR-335-5p* was reported across different cancers and this miR has been identified as a tumor suppressor, favorable biomarker, or therapeutic target in some of those cancers [201–206]. The studied mechanisms of action of *hsa-miR-335-5p* are mainly mediated by cell signaling [201, 203, 204] and cytoskeleton remodeling [202, 205] proteins, but there is only one report of tumor suppression mediated through a down-regulation of the metabolic enzyme LDHB [207]. How miR-335-5p-induced LDHB down-regulation inhibits cancer has not been clarified yet. However, the ability of LDHB to convert lactate to pyruvate was shown to be important for oxidative phosphorylation-dependent CCs to obtain energy in glucose-deprived conditions and upon cooperativeness with lactate-producing glycolytic CCs [208], while LDHB activity proved to be necessary for autophagy-promoted cell proliferation in both oxidative and glycolytic CCs [208]. Several studies demonstrate the *hsa-miR-26b-5p* capacity to reduce tumor progression, cell proliferation, and metastasis in different cancer types [209–212]. These studies focus on growth-factor and cell-signaling mediated mechanisms to explain *hsa-miR-26b-5p* function [210, 212]. Nevertheless, one study of bladder cancer suggested that a gene down-regulated by that miR, PLOD2, might potentially up-regulate the glycolytic enzyme hexokinase 2, contributing to a glucose-promoted growth in CCs [209], although more studies are needed to validate this result. Also, this miR has as potential targets genes encoding enzymes, like ACSL3 and ACADM [138], which are involved in acetyl-CoA and fatty acid metabolism, two processes essential in cancer. Furthermore, in another study of breast cancer, *hsa-miR-26b-5p* silences the expression of SLC7A11 [213]. That protein is a member of the glutamate–cystine antiporter that allows cystine uptake-induced

glutathione and protein synthesis and, consequently, promotes cancer growth in the pancreas [214]. Therefore, this might be a metabolic mechanism for *hsa-miR-26b-5p* that might attract the interest of wet-lab scientists for experimental validation. *hsa-let-7b-5p* (known as let-7) was reported as a growth suppressor in different cancer types [215, 216], and a link has already been established between let-7 and the reduction in glucose uptake through the inhibition of several components of the insulin-PI3K-mTOR pathway [217]. In this way, it can be postulated that let-7 decreases cancer growth through the repression of glucose uptake in glycolytic cancer cells. Overall, the detection of the abovementioned miRs as potential therapeutic strategies against cancer in this study calls the attention of wet-lab scientists to the need to further understand the role of those miRs in cancer metabolism. On the other hand, the fact that many of the identified miRs (even those not discussed here) are reported as cancer suppressors contributes to the validation of the reconstructed metabolic models. Furthermore, some miRs here presented (Figure 2.3-B and Figure A.6-B) are not directly reported as cancer suppressors and may be interesting to study, for example, *hsa-miR-6499-3p*, *hsa-miR-8485*, and *hsa-miR-6849-3p*.

2.4 Materials and Methods

2.4.1 Transcriptomics data collection and gene expression analysis

Gene expression datasets, from either RNA-seq or microarray experiments, were retrieved from *Gene Expression Omnibus* [107] (*GEO*), *Array Express* [218] (*AE*), and/or *European Genome-phenome Archive* [219] (*EGA*) databases. All RNA-seq raw data were pre-processed with the same pipeline. A read quality evaluation step with *fastQC* [220] was followed by read filtering with *Trimmomatic* [221], which kept reads with at least 36 bp length and an average quality score of at least 24. Adapter/primer sequences and other contaminants were also removed when present. *STAR* [222] was then used to align reads to a reference human genome (version GRCh38 from Ensembl) and aligned reads were counted with *HTSeq* [223] (annotation release GRCh38.99). Raw counts were normalized with the *GeTMM* [224] method, which combines gene-length correction with TMM normalization, allowing both intra- and inter-sample comparison.

Microarray raw data was analyzed according to the microarray platform. Affymetrix raw datasets were processed with Robust Multichip Average (RMA) normalization method, from the *oligo* [225] *R* [226] package, while Illumina and Agilent data underwent logarithm and quantile normalization with the

limma [227] package. Other R packages used were *beadarray* [228], *ArrayExpress* [229], and *GEOquery* [230]. Donors or cell lines unmatched between different conditions (e.g., present in CCs but not in CSCs) were excluded from both RNA-seq and microarray pre-processing and normalized expression values of technical replicates were averaged.

The complete RNA-seq data-analysis pipeline, which was based on Bash [231] language and implemented on docker containers, together with the microarray analysis pipeline, implemented in *R* [226], are available in GitHub (`https://github.com/BioSystemsUM/bRNAsPipe`). A list of all datasets used, with respective identifiers, is provided in Table C.2.

2.4.2 Reconstruction of genome-scale metabolic models and task gap-filling

The *in silico* procedures of this and the following sections were overall based on the use of the *Troppo* [232] Python package developed in-house, following pre-processing pipelines and parametrizations similar to the study of *Vieira et al.* [126]. The flow diagram for the overall study methodology is shown in Figure A.7. Genome-scale metabolic models were built for each donor or cell line of each study (gene expression dataset). The best transcriptomics integration strategies and reconstruction algorithms identified for one sample of an RNA-seq dataset and one sample a microarray dataset (selection procedure explained in sections below) were used to reconstruct models for all samples of RNA-seq and microarray datasets, respectively. Models had all exchange reactions closed except for those referring to metabolites of Ham's medium, considered to be able to enter the cell (Table C.5). Models were gap-filled for growth in Ham's medium. We assessed whether all models could fulfill 57 metabolic tasks essential for human cell viability (retrieved from `https://github.com/SysBioChalmers/Human-GEM/tree/master/data/metabolicTasks`). When those tasks were not accomplished, models were gap-filled. This analysis was done in Python and utilized different modules, like *CoBAMP* [233], *Troppo* [232], *COBRApy* [234], and *Pandas* [235]. Only models of cell lines/donors that were able to grow (produce biomass) during Flux Balance Analysis (FBA), and to perform all 57 essential tasks, were considered successfully reconstructed. When a model of a specific donor/cell line was not successfully reconstructed for one of the cell types (e.g., successful reconstruction for CC but not for CSC) all models of that donor were excluded, so that subsequent analyses were done only on models with matched donors.

2.4.3 Parsimonious flux balance analysis

A parsimonious Flux Balance Analysis (pFBA) simulation was also done to identify the distribution of fluxes that maximized the biomass reaction, while minimizing the sum of the absolute value of fluxes. After doing pFBA for all models, the difference of the absolute values of reaction fluxes between CSCs and CCs was determined for each donor/cell line. The median value of those differences was then calculated to obtain a value per reaction and tissue. For each tissue, reactions were then sorted from those with more flux in CSCs to those with more flux in CCs. As reactions can be classified in different metabolic subsystems, a procedure similar to a ranked gene-set enrichment analysis [236], but with reactions instead of genes, was performed. This was done with the mean-rank gene set test from the *limma* [227] package in *R*. Subsystems with a p-value below 0.05 were identified.

2.4.4 Simulation of lethal/essential genes and metabolites

To find essential genes, we simulated the knockout of one gene at a time by excluding its corresponding reactions from reconstructed models in compliance with the Gene-Protein-Reaction (GPR) rules of the models. To identify Essential Metabolites (EMs), we simulated the removal of one metabolite at a time by constraining to zero the flux of irreversible reactions that used the metabolite as a substrate and of reversible reactions in the direction where the metabolite is a substrate.

Gene or metabolite knockouts that produced biomass flux lower than 0.1% of biomass flux in wildtype were considered essential genes or metabolites. Then, those that were essential in at least 50% of the donors/cell lines of one cell type and tissue were considered essential for that cell type and tissue. Essential genes and metabolites were split into two groups: those specific to CSCs and those common to both CCs and CSCs.

2.4.5 Strategies for transcriptomics data integration

To reduce the weight that highly-expressed genes have in comparison with lowly-expressed genes, each dataset of normalized gene expression data has gone through min-max scaling: $(x - min)/(max - min)$, where x is a specific expression value and $max - min$ is the expression range of a gene.

Three gene threshold approaches, first introduced by *Richelle et al.* [153], were assessed: a *global* strategy (which uses a global threshold), a *local 1* strategy

(using the combination of one global threshold and one local threshold), and a *local 2* strategy (using the combination of two global thresholds and a local threshold). Genes below the global threshold in *local 1* and below the lower global threshold in *local 2* were regarded as inactive (i.e., *off* genes) and genes above the higher global threshold in *local 2* were considered active (i.e., *on genes*). Those genes above the global threshold in *local 1* and between the two global thresholds in *local 2* (i.e., *maybe on* genes) were subsequently included in the *on* or *off* groups, if their expression values were above or below the local thresholds, respectively [153] (Figure A.9). For each of those strategies, the gene score was computed as $log(value/threshold)$. In the case of local thresholds, the set of tested thresholds included 10^{th}, 25^{th}, 50^{th}, 75^{th} and 90^{th} percentiles of samples of each gene. Global thresholds were the average of local thresholds of the same percentile.

Additionally, two other *localB* strategies, where genes defined as *on*/*off* by global thresholds were given higher/lower gene scores, respectively, than those defined by local threshold, were also tested (Figure A.9). In *local2B*, *on* and *off* genes determined by the local threshold had scores between 0 and 1, and -1 and 0, respectively, while *on* and *off* genes defined by global thresholds had scores above 1 and below -1, respectively. Similarly, *local1B off* genes determined by the global threshold had scores inferior to -1, while the remaining genes had a score of at least -1. Another tested pre-processing decision was whether to use all genes or just the metabolic genes in the thresholding calculation.

Besides these thresholding approaches, two gene to reaction scores conversion strategies were evaluated: *min-max* and *min-sum*. In both strategies, reactions catalyzed by enzyme complexes obtained reaction scores equal to the minimum of the scores of the genes encoding those enzymes. Regarding the reaction scores of isoenzymes, in *min-max*, these were defined as the maximum of the gene scores, while in *min-sum* they were set as the sum of the gene scores (Figure A.9).

2.4.6 Assessment and selection of best strategies for transcriptomics data integration

To evaluate the best pre-processing decisions among the options described above, we ran an analysis in two steps. First, we split biological samples into groups of those of the same study and cell type, which are expected to present similar reaction scores. The average Euclidean distance between reaction scores of samples in each sample group was determined. To find the probability of

getting those average distances or lower by chance, randomly simulated sample groups of the same size were created in 1000 different simulations, and the same average Euclidean distance calculation was computed for simulated groups. Subsequently, the following metric was calculated for each group: the proportion of simulations where the simulated average distance was lower than or equal to the observed average distance in the real group (Figure A.8). Results for all sample groups are shown in Figure A.2.

The first step allowed to narrow down the number of strategies to test in the second step, as the latter is more computationally expensive. The second step involved the comparison of computationally simulated essential genes with experimental verified lethal genes, for the strategies with the three best results (with 3 lowest values) in the abovementioned metric, to further select the one with the best performance. Note that, because experimental gene lethality information was not available for all cell types (for all sample groups), only the three best performing strategies in one cell line of a group of RNA-seq samples and one cell line of a group of microarray samples were further assessed.

Reaction scores of each of those strategies were used to reconstruct cell-specific models with both FASTCORE and INIT algorithms. Models were gap-filled for essential tasks and biomass growth. Essential genes in reconstructed models were identified and the Mathews Correlation Coeficient (MCC) score was applied to compare the simulated essential genes with experimental verified lethal genes. The strategies rendering the best MCC values for the abovementioned microarray and RNA-seq sample groups were identified (Figure A.8). This second step enabled to test the influence of reconstruction algorithms, which the first step could not accomplish.

The best identified strategies in the last step, for the microarray and RNA-seq sample groups, were applied to pre-process data of all microarray and RNA-seq studies, respectively.

2.4.7 Detection of potential antimetabolites

Antimetabolites are drugs that can counteract the effect of EMs by competitive inhibition of their targets. To find antimetabolites, genes of active reactions where EMs were substrates were first identified for each model, and those in at least 50% of the donors/cell lines of one cell type and tissue were attributed to that cell type and tissue. Then, the complete DrugBank database [114] (`https://go.drugbank.com/`) was parsed. From the aforementioned genes, those that were reported in the database as coding proteins (targets, enzymes, car-

riers, and transporters) that interact with antimetabolites, were detected. We identified antimetabolites associated with genes that coded proteins targeted by EMs specific to CSCs and for those common to both CCs and CSCs.

2.4.8 Prediction of transcription factors and miRNAs that may potentially affect cell survival

A Flux Variability Analysis (FVA) was performed for each donor/cell line considering different fractions of biomass (from 0 to 90%, in intervals of 10%). The minimum and maximum fluxes of each reaction for each value of fraction of biomass were averaged. Reactions where the average flux directly (Pearson correlation > 0.7) correlated with the fraction of biomass (flux increases with biomass) were detected for each donor/cell line. Then, from those reactions, we selected the ones common in at least 50% of the donors/cell lines, for each cell type and tissue and, subsequently, identified the genes coding for those reactions (genes directly correlated with biomass), using GPR rules.

The genes directly correlated with the biomass fraction in each tissue were overlapped with genes whose expression is known to be downregulated when a TF is knocked-out, to identify TFs that may potentially cause cell death when inhibited. To find such TFs, we queried a database of human gene expression profiles for knockdown/knockout of TFs, the knockTF database [237] (`http://www.licpathway.net/KnockTF/`) and specifically identified genes whose expression significantly decreases ($\log_2$(fold-change) $<$ -1.5 and FDR $<$ 0.05) once the TF is knocked-out.

Similarly, genes directly correlated with biomass were overlapped with genes whose expression is known to be affected by miRNAs (miRs), to identify miRs that may potentially cause cell death by interfering with mRNA translation of those genes. We collected miRs that targeted human genes in a database of experimentally validated miRNA targets, the miRTarBase [238] (`http://mirtarbase.cuhk.edu.cn/php/index.php`). Then, the top 10 miRs (can be more than 10 if two or more miRs have the same number of targets) that targeted more genes directly correlated with biomass were identified for each tissue. MiRs with 10 or more known targets in the database were included. Note that genes directly correlated with biomass were split into those specific to CSCs and those common to both CCs and CSCs, and TFs/miRs targeting both groups are shown.

Data Availability:

Datasets used and respective identifiers are shown in Table C.2.

The code for the present work is deposited at `https://github.com/BioSystemsUM/human_ts_models/tree/master/projects/csc_devel` and `https://github.com/BioSystemsUM/bRNAsPipe`

Chapter 3

Reconstruction of Cell-specific Models Capturing the Interplay Between Metabolism and Epigenetics in Cancer

3.1 Introduction

In the past decades, there has been an increase in the incidence of early-onset cases of cancer [239]. Changes in lifestyle, environment, and diet, together with genetic susceptibilities, have contributed to genetic mutations that trigger an imbalance of cell differentiation, survival, and/or proliferation, promoting cancer onset and development [28, 239]. In addition to genetic mutations, which directly affect the DNA sequence, the de-regulation of epigenetic mechanisms, which control the attachment of chemical groups to DNA, histones, and nucleosome-positioning protein complexes, can also induce cancerous phenotypes. In particular, the unbalance in epigenetic modifications may change the chromatin accessibility to transcriptional complexes, and subsequently, induce aberrant gene expression profiles without affecting the genomic sequence [28].

Another fundamental feature of cancer is its metabolic rewiring, as cancer cells are forced to adapt their metabolism to generate enough energy and elementary metabolites for the synthesis of new cellular membranes, proteins, or nucleic acids necessary for cell proliferation [12, 144]. Furthermore, given that distinct metabolites are also substrates or cofactors of epigenetic regulators [145], alterations in their availability, as a consequence of metabolic reprogramming, can induce a cancerous phenotype through epigenetic deregulation [28, 146], whereas, on the other end, epigenetic alterations on genes encoding metabolic enzymes may contribute to the metabolic shift characteristic of cancer. Therefore, there is an urge to investigate the cross-talk between cancer, epigenetics, and metabolism to develop new and efficient therapeutic strategies against the

disease.

Genome-Scale Metabolic Models (GSMMs) are mathematical representations of all metabolic reactions of a cell, where reactions catalyzed by enzymes are mapped to associated genes and/or proteins [48]. By assuming the steady state (that metabolite concentrations do not change over time), it has been possible to utilize these *in silico* constructs in the prediction of metabolic phenotypes. In detail, the product of a matrix with the stoichiometric coefficients (where columns and rows represent respectively the reactions and metabolites) and a vector of reaction fluxes (rates) is assumed to be zero upon the steady state. This results in a solvable system of linear equations, the solution of which comprises fluxes of all metabolic reactions represented in the system [81]. The development of methodologies for omics data acquisition over the years has favored the reconstruction of this type of model under specific biological contexts by integration of omics data of a particular tissue or cell type. In particular, a variety of healthy cells and morbidities [64–66, 129, 130], including cancer [63, 126, 240] have been modeled with context-specific GSMMs.

The main disadvantage of traditional constraint-based GSMMs is that, unless some nutrient uptake fluxes are known, no finite flux distribution can be obtained. Unlike traditional GSMMs, Genome Scale Metabolic Models enhanced with Enzymatic Constraints using Kinetic and Omics data (GECKOs) do not require nutrient-uptake rates to produce finite flux values during simulations, as they integrate both enzymatic kinetic information and concentration, serving as additional constraints to the flux solution space. Specifically, enzymes are added as pseudo-metabolites that although represented as substrates, do not affect the mass balance of the reactions they catalyze, and pseudo-uptake reactions for each enzyme are included to guarantee enzyme mass balance. This results in an extended version of the abovementioned matrix of stoichiometric coefficients, where additional rows representing the enzyme mass balance and columns depicting enzyme usage reactions are introduced. The catalytic information is introduced in the form of the inverse of turnover number (k_{cat}) values as coefficients to the enzymes in metabolic reactions, whereas enzyme concentration is used as the upper bound of each enzyme usage reaction. When no proteomic data is available to limit the flux of each enzyme usage reaction, an enzyme usage reaction of the pool of all enzymes is introduced instead and each enzyme is drawn from the enzyme pool [137].

Few studies have attempted to use GSMMs to modulate the interaction of metabolism and epigenetics. An old study from 2014 [241], integrated the decrease in gene expression observed upon mutation of histone tails, which are

often mutated in cancer and are targeted by epigenetic marks, into a yeast model to modulate the effect of those mutations on the rate of production/consumption of acetyl-CoA, a substrate for histone acetylation. In the following year, an analysis was published [242] where metabolic models reconstructed for different time points based on time-course transcriptomics data provided simulations that were compared with ChIP-seq data for a histone-acetylation mark, to capture the differentiation of primary human monocytes to macrophages. The authors observed that enhancers of metabolic genes under high regulatory load (close to histones with high levels of the acetylation mark) were mainly associated to transport reactions and other metabolic pathway entry points in comparison with other metabolic genes, suggesting that the former are critical epigenetic-regulatory control points for the metabolic reprogramming during monocyte to macrophage differentiation [242]. In another study from 2017, Chandrasekaran et al. [243] tried to predict in which of the two states murine pluripotent stem cells go through during embryonic development, preceding (naïve state) or succeeding (primed state) the implantation of the embryo in the uterus, was producing more S-Adenosyl-Methionine (SAM), a substrate for methylation. Using a semi-dynamic modeling approach, the authors suggested that histone methylation was more intense in the primed cells, which was experimentally verified afterward [243]. Most recently, Shen et al. successfully predicted the increase or decrease in protein acetylation levels in human cells in the presence of different nutrient sources. Furthermore, through the inclusion of one reaction representing the overall protein acetylation, cancer cell lines that were more sensitive to *vorinostat*, a deacetylase inhibitor used in cancer treatment, were estimated to have higher acetylation levels, suggesting that GSMMs could be used to identify cancer cells more responsive to treatments with deacetylase inhibitors [244].

Although those studies represent important steps toward the modulation of the interplay between metabolism and epigenetics, they all focus on histone modifications, particularly acetylation. The only study that addresses methylation dwells on histone methylation in murine cells and simply uses the flux of SAM as a surrogate for methylation. In the present work, we reconstruct models for 31 human cancer cell lines which included DNA methylation and demethylation reactions described in the literature, as well as DNA methylation levels estimated from experimental data. Furthermore, these models are GECKOs, which present the advantage of providing more accurate flux distributions than traditional GSMMs when experimental flux values are unavailable.

3.2 Results

In this study, GECKO models containing DNA methylation and demethylation reactions were reconstructed for different cancer cell lines. Those reactions, which included DNA containing modified cytosines were retrieved from literature, adapted for charge and mass balance, and were first introduced on the generic GSMM *Human1* [75] (more details on *Creation of the generic DNA methylation model* section of Materials and Methods). The complete list of reactions introduced is shown in Table D.1 and a simplified visual representation of how those reactions integrate with the model is presented in Figure 3.1.

In a nutshell, the DNA methylation process starts when SAM is produced in the one-carbon cycle in the cytoplasm through reaction *MAR03875* and, once inside the nucleus, it is used as a substrate of DNA methylation through reaction *MAR08641* (Figure 3.1). DNA can then be demethylated using different pathways (Figure 3.1). DNA-5-methylcytosines (DNA5mC) can be successively oxidized to DNA-5-hydroxymethylcytosines (DNA5hmC), DNA-5-formylcytosines (DNA5fCs), and DNA-5-carboxylcytosines (DNA5CaC) or converted to thymines, i.e. DNA-5-methyluracils (DNA5mU). DNA5fC and DNA5CaC can be transformed back to unmethylated cytosines in DNA (through the enzyme non-catalyzed reactions *consdirectDNA5fC* and *consdirectDNA5CaC*), or like DNA5mU, they can be replaced by unmethylated-cytosines through the cellular Base-Excision Repair (BER) mechanism. The BER starts with the excision of the modified cytosine (reactions *prodAPsite3* and *prodAPsite4*) or the mismatched thymine (*prodAPsite1* reaction) using DNA-glycosylases that cleave the bound between the base and the deoxyribose creating an apyrimidinic site (APsite). An endonuclease then cuts the phosphate backbone at the APsite, leaving a nick and a deoxyribo-5'-phosphate (dRP) to which the excised base was connected (in reaction *proddRPsite*). A new unmethylated cytosine is inserted afterward while the dRP is still hung by its 3' side to the phosphate backbone (in *prodhangdRPsite* reaction). The dRP is excised by a dRP lyase, creating a nick in DNA strand (in *prodDNAnick* reaction), which is then ligated by a DNA ligase, restoring the unmethylated DNA (in *ligate DNA* reaction) (see Figure 3.1).

Since the ratio of DNA5mC, DNA5hmC, and DNA5fC in relation to unmethylated DNA can be estimated, a pseudo-reaction representing the composition of total DNA (DNAtot) in terms of those species was also introduced in the model (*prodDNAtot* reaction), and the original biomass reaction was replaced by an equivalent one (*adaptbiomass* reaction) where DNA was substituted

by DNAtot (Figure 3.1).

The final adapted generic model was able to produce biomass and none of the introduced (de)/methylation reactions were blocked. Context-specific GSMMs were then built for different NCI-60 cell lines through the integration of transcriptomics data, and a procedure from Robinson et al. [75] was used to convert those traditional GSMMs into GECKO models.

3.2.1 Reconstruction of cell-specific metabolic models

This study tested two strategies previously applied to build GSMMs of NCI-60 cell lines. One of those, introduced by Richelle et al. [92] in MATLAB and here implemented in Python, is based on the inclusion of cell-type specific metabolic tasks. Initially, gene scores resulting from the preprocessing of transcriptomics data were converted to reaction scores for each cell line, using Gene-Protein-Reaction (GPR) rules. The highest reaction scores in a cell line were then attributed to all reactions necessary for each generic metabolic task, deeming a generic task as a metabolic task done by all cell types. Also, the same procedure was applied for reactions necessary for cell-type specific tasks, as long as those tasks are done by the specific cell type under consideration. Reactions necessary for a task were identified as the ones carrying flux after the implementation of the task-associated flux constraints on the generic model upon minimization of the sum of all fluxes. In order to determine whether a cell type performs a certain cell-specific task, a metabolic score was calculated for each task and cell type combination (see more details in *Reconstruction of cell line-specific traditional GSMMs* section of Materials and Methods). Afterward, the reconstruction algorithm FASTCORE was applied to build the cell-specific models, because, like other MBA-based methods and unlike iMAT-based methods (such INIT), it preserves almost all tasks after reconstruction [92].

The second strategy consisted of directly using a version of the tINIT algorithm already implemented in MATLAB by Robinson et al. [75], which preserves the generic metabolic tasks (see more details in *Reconstruction of cell line-specific traditional GSMMs* section of Materials and Methods).

For the selection of the best reconstruction and simulation strategies, models were initially built for only 40 to 42 of the NCI-60 cell lines, due to the lack of transcriptomics data, DNA methylation measurements, and metabolite uptake rates for some cell lines. The exact number varied with the reconstruction strategy applied depending on the number of infeasible models (around 1 or 2). However, after the selection of the best strategy, the number of models

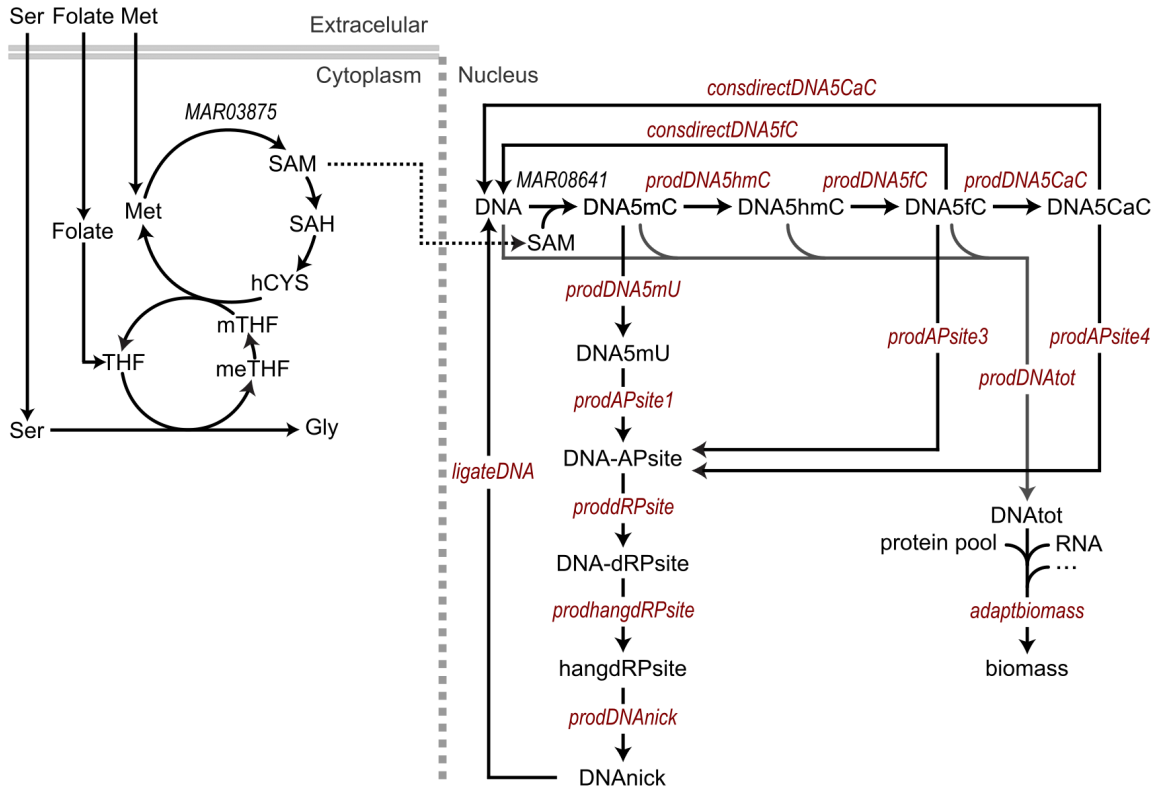

Figure 3.1: Visual representation of reactions contributing to DNA methylation and demethylation. Note that this is just a simplified scheme, as it does not reflect the stoichiometric proportions and excludes many metabolites and transport reactions (for all complete reactions see Table D.1). Reaction identifiers are in italic. Those in red color were added to the original *Human1* (version 1.12) generic model. Gly: Glycine; Ser: Serine; Met: Methionine; THF: Tetrahydrofolate; mTHF: methyl-THF; meTHF: 5,10-methylene-THF; SAM: S-Adenosyl-Methionine; SAH: S-Adenosyl-Homocysteine; hCYS: Homocysteine; DNA5mC: DNA-5-methylcytosine (i.e. methylated DNA); DNA5hmC: DNA-5-hydroxymethylcytosine; DNA5fC: DNA-5-formylcytosine; DNA5CaC: DNA-5-carboxylcytosine; DNA5mU: DNA-5-methyluracil.

used in subsequent simulations was reduced to 31, due to lack of another data type needed for model integration with degree of DNA methylation, as will be explained further bellow.

3.2.2 Detection and validation of the best reconstruction and simulation pipelines

Since no flux distribution can be obtained from unconstrained traditional GSMMs, the uptake rates of metabolites in Ham's media were loosely constrained (from -1000 to 1000) and those of other metabolites were closed (set from 0 to 1000). Then, a parsimonious Flux Balance Analysis (pFBA) was applied and the resulting simulated fluxes of exchange reactions of 26 metabolites were compared with experimentally measured ones. Although there was a small correlation between simulated and measured fluxes, 0.33-0.51 and 0.34-0.49 of Pearson and Spearman correlation respectively (with a p-value of zero), most simulated values did not match measured ones, i.e. the logarithm of their absolute values were higher or lower than $\pm$ 1 of $\log_{10}(|\text{measured value}|)$ (most data points fell outside the pink area of the graphs in Figure 3.2). Note that the absolute values of the fluxes were logarithmized as the majority presented small values (close to zero) (Figure 3.2-E). Richelle's approach (using FASTCORE) (Figure 3.2-A,B) showed a higher percentage of matching values than Robinson's strategy (using tINIT) (Figure 3.2-C,D), and the integration of tasks was slightly detrimental to the correlation values in both reconstruction methodologies (Figure 3.2-A,C versus B,D). The predicted flux values of biomass were much higher than the measured ones and the relative errors of predicted growth rates were high (Figure B.1).

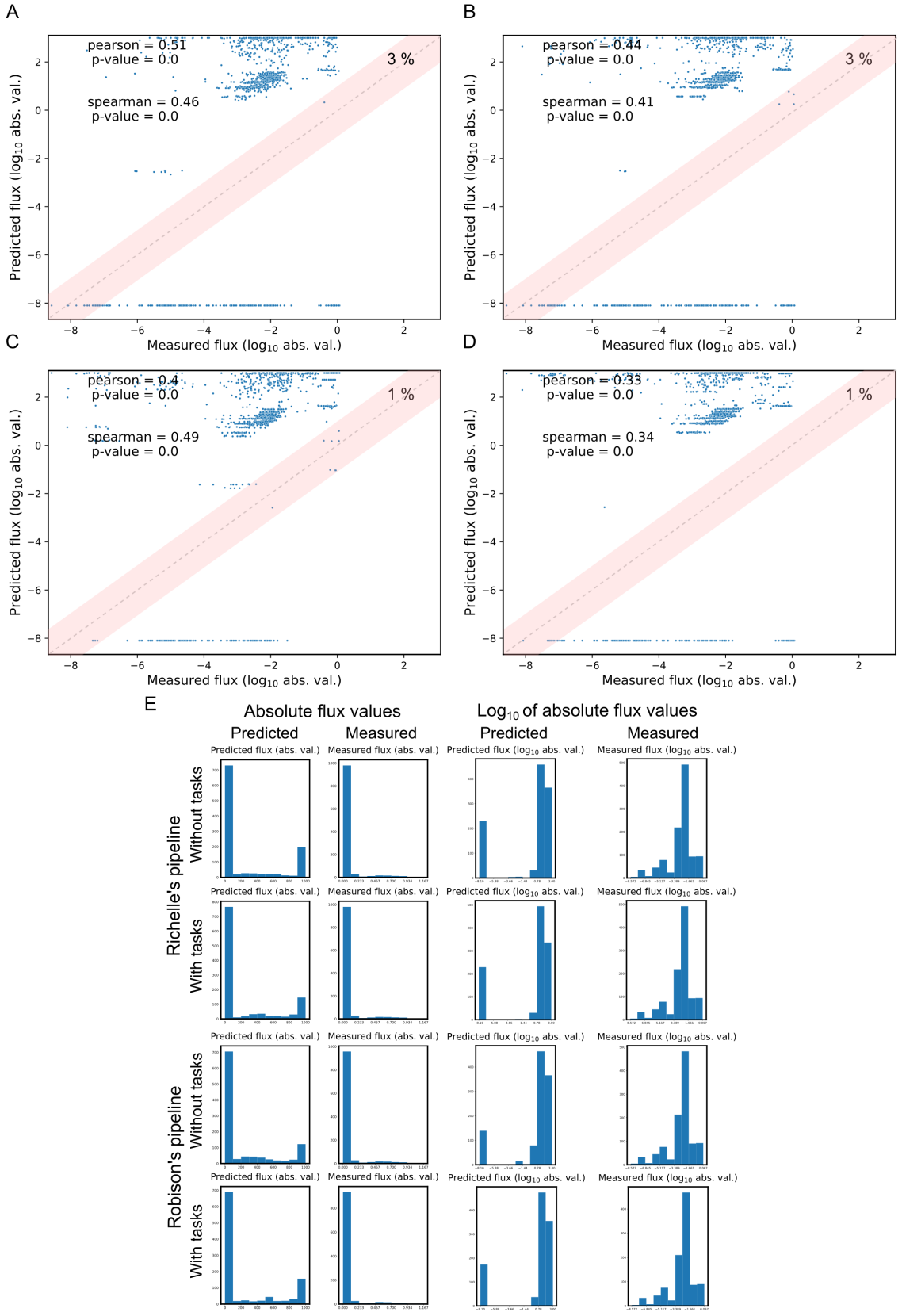

Figure 3.2: Comparison of measured and simulated exchange fluxes produced by traditional GSMMs where uptake/secretion rates of metabolites in Ham's media were loosely constrained (from -1000 to 1000). *(continues)*

Figure 3.2 *(continued from previous page)*: **A-D**: scatter plots with $\log_{10}$ of absolute values of simulated and measured fluxes of exchange reactions of 26 metabolites. Value at top right of each graph is the percentage of data points that are inside the pink area (where $\log_{10}(|\text{predicted value}|)$ is within $\log_{10}(|\text{measured value}|) \pm 1$). Either Richelle's pipeline using FASTCORE (**A**, **B**) and Robinson's pipeline using tINIT (**C**, **D**) were applied to reconstruct the models employed in the simulation. The effect of the integration (**B**, **D**) or not (**A**, **C**) of all tissue-specific metabolic tasks in those models was also assessed. **E**: histograms with the distribution of absolute values of measured and simulated fluxes before and after logarithmization. Data points forming a line at the bottom of **A-D** correspond to metabolites with a predicted flux of zero, which are shown in the graphs as holding the lowest absolute measured value (besides zero), as the logarithm of zero is undefined. In the distributions of logarithmized absolute values (**E**), the values of those metabolites fall in the lowest bin, creating an oddly tall bin at the beginning.

Since constraints in the uptake/secretion rates of three metabolites (glucose, lactate and threonine) with experimentally measured values had been previously reported as sufficient to generate small growth rate prediction errors for the models of eleven of the NCI-60 cell lines [75], we decided to test the effect of those constraints here. The absolute values of the fluxes were again logarithmized because many presented small values (close to zero) (Figure 3.3-E), and only 23 metabolites were taken into account, as the three metabolites whose fluxes were constrained were excluded from the analysis to prevent bias. Overall, there was an increase in the percentage of simulated fluxes whose values were similar to the measured ones (the $\log_{10}(|\text{simulated value}|)$ was within $\log_{10}(|\text{measured value}|) \pm 1$, as 65-74% of data points are inside the pink area of the graphs in Figure 3.3) in relation to the loosely constrained models (Figure 3.2), which was coupled with an improvement in the correlation between simulated and measured values (the Pearson and Spearman correlations enhanced to 0.47-0.62 and 0.54-0.59). Furthermore, the correlations and the percentages of biomass flux values in close proximity to the measured ones greatly increased, while the relative errors of predicted growth rates reduced in comparison with the loosely constrained models (Figure B.2 versus Figure B.1). Unlike the loosely constrained models, the integration of cell-specific tasks provided a slight improvement when using Richelle's strategy (Figure 3.3-A versus B) and the overall best-performing reconstruction strategy was, in this case, Robinson's approach (using tINIT) (Figure 3.3-A,B versus C,D).

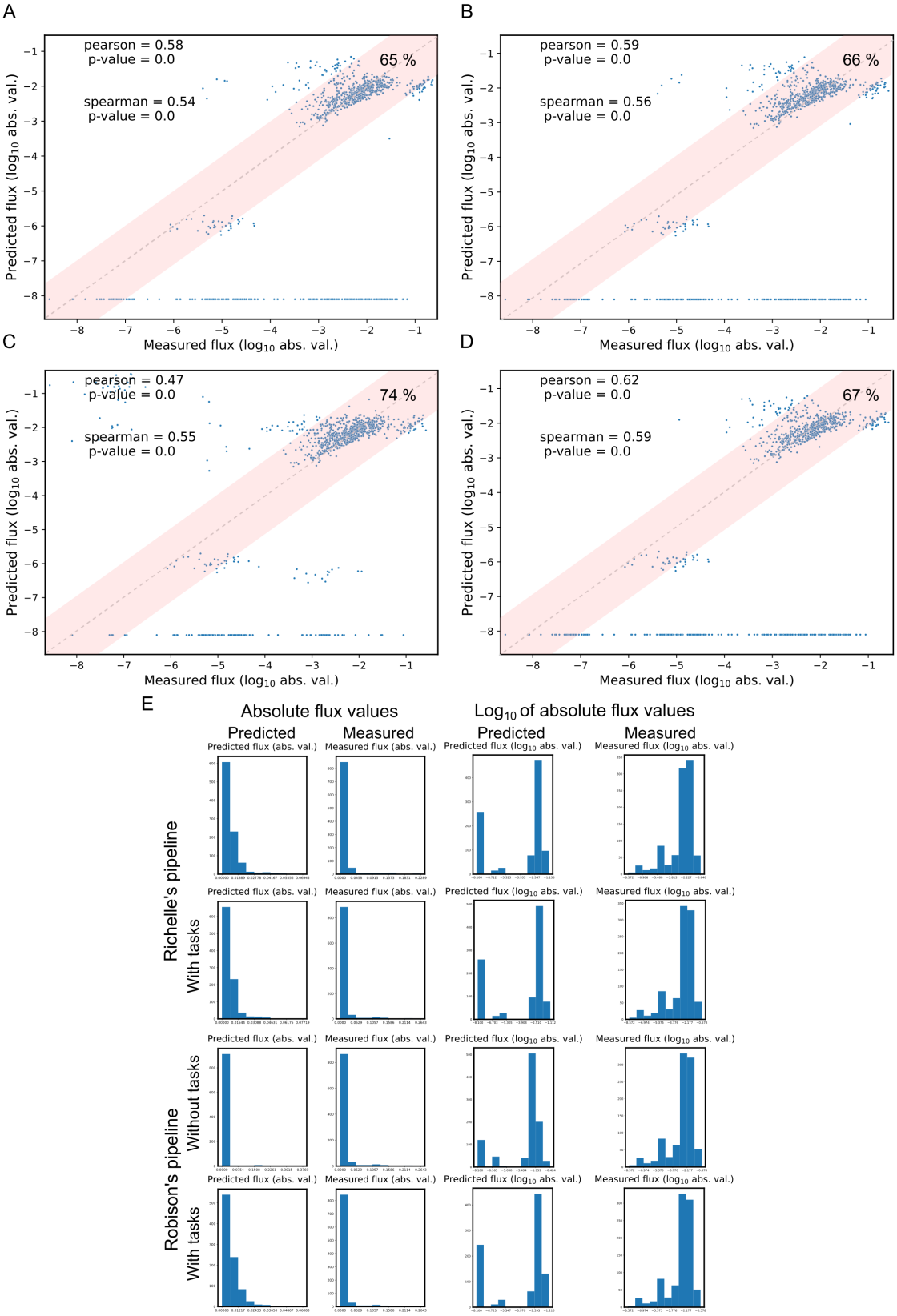

Figure 3.3: Comparison of measured and simulated exchange fluxes produced by traditional GSMMs where uptake/secretion rates of three metabolites (glucose, lactate and threonine) were constrained with measured fluxes *(continues)*.

Figure 3.3 *(continued from previous page)*: **A-D** are scatterplots with $\log_{10}$ of absolute values of simulated and measured fluxes of exchange reactions of 23 metabolites (the three metabolites whose fluxes were constrained are excluded). Value at top right of each graph is the percentage of data points that are inside the pink area (where $\log_{10}(|\text{predicted value}|)$ is within $\log_{10}(|\text{measured value}|) \pm 1$). Either Richelle's pipeline using FASTCORE (**A**, **B**) and Robinson's pipeline using tINIT (**C**, **D**) were applied to reconstruct the models employed in the simulation. The effect of the integration (**B**, **D**) or not (**A**, **C**) of all tissue-specific metabolic tasks in those models was also assessed. **E**: histograms with the distribution of absolute values of measured and simulated fluxes before and after logarithmization. Data points forming a line at the bottom of **A-D** correspond to metabolites with a predicted flux of zero, which are shown in the graphs as holding the lowest absolute measured value (besides zero), as the logarithm of zero is undefined. In the distributions of logarithmized absolute values (E), the values of those metabolites fall in the lowest bin, creating an oddly tall bin at the beginning.

With this dataset, good simulations were obtained by limiting the fluxes of three exometabolites with experimental data. However, one of the purposes of this study is to present a pipeline that can be adopted in the future to different datasets, creating models that depict the interplay of metabolism and DNA methylation in other biological contexts, for most of which such experimentally measured metabolite uptake/secretion rates are unknown. Therefore, we assessed whether GECKO models without constraints on exchange metabolite uptake rates could be enough to make accurate predictions. Although an enzymatic pFBA with GECKO models in which the only constrain was the limitation of the protein pool uptake (with estimated cell-specific total protein concentrations) provided smaller correlations, it predicted more fluxes in close agreement with measured values (69-77% in Figure 3.4) than both the traditional loosely constrained GSMMs (1-3% in Figure 3.2) and those constrained with the three exometabolites uptake rates (65-74% in Figure 3.3). Robinson's strategy (using tINIT) was the reconstruction approach that gave the best results with enzyme-constrained models (Figure 3.4-A,B versus C,D) and the inclusion of tasks was detrimental (Figure 3.4-A,C versus B,D). Hence, subsequent simulations were performed with GECKO models reconstructed with Robinson's approach and excluding reactions necessary for tissue-specific tasks.

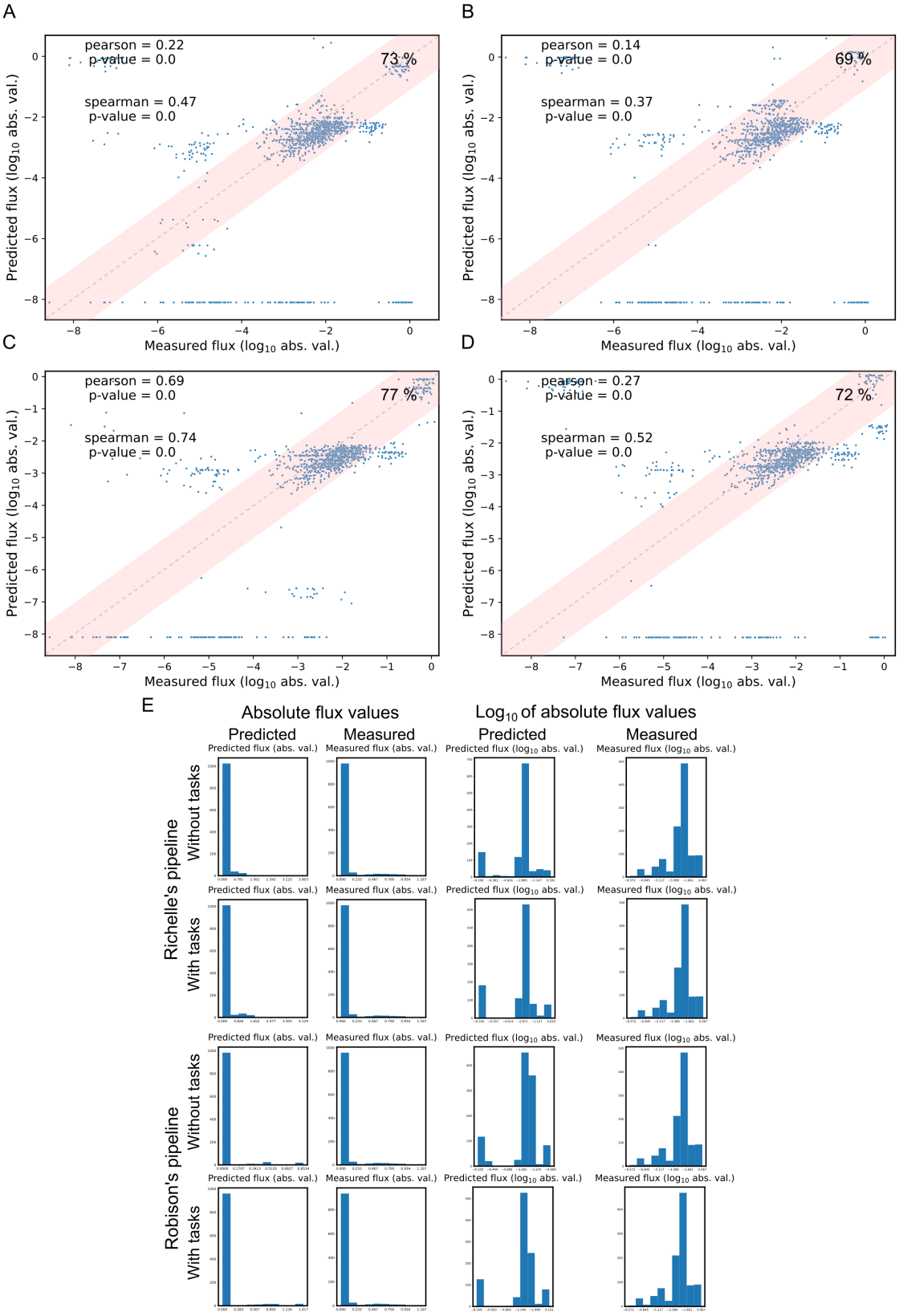

Figure 3.4: Comparison of measured and simulated exchange fluxes produced by GECKO models limited by total protein concentration *(continues)*.

Figure 3.4 *(continued from previous page)*: **A-D** are scatter plots with $\log_{10}$ of absolute values of simulated and measured fluxes of exchange reactions of 26 metabolites. Value at top right of each graph is the percentage of data points that are inside the pink area (where $\log_{10}(|\text{predicted value}|)$ is within $\log_{10}(|\text{measured value}|) \pm 1$). Either Richelle's pipeline using FASTCORE (**A**, **B**) and Robinson's pipeline using tINIT (**C**, **D**) were applied to reconstruct the models employed in the simulation. The effect of the integration (**B**, **D**) or not (**A**, **C**) of all tissue-specific metabolic tasks in those models was also assessed. **E**: histograms with the distribution of absolute values of measured and simulated fluxes before and after logarithmization. Data points forming a line at the bottom of **A-D** correspond to metabolites with a predicted flux of zero, which are shown in the graphs as holding the lowest absolute measured value (besides zero), as the logarithm of zero is undefined. In the distributions of logarithmized absolute values (E), the values of those metabolites fall in the lowest bin, creating an oddly tall bin at the beginning.

Even though 100% of $\log_{10}(|\text{biomass flux}|)$ values predicted with GECKO models lay within $\pm$ 1 of $\log_{10}(|\text{measured value}|)$ (Figure 3.5) and the relative error in prediction of growth rates is in agreement with previously reported values for eleven of the NCI-60 cell lines [75], there is no significant correlation (p-value > 0.05) between simulated and real biomass flux values as the simulated values were underestimated (i.e. most data points are beneath the diagonal line in Figure 3.5-A-D).

One possible explanation for this is the assignment of default values to two parameters influencing the limitation given to the total protein uptake flux. Those parameters are σ, which accounts for the level of enzyme saturation *in vivo*, and f, the mass fraction of enzymes that are accounted for in the model out of all proteins present in the cell. These parameters can change with the cell type and are unknown for NCI-60 cell lines. Another factor that could have contributed to the underestimation of biomass flux is an incorrect assessment of the real value of total protein concentration.

In fact, when biomass fluxes together with the total protein concentration were constrained with experimental values and an FBA with minimization of total protein uptake reaction was performed on GECKOs reconstructed with the best strategy (Robinson's pipeline and without tasks), only two models were feasible, reinforcing that the aforementioned parameters or total protein concentrations are not correct. Therefore, we did a similar simulation where biomass fluxes were limited with experimentally measured flux rates, but without limiting the total enzyme pool uptake rate.

As expected, the limitation of the biomass flux with bounds determined

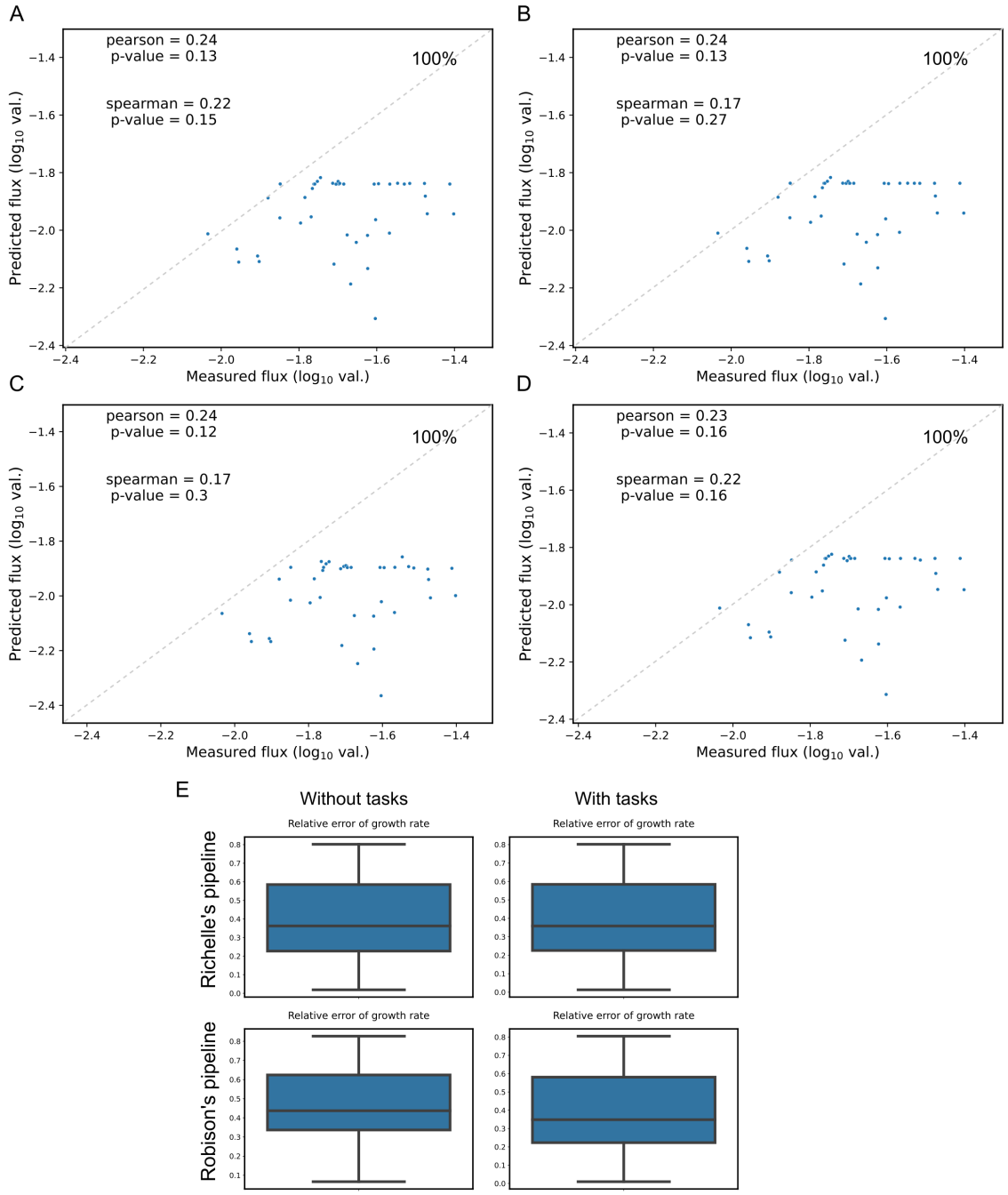

Figure 3.5: Comparison of measured and simulated growth rates produced by GECKO models limited by total protein concentration. **A-D** are scatter plots with $\log_{10}$ of values of simulated and measured growth rates. Value at top right of each graph is the percentage of data points where $\log_{10}(|\text{predicted value}|)$ is within $\log_{10}(|\text{measured value}|) \pm 1$. Either Richelle's pipeline using FASTCORE (**A**, **B**) and Robinson's pipeline using tINIT (**C**, **D**) were applied to reconstruct the models employed in the simulation. The effect of the integration (**B**, **D**) or not (**A**, **C**) of all tissue-specific metabolic tasks in those models was also assessed. **E**: Relative errors of predicted growth rates.

from experimental values lead the biomass simulated flux to be closer to the mean measured flux (Figure 3.6-B versus Figure 3.5-C), improving the relative error of the growth rate (Figure 3.6-D versus Figure 3.5-E). Regarding the fluxes of the 26 exometabolites, the restriction of growth rates gave as good results as without the constraints on biomass (Figure 3.6-A versus Figure 3.4-C). Furthermore, the percentage of simulated flux values within close proximity to measured ones in GECKOs with a constraint on biomass (77% in Figure 3.6) is higher than with traditional GSMMs with a constraint on biomass (71% in Figure B.3). Hence, subsequent simulations were done with GECKO models reconstructed with Robinson's approach and limited by experimental growth rates while minimizing the total enzyme usage.

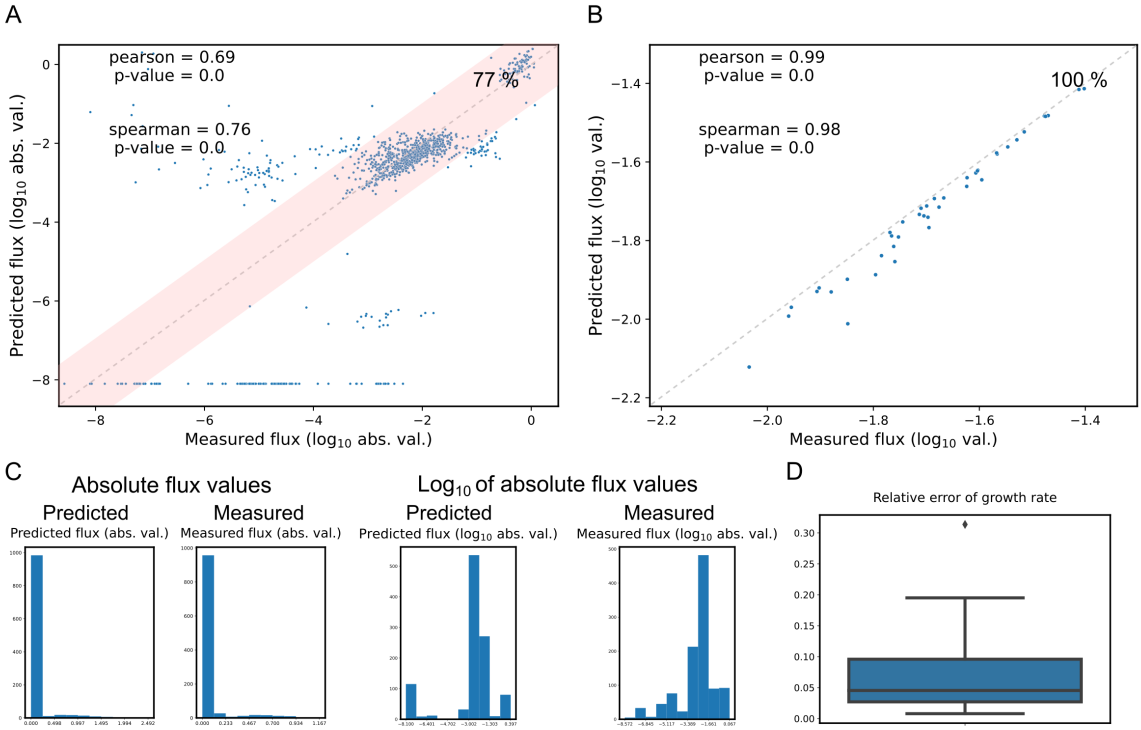

Figure 3.6: Comparison of measured and simulated exchange fluxes produced by GECKO models constrained with measured growth rates. **A**: a scatter plot with $\log_{10}$ of values of simulated and measured fluxes of exchange reactions of 26 metabolites. Value at top right of **A** and **B** is the percentage of data points where $\log_{10}(|\text{predicted value}|)$ is within $\log_{10}(|\text{measured value}|) \pm 1$ (in **A**, it corresponds to the pink area). **B**: a scatter plot with $\log_{10}$ of values of simulated and measured growth rates. The correlation coefficients are not exactly one, because experimentally determined upper and lower bounds were used to constraint simulated biomass fluxes, while the value of the measured biomass in the graph is the average of those bounds. **C**: histograms with the distribution of absolute values of measured and simulated fluxes before and after logarithmization. **D**: Relative errors of predicted growth rates. Data points forming a line at the bottom of **A** correspond to metabolites with a predicted flux of zero, which are shown in the graphs as holding the lowest absolute measured value (besides zero), as the logarithm of zero is undefined. In **C**, the values of those metabolites fall in the lowest bin, creating an oddly tall bin at the beginning. Models used were reconstructed with Robinson's pipeline (using tINIT) and without tissue-specific tasks.

3.2.3 Integration of models with cell line-specific DNA methylation levels and generic DNA methylation flux rules

The overall degree of protein acetylation of different human cell lines has been previously predicted in a study using traditional GSMMs, in which the

simulated flux of a pseudo-reaction of global protein acetylation was shown to correlate with the amount of one type of histone acetylation mark that functions as an epigenetic regulator [244]. Therefore, in this study, we assessed whether an equivalent correlation could be observed between the simulated DNA methylation flux and the degree of DNA methylation estimated with experimental data (details on the estimation procedure in *Comparison of fluxes of reactions involved in DNA (de)/methylation and the degree of DNA methylation* section of Material and Methods).

Results in Figure 3.7-A demonstrated that no strong correlation was observed between the actual global DNA methylation level (details of its estimation in *Calculation of the composition of total DNA* section of Materials and Methods) and the simulated flux of the DNA methylation reaction (*MAR08641*) or of the reaction that produces the cytoplasmatic SAM (*MAR03875*), which is one of the substrates of DNA methylation. Note that the values in the scatter plots were logarithmized because many of the simulated flux values of reactions *MAR03875* and *MAR08641* were close to zero (Figure 3.7-B). Only a weak, but significant (p-value ≤ 0.05) correlation, with just the Spearman (not with the Pearson) method, was observed between the flux of each mentioned reaction and the global DNA methylation. The genomic region which gave the best significant correlations for the DNA methylation reaction (*MAR08641*), although still weak (0.36 and 0.52 of Pearson and Spearman coefficients respectively), was the one comprising 1000bp upstream of the genes' Transcription Start Sites (TSS), i.e. gene promoters (see reaction *MAR08641* in Figure 3.7-A). In addition, it was observed that the overall correlation values are slightly higher for the DNA methylation reaction (*MAR08641*) than for the one producing SAM (*MAR03875*).

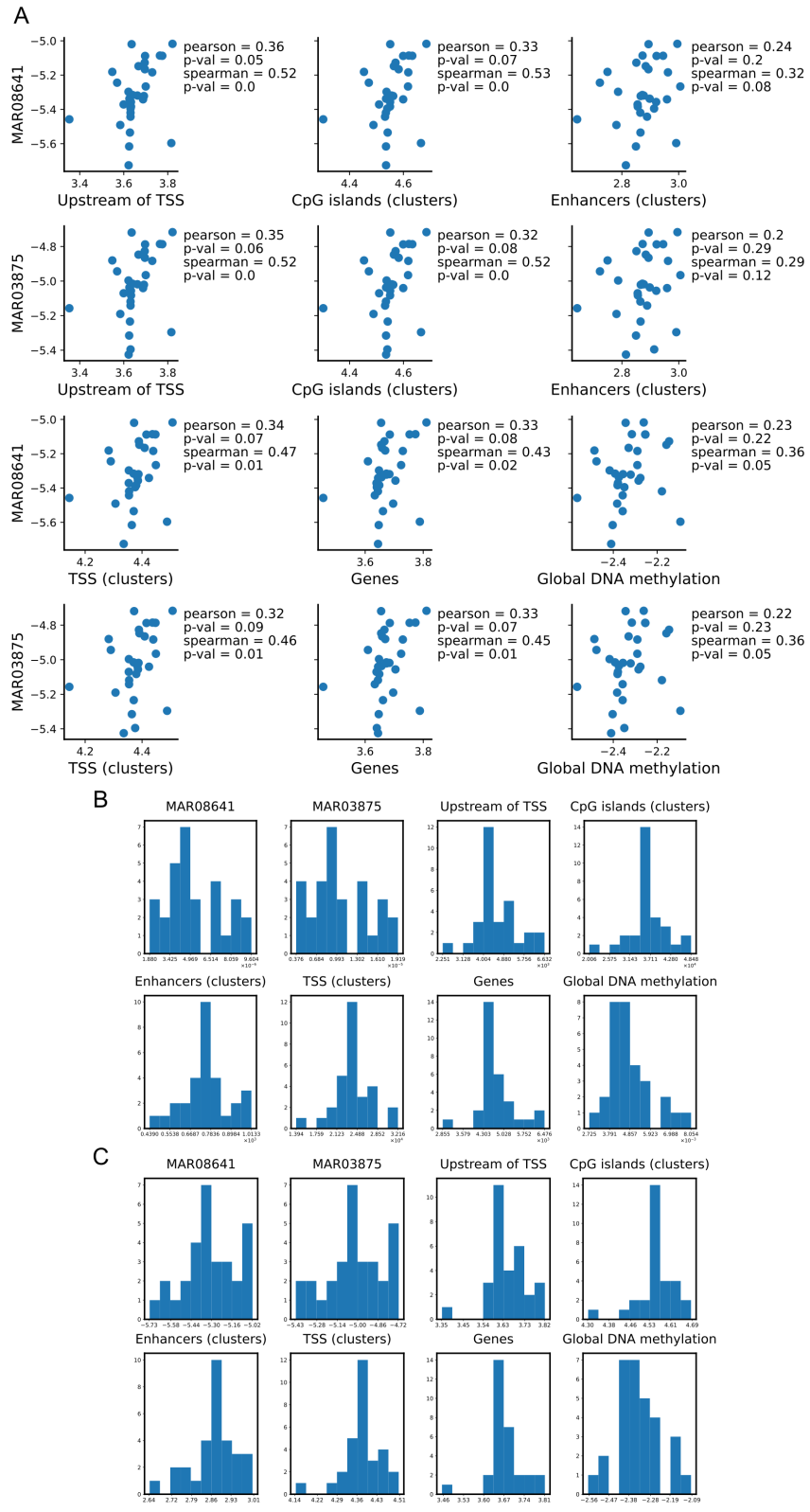

Figure 3.7: Comparison of simulated fluxes of reactions involved in DNA methylation and the estimated degree of DNA methylation. **A**: scatter plots with $\log_{10}$ values of simulated fluxes versus the experimentally estimated degree of methylation across all genome or in close proximity to different genomic features *(continues)*.

Figure 3.7 *(continued from previous page)*: The procedure and the description of datasets applied to estimate the level of DNA methylation is detailed in *Comparison of fluxes of reactions involved in DNA (de)/methylation and the degree of DNA methylation* section of Materials and methods. The vertical labels are identifiers of reactions shown in Figure 3.1. Zero flux values were replaced with a very small value (1×10^{-15}) because $log_{10}(0)$ is undetermined. **B**, **C**: histograms with the distribution of simulated flux values and estimated DNA methylation levels before (**A**) and after (**B**) logarithmization. Only the 30 cell lines for which there was experimental data across all types of genomic intervals were here used.

Since the simulated methylation fluxes were not able to strongly predict the degree of DNA methylation, we switched our focus to understanding how metabolic mechanisms and metabolic shifts are related to the overall degree of DNA methylation in cancer, which is the ultimate goal of the present work. For that purpose, the degree of DNA methylation across the genome was integrated with the models. Specifically, the stochiometric coefficients of the pseudo-reaction *prodDNAtot*, which represents the composition of total DNA in terms of DNA cytosine (de)/methylation marks, were modified based on published cell/tissue-specific DNA5mC and DNA5hmC datasets (more details on *Calculation of the composition of total DNA* section of Materials and Methods). Note that values of DNA5hmC levels were not available for all cell lines. Hence, from then on simulations were made with models for only 31 of the 41 cell lines.

Another interesting observation from the previous simulations is that while the DNA methylation reaction always carried flux for any of the different cell lines, none of the DNA demethylation reactions included in the model was able to do the same, because the formation of unmethylated DNA necessary for the biomass production was directly obtained through the DNA polymerization reaction of the individual nucleotides instead of the DNA demethylation reactions. However, it is known that the extent to which the DNA is methylated depends on the balance between the rates of the reactions of methylation and demethylation, which produces a dynamic DNA methylation turnover steady-state [245]. In fact, variations in the proportion of those rates can originate methylation deregulation like the hypermethylation (i.e. silencing) of tumor suppressor genes and hypomethylation (i.e. activation) of pro-metastatic genes observed in cancer cells [245–247]. Hence, to guarantee that the simulations can reflect the dynamic DNA methylation turnover state, the flux of certain DNA demethylation reactions was forced to be positive in the subsequent simulations by constraining those reactions in each model with reaction rate ratios

previously described in the literature, as long as the imposed constraints would not produce an infeasible flux distribution. Those rate ratios are described in Table D.8 and will herein forth be called *methylation flux rules*.

As would be expected, there was an improvement in the correlation between the simulated fluxes of reactions *MAR08641* or *MAR03875* and the estimated degree of DNA methylation after adapting the composition of total DNA with cell-specific information (compare Figure 3.8 with Figure 3.7). This time, the correlation between the DNA methylation flux (*MAR08641*) and the global DNA methylation was strong (0.62 and 0.73 of Pearson and Spearman coefficients respectively) and significant (p-value ≤ 0.05), and the *Upstream of TSS* was again the genomic feature that gave the best correlations (Figure 3.8). Also, the integration of the abovementioned *methylation flux rules* for models where their inclusion provided feasible flux distributions enabled the activation of some DNA demethylation reactions in those models, whereas neither a positive nor negative correlation was observed between simulated fluxes of any DNA demethylation reaction and the degree of DNA methylation (Figures B.4 and B.5). Note that, although the number of models used was lower than in previous simulations because the cell-specific methylation ratios could only be estimated for 31 cell lines, the correlations between measured and simulated fluxes and the percentage of simulated flux values in close proximity to the real ones for exchange reactions or biomass, as well as the relative errors of biomass were as good as without the cell-specific methylation ratios and the *methylation flux rules* (Figure B.6 versus Figure 3.6).

3.2.4 Analysis of active pathways and protein usage in cell-line-specific models

After model reconstruction and integration with methylation data, the results from simulated flux value distributions of each cell line were analyzed. From Figure 3.9-A, it was possible to observe that among the central carbon metabolism pathways, *Glycolysis or Gluconeogenesis* is the one carrying the most flux for all cell lines, while the flux through the pathways directly or indirectly related with DNA (de)/methylation (*Folate metabolism*, *Cysteine and methionine metabolism* and *DNA methylation or demethylation*) is very low. However, simulations suggest that these cell lines utilize a higher mass of enzymes to activate reactions associated with *Cysteine and methionine metabolism* than those involved in *Glycolysis or Gluconeogenesis* (Figure 3.9-B). Furthermore, the top five pathways with the most flux were identified as *Glycolysis or Gluconeogenesis*, *Oxidative phos-*

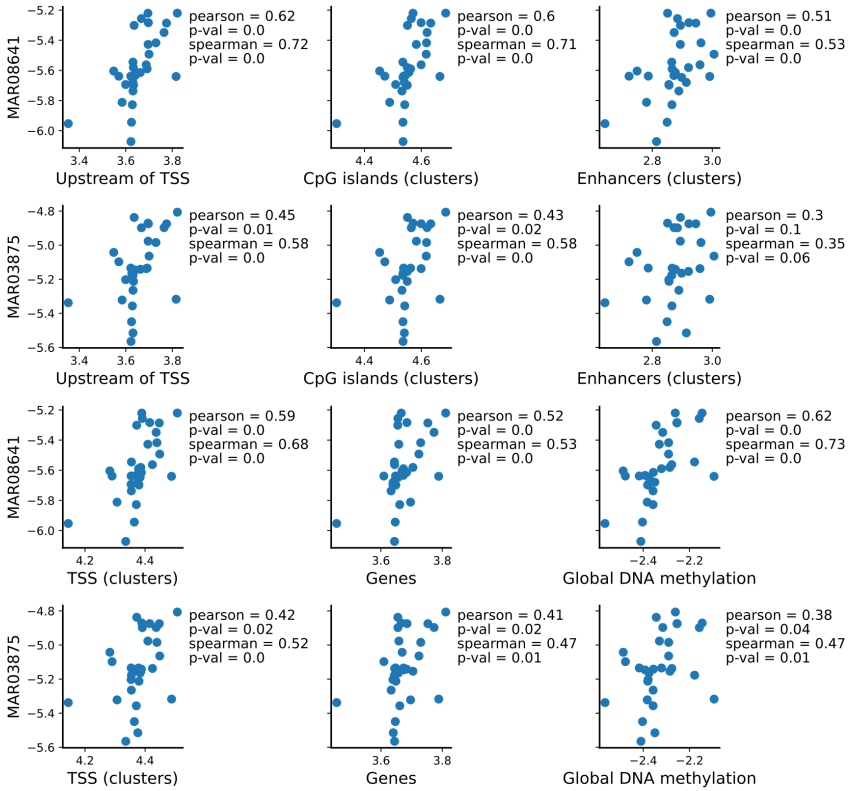

Figure 3.8: Comparison of simulated fluxes of reactions involved in DNA methylation and the estimated degree of DNA methylation for models with *methylation flux rules* and cell-specific methylation ratios. Scatter plots with $\log_{10}$ values of simulated fluxes versus the experimentally estimated degree of methylation across all genome or in close proximity to different genomic features. The procedure and the description of datasets applied to estimate the level of DNA methylation is detailed in *Comparison of fluxes of reactions involved in DNA (de)/methylation and the degree of DNA methylation* section of *Materials and methods*. The vertical labels are identifiers of reactions shown in Figure 3.1. Zero flux values were replaced with a very small value (1×10^{-15}) because $log_{10}(0)$ is undetermined. Only the 30 cell lines for which there was experimental data across all types of genomic intervals were here used.

phorylation, *Purine metabolism*, *Fatty-acid biosynthesis (even-chain)* and *Aminoacyl-tRNA biosynthesis*. With respect to protein mass, the top five scoring pathways are *Cholesterol biosynthesis 2*, *Glycerophospholipid metabolism*, *Cholesterol biosynthesis 1 (Bloch pathway)*, *Acylglycerides metabolism* and *Aminoacyl-tRNA biosynthesis* (Figure 3.10).

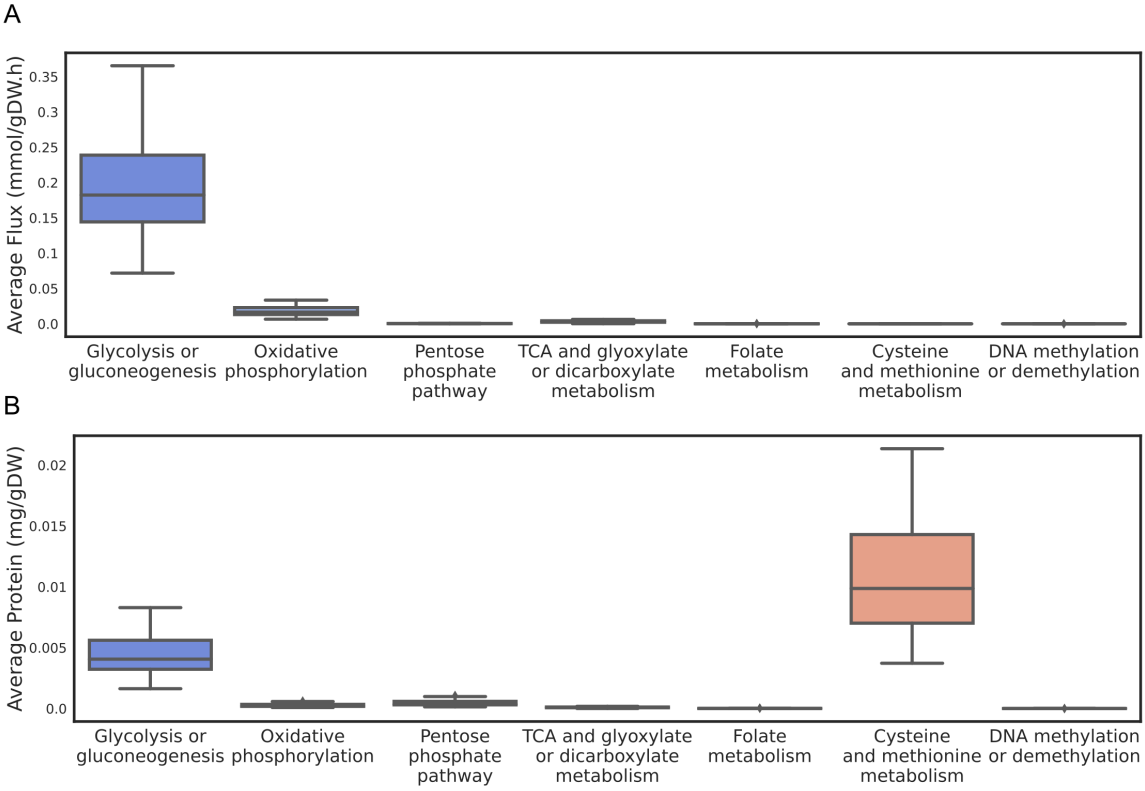

Figure 3.9: Flux values and protein usage in pathways related with central carbon metabolism and DNA (de)/methylation. **A**: Boxplots show flux values across all cell lines. **B**: Boxplots show amount of protein spent across all cell lines.

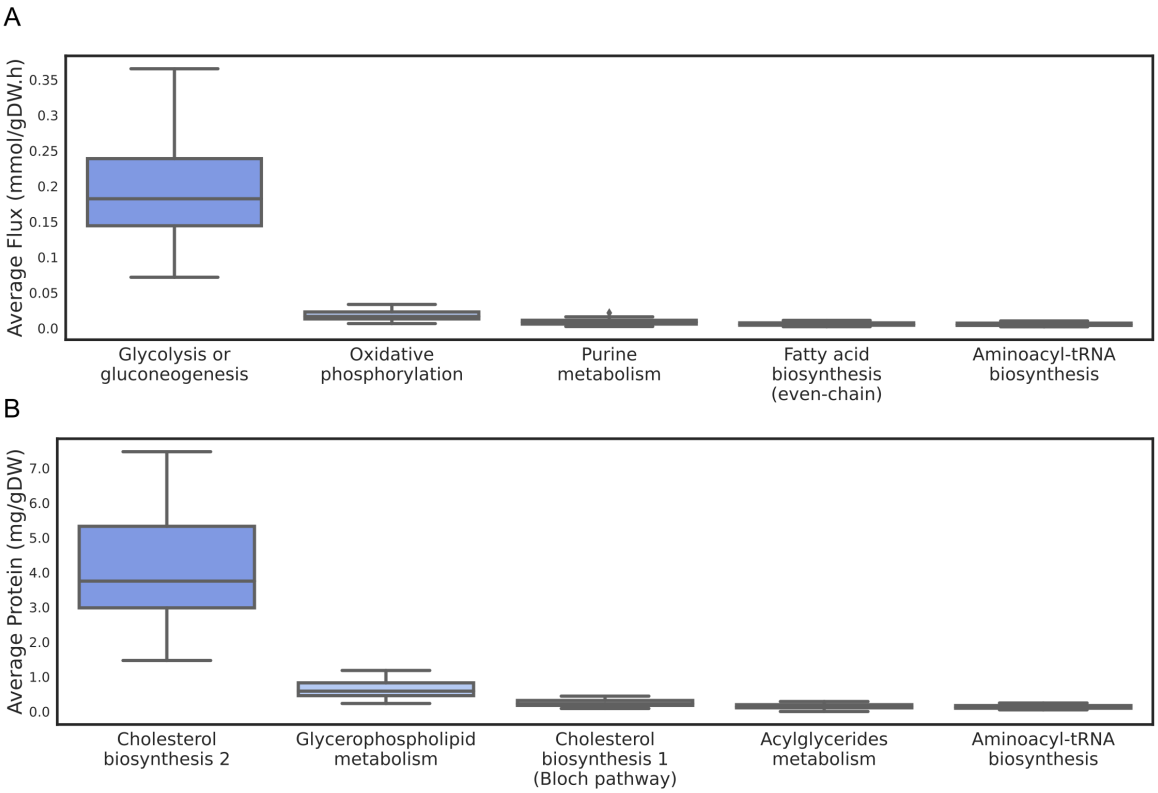

Figure 3.10: Top five pathways with highest flux values or protein usage. **A**: Boxplots show flux values across all cell lines. **B**: Boxplots show amount of protein spent across all cell lines. Top five pathways were selected based on the median across all cell lines of average values of each pathway.

3.3 Discussion

To the best of our knowledge, this is the first time that models were built to simulate the interaction between metabolism and DNA methylation. Past studies have tackled the interplay between metabolism and epigenetics, but most focused on histone acetylation [241, 242, 244]. Besides, the only study covering methylation [243] dealt with methylation of histones on murine cells and used the flux of the substrate for methylation (SAM) as a surrogate for methylation, instead of introducing protein methylation reactions.

In this study, we developed models for 31 cancer cell lines of Human that capture the interaction between metabolism and DNA methylation, through the integration of reactions related to DNA methylation and demethylation obtained from a review of information deposited in databases and reported in the literature, and the integration of cell-type specific DNA methylation levels.

The included reactions were curated to guarantee mass and charge balance and their integration with the remaining metabolic network of the generic model *Human1*.

Furthermore, past studies trying to mimic the relationship between metabolism and epigenetics utilized traditional GSMMs, whereas the present study also employs GECKO models. When no metabolite uptake rates are known traditional constraint-based GSMMs cannot provide a finite flux distribution, because the solution is unbounded. For example, if uptake reactions of nutrients have bounds from -∞ to zero (in a model including reversible reactions) and the metabolic objective is to maximize biomass flux, the maximum growth rate will be infinite [75]. High finite values can be assigned to bounds of those exchange reactions (e.g. $\pm$ 1000) to obtain flux distributions that enable the assessment of flux allocation to distinct metabolic pathways [126, 240], the identification of metabolic tasks [92], or detection of essential reactions and genes [126, 240]. However, such models perform poorly in a quantitative evaluation of flux values, i.e. simulated fluxes are different from the experimentally obtained fluxomics data [126]. Therefore, to improve the predictability of these models, it is important to constrain exchange fluxes, with real flux values [75], which are not always available. On the other hand, GECKO models do not require the definition of specific uptake rates to provide accurate flux distributions, because they introduce enzyme kinetic and concentration data that is sufficient to reduce the flux solution space.

Even though uptake/production rates of some exometabolites were available and therefore could be used to make acceptable predictions with traditional GSMMs, the use of GECKO models provided better results than traditional GSMMs constrained with flux values of three metabolites (previously reported as sufficient to generate small growth rate prediction errors in eleven of the same cell lines), as the percentage of simulated fluxes with values close to the measured ones was higher with GECKO models than with the traditional models constrained with the three fluxes (compare percentages between Figure 3.4 and Figure 3.3). Additionally, the use of cell-specific GECKO models is useful in presenting a proof-of-concept pipeline that can be adopted in the future to other datasets for the production of models portraying the interplay of metabolism and DNA methylation in other biological contexts, for most of which such experimentally measured uptake/secretion rates may be unknown. Also, note that even though a constraint, the limitation of biomass flux with measured growth rates, had to be eventually applied to GECKO models, that still requires gathering less experimental data than when applying measured fluxes of three different

metabolites (as in constrained traditional models).

Interestingly, the assessment of the best model reconstruction pipeline in this study suggested that the optimal strategy to apply depends on the decision to either include or exclude experimental flux information, as the optimal approach for traditional GSMMs constrained with specific flux values was tINIT without cell line-specific task integration, while for loosely constrained fluxes the approach providing a higher percentage of matching values was FASTCORE, which gave the highest correlation values when combined with the non-integration of reactions essential for cell line-specific tasks.

In this study, the stoichiometric coefficients of modified methyl-cytosines in the reaction representing the total DNA composition were adapted in accordance with experimentally measured DNA methylation levels specific to each cell line, so that the simulations could predict metabolic phenotypes associated with DNA methylation that were cell-type specific. Naturally, one of the consequences of this adaptation was the increase in correlation between fluxes of the reactions of DNA methylation and of the production of SAM and the experimentally observed DNA methylation levels. Notably, those correlations were higher in the region *Upstream of TSS* of genes (i.e. gene promoters) than in other genomic regions, including regions surrounding upstream and downstream of the TSS (the *TSS (clusters)*), suggesting that the variation in global DNA methylation across these cell lines can be in its majority attributed to alteration of methylation in gene promoters. Given that Ghandi et al. [248] observed a negative correlation between gene expression and promoter methylation of many genes in these cell lines, it is possible to hypothesize that variation in global DNA methylation across cell lines might primarily be associated with variations in down-regulation of gene expression. Furthermore, another study with the same cell lines reported that the most significant correlation between DNA methylation and gene expression was an inverse correlation for epithelial and mesenchymal genes, more expressive for the former than the latter (-0.639 versus -0.525 correlation), and although to a smaller degree, with tumor suppressors as well [249]. This suggests that the increase in global DNA methylation across different cell lines is mostly associated with endothelial-to-mesenchymal transition and silencing of tumor suppressor genes, and therefore, it would be expected that the proliferation ability of these cell lines increases with the global DNA methylation levels. To test this, we compared the growth rate of the cell lines and their global DNA methylation levels and, in fact, a significant but moderate correlation could be observed Figure B.7.

Another important observation was that the inclusion of the aforementioned

methylation flux rules based on reported values of ratios of DNA (de)/methylation reaction rates in the model simulations guaranteed flux through the DNA demethylation reactions at least in some models. This is relevant because it is known that there is a dynamic DNA methylation turnover steady-state, where both methylation and demethylation reactions are active [245]. Nevertheless, the *methylation flux rules* could only be applied to some of the cell lines without affecting the feasibility of the flux distribution, maybe because the metabolite pools (for e.g. protein pools) are not adapted to the cell type or because those flux rules are general rules that might not apply to all cell lines. In that sense, future *in silico* studies could try to integrate more cell-detailed experimentally determined flux rules as soon as they become available.

The final models here created, which represented 31 different cell lines, provided a high percentage of simulated flux values in close proximity to corresponding experimental values (around 77% in Figure B.6) for exchange reactions of 26 metabolites, which in itself serves as good validation criteria. Furthermore, the flux and protein mass distribution across metabolic pathways agreed with the reported experimental evidence.

Glycolysis or Gluconeogenesis was the subsystem among the central carbon metabolism pathways to carry more flux, and it was the most active metabolic pathway. This is expected because aerobic glycolysis is one of the characteristics of cancer cells, as it allows them to quickly obtain energy and elementary metabolites for fast growth [13, 16].

Although Glycolysis is more active than *Oxidative phosphorylation* across the models of all cell lines, as expected in cancer cells, the latter is still the second pathway with the most flux. A possible explanation for this is that even though cancer cells prefer in general aerobic glycolysis, they use Oxidative Phosphorylation (OXPHOS) to produce at least some level of Reactive Oxygen Species (ROS), as a moderate amount of ROS is beneficial for tumorigenesis, resistance to chemotherapy and cancer progression [25]. Another putative reason could be the increased levels of citrate in cancer cells, due to the conversion of α-KG to citrate induced by oncogenes, which indirectly would allow the production of some energy through the *Oxidative phosphorylation* by feeding the Tricarboxylic acid cycle (TCA cycle) [16]. However, in that case, it would be expected that the TCA cycle was also among the most activated pathways, which is not the case.

In addition, the fact that *Purine metabolism*, *Fatty acid biosynthesis* and *Aminoacyl-tRNA biosynthesis* are the third, fourth and fifth pathways with the most flux in the models could also be anticipated because the first pathway is

necessary for the synthesis of nucleotides fundamental for the production of new DNA molecules, the second is essential for the formation of new cell membranes and the last allows protein production through mRNA translation, all of which are important factors for fast-dividing cells.

The first and third pathways to use most protein mass are related to *Cholesterol biosynthesis*, which is also in accordance with the literature. Cholesterol biosynthesis is often enhanced in cancer as on one hand, it can activate mTORC1 signaling, which in turn promotes cell proliferation, invasion, and metastasis, while on the other it alters lipid rafts composition, promoting the loss of integrin-mediated cell adhesion, and consequently contributing to cancer aggressiveness [41, 251]. The second and fourth pathways to use most protein mass, the *Glycerophospholipid metabolism* and *Acylglycerides metabolism*, are expected to be activated in cancer cells as well because those metabolites are part of new cell membranes needed for intense cell proliferation.

With respect to pathways directly or indirectly related to DNA (de)/methylation, the flux is reduced. This is expected, since it is possible to anticipate that a small DNA methylation rate is enough to methylate less than 1% of the genome (the average percentage of methylation of the human genome). Nevertheless, the use of protein mass in one of those pathways, *Cysteine and methionine* metabolism is elevated, even more than in glycolysis, suggesting that although holding a small amount of flux, it is an important pathway.

The models developed in this study could have different applications. One possible use is to identify metabolic pathways in which variation in flux or protein usage follows the change in global DNA methylation levels across the different cell lines, to identify metabolic shifts that could explain the observed variation in DNA methylation. Another related application could be the selection of a set of reactions whose flux variations across the different cell lines are inversely correlated with the flux of biomass reaction. The set of genes mapped to the identified reactions would then be intersected with a list of genes in which promoter methylation accompanies cell growth. This way, it would be possible to find genes that upon methylation would promote metabolic shifts (silence metabolic pathways) necessary for cancer growth. Furthermore, the same generic model and reconstruction pipeline validated in this study could be applied to other datasets to understand how the interaction between metabolism and DNA methylation explains other biological questions. For example, the reconstruction of models of cancer cells in the presence versus absence of epigenetic modulators traditionally used in cancer treatment, or even to study the effect of metabolic diseases, like obesity, insulin resistance, or dyslipidemia in

DNA methylation [252].

3.4 Materials and Methods

3.4.1 Creation of the generic DNA methylation model

Alterations were made to the model Human1 (version 1.12) to create a generic model depicting DNA methylation and demethylation, which was made available to the public. Overall, the reaction of DNA methylation in the cytoplasm was removed (as there is no DNA in the cytoplasm), and the gene rule of the equivalent reaction in the nucleus was updated (explanation in Table D.4). Reactions and corresponding GPR rules involved in DNA (de)/methylation were obtained by literature curation (see explanation in Table D.1 and Table D.4). Some transport reactions were added, and two reactions that occur when cytosine is inside the DNA were assumed to take place also when it is in its monomeric form (*consdirect5fC* and *consdirect5CaC* reactions in Table D.1), to guarantee flux through the remaining DNA (de)/methylation reactions (i.e. to prevent their blockage). Metabolites taking part in the added reactions are described in Table D.5. A pseudo-reaction representing the average total human DNA composition in the nucleus (*prodDNAtot*) in terms of DNA cytosine (de)/methylation marks, like DNA-5-methylcytosine (DNA5mC), DNA-5-hydroxymethylcytosine (DNA5hmC) and DNA-5-formylcytosine (DNA5fC) was introduced (how the composition was determined is shown in Tables D.2 and D.3). Moreover, the generic biomass reaction was replaced by a similar reaction (*adaptbiomass* reaction) where the DNA was changed into the pseudo-metabolite (DNAtot) representing the total DNA harboring all DNA methylation and demethylation marks (in Table D.1). All introduced reactions were corrected for charge and mass balance (Table D.6). All blocked reactions and associated genes and metabolites were removed.

3.4.2 Reconstruction of cell line-specific traditional GSMMs

Cell line-specific GSMMs were built for different NCI-60 cell lines through the integration of transcriptomics data from Cancer Cell Line Encyclopedia (CCLE) deposited in DepMap repository in 2019. This version of the transcriptomics dataset was chosen because it was produced by the same study, Ghandi et al. [248], which the Reduced-Representation Bisulfite Sequencing (RRBS) dataset used in the present work was retrieved from. The approach presented in Richelle et. al [92] to build cell type-specific models was here implemented

in Python and made available to the public. Gene scores were determined from cell-specific gene expression data using the following expression:

$$genescore = 5 * log\left(1 + \frac{expression\ level}{threshold}\right) \tag{3.1}$$

where the threshold is the mean value of gene expression over all samples unless it is lower than the 25^{th} or higher than the 75^{th} percentiles of the gene expression value distribution, in which cases the threshold value is considered to be the same as the mentioned percentiles.

Reaction scores were subsequently calculated from gene scores taking the GPR rules into account so that: the score of a reaction catalyzed by an enzyme complex was the minimum score of all genes associated with the complex (*AND* rule) and that of a reaction catalyzed by isozymes was the maximum score of all genes encoding the isozymes (*OR* rule). The highest reaction scores of a cell line were attributed to the reactions considered necessary for the generic metabolic tasks (a.k.a. essential metabolic tasks), while the same procedure was applied to necessary reactions of other tasks if those are performed in that specific cell type. This was done to give the reconstruction algorithm a higher probability of building a cell type-specific model that can pass all the generic tasks and tasks specific to that cell type. A reaction was considered necessary for a task if it was carrying flux upon the inclusion of the task-associated flux constraints on the generic model followed by a minimization of the sum of all fluxes. In order to determine whether a task was done in a certain cell type, a metabolic score consisting of the average of the scores of the reactions (previously identified as) required for that task was calculated. When the task metabolic score was higher than $5 * log(2)$ (the gene/reaction score to which the expression level is equal to the aforementioned threshold) the task was considered to be done in that specific cell type. Then, the FASTCORE algorithm from Troppo package [232] was run to obtain the cell type-specific models. Note that three DNA demethylation tasks (each one corresponding to a distinct demethylation pathway) were created and added to the original list of tissue-specific tasks so that each final cell-line model could have all reactions necessary for the DNA demethylation pathway done by that specific cell type. We also included two DNA demethylation-associated reactions (*consdirectDNA5fC* and *consdirectDNA5CaC* in Figure 3.1) that were non-necessary for enzyme-catalyzed DNA demethylation to occur (demethylation could happen in the generic model without them) because they were not associated with any gene (were not catalyzed by an enzyme), and therefore were always excluded from the reconstructed models (due

to the lack of associated reactions scores) although they could happen without the presence of enzymes. For comparison purposes, an equivalent analysis was done without including the reactions necessary for cell type-specific tasks.

The second approach to building context-specific models consisted of the application of a version of the tINIT algorithm run in MATLAB that had already been implemented by Robinson et al. [75]. That version tries to include reactions with scores above a threshold while removing those below the threshold and keeping the connectivity of the model (i.e. making sure all reactions carry flux). The gene-to-reaction scores conversion strategy applied was again the minimum and maximum of gene scores for complexes and isozymes, respectively [75]. The final models also kept the ability to perform the generic metabolic tasks, which consisted of the previously reported 57 essential metabolic tasks [75]. Furthermore, reactions previously identified as necessary only for DNA demethylation tasks (in Richelle's approach using FASTCORE) were included after reconstruction if the task metabolic score was above the threshold for the particular tissue. The two non-catalyzed DNA demethylation-associated reactions *consdirectDNA5fC* and *consdirectDNA5CaC* were also introduced in all models due to the reasons explained above for Richelle's approach. When testing the inclusion of tissue-specific tasks with this approach, the reactions needed for tissue-specific tasks were included as well.

3.4.3 Generation of cell line-specific GECKO models from traditional GSMMs

GECKO models were created from traditional GSMMs using a MATLAB script produced by Robinson et al. [75], which pipeline was first described in Sanchez et al. [137]. In that pipeline, enzymes are introduced as pseudo-substrates in the reactions they catalyze and the stoichiometric coefficients are the inverse of the turnover numbers of the corresponding enzyme-metabolic substrate pairs. Reversible reactions are split into two irreversible reactions in opposite directions, and isozymes are separated into different reactions, each catalyzed by one of the isozymes. Furthermore, for each original un-split isozyme-associated reaction, a new pseudo-reaction, named *arm* reaction, is added. The only product of an *arm* reaction is an intermediary pseudo-metabolite which is used as a substrate by each of the isozyme-split reactions so that the flux bounds of each original un-split reaction can still be applied. For reactions catalyzed by complexes, each enzyme of the complex is introduced as a substrate and the stoichiometric coefficient is in that case the product of the inverse of

its turnover number and the stoichiometric coefficient of the enzyme inside the complex.

Finally, supply reactions for each enzyme known as protein *draw* reactions, are added, where each reaction consumes a proportion (based on the enzyme molecular weight) of a total protein pool, which in turn is supplied by another included boundary reaction called *protein pool exchange* reaction [75, 137].

3.4.4 Detection and validation of the best reconstruction and simulation pipelines

For the assessment of the best type of models (traditional GSMMs or GECKOs) and the selection of the most suitable model reconstruction and simulation strategies, values of simulated uptake/secretion rates of 26 metabolites across different cell lines were compared with corresponding experimentally measured ones originally obtained from Jain et al. [253]. The 26 metabolites chosen were the ones previously utilized to validate the reconstruction of eleven NCI-60 cell lines in Robinson et al. [75] article. The same comparison was made between the simulated fluxes of biomass reaction and measured rates of cell growth retrieved from Zielinski et al. [254]. The strategies and model types giving the best percentage of simulated flux values in close proximity to measured ones were identified. For simulations with traditional GSMMs, parsimonious Flux Balance Analyses (pFBAs) were carried out whereby the minimization of the sum of all fluxes took place after constraining the biomass flux with either, the objective value of an FBA whose metabolic objective was maximization of biomass flux or with experimentally measured growth rates. For GECKO models, the simulations were accomplished through either an enzymatic pFBA where the minimization of the flux of the total protein uptake reaction followed the maximization of the flux of biomass reaction, or by an FBA whose metabolic objective was to minimize the total protein uptake upon limitation of the flux of biomass reaction with measured growth rates. The last strategy was selected to be applied to all subsequent simulations, as it was the one to give the best results. Model manipulation and simulation were done with the *MEWpy* [255] and *COBRApy* [234] python modules.

3.4.5 Calculation of the composition of total DNA

The overall composition of the total DNA in terms of modified-cytosine species involved in methylation and demethylation was initially estimated for a generic human cell based on general knowledge of the human genome (see

Table D.2), and the average level of DNA5hmC sites across different healthy human tissues (Table D.3) obtained from a Chemical-assistant C-to-T conversion of 5hmC sequencing (hmC-CATCH) experiment, which values were kindly provided by the authors of He et al. [256].

Those estimations were initially integrated into the total DNA composition reaction, the *prodDNAtot* reaction, of the generic model before the reconstruction of the context-specific models. However, simulations with the cell-line-specific models have later been performed with a cell-line-specific *prodDNAtot* reaction. The stoichiometric coefficient of the DNA5hmC in *prodDNAtot* reaction of a specific cell line was the estimated ratio of DNA with 5hmCs of the healthy tissue to which that cell line corresponds (calculation shown in Table D.3). The estimation of the stoichiometric coefficients of the remaining cytosine species was grounded on the results of a Reduced-Representation Bisulfite Sequencing (RRBS) experiment obtained from the same study that produced the transcriptomics data used in the model reconstruction [248]. The output of bisulfite sequencing is the proportion of all cytosines that have remained unconverted (i.e. were not converted to Uracil) during the bisulfite treatment, and it is generally used as a proxy for the ratio of cytosines that are methylated. However, in reality, not only DNA5mCs but also DNA5hmCs are not converted to Uracil [257] upon treatment with bisulfite, while aside from the fully unmethylated cytosines also the DNA5fCs are converted to Uracil [258] during the process. So, the assumption that the bisulfite sequencing signal is the ratio of DNA that is methylated could lead to imprecise estimations, as in reality, it represents the ratio (DNA5mC + DNA5hmC)/(unmethylated DNA-5-cytosine + DNA5fC). Fortunately, in this case, the ratio of DNA with DNA5hmCs and DNA5fCs could be calculated (from the hmC-CATCH results and literature, respectively), and therefore, there was no need to use the bisulfite sequencing signal directly as a proxy (calculations shown in formulas of excel Table D.7).

3.4.6 Comparison of fluxes of reactions involved in DNA (de)/methylation and the degree of DNA methylation

The correlation between simulated fluxes of important reactions involved in DNA methylation and demethylation and the overall level of DNA methylation across different cell lines (represented by the stoichiometric coefficients of DNA5mC calculated above) was assessed in this study. In addition to the global methylation state, the comparison to the methylation levels of CpGs at specific genomic regions was also analyzed. For this, the RRBS signal of CpGs within

1000 bp-length genomic intervals *Upstream of the TSS* of genes (gene promoters), retrieved from the same abovementioned RRBS study [248], was added across all genes for each cell line. The sum of the signal was then directly used as a proxy for the DNA methylation level, in contrast to the procedure applied above in the calculation of stoichiometric coefficients, because in this case, the number of CpGs containing DNA5hmC and DNA5fC within the particular genomic regions was unknown.

This same analysis strategy was further applied to other datasets of the same study where the level of methylation in methylation clusters (i.e. regions where CpG sites have similar methylation changes across different cell lines) was retrieved for genomic intervals centered around TSS (from 3000 bp upstream to 2000 bp downstream), CpG islands, and enhancers (from 2000 bp upstream to 2000 bp downstream).

Furthermore, the methylation level of genes was also roughly determined by adding the methylation values of all genes in each cell line from a dataset deposited at CellMiner [259] database which resulted from a DNA methylation array experiment [249]. In that case, the average gene methylation values (i.e. average of beta values) were given by the ratio of the intensity of the probes for methylated DNA and the intensity of all probes (those detecting methylated and unmethylated DNA) annotated to that gene.

3.4.7 Analysis of active pathways and protein usage

To compare the simulated flux distribution across the different metabolic pathways and cell lines, a generic GECKO model was first created from the generic traditional GSMM where each reaction was associated with a metabolic subsystem. The flux of all reactions of the same metabolic subsystem in each cell-line-specific model was added and divided by the number of reactions attributed to that subsystem in the generic GECKO model to correct for the bias that subsystems with more reactions have a higher chance to have more active reactions (and therefore carry more flux). Since each isozyme-associated reaction of a traditional GSMM is split into different reactions in a GECKO model (each associated with one of the isozymes) that consume the same pseudo-metabolite of an *arm* reaction (mentioned above), the flux of the *arm* reaction is the sum of the fluxes of the other split-reactions. Therefore, for reactions associated with isozymes only the *arm* reactions were considered in the analysis. Also, the abovementioned protein *draw* reactions and the *protein pool exchange* reaction were naturally excluded, as they were not associated with any subsystem.

For the estimation of protein usage in each metabolic subsystem, instead of directly quantifying the amount of enzyme spent in each *draw* reaction, the amount of protein used in each enzyme-reaction combination was calculated instead, because the same enzyme can participate in different reactions of distinct metabolic subsystems. The flux of each reaction (in mmol.gDW^{-1}.h^{-1}) was divided by the k_{cat} of each enzyme-reaction combination (in h^{-1}), and then multiplied by the molecular weight of the enzyme (in KDa, i.e. 1g.mmol^{-1}) and 1000, to obtain the amount of the enzyme used in the reaction (in mg/gDW). All reactions that do not use any enzyme as a pseudo-substrate (the *arm* reactions and non-catalyzed reactions) were excluded. Then, the sum of all protein usage values of each metabolic subsystem was divided by the number of enzyme-reaction combinations attributed to that subsystem in the generic GECKO model to correct for the bias that subsystems with more reactions and with reactions containing more enzymes have the tendency to use more protein.

Code Availability:

The code produced and models built in the present work are deposited at: `https://gitfront.io/r/user-9348496/UrviDdiW596W/epigen/`

Chapter 4

Conclusion

Two studies were presented in this thesis where GSMMs, traditional GSMMs and GECKO models, were reconstructed for specific tissues/cell-lines of human cancer cells so they could be applied to study the role that metabolism plays in cancer, extending the domains of metabolic modeling to other less studied areas, including cancer stem cells and epigenetics. In the next sections, we briefly review the main contributions of each study, but also their limitations and relevant future directions for research.

4.1 Reconstruction of Tissue-Specific Genome-Scale Metabolic Models for Human Cancer Stem Cells

In the first study, we built GSMMs of CSCs and CCs of ten different tissues in humans, by integrating gene expression of distinct datasets with the generic GSMM of human cells, *Human1*. After pre-processing published RNA-seq and microarray datasets, the best parameters to integrate each type of transcriptomics data (RNA-seq or microarray) were identified and applied to all respective datasets. Models were adapted for medium composition and gapfilled in order to produce biomass and perform essential tasks.

The comparison of simulated essential genes and lethal genes previously identified in a gene knockout experiment revealed a similar correlation coefficient to the one reported in a previous study with models of human cancer cells, showing evidences of the validity of our models. Metabolic pathways predicted, from flux simulations, to be more active in CSCs than CCs, were in line with the available literature. In addition, the study brought attention to new possible metabolic mechanisms of already known cancer targets and tumor suppressors, like FLI1, *hsa-miR-335-5p*, *hsa-miR-26b-5p*, and *hsa-let-7b-5p*. Besides, new

potential tumor suppressors for treatment of CSCs and bulk cancer, namely *hsa-miR-6499-3p*, *hsa-miR-8485*, and *hsa-miR-6849-3p*, as well as *ribavirin* as an antimetabolite for the treatment of lung CSCs, were identified.

One limitation of the study relates to the ambiguity of the term CSC, since both actively dividing cells and quiescent/slow-cycling cells can be considered CSCs. While, for the first definition, the assumption of growth as the main cellular objective holds, it may not be the case for the latter. Hence, we suggest that future studies try to experimentally assess which is the case. Furthermore, one restriction of the study is that in some cases it uses only one model for each condition (cell type and tissue). As more data becomes available prospective studies could include many more models of the same condition obtained from different datasets to account for the bias of experimental platforms and donor variability.

Another problem is that the study does not include models of normal stem and differentiated cells due to the difficulty in finding datasets with all four cell types. This is a drawback because some of the potential targets suggested for CCs and CSCs might also impact normal cells. So, future studies should focus on including those two cell types if new experimental data becomes available. Nevertheless, existing literature might guide wet-lab scientists in choosing the best candidates for experimental validation from those suggested in the study.

Additionally, although constraints can be defined over exchange fluxes according to media composition, the exact nutrient consumption rates are often unknown, hampering the definition of accurate bounds for exchange reactions, which affects the phenotype prediction of cancer models [126]. Therefore, future studies should attempt to include experimentally measured uptake/secretion rates or reconstruct enzyme-constrained GECKO models.

4.2 Reconstruction of Cell-specific Models Capturing the Interplay Between Metabolism and Epigenetics in Cancer

In the second study, and for the first time, we successfully built traditional GSMMs and GECKO models for human cancer cell lines comprising the DNA methylation and demethylation machinery. First, reactions (in)/directly involved in DNA (de)/methylation were retrieved from literature and databases, curated, and integrated with the most updated version of the generic GSMM of human cells, the *Human1* version 12.0. Transcriptomics data of different NCI-60 cell lines were then used to create cell-line specific traditional GSMMs, which in

turn served as draft for the construction of cell-specific GECKO models through the incorporation of enzyme kinetics information. Different reconstruction and simulation strategies were tested with both types of models, and the ones providing the best simulations were identified.

Models were validated through the observation that simulated exchange rates were similar to their experimentally measured values. GECKO models performed better than GSMMs in that regard, and were able to give small relative errors for growth rates, which were similar to the one reported in a previous study for 11 of these cell lines. Furthermore, the simulated flux and protein mass distribution across different metabolic pathways of GECKO models were in line with what was reported in the literature for cancer cells. Additionally, cell-line-specific DNA methylation levels estimated based on experimental data were introduced in GECKOs to understand how metabolism affects DNA methylation across different cell lines, whereas DNA methylation flux rules were added to force flux through demethylation reactions and consequently simulate the experimentally observable dynamic DNA methylation turnover steady-state, where both DNA demethylation and methylation reactions are active.

The biggest advantage of the GECKO models reconstructed in the study is that they incorporate enzymatic constraints, and therefore, do not need to be limited with exact values of many exchange fluxes to guarantee accurate flux distributions. Besides, it was observed that even the use of constraints on three exchange fluxes of traditional GSMMs was not sufficient to give as good results as with GECKO models.

Nevertheless, those models only mirror active DNA demethylation mechanisms, whereas passive DNA demethylation, i.e. the dilution of DNA methylation signal as a result of fast cellular division in the absence of functional DNA methylation maintenance machinery, is not taken into account [260].

Another problem is that active DNA demethylation is not achieved with flux simulations for most of the cell lines modeled, because the flux rules applied are previously reported generalized flux ratios that may not in fact apply to certain cell lines, causing the simulations to become infeasible. Hence, it is expected that prospective studies will incorporate newly discovered cell-specific DNA methylation rules.

4.3 Limitations of both studies and future directions

A common limitation to both studies is the use of transcriptomics data to reconstruct models of specific tissues or cells. Despite often being more com-

prehensive in terms of its coverage, providing data related with a high number of genes and related reactions, transcriptomics can lead to incorrect predictions due to the limited correlation between gene and protein expression levels. We expect, however, that future protein quantification techniques may simplify the acquisition of more comprehensive proteomics datasets that can be applied to the reconstruction of models covering all cellular metabolic reactions.

The lack of knowledge about the composition of biomass in each tissue/cell line is another problem, aggravated by the fact that biomass composition varies with environmental conditions [261, 262]. The generalized one used in these studies is only an approximation that can lead to limited accuracy in the model predictions. Therefore, prospective investigations should integrate experimentally measured cell-specific biomass composition values.

It is also important to note that the reconstructed models are metabolic, while some of the suggested targets in the first study and targets that might be discovered in the future with the models in the second study, may have opposite effects through regulatory and signaling pathways. Furthermore, DNA methylation is strongly affected by other regulatory mechanisms, such as other epigenetic modifications [263]. Therefore, it will be important to integrate regulatory and signaling networks into future metabolic models of CSCs and of CCs with epigenetic machinery.

Bibliography

[1] J. Ferlay, M. Colombet, and F. Bray, "Cancer incidence in five continents," *CI5plus: IARC CancerBase No. 9 [Internet]. Lyon, France: International Agency for Research on cancer. Available at: https://gco.iarc.fr/*, 2018.

[2] G. Cooper, "The cell: a molecular approach. 2nd edition.," *Sunderland (MA): Sinauer Associates. The Development and Causes of Cancer. Available from: https://www.ncbi.nlm.nih.gov/books/NBK9963/*, 2000.

[3] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: The next generation," *Cell*, vol. 144, pp. 646–674, 2011.

[4] E. A. Mroz and J. W. Rocco, "The challenges of tumor genetic diversity," *Cancer*, vol. 123, pp. 917–927, 3 2017.

[5] E. M. D. Francesco, F. Sotgia, and M. P. Lisanti, "Cancer stem cells (cscs): Metabolic strategies for their identification and eradication," *Biochemical Journal*, vol. 475, pp. 1611–1634, 2018.

[6] K. Dzobo, D. A. Senthebane, C. Ganz, N. E. Thomford, A. Wonkam, and C. Dandara, "Advances in therapeutic targeting of cancer stem cells within the tumor microenvironment: An updated review," *Cells*, vol. 9, 2020.

[7] E. Batlle and H. Clevers, "Cancer stem cells revisited," *Nature Medicine*, vol. 23, pp. 1124–1134, 2017.

[8] F. Ferraro, C. L. Celso, and D. Scadden, "Adult stem cells and their niches," *Advances in Experimental Medicine and Biology*, vol. 695, pp. 155–168, 2010.

[9] C. L. Chaffer and R. A. Weinberg, "A perspective on cancer cell metastasis," *Science*, vol. 331, pp. 1559–1564, 2011.

[10] S. Tanabe, S. Quader, H. Cabral, and R. Ono, "Interplay of emt and csc in cancer and the potential therapeutic strategies," *Frontiers in Pharmacology*, vol. 11, pp. 1–8, 2020.

[11] A. Agliano, A. Calvo, and C. Box, "The challenge of targeting cancer stem cells to halt metastasis," *Seminars in Cancer Biology*, vol. 44, pp. 25–42, 2017.

[12] J. Peixoto and J. Lima, "Metabolic traits of cancer stem cells," *Disease Models & Mechanisms*, vol. 11, pp. 1–13, 2018.

[13] S. Agnihotri and G. Zadeh, "Metabolic reprogramming in glioblastoma: the influence of cancer metabolism on epigenetics and unanswered questions," *Neuro-Oncology*, vol. 18, pp. 160–172, 2016.

[14] O. Warburg, "On the origin of cancer cells," *Science*, vol. 123, pp. 309–314, 2 1956.

[15] O. Warburg, F. Wind, and E. Negelein, "The metabolism of tumors in the body," *Journal of General Physiology*, vol. 8, pp. 519–530, 3 1927.

[16] N. N. Pavlova and T. C. B., "The emerging hallmarks of cancer metabolism," *Cell Metabolism*, vol. 23, pp. 27–47, 2016.

[17] J. H. Park, W. Y. Pyun, and H. H. Park, "Cancer metabolism: Phenotype, signaling and therapeutic targets," *Cells*, vol. 9, 2020.

[18] X. Zhu, H. hui Chen, C. yi Gao, X. xin Zhang, J. xin Jiang, Y. Zhang, J. Fang, F. Zhao, Z. gang Chen, H. hui Chen, C. yi Gao, X. xin Zhang, J. xin Jiang, Y. Zhang, and Z. gang Chen, "Energy metabolism in cancer stem cells," *World journal of stem cells*, vol. 12, pp. 448–461, 2020.

[19] Y. Kim, E. Y. Kim, Y. M. Seo, T. K. Yoon, W. S. Lee, and K. A. Lee, "Function of the pentose phosphate pathway and its key enzyme, transketolase, in the regulation of the meiotic cell cycle in oocytes," *Clinical and Experimental Reproductive Medicine*, vol. 39, pp. 58–67, 2012.

[20] A. Stincone, A. Prigione, T. Cramer, M. M. C. Wamelink, K. Campbell, E. Cheung, V. Olin-Sandoval, N. maria Grüning, A. Krüger, M. T. Alam, M. A. Keller, M. Breitenbach, K. M. Brindle, J. D. Rabinowitz, and M. Ralser, "The return of metabolism: biochemistry and physiology of the pentose phosphate pathway," *Biological Reviews*, vol. 90, pp. 927–963, 8 2015.

[21] W. Wu, L. Chen, Y. Wang, J. Jin, X. Xie, and J. Zhang, "Hyaluronic acid predicts poor prognosis in breast cancer patients: A protocol for systematic review and meta analysis," *Medicine*, vol. 99, p. e20438, 2020.

[22] A. Rosenzweig, J. Blenis, and A. P. Gomes, "Beyond the warburg effect: How do cancer cells regulate one-carbon metabolism?," *Frontiers in Cell and Developmental Biology*, vol. 6, pp. 1–7, 2018.

[23] G. S. Ducker and J. D. Rabinowitz, "One-carbon metabolism in health and disease," *Cell Metabolism*, vol. 25, pp. 27–42, 2017.

[24] J. G. Ryall, T. Cliff, S. Dalton, and V. Sartorelli, "Metabolic reprogramming of stem cell epigenetics," *Cell Stem Cell*, vol. 17, pp. 651–662, 2015.

[25] Z. G. Movahed, M. Rastegari-Pouyani, and M. hossein Mohammadi, "Cancer cells change their glucose metabolism to overcome increased ros: One step from cancer cell to cancer stem cell?," *Biomedicine & Pharmacotherapy*, vol. 112, pp. 1–15, 2019.

[26] K. Kurmi and M. C. Haigis, "Nitrogen metabolism in cancer and immunity," *Trends in Cell Biology*, vol. 30, pp. 408–424, 2020.

[27] F. Pietrocola, L. Galluzzi, J. M. B.-S. Pedro, F. Madeo, and G. Kroemer, "Acetyl coenzyme a: A central metabolite and second messenger," *Cell Metabolism*, vol. 21, pp. 805–821, 2015.

[28] P. Saggese, A. Sellitto, C. A. Martinez, G. Giurato, G. Nassa, F. Rizzo, R. Tarallo, and C. Scafoglio, "Metabolic regulation of epigenetic modifications and cell differentiation in cancer," *Cancers*, vol. 12, pp. 1–24, 2020.

[29] H. Zhao and T. Chen, "Tet family of 5-methylcytosine dioxygenases in mammalian development," *Journal of Human Genetics*, vol. 58, pp. 421–427, 2013.

[30] D. R. Menon, H. Hammerlindl, J. Torrano, and H. Schaider, "Epigenetics and metabolism at the crossroads of stress-induced plasticity, stemness and therapeutic resistance in cancer," *Theranostics*, vol. 10, pp. 6261–6277, 2020.

[31] X. Hu, D. Xiang, Y. Xie, L. Tao, Y. Zhang, Y. Jin, L. Pinello, and Y. Wan, "Lsd1 suppresses invasion , migration and metastasis of luminal breast cancer cells via activation of gata3 and repression of trim37 expression," *Oncogene*, vol. 38, pp. 7017–7034, 2019.

[32] B. Majello, F. Gorini, C. Sacca, and S. Amente, "Expanding the role of the histone lysine-specific demethylase lsd1 in cancer," *Cancers*, vol. 11, pp. 1–15, 2019.

[33] P. Jagust, B. de Luxán-Delgado, B. Parejo-Alonso, and P. Sancho, "Metabolism-based therapeutic strategies targeting cancer stem cells," *Frontiers in Pharmacology*, vol. 10, pp. 1–26, 3 2019.

[34] H. Ye, B. Adane, N. Khan, T. Sullivan, M. Minhajuddin, M. Gasparetto, B. Stevens, S. Pei, M. Balys, J. M. Ashton, D. J. Klemm, C. M. Woolthuis, A. W. Stranahan, C. Y. Park, and C. T. Jordan, "Leukemic stem cells evade chemotherapy by metabolic adaptation to an adipose tissue niche," *Cell Stem Cell*, vol. 19, pp. 23–37, 2017.

[35] J. Liao, P. pan Liu, G. Hou, J. Shao, J. Yang, K. Liu, W. Lu, and S. Wen, "Regulation of stem-like cancer cells by glutamine through β-catenin pathway mediated by redox signaling," *Molecular Cancer*, vol. 16, pp. 1–13, 2017.

[36] D. Li, Z. Fu, R. Chen, X. Zhao, Y. Zhou, J. Zhou, Z. Li, Y. Liu, and R. Chen, "Inhibition of glutamine metabolism counteracts pancreatic cancer stem cell features and sensitizes cells to radiotherapy," *Oncotarget*, vol. 6, pp. 331151–31163, 2015.

[37] J. Kasznicki, A. Sliwinska, and J. Drzewoski, "Metformin in cancer prevention and therapy," *Annals of Translational Medicine*, vol. 2, pp. 1–11, 2014.

[38] J. H. Kim, K. J. Lee, Y. Seo, J. hee Kwon, J. P. Yoon, J. Y. Kang, H. J. Lee, S. J. Park, S. P. Hong, J. H. Cheon, and W. H. Kim, "Effects of metformin on colorectal cancer stem cells depend on alterations in glutamine metabolism," *Scientific Reports*, vol. 8, pp. 1–13, 2018.

[39] Y. Li, W. Ding, C. Y. Li, and Y. Liu, "Hlh-11 modulates lipid metabolism in response to nutrient availability," *Nature Communications*, vol. 11, pp. 1–13, 2020.

[40] M. Visweswaran, F. Arfuso, S. Warrier, and A. Dharmarajan, "Aberrant lipid metabolism as an emerging therapeutic strategy to target cancer stem cells," *Stem Cells*, vol. 38, pp. 6–14, 2020.

[41] T. Murai, "The role of lipid rafts in cancer cell adhesion and migration," *International Journal of Cell Biology*, vol. 2012, pp. 1–6, 2012.

[42] Y. Lazebnik, "Can a biologist fix a radio?—or, what i learned while studying apoptosis," *Cancer Cell*, vol. 2, pp. 179–182, 9 2002.

[43] L. You, "Toward computational systems biology," *Cell Biochemistry and Biophysics*, vol. 40, pp. 167–184, 2004.

[44] H. V. Westerhoff and B. O. Palsson, "The evolution of molecular biology into systems biology," *Nature Biotechnology*, vol. 22, pp. 1249–1252, 2004.

[45] K. Smallbone, E. Simeonidis, D. S. Broomhead, and D. B. Kell, "Something from nothing - bridging the gap between constraint-based and kinetic modelling," *FEBS Journal*, vol. 274, pp. 5576–5585, 2007.

[46] R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick, K. McKenney, G. Sutton, W. FitzHugh, C. Fields, J. D. Gocayne, J. Scott, R. Shirley, L. I. Liu, A. Glodek, J. M. Kelley, J. F. Weidman, C. A. Phillips, T. Spriggs, E. Hedblom, M. D. Cotton, T. R. Utterback, M. C. Hanna, D. T. Nguyen, D. M. Saudek, R. C. Brandon, L. D. Fine, J. L. Fritchman, J. L. Fuhrmann, N. S. Geoghagen, C. L. Gnehm, L. A. McDonald, K. V. Small, C. M. Fraser, H. O. Smith, and J. C. Venter, "Whole-genome random sequencing and assembly of haemophilus influenzae rd," *Science*, vol. 269, pp. 496–512, 1995.

[47] J. S. Edwards and B. O. Palsson, "Systems properties of the haemophilus influenzae rd metabolic genotype," *Journal of Biological Chemistry*, vol. 274, pp. 17410–17416, 1999.

[48] C. Gu, G. B. Kim, W. J. Kim, H. U. Kim, and S. Y. Lee, "Current status and applications of genome-scale metabolic models," *Genome Biology*, vol. 20, pp. 1–18, 2019.

[49] P. Maia, M. Rocha, and I. Rocha, "In silico constraint-based strain optimization methods: the quest for optimal cell factories," *Microbiology and Molecular Biology Reviews*, vol. 80, pp. 45–67, 2016.

[50] E. M. Blais, A. K. Chavali, and J. A. Papin, "Linking genome-scale metabolic modeling and genome annotation," *Methods in Molecular Biology*, vol. 985, pp. 61–83, 2013.

[51] I. Thiele and B. Palsson, "A protocol for generating a high-quality genome-scale metabolic reconstruction," *Nature Protocols*, vol. 5, pp. 93–121, 2010.

[52] Y. Zhang, J. Cai, X. Shang, B. Wang, S. Liu, X. Chai, T. Tan, Y. Zhang, and T. Wen, "A new genome-scale metabolic model of corynebacterium

glutamicum and its application," *Biotechnology for Biofuels*, vol. 10, pp. 1–16, 2017.

[53] S. Ferreira, R. Pereira, F. Liu, P. Vilaça, and I. Rocha, "Discovery and implementation of a novel pathway for n-butanol production via 2-oxoglutarate," *Biotechnology for Biofuels*, vol. 12, pp. 1–14, 2019.

[54] R. Viana, O. Dias, D. Lagoa, M. Galocha, I. Rocha, and M. C. Teixeira, "Genome-scale metabolic model of the human pathogen candida albicans: A promising platform for drug target prediction," *Journal of Fungi*, vol. 6, pp. 1–19, 9 2020.

[55] C. Liao, T. Wang, S. Maslov, and J. B. Xavier, "Modeling microbial cross-feeding at intermediate scale portrays community dynamics and species coexistence," *PLoS Computational Biology*, vol. 16, pp. 1–23, 2020.

[56] S. Magnúsdóttir, A. Heinken, L. Kutt, D. A. Ravcheev, E. Bauer, A. Noronha, K. Greenhalgh, C. Jäger, J. Baginska, P. Wilmes, R. M. Fleming, and I. Thiele, "Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota," *Nature Biotechnology*, vol. 35, pp. 81–89, 2017.

[57] R. A. Rienksma, P. J. Schaap, V. A. D. Santos, and M. Suarez-Diez, "Modeling host-pathogen interaction to elucidate the metabolic drug response of intracellular mycobacterium tuberculosis," *Frontiers in Cellular and Infection Microbiology*, vol. 9, pp. 1–14, 2019.

[58] M. P. Pacheco, T. Bintener, D. Ternes, D. Kulms, S. Haan, E. Letellier, and T. Sauter, "Identifying and targeting cancer-specific metabolism with network-based drug target prediction," *EBioMedicine*, vol. 43, pp. 98–106, 2019.

[59] J. R. et al. F.S. Collins, E.S. Lander, "Finishing the euchromatic sequence of the human genome," *Nature*, vol. 431, pp. 931–945, 10 2004.

[60] N. C. Duarte, S. A. Becker, N. Jamshidi, I. Thiele, M. L. Mo, T. D. Vo, R. Srivas, and B. Palsson, "Global reconstruction of the human metabolic network based on genomic and bibliomic data," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, pp. 1777–1782, 2007.

[61] H. Ma, A. Sorokin, A. Mazein, A. Selkov, E. Selkov, O. Demin, and I. Goryanin, "The edinburgh human metabolic network reconstruction and its functional analysis," *Molecular Systems Biology*, vol. 3, pp. 1–8, 2007.

[62] C. Gille, C. Bölling, A. Hoppe, S. Bulik, S. Hoffmann, K. Hübner, A. Karlstädt, R. Ganeshan, M. König, K. Rother, M. Weidlich, J. Behre, and H. G. Holzhütter, "Hepatonet1: a comprehensive metabolic reconstruction of the human hepatocyte for the analysis of liver physiology," *Molecular Systems Biology*, vol. 6, 2010.

[63] R. Agren, S. Bordel, A. Mardinoglu, N. Pornputtapong, I. Nookaew, and J. Nielsen, "Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using init," *PLoS Computational Biology*, vol. 8, pp. 1–9, 2012.

[64] A. Mardinoglu, R. Agren, C. Kampf, A. Asplund, M. Uhlen, and J. Nielsen, "Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease," *Nature Communications*, vol. 5, pp. 1–11, 2014.

[65] A. Mardinoglu, R. Agren, C. Kampf, A. Asplund, I. Nookaew, P. Jacobson, A. J. Walley, P. Froguel, L. M. Carlsson, M. Uhlen, and J. Nielsen, "Integration of clinical data with a genome-scale metabolic model of the human adipocyte," *Molecular Systems Biology*, vol. 9, pp. 1–16, 2013.

[66] L. Väremo, C. Scheele, C. Broholm, A. Mardinoglu, C. Kampf, A. Asplund, I. Nookaew, M. Uhlén, B. K. Pedersen, and J. Nielsen, "Proteome- and transcriptome-driven reconstruction of the human myocyte metabolic network and its use for identification of markers for diabetes," *Cell Reports*, vol. 11, pp. 921–933, 5 2015.

[67] I. Thiele, N. Swainston, R. M. Fleming, A. Hoppe, S. Sahoo, M. K. Aurich, H. Haraldsdottir, M. L. Mo, O. Rolfsson, M. D. Stobbe, S. G. Thorleifsson, R. Agren, C. Bölling, S. Bordel, A. K. Chavali, P. Dobson, W. B. Dunn, L. Endler, D. Hala, M. Hucka, D. Hull, D. Jameson, N. Jamshidi, J. J. Jonsson, N. Juty, S. Keating, I. Nookaew, N. L. Novère, N. Malys, A. Mazein, J. A. Papin, N. D. Price, E. Selkov, M. I. Sigurdsson, E. Simeonidis, N. Sonnenschein, K. Smallbone, A. Sorokin, J. H. V. Beek, D. Weichart, I. Goryanin, J. Nielsen, H. V. Westerhoff, D. B. Kell, P. Mendes, and B. O. Palsson, "A community-driven global reconstruction of human metabolism," *Nature Biotechnology*, vol. 31, pp. 419–425, 2013.

[68] K. Smallbone, "Striking a balance with recon 2.1," *arXiv*, vol. 1311, pp. 14–17, 2013.

[69] N. Swainston, K. Smallbone, H. Hefzi, P. D. Dobson, J. Brewer, M. Hanscho, D. C. Zielinski, K. S. Ang, N. J. Gardiner, J. M. Gutierrez, S. Kyriakopoulos, M. Lakshmanan, S. Li, J. K. Liu, V. S. Martínez, C. A. Orellana, L.-E. Quek, A. Thomas, J. Zanghellini, N. Borth, D.-Y. Lee, L. K. Nielsen, D. B. Kell, N. E. Lewis, and P. Mendes, "Recon 2.2: from reconstruction to model of human metabolism," *Metabolomics*, vol. 12, pp. 1–7, 7 2016.

[70] J. Y. Ryu, H. U. Kim, and S. Y. Lee, "Framework and resource for more than 11,000 gene-transcript-protein-reaction associations in human metabolism," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 114, pp. E9740–E9749, 2017.

[71] E. Brunk, S. Sahoo, D. C. Zielinski, A. Altunkaya, N. Mih, F. Gatto, A. Nilsson, G. A. Preciat, M. K. Aurich, A. Sastry, A. D. Danielsdottir, A. Heinken, A. Noronha, P. W. Rose, S. K. Burley, M. T. Ronan, J. Nielsen, I. Thiele, and B. O. Palsson, "Recon3d: A resource enabling a three-dimensional view of gene variation in human metabolism," *Nature Biotechnology*, vol. 36, pp. 272–281, 2018.

[72] A. Noronha, A. D. Daníelsdóttir, P. Gawron, F. Jóhannsson, S. Jónsdóttir, S. Jarlsson, J. P. Gunnarsson, S. Brynjólfsson, R. Schneider, I. Thiele, and R. M. T. Fleming, "Reconmap: an interactive visualization of human metabolism," *Bioinformatics*, vol. 33, pp. 605–607, 12 2017.

[73] A. Noronha, J. Modamio, Y. Jarosz, E. Guerard, N. Sompairac, G. Preciat, A. D. Daníelsdóttir, M. Krecke, D. Merten, H. S. Haraldsdóttir, A. Heinken, L. Heirendt, S. Magnúsdóttir, D. A. Ravcheev, S. Sahoo, P. Gawron, L. Friscioni, B. Garcia, M. Prendergast, A. Puente, M. Rodrigues, A. Roy, M. Rouquaya, L. Wiltgen, A. Žagare, E. John, M. Krueger, I. Kuperstein, A. Zinovyev, R. Schneider, R. M. Fleming, and I. Thiele, "The virtual metabolic human database: integrating human and gut microbiome metabolism with nutrition and disease," *Nucleic Acids Research*, vol. 47, pp. D614–D624, 1 2019.

[74] I. Thiele, S. Sahoo, A. Heinken, J. Hertel, L. Heirendt, M. K. Aurich, and R. M. Fleming, "Personalized whole-body models integrate metabolism, physiology, and the gut microbiome," *Molecular Systems Biology*, vol. 16, pp. 1–24, 5 2020.

[75] J. L. Robinson, P. Kocabaş, H. Wang, P. etienne Cholley, D. Cook, A. Nilsson, M. Anton, R. Ferreira, I. Domenzain, V. Billa, A. Limeta, A. Hedin, J. Gustafsson, E. J. Kerkhoven, L. T. Svensson, B. O. Palsson, A. Mardinoglu, L. Hansson, M. Uhlén, and J. Nielsen, "An atlas of human metabolism," *Science Signaling*, vol. 13, pp. 1–11, 3 2020.

[76] E. M. Blais, K. D. Rawls, B. V. Dougherty, Z. I. Li, G. L. Kolling, P. Ye, A. Wallqvist, and J. A. Papin, "Reconciled rat and human metabolic networks for comparative toxicogenomics and biomarker predictions," *Nature Communications*, vol. 8, pp. 1–15, 4 2017.

[77] M. Ederer, R. Schlatter, J. Witt, R. Feuer, J. Bóna-Lovász, S. Henkel, and O. Sawodny, "An introduction to kinetic, constraint-based and boolean modeling in systems biology," *Proceedings of the IEEE International Conference on Control Applications*, pp. 129–134, 2010.

[78] M. W. Covert, I. Famili, and B. O. Palsson, "Identifying constraints that govern cell behavior: a key to converting conceptual to computational models in biology?," *Biotechnology and Bioengineering*, vol. 84, pp. 763–772, 12 2003.

[79] M. Yasemi and M. Jolicoeur, "Modelling cell metabolism: A review on constraint-based steady-state and kinetic approaches," *Processes*, vol. 9, pp. 1–38, 2021.

[80] J. Boyle, "Enzymes - chapter 6," *Lehninger principles of biochemistry (4th ed.)*, vol. 33, pp. 203–2017, 1 2005.

[81] S. Volkova, M. R. Matos, M. Mattanovich, and I. M. de Mas, "Metabolic modelling as a framework for metabolomics data integration and analysis," *Metabolites*, vol. 10, pp. 1–27, 2020.

[82] A.-M. Reimers and A. C. Reimers, "The steady-state assumption in oscillating and growing systems," *Journal of Theoretical Biology*, vol. 406, pp. 176–186, 10 2016.

[83] C. Angione, "Human systems biology and metabolic modelling: A review-from disease metabolism to precision medicine," *BioMed Research International*, vol. 2019, 2019.

[84] R. Gunawan and S. Hutter, "Assessing and resolving model misspecifications in metabolic flux analysis," *Bioengineering*, vol. 4, pp. 1–17, 5 2017.

[85] J. D. Orth, I. Thiele, and B. O. Palsson, "What is flux balance analysis?," *Nature Biotechnology*, vol. 28, pp. 245–248, 2010.

[86] Y. Chen, B. O. McConnell, V. G. Dhara, H. M. Naik, C.-T. Li, M. R. Antoniewicz, and M. J. Betenbaugh, "An unconventional uptake rate objective function approach enhances applicability of genome-scale models for mammalian cells," *npj Systems Biology and Applications*, vol. 5, pp. 1–11, 12 2019.

[87] D. Segre, D. Vitkup, and G. M. Church, "Analysis of optimality in natural and perturbed metabolic networks," *Proceedings of the National Academy of Sciences*, vol. 99, pp. 15112–15117, 11 2002.

[88] A. L. Knorr, R. Jain, and R. Srivastava, "Bayesian-based selection of metabolic objective functions," *Bioinformatics*, vol. 23, pp. 351–357, 2 2007.

[89] E. Gonçalves, R. Pereira, I. Rocha, and M. Rocha, "Optimization approaches for the in silico discovery of optimal targets for gene over/underexpression," *Journal of Computational Biology*, vol. 19, pp. 102–114, 2 2012.

[90] N. E. Lewis, K. K. Hixson, T. M. Conrad, J. A. Lerman, P. Charusanti, A. D. Polpitiya, J. N. Adkins, G. Schramm, S. O. Purvine, D. Lopez-Ferrer, K. K. Weitz, R. Eils, R. König, R. D. Smith, and B. Palsson, "Omic data from evolved e. coli are consistent with computed optimal growth from genome-scale models," *Molecular Systems Biology*, vol. 6, pp. 1–13, 2010.

[91] T. Shlomi, O. Berkman, and E. Ruppin, "Regulatory on/off minimization of metabolic flux changes after genetic perturbations," *Proceedings of the National Academy of Sciences*, vol. 102, pp. 7695–7700, 5 2005.

[92] A. Richelle, A. W. T. Chiang, C. chung Kuo, and N. E. Lewis, "Increasing consensus of context-specific metabolic models by integrating data-inferred cell functions," *PLOS Computational Biology*, vol. 15, pp. 1–19, 4 2019.

[93] J. Zanghellini, D. E. Ruckerbauer, M. Hanscho, and C. Jungreuthmayer, "Elementary flux modes in a nutshell: Properties, calculation and applications," *Biotechnology Journal*, vol. 8, pp. 1009–1016, 2013.

[94] R. Mahadevan and C. H. Schilling, "The effects of alternate optimal solutions in constraint-based genome-scale metabolic models," *Metabolic Engineering*, vol. 5, pp. 264–276, 2003.

[95] T. Chen, Z. Xie, and Q. Ouyang, "Expanded flux variability analysis on metabolic network of escherichia coli," *Science Bulletin*, vol. 54, pp. 2610–2619, 8 2009.

[96] J. L. Reed, "Genome-scale in silico models of e. coli have multiple equivalent phenotypic states: Assessment of correlated reaction subsets that comprise network states," *Genome Research*, vol. 14, pp. 1797–1805, 9 2004.

[97] K. Jensen, V. Broeken, A. S. L. Hansen, N. Sonnenschein, and M. J. Herrgård, "Optcouple: Joint simulation of gene knockouts, insertions and medium modifications for prediction of growth-coupled strain designs," *Metabolic Engineering Communications*, vol. 8, p. e00087, 6 2019.

[98] T. Wilhelm, J. Behre, and S. Schuster, "Analysis of structural robustness of metabolic networks," *Systems Biology*, vol. 1, pp. 114–120, 6 2004.

[99] S. R. Estévez and Z. Nikoloski, "Generalized framework for context-specific metabolic model extraction methods," *Frontiers in Plant Science*, vol. 5, pp. 1–12, 2014.

[100] D. Machado and M. Herrgård, "Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism," *PLoS Computational Biology*, vol. 10, 2014.

[101] J. S. Cho, C. Gu, T. H. Han, J. Y. Ryu, and S. Y. Lee, "Reconstruction of context-specific genome-scale metabolic models using multiomics data to study metabolic rewiring," *Current Opinion in Systems Biology*, vol. 15, pp. 1–11, 6 2019.

[102] E. W. Sayers, M. Cavanaugh, K. Clark, J. Ostell, K. D. Pruitt, and I. Karsch-Mizrachi, "Genbank," *Nucleic Acids Research*, vol. 48, pp. D84–D86, 2020.

[103] P. W. Harrison, A. Ahamed, R. Aslam, B. T. F. Alako, J. Burgin, N. Buso, M. Courtot, J. Fan, D. Gupta, M. Haseeb, S. Holt, T. Ibrahim, E. Ivanov, S. Jayathilaka, V. B. Kadhirvelu, M. Kumar, R. Lopez, S. Kay, R. Leinonen, X. Liu, C. O'Cathail, A. Pakseresht, Y. Park, S. Pesant, N. Rahman, J. Rajan, A. Sokolov, S. Vijayaraja, Z. Waheed, A. Zyoud,

T. Burdett, and G. Cochrane, "The european nucleotide archive in 2020," *Nucleic Acids Research*, vol. 49, pp. D82–D85, 1 2021.

[104] J. Mashima, Y. Kodama, T. Fujisawa, T. Katayama, Y. Okuda, E. Kaminuma, O. Ogasawara, K. Okubo, Y. Nakamura, and T. Takagi, "Dna data bank of japan," *Nucleic Acids Research*, vol. 45, pp. D25–D31, 2017.

[105] A. P. Heath, V. Ferretti, S. Agrawal, M. An, J. C. Angelakos, R. Arya, R. Bajari, B. Baqar, J. H. B. Barnowski, J. Burt, A. Catton, B. F. Chan, F. Chu, K. Cullion, T. Davidsen, P.-M. Do, C. Dompierre, M. L. Ferguson, M. S. Fitzsimons, M. Ford, M. Fukuma, S. Gaheen, G. L. Ganji, T. I. Garcia, S. S. George, D. S. Gerhard, F. Gerthoffert, F. Gomez, K. Han, K. M. Hernandez, B. Issac, R. Jackson, M. A. Jensen, S. Joshi, A. Kadam, A. Khurana, K. M. J. Kim, V. E. Kraft, S. Li, T. M. Lichtenberg, J. Lodato, L. Lolla, P. Martinov, J. A. Mazzone, D. P. Miller, I. Miller, J. S. Miller, K. Miyauchi, M. W. Murphy, T. Nullet, R. O. Ogwara, F. M. Ortuño, J. Pedrosa, P. L. Pham, M. Y. Popov, J. J. Porter, R. Powell, K. Rademacher, C. P. Reid, S. Rich, B. Rogel, H. Sahni, J. H. Savage, K. A. Schmitt, T. J. Simmons, J. Sislow, J. Spring, L. Stein, S. Sullivan, Y. Tang, M. Thiagarajan, H. D. Troyer, C. Wang, Z. Wang, B. L. West, A. Wilmer, S. Wilson, K. Wu, W. P. Wysocki, L. Xiang, J. T. Yamada, L. Yang, C. Yu, C. K. Yung, J. C. Zenklusen, J. Zhang, Z. Zhang, Y. Zhao, A. Zubair, L. M. Staudt, and R. L. Grossman, "The nci genomic data commons," *Nature Genetics*, vol. 53, pp. 257–262, 3 2021.

[106] M. Uhlen, L. Fagerberg, B. M. Hallstrom, C. Lindskog, P. Oksvold, A. Mardinoglu, A. Sivertsson, C. Kampf, E. Sjostedt, A. Asplund, I. Olsson, K. Edlund, E. Lundberg, S. Navani, C. A.-K. Szigyarto, J. Odeberg, D. Djureinovic, J. O. Takanen, S. Hober, T. Alm, P.-H. Edqvist, H. Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J. M. Schwenk, M. Hamsten, K. von Feilitzen, M. Forsberg, L. Persson, F. Johansson, M. Zwahlen, G. von Heijne, J. Nielsen, and F. Ponten, "Tissue-based map of the human proteome," *Science*, vol. 347, pp. 1260419–1260419, 1 2015.

[107] E. Clough and T. Barrett, "The gene expression omnibus database," *Methods in Molecular Biology*, vol. 1418, pp. 93–110, 2016.

[108] M. Krupp, J. U. Marquardt, U. Sahin, P. R. Galle, J. Castle, and A. Teufel, "Rna-seq atlas — a reference database for gene expression profiling in

normal tissue by next-generation sequencing," *Bioinformatics*, vol. 28, pp. 1184–1185, 2012.

[109] N. Kolesnikov, E. Hastings, M. Keays, O. Melnichuk, Y. A. Tang, E. Williams, M. Dylag, N. Kurbatova, M. Brandizi, T. Burdett, K. Megy, E. Pilicheva, G. Rustici, A. Tikhonov, H. Parkinson, R. Petryszak, U. Sarkans, and A. Brazma, "Arrayexpress update-simplifying data submissions," *Nucleic Acids Research*, vol. 43, pp. D1113–D1116, 2015.

[110] J. Lonsdale, J. Thomas, M. Salvatore, R. Phillips, E. Lo, S. Shad, R. Hasz, G. Walters, F. Garcia, N. Young, B. Foster, M. Moser, E. Karasik, B. Gillard, K. Ramsey, S. Sullivan, J. Bridge, H. Magazine, J. Syron, J. Fleming, L. Siminoff, H. Traino, M. Mosavel, L. Barker, S. Jewell, D. Rohrer, D. Maxim, D. Filkins, P. Harbach, E. Cortadillo, B. Berghuis, L. Turner, E. Hudson, K. Feenstra, L. Sobin, J. Robb, P. Branton, G. Korzeniewski, C. Shive, D. Tabor, L. Qi, K. Groch, S. Nampally, S. Buia, A. Zimmerman, A. Smith, R. Burges, K. Robinson, K. Valentino, D. Bradbury, M. Cosentino, N. Diaz-Mayoral, M. Kennedy, T. Engel, P. Williams, K. Erickson, K. Ardlie, W. Winckler, G. Getz, D. DeLuca, D. MacArthur, M. Kellis, A. Thomson, T. Young, E. Gelfand, M. Donovan, Y. Meng, G. Grant, D. Mash, Y. Marcus, M. Basile, J. Liu, J. Zhu, Z. Tu, N. J. Cox, D. L. Nicolae, E. R. Gamazon, H. K. Im, A. Konkashbaev, J. Pritchard, M. Stevens, T. Flutre, X. Wen, E. T. Dermitzakis, T. Lappalainen, R. Guigo, J. Monlong, M. Sammeth, D. Koller, A. Battle, S. Mostafavi, M. McCarthy, M. Rivas, J. Maller, I. Rusyn, A. Nobel, F. Wright, A. Shabalin, M. Feolo, N. Sharopova, A. Sturcke, J. Paschal, J. M. Anderson, E. L. Wilder, L. K. Derr, E. D. Green, J. P. Struewing, G. Temple, S. Volpi, J. T. Boyer, E. J. Thomson, M. S. Guyer, C. Ng, A. Abdallah, D. Colantuoni, T. R. Insel, S. E. Koester, A. R. Little, P. K. Bender, T. Lehner, Y. Yao, C. C. Compton, J. B. Vaught, S. Sawyer, N. C. Lockhart, J. Demchok, and H. F. Moore, "The genotype-tissue expression (gtex) project," *Nature Genetics*, vol. 45, pp. 580–585, 2013.

[111] T. S. K. Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. H. Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. I. Ramachandra, V. Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady, and

A. Pandey, "Human protein reference database - 2009 update," *Nucleic Acids Research*, vol. 37, pp. 767–772, 2009.

[112] M. S. Kim, S. M. Pinto, D. Getnet, R. S. Nirujogi, S. S. Manda, R. Chaerkady, A. K. Madugundu, D. S. Kelkar, R. Isserlin, S. Jain, J. K. Thomas, B. Muthusamy, P. Leal-Rojas, P. Kumar, N. A. Sahasrabuddhe, L. Balakrishnan, J. Advani, B. George, S. Renuse, L. D. N. Selvan, A. H. Patil, V. Nanjappa, A. Radhakrishnan, S. Prasad, T. Subbannayya, R. Raju, M. Kumar, S. K. Sreenivasamurthy, A. Marimuthu, G. J. Sathe, S. Chavan, K. K. Datta, Y. Subbannayya, A. Sahu, S. D. Yelamanchi, S. Jayaram, P. Rajagopalan, J. Sharma, K. R. Murthy, N. Syed, R. Goel, A. A. Khan, S. Ahmad, G. Dey, K. Mudgal, A. Chatterjee, T. C. Huang, J. Zhong, X. Wu, P. G. Shaw, D. Freed, M. S. Zahari, K. K. Mukherjee, S. Shankar, A. Mahadevan, H. Lam, C. J. Mitchell, S. K. Shankar, P. Satishchandra, J. T. Schroeder, R. Sirdeshmukh, A. Maitra, S. D. Leach, C. G. Drake, M. K. Halushka, T. S. Prasad, R. H. Hruban, C. L. Kerr, G. D. Bader, C. A. Iacobuzio-Donahue, H. Gowda, and A. Pandey, "A draft map of the human proteome," *Nature*, vol. 509, pp. 575–581, 2014.

[113] K. Haug, R. M. Salek, P. Conesa, J. Hastings, P. de Matos, M. Rijnbeek, T. Mahendraker, M. Williams, S. Neumann, P. Rocca-Serra, E. Maguire, A. González-Beltrán, S.-A. Sansone, J. L. Griffin, and C. Steinbeck, "Metabolights—an open-access general-purpose repository for metabolomics studies and associated meta-data," *Nucleic Acids Research*, vol. 41, pp. D781–D786, 1 2013.

[114] D. S. Wishart, Y. D. Feunang, A. Marcu, A. C. Guo, K. Liang, R. Vázquez-Fresno, T. Sajed, D. Johnson, C. Li, N. Karu, Z. Sayeeda, E. Lo, N. Assempour, M. Berjanskii, S. Singhal, D. Arndt, Y. Liang, H. Badran, J. Grant, A. Serra-Cayuela, Y. Liu, R. Mandal, V. Neveu, A. Pon, C. Knox, M. Wilson, C. Manach, and A. Scalbert, "Hmdb 4.0: The human metabolome database for 2018," *Nucleic Acids Research*, vol. 46, pp. D608–D617, 2018.

[115] M. Sud, E. Fahy, D. Cotter, K. Azam, I. Vadivelu, C. Burant, A. Edison, O. Fiehn, R. Higashi, K. S. Nair, S. Sumner, and S. Subramaniam, "Metabolomics workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools," *Nucleic Acids Research*, vol. 44, pp. D463–D470, 2016.

[116] Z. Zhang, T. Shen, B. Rui, W. Zhou, X. Zhou, C. Shang, C. Xin, X. Liu, G. Li, J. Jiang, C. Li, R. Li, M. Han, S. You, G. Yu, Y. Yi, H. Wen, Z. Liu, and X. Xie, "Cecafdb: a curated database for the documentation, visualization and comparative analysis of central carbon metabolic flux distributions explored by 13c-fluxomics," *Nucl. Acids Res.*, vol. 43, pp. 549–557, 2015.

[117] S. A. Becker and B. O. Palsson, "Context-specific metabolic networks are consistent with experiments," *PLoS Computational Biology*, vol. 4, p. e1000082, 5 2008.

[118] A. Bordbar, M. L. Mo, E. S. Nakayasu, A. C. Schrimpe-Rutledge, Y. M. Kim, T. O. Metz, M. B. Jones, B. C. Frank, R. D. Smith, S. N. Peterson, D. R. Hyduke, J. N. Adkins, and B. O. Palsson, "Model-driven multi-omic data analysis elucidates metabolic immunomodulators of macrophage activation," *Molecular Systems Biology*, vol. 8, pp. 1–12, 2012.

[119] B. J. Schmidt, A. Ebrahim, T. O. Metz, J. N. Adkins, B. Palsson, and D. R. Hyduke, "Gim3e: Condition-specific models of cellular metabolism developed from metabolomics and expression data," *Bioinformatics*, vol. 29, pp. 2900–2908, 2013.

[120] C. Colijn, A. Brandes, J. Zucker, D. S. Lun, B. Weiner, M. R. Farhat, T.-Y. Cheng, D. B. Moody, M. Murray, and J. E. Galagan, "Interpreting expression data with metabolic flux models: Predicting mycobacterium tuberculosis mycolic acid production," *PLoS Computational Biology*, vol. 5, p. e1000489, 8 2009.

[121] T. Shlomi, M. N. Cabili, M. J. Herrgård, B. Palsson, and E. Ruppin, "Network-based prediction of human tissue-specific metabolism," *Nature Biotechnology*, vol. 26, pp. 1003–1010, 2008.

[122] R. Agren, A. Mardinoglu, A. Asplund, C. Kampf, M. Uhlen, and J. Nielsen, "Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling," *Molecular Systems Biology*, vol. 10, pp. 1–13, 3 2014.

[123] L. Jerby, T. Shlomi, and E. Ruppin, "Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism," *Molecular Systems Biology*, vol. 6, pp. 1–9, 1 2010.

[124] Y. Wang, J. A. Eddy, and N. D. Price, "Reconstruction of genome-scale metabolic models for 126 human tissues using mcadre," *BMC Systems Biology*, vol. 6, 2012.

[125] N. Vlassis, M. P. Pacheco, and T. Sauter, "Fast reconstruction of compact context-specific metabolic network models," *PLoS Computational Biology*, vol. 10, p. e1003424, 1 2014.

[126] V. Vieira, J. Ferreira, and M. Rocha, "A pipeline for the reconstruction and evaluation of context-specific human metabolic models at a large-scale," *PLoS Computational Biology*, vol. 18, 6 2022.

[127] A. Schultz and A. A. Qutub, "Reconstruction of tissue-specific metabolic networks using corda," *PLOS Computational Biology*, vol. 12, p. e1004808, 3 2016.

[128] C. J. Joshi, S. M. Schinn, A. Richelle, I. Shamie, E. J. O'Rourke, and N. E. Lewis, "Standep: Capturing transcriptomic variability improves context-specific metabolic models," *PLoS Computational Biology*, vol. 16, pp. 1–24, 2020.

[129] S. McGarrity, Ósk Anuforo, H. Halldórsson, A. Bergmann, S. Halldórsson, S. Palsson, H. H. Henriksen, P. I. Johansson, and Óttar Rolfsson, "Metabolic systems analysis of lps induced endothelial dysfunction applied to sepsis patient stratification," *Scientific Reports*, vol. 8, pp. 1–14, 2018.

[130] S. Aller, A. Scott, M. Sarkar-Tyson, and O. S. Soyer, "Integrated human-virus metabolic stoichiometric modelling predicts host-based antiviral targets against chikungunya, dengue and zika viruses," *Journal of the Royal Society Interface*, vol. 15, pp. 1–12, 2018.

[131] J. Y. Ryu, H. U. Kim, and S. Y. Lee, "Reconstruction of genome-scale human metabolic models using omics data," *Integrative Biology (United Kingdom)*, vol. 7, pp. 859–868, 2015.

[132] J. E. Lewis, T. E. Forshaw, D. A. Boothman, C. M. Furdui, and M. L. Kemp, "Personalized genome-scale metabolic models identify targets of redox metabolism in radiation-resistant tumors," *Cell Systems*, vol. 12, pp. 68–81.e11, 2021.

[133] L. Ma, Y. Tao, A. Duran, V. Llado, A. Galvez, J. F. Barger, E. A. Castilla, J. Chen, T. Yajima, A. Porollo, M. Medvedovic, L. M. Brill, D. R. Plas,

S. J. Riedl, M. Leitges, M. T. Diaz-Meco, A. D. Richardson, and J. Moscat, "Control of nutrient stress-induced metabolic reprogramming by pkcζ in tumorigenesis," *Cell*, vol. 152, pp. 599–611, 2013.

[134] T. Shlomi, T. Benyamini, E. Gottlieb, R. Sharan, and E. Ruppin, "Genome-scale metabolic modeling elucidates the role of proliferative adaptation in causing the warburg effect," *PLoS Computational Biology*, vol. 7, pp. 1–8, 2011.

[135] C. Damiani, D. Maspero, M. D. Filippo, R. Colombo, D. Pescini, A. Graudenzi, H. V. Westerhoff, L. Alberghina, M. Vanoni, and G. Mauri, "Integration of single-cell rna-seq data into population models to characterize cancer metabolism," *PLoS Computational Biology*, vol. 15, pp. 1–25, 2019.

[136] E. J. O'Brien, J. A. Lerman, R. L. Chang, D. R. Hyduke, and B. Palsson, "Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction," *Molecular systems biology*, vol. 9, 2013.

[137] B. J. Sánchez, C. Zhang, A. Nilsson, P. Lahtvee, E. J. Kerkhoven, and J. Nielsen, "Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints," *Molecular Systems Biology*, vol. 13, p. 935, 2017.

[138] Y. Chen and X. Wang, "Mirdb: An online database for prediction of functional microrna targets," *Nucleic Acids Research*, vol. 48, pp. D127–D131, 2020.

[139] I. Massaiu, L. Pasotti, N. Sonnenschein, E. Rama, M. Cavaletti, P. Magni, C. Calvio, and M. J. Herrgård, "Integration of enzymatic data in bacillus subtilis genome-scale metabolic model improves phenotype predictions and enables in silico design of poly-γ-glutamic acid production strains," *Microbial Cell Factories*, vol. 18, pp. 1–20, 1 2019.

[140] S. Chandrasekaran and N. D. Price, "Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in escherichia coli and mycobacterium tuberculosis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, pp. 17845–17850, 2010.

[141] S. Imam, S. Schäuble, A. N. Brooks, N. S. Baliga, and N. D. Price, "Data-driven integration of genome-scale regulatory and metabolic network models," *Frontiers in Microbiology*, vol. 6, pp. 1–10, 2015.

[142] H. Iwasaki and T. Suda, "Cancer stem cells and their niche," *Cancer Science*, vol. 100, pp. 1166–1172, 2009.

[143] A. Z. Ayob and T. S. Ramasamy, "Cancer stem cells as key drivers of tumour progression," *Journal of Biomedical Science*, vol. 25, pp. 1–18, 2018.

[144] H. A. Coller, "Is cancer a metabolic disease?," *American Journal of Pathology*, vol. 184, pp. 4–17, 2014.

[145] S. Fernández-Arroyo, E. Cuyàs, J. Bosch-Barrera, T. Alarcón, J. Joven, and J. A. Menendez, "Activation of the methylation cycle in cells reprogrammed into a stem cell-like state," *Oncoscience*, vol. 2, pp. 958–967, 1 2016.

[146] A. M. Intlekofer and L. W. S. Finley, "Metabolic signatures of cancer cells and stem cells," *Nature Metabolism*, vol. 1, pp. 177–188, 2019.

[147] H. Fouladiha and S.-A. Marashi, "Biomedical applications of cell- and tissue-specific metabolic network models," *Journal of Biomedical Informatics*, vol. 68, pp. 35–49, 4 2017.

[148] I. Larsson, M. Uhlén, C. Zhang, and A. Mardinoglu, "Genome-scale metabolic modeling of glioblastoma reveals promising targets for drug development," *Frontiers in Genetics*, vol. 11, pp. 1–12, 4 2020.

[149] O. Folger, L. Jerby, C. Frezza, E. Gottlieb, E. Ruppin, and T. Shlomi, "Predicting selective drug targets in cancer through metabolic networks," *Molecular Systems Biology*, vol. 7, pp. 1–10, 2011.

[150] K. Yizhak, S. E. L. Dévédec, V. M. Rogkoti, F. Baenke, V. C. Boer, C. Frezza, A. Schulze, B. Water, and E. Ruppin, "A computational study of the warburg effect identifies metabolic targets inhibiting cancer migration," *Molecular Systems Biology*, vol. 10, pp. 1–12, 2014.

[151] Y. He, Y. Wang, B. Zhang, Y. Li, L. Diao, L. Lu, and J. Yao, "Revealing the metabolic characteristics of human embryonic stem cells by genome-scale metabolic modeling," *FEBS Letters*, vol. 592, pp. 3670–3682, 2018.

[152] W. Hur, J. Y. Ryu, H. U. Kim, S. W. Hong, and E. B. Lee, "Systems approach to characterize the metabolism of liver cancer stem cells expressing cd133," *Scientific Reports*, vol. 7, pp. 1–11, 2017.

[153] A. Richelle, C. Joshi, and N. E. Lewis, "Assessing key decisions for transcriptomic data integration in biochemical networks," *PLoS Computational Biology*, vol. 15, pp. 1–18, 2019.

[154] D. Ciavardelli, C. Rossi, D. Barcaroli, S. Volpe, A. Consalvo, M. Zucchelli, A. D. Cola, E. Scavo, R. Carollo, D. D'Agostino, F. Forlì, S. D'Aguanno, M. Todaro, G. Stassi, C. D. Ilio, V. D. Laurenzi, and A. Urbani, "Breast cancer stem cells rely on fermentative glycolysis and are sensitive to 2-deoxyglucose treatment," *Cell Death and Disease*, vol. 5, pp. 1–12, 2014.

[155] J. Liao, F. Qian, N. Tchabo, P. Mhawech-Fauceglia, A. Beck, Z. Qian, X. Wang, W. J. Huss, S. B. Lele, C. D. Morrison, and K. Odunsi, "Ovarian cancer spheroid cells with stem cell-like properties contribute to tumor generation, metastasis and chemotherapy resistance through hypoxia-resistant metabolism," *PLoS ONE*, vol. 9, pp. 1–13, 2014.

[156] X. Zhang, A. D. Milito, A. Demiroglu-Zergeroglu, J. Gullbo, P. D'Arcy, and S. Linder, "Eradicating quiescent tumor cells by targeting mitochondrial bioenergetics," *Trends in Cancer*, vol. 2, pp. 657–663, 2016.

[157] X. qun Ye, Q. Li, G. hui Wang, F. fen Sun, G. jun Huang, X. wu Bian, and S. cang Yu, "Mitochondrial and energy metabolism-related properties as novel indicators of lung cancer stem cells," *International Journal of Cancer*, vol. 129, pp. 820–831, 2011.

[158] E. Vlashi, C. Lagadec, L. Vergnes, K. Reune, P. Frohnen, M. Chan, Y. Alhiyari, M. B. Dratver, and F. Pajonk, "Metabolic differences in breast cancer stem cells and differentiated progeny," *Breast Cancer Res. Treat.*, vol. 146, pp. 525–534, 2015.

[159] A. Pastò, C. Bellio, G. Pilotto, V. Ciminale, M. Silic-Benussi, G. Guzzo, A. Rasola, C. Frasson, G. Nardo, E. Zulato, M. O. Nicoletto, M. Manicone, S. Indraccolo, and A. Amadori, "Cancer stem cells from epithelial ovarian cancer patients privilege oxidative phosphorylation, and resist glucose deprivation," *Oncotarget*, vol. 5, pp. 4305–4319, 2014.

[160] M. Janiszewska, M. L. Suvà, N. Riggi, R. H. Houtkooper, J. Auwerx, V. Clément-Schatlo, I. Radovanovic, E. Rheinbay, P. Provero, and I. Stamenkovic, "Imp2 controls oxidative phosphorylation and is crucial for preserving glioblastoma cancer stem cells," *Genes & Development*, vol. 26, pp. 1926–1944, 9 2012.

[161] G. Liu, Q. Luo, H. Li, Q. Liu, and Y. Ju, "Increased oxidative phosphorylation is required for stemness maintenance in liver cancer stem cells from hepatocellular carcinoma cell line hcclm3 cells," *International journal of molecular sciences*, vol. 21, pp. 1–13, 2020.

[162] S. Pavlides, D. Whitaker-Menezes, R. Castello-Cros, N. Flomenberg, A. K. Witkiewicz, P. G. Frank, M. C. Casimiro, C. Wang, P. Fortina, S. Addya, R. G. Pestell, U. E. Martinez-Outschoorn, F. Sotgia, and M. P. Lisanti, "The reverse warburg effect: Aerobic glycolysis in cancer associated fibroblasts and the tumor stroma," *Cell Cycle*, vol. 8, pp. 3984–4001, 2009.

[163] C. M. Sousa, D. E. Biancur, X. Wang, C. J. Halbrook, M. H. Sherman, L. Zhang, D. Kremer, R. F. Hwang, A. K. Witkiewicz, H. Ying, J. M. Asara, R. M. Evans, L. C. Cantley, C. A. Lyssiotis, and A. C. Kimmelman, "Pancreatic stellate cells support tumour metabolism through autophagic alanine secretion," *Nature*, vol. 536, pp. 479–483, 2016.

[164] E. Dornier, N. Rabas, L. Mitchell, D. Novo, S. Dhayade, S. Marco, G. MacKay, D. Sumpton, M. Pallares, C. Nixon, K. Blyth, I. R. MacPherson, E. Rainero, and J. C. Norman, "Glutaminolysis drives membrane trafficking to promote invasiveness of breast cancer cells," *Nature Communications*, vol. 8, pp. 1–14, 2017.

[165] X. Xu, L. Wang, Q. Zang, S. Li, L. Li, Z. Wang, J. He, B. Qiang, W. Han, R. Zhang, X. Peng, and Z. Abliz, "Rewiring of purine metabolism in response to acidosis stress in glioma stem cells," *Cell Death and Disease*, vol. 12, 2021.

[166] W. C. Gao, Y. T. Xu, T. Chen, Z. G. Du, X. J. Liu, Z. Q. Hu, D. Wei, C. F. Gao, W. Zhang, and Q. Q. Li, "Targeting oxidative pentose phosphate pathway prevents recurrence in mutant kras colorectal carcinomas," *PLoS Biology*, vol. 17, pp. 1–28, 2019.

[167] S. Monick, V. Mohanty, M. Khan, G. Yerneni, R. Kumar, J. Cantu, S. Ichi, G. Xi, B. R. Singh, T. Tomita, and C. S. Mayanil, "A phenotypic switch of differentiated glial cells to dedifferentiated cells is regulated by folate receptor α," *Stem Cells*, vol. 37, pp. 1441–1454, 2019.

[168] E. A. Ananieva and A. C. Wilkinson, "Branched-chain amino acid metabolism in cancer," *Current Opinion in Clinical Nutrition and Metabolic Care*, vol. 21, pp. 64–70, 2018.

[169] M. A. B. Melone, A. Valentino, S. Margarucci, U. Galderisi, A. Giordano, and G. Peluso, "The carnitine system and cancer metabolic plasticity," *Cell Death and Disease*, vol. 9, pp. 1–12, 2018.

[170] N. Liu, S. Li, N. Wu, and K. S. Cho, "Acetylation and deacetylation in cancer stem-like cells," *Oncotarget*, vol. 8, pp. 89315–89325, 2017.

[171] A. Valentino, A. Calarco, A. D. Salle, M. Finicelli, S. Crispi, R. A. Calogero, F. Riccardo, A. Sciarra, A. Gentilucci, U. Galderisi, S. Margarucci, and G. Peluso, "Deregulation of micrornas mediated control of carnitine cycle in prostate cancer: molecular basis and pathophysiological consequences," *Oncogene*, vol. 36, pp. 6030–6040, 2017.

[172] N. Albarakati, D. Khayyat, A. Dallol, J. Al-Maghrabi, and T. Nedjadi, "The prognostic impact of gstm1/gstp1 genetic variants in bladder cancer," *BMC Cancer*, vol. 19, pp. 1–11, 2019.

[173] J. E. Klaunig and L. M. Kamendulis, "The role of oxidative stress in carcinogenesis," *Annual Review of Pharmacology and Toxicology*, vol. 44, pp. 239–267, 2004.

[174] Z. Ye and J. M. Parry, "Meta-analysis of 20 case-control studies on the n-acetyltransferase 2 acetylation status and colorectal cancer risk.," *Medical science monitor : international medical journal of experimental and clinical research*, vol. 8, pp. CR558–65, 8 2002.

[175] S. Lizard-Nacol, B. Coudert, P. Colosetti, J. M. Riedinger, P. Fargeot, and P. Brunet-Lecomte, "Glutathione s-transferase m1 null genotype: Lack of association with tumour characteristics and survival in advanced breast cancer," *Breast Cancer Research*, vol. 1, pp. 81–87, 1999.

[176] X. Yin and J. Chen, "Is there any association between glutathione s-transferases m1 and glutathione s-transferases t1 gene polymorphisms and endometrial cancer risk? a meta-analysis," *International Journal of Preventive Medicine*, vol. 8, pp. 1–6, 2017.

[177] A. Agudo, N. Sala, G. Pera, G. Capellá, A. Berenguer, N. García, D. Palli, H. Boeing, G. D. Giudice, C. Saieva, F. Carneiro, F. Berrino, C. Sacerdote, R. Tumino, S. Panico, G. Berglund, H. Simán, R. Stenling, G. Hallmans, C. Martínez, P. Amiano, A. Barricarte, C. Navarro, J. R. Quirós, N. Allen, T. Key, S. Bingham, K. T. Khaw, J. Linseisen, G. Nagel, K. Overvad, A. Tjonneland, A. Olsen, H. B. Bueno-De-Mesquita, H. C. Boshuizen, P. H. Peeters, M. E. Numans, F. Clavel-Chapelon, M. C. Boutron-Ruault, A. Trichopoulou, E. Lund, H. Bläker, M. Jenab, P. Ferrari, T. Norat, E. Riboli, and C. A. González, "No association between polymorphisms in cyp2e1, gstm1, nat1, nat2 and the risk of gastric adenocarcinoma in the european prospective investigation into cancer and nutrition," *Cancer Epidemiology Biomarkers and Prevention*, vol. 15, pp. 1043–1045, 2006.

[178] D. M. Townsend and K. D. Tew, "The role of glutathione-s-transferase in anti-cancer drug resistance," *Oncogene*, vol. 22, pp. 7369–7375, 10 2003.

[179] K. Hama, Y. Fujiwara, T. Hayama, T. Ozawa, K. Nozawa, K. Matsuda, Y. Hashiguchi, and K. Yokoyama, "Very long-chain fatty acids are accumulated in triacylglycerol and nonesterified forms in colorectal cancer tissues," *Scientific Reports*, vol. 11, pp. 1–10, 2021.

[180] A. Mika, J. Kobiela, A. Czumaj, M. Chmielewski, P. Stepnowski, and T. Sledzinski, "Hyper-elongation in colorectal cancer tissue - cerotic acid is a potential novel serum metabolic marker of colorectal malignancies," *Cellular Physiology and Biochemistry*, vol. 41, pp. 722–730, 2017.

[181] M. Hilvo, C. Denkert, L. Lehtinen, B. Müller, S. Brockmöller, T. Seppänen-Laakso, J. Budczies, E. Bucher, L. Yetukuri, S. Castillo, E. Berg, H. Nygren, M. Sysi-Aho, J. L. Griffin, O. Fiehn, S. Loibl, C. Richter-Ehrenstein, C. Radke, T. Hyötyläinen, O. Kallioniemi, K. Iljin, and M. Orešič, "Novel theranostic opportunities offered by characterization of altered membrane lipid metabolism in breast cancer progression," *Cancer Research*, vol. 71, pp. 3236–3245, 2011.

[182] R. Safi, T. Mohsen-Kanson, G. Nemer, B. Dekmak, N. Rubeiz, M. El-Sabban, D. Nassar, A. Eid, O. Abbas, A. G. Kibbi, and M. Kurban, "Loss of ferrochelatase is protective against colon cancer cells: ferrochelatase a possible regulator of the long noncoding rna h19," *Journal of Gastrointestinal Oncology*, vol. 10, pp. 859–868, 2019.

[183] S. Bazzocco, H. Dopeso, F. Carton-Garcia, I. Macaya, E. Andretta, F. Chionh, P. Rodrigues, M. Garrido, H. Alazzouzi, R. Nieto, A. Sanchez, S. Schwartz, J. Bilic, J. M. Mariadason, and D. Arango, "Highly expressed genes in rapidly proliferating tumor cells as new targets for colorectal cancer treatment," *Clinical Cancer Research*, vol. 21, pp. 3695–3704, 2015.

[184] G. T. Yiang, T. Y. Chen, C. Chen, Y. T. Hung, K. C. Hsueh, T. K. Wu, Y. R. Pan, Y. C. Chien, C. H. Chen, Y. Yu, and C. W. Wei, "Antioxidant vitamins promote anticancer effects on low-concentration methotrexate-treated glioblastoma cells via enhancing the caspase-3 death pathway," *Food Science and Nutrition*, vol. 9, pp. 3308–3316, 2021.

[185] H. Tan, L. He, and Z. Cheng, "Inhibition of eif4e signaling by ribavirin selectively targets lung cancer and angiogenesis," *Biochemical and Biophysical Research Communications*, vol. 529, pp. 519–525, 2020.

[186] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, and E. E. Bolton, "Pubchem in 2021: new data content and improved web interfaces," *Nucleic Acids Research*, vol. 49, pp. D1388–D1395, 2021.

[187] M. N. Scheiber, P. M. Watson, T. Rumboldt, C. Stanley, R. C. Wilson, V. J. Findlay, P. E. Anderson, and D. K. Watson, "Fli1 expression is correlated with breast cancer cellular growth, migration, and invasion and altered gene expression," *Neoplasia (United States)*, vol. 16, pp. 801–813, 2014.

[188] N. Ramani, P. P. Aung, W. J. Hwu, P. Nagarajan, M. T. Tetzlaff, J. L. Curry, D. Ivan, V. G. Prieto, and C. A. Torres-Cabala, "Aberrant expression of fli-1 in melanoma," *Journal of Cutaneous Pathology*, vol. 44, pp. 790–793, 2017.

[189] J. Zhang, H. Guo, H. Zhang, and H. Wang, "Putative tumor suppressor mir-145 inhibits colon cancer cell growth by targeting oncogene friend leukemia virus integration 1," *Cancer*, vol. 117, pp. 86–95, 2011.

[190] H. V. Erkizan, V. N. Uversky, and J. A. Toretsky, "Oncogenic partnerships: Ews-fli1 protein interactions initiate key pathways of ewing's sarcoma," *Clinical Cancer Research*, vol. 16, pp. 4077–4083, 8 2010.

[191] O. Awad, J. T. Yustein, P. Shah, N. Gul, V. Katuri, A. O'Neill, Y. Kong, M. L. Brown, J. A. Toretsky, and D. M. Loeb, "High aldh activity identifies chemotherapy-resistant ewing's sarcoma stem cells that retain sensitivity to ews-fli1 inhibition," *PLoS ONE*, vol. 5, pp. 1–17, 2010.

[192] W. Song, L. Hu, W. Li, G. Wang, Y. Li, L. Yan, A. Li, and J. Cui, "Oncogenic fli-1 is a potential prognostic marker for the progression of epithelial ovarian cancer," *BMC Cancer*, vol. 14, pp. 1–9, 2014.

[193] B. Miao, A. S. Bauer, K. Hufnagel, Y. Wu, M. Trajkovic-Arsic, A. C. Pirona, N. Giese, J. Taipale, J. T. Siveke, J. D. Hoheisel, and S. Lueong, "The transcription factor fli1 promotes cancer progression by affecting cell cycle regulation," *International Journal of Cancer*, vol. 147, pp. 189–201, 2020.

[194] S. H. Issaq, A. Mendoza, R. Kidner, T. I. Rosales, D. Y. Duveau, C. M. Heske, J. M. Rohde, M. B. Boxer, C. J. Thomas, R. J. DeBerardinis, and L. J. Helman, "Ews-fli1-regulated serine synthesis and exogenous serine

are necessary for ewing sarcoma cellular proliferation and tumor growth," *Molecular Cancer Therapeutics*, vol. 19, pp. 1520–1529, 2020.

[195] S. Schlottmann, H. V. Erkizan, J. S. Barber-Rotenberg, C. Knights, A. Cheema, A. Üren, M. L. Avantaggiati, and J. A. Toretsky, "Acetylation increases ews-fli1 dna binding and transcriptional activity," *Frontiers in Oncology*, vol. 2, pp. 1–12, 2012.

[196] E. V. Abel, M. Goto, B. Magnuson, S. Abraham, N. Ramanathan, E. Hotaling, A. A. Alaniz, C. Kumar-Sinha, M. L. Dziubinski, S. Urs, L. Wang, J. Shi, M. Waghray, M. Ljungman, H. C. Crawford, and D. M. Simeone, "Hnf1a is a novel oncogene that regulates human pancreatic cancer stem cell properties," *eLife*, vol. 7, pp. 1–35, 2018.

[197] S. Fujino, N. Miyoshi, A. Ito, M. Yasui, C. Matsuda, M. Ohue, M. Uemura, T. Mizushima, Y. Doki, and H. Eguchi, "Hnf1a regulates colorectal cancer progression and drug resistance as a downstream of pou5f1," *Scientific Reports*, vol. 11, pp. 1–15, 2021.

[198] Z. Yuan, M. Ye, J. Qie, and T. Ye, "Foxa1 promotes cell proliferation and suppresses apoptosis in hcc by directly regulating mir-212-3p/foxa1/agr2 signaling pathway," *OncoTargets and Therapy*, vol. Volume 13, pp. 5231–5240, 6 2020.

[199] Y. Lu, D. Xu, J. Peng, Z. Luo, C. Chen, Y. Chen, H. Chen, M. Zheng, P. Yin, and Z. Wang, "Hnf1a inhibition induces the resistance of pancreatic cancer cells to gemcitabine by targeting abcb1," *EBioMedicine*, vol. 44, pp. 403–418, 2019.

[200] S. A. Camolotto, S. Pattabiraman, T. L. Mosbruger, A. Jones, V. K. Belova, G. Orstad, M. Streiff, L. Salmond, C. Stubben, K. H. Kaestner, and E. L. Snyder, "Foxa1 and foxa2 drive gastric differentiation and suppress squamous identity in nkx2-1-negative lung cancer," *eLife*, vol. 7, pp. 1–28, 11 2018.

[201] X. Wang, H. Xiao, D. Wu, D. Zhang, and Z. Zhang, "mir-335-5p regulates cell cycle and metastasis in lung adenocarcinoma by targeting ccnb2," *OncoTargets and Therapy*, vol. 13, pp. 6255–6263, 2020.

[202] J. Wang, Z. He, J. Xu, P. Chen, and J. Jiang, "Long noncoding rna linc00941 promotes pancreatic cancer progression by competitively binding mir-335-5p to regulate rock1-mediated limk1/cofilin-1 signaling," *Cell Death and Disease*, vol. 12, pp. 1–5, 2021.

[203] Y. Gao, Y. Wang, X. Wang, C. Zhao, F. Wang, J. Du, H. Zhang, H. Shi, Y. Feng, D. Li, J. Yan, Y. Yao, W. Hu, R. Ding, M. Zhang, L. Wang, C. Huang, and J. Zhang, "mir-335-5p suppresses gastric cancer progression by targeting mapk10," *Cancer Cell International*, vol. 21, pp. 1–12, 2021.

[204] W. Du, H. Tang, Z. Lei, J. Zhu, Y. Zeng, Z. Liu, and J. A. Huang, "Mir-335-5p inhibits tgf-β1-induced epithelial-mesenchymal transition in non-small cell lung cancer via rock1," *Respiratory Research*, vol. 20, pp. 1–11, 2019.

[205] H. Liang, C. Zhang, H. Guan, J. Liu, and Y. Cui, "Lncrna dancr promotes cervical cancer progression by upregulating rock1 via sponging mir-335-5p," *Journal of Cellular Physiology*, vol. 234, pp. 7266–7278, 2019.

[206] Q. Jia, L. Ye, S. Xu, H. Xiao, S. Xu, Z. Shi, J. Li, and Z. Chen, "Circular rna 0007255 regulates the progression of breast cancer through mir-335-5p/six2 axis," *Thoracic Cancer*, vol. 11, pp. 619–630, 2020.

[207] D. Zhang and N. Yang, "mir-335-5p inhibits cell proliferation, migration and invasion in colorectal cancer through downregulating ldhb," *Journal of B.U.ON.*, vol. 24, pp. 1128–1136, 2019.

[208] L. Brisson, P. Bański, M. Sboarina, C. Dethier, P. Danhier, M. J. Fontenille, V. F. V. Hée, T. Vazeille, M. Tardy, J. Falces, C. Bouzin, P. E. Porporato, R. Frédérick, C. Michiels, T. Copetti, and P. Sonveaux, "Lactate dehydrogenase b controls lysosome activity and autophagy in cancer," *Cancer Cell*, vol. 30, pp. 418–431, 2016.

[209] K. Miyamoto, N. Seki, R. Matsushita, M. Yonemori, H. Yoshino, M. Nakagawa, and H. Enokida, "Tumour-suppressive mirna-26a-5p and mir-26b-5p inhibit cell aggressiveness by regulating plod2 in bladder cancer," *British Journal of Cancer*, vol. 115, pp. 354–363, 2016.

[210] J. Li, X. Li, X. Kong, Q. Luo, J. Zhang, and L. Fang, "Mirna-26b inhibits cellular proliferation by targeting cdk8 in breast cancer," *International Journal of Clinical and Experimental Medicine*, vol. 7, pp. 558–565, 2014.

[211] F. Fan, J. Lu, W. Yu, Y. Zhang, S. Xu, L. Pang, and B. Zhu, "Microrna-26b-5p regulates cell proliferation, invasion and metastasis in human intrahepatic cholangiocarcinoma by targeting s100a7," *Oncology Letters*, vol. 15, pp. 386–392, 2018.

[212] G. Duan, C. Ren, Y. Zhang, and S. Feng, "Microrna-26b inhibits metastasis of osteosarcoma via targeting ctgf and smad1," *Tumor Biology*, vol. 36, pp. 6201–6209, 2015.

[213] J. Liu, J. Liu, L. Chu, Y. Wang, Y. Duan, L. Feng, C. Yang, L. Wang, and D. Kong, "Novel peptide-dendrimer conjugates as drug carriers for targeting nonsmall cell lung cancer," *International Journal of Nanomedicine*, vol. 6, pp. 59–69, 2011.

[214] M. Lo, V. Ling, Y. Z. Wang, and P. W. Gout, "The xc- cystine/glutamate antiporter: a mediator of pancreatic cancer growth with a role in drug resistance," *British Journal of Cancer*, vol. 99, pp. 464–472, 2008.

[215] Y. Kuang, H. Xu, F. Lu, J. Meng, Y. Yi, H. Yang, H. Hou, H. Wei, and S. Su, "Inhibition of microrna let-7b expression by kdm2b promotes cancer progression by targeting ezh2 in ovarian cancer," *Cancer Science*, vol. 112, pp. 231–242, 2021.

[216] A. Buonfiglioli, I. E. Efe, D. Guneykaya, A. Ivanov, Y. Huang, E. Orlowski, C. Krüger, R. A. Deisz, D. Markovic, C. Flüh, A. G. Newman, U. C. Schneider, D. Beule, S. A. Wolf, O. Dzaye, D. H. Gutmann, M. Semtner, H. Kettenmann, and S. Lehnardt, "let-7 microrna regulate microglial function and suppress glioma growth through toll-like receptor 7," *Cell Reports*, vol. 29, pp. 3460–3471, 2019.

[217] H. Zhu, N. Shyh-Chang, and A. Segre, "The lin28/let-7 axis regulates glucose metabolism," *Cell*, vol. 147, pp. 81–94, 2011.

[218] P. Rocca-Serra, A. Brazma, H. Parkinson, U. Sarkans, M. Shojatalab, S. Contrino, J. Vilo, N. Abeygunawardena, G. Mukherjee, E. Holloway, M. Kapushesky, P. Kemmeren, G. G. Lara, A. Oezcimen, and S.-A. Sansone, "Arrayexpress: a public database of gene expression data at ebi," *Comptes Rendus Biologies*, vol. 326, pp. 1075–1078, 10 2003.

[219] I. Lappalainen, J. Almeida-King, V. Kumanduri, A. Senf, J. D. Spalding, S. Ur-Rehman, G. Saunders, J. Kandasamy, M. Caccamo, R. Leinonen, B. Vaughan, T. Laurent, F. Rowland, P. Marin-Garcia, J. Barker, P. Jokinen, A. C. Torres, J. R. D. Argila, O. M. Llobet, I. Medina, M. S. Puy, M. Alberich, S. D. L. Torre, A. Navarro, J. Paschall, and P. Flicek, "The european genome-phenome archive of human data consented for biomedical research," *Nature Genetics*, vol. 47, pp. 692–695, 2015.

[220] S. Andrews, "Fastqc: A quality control tool for high throughput sequence data [online].," *Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc*, 2010.

[221] A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: A flexible trimmer for illumina sequence data," *Bioinformatics*, vol. 30, pp. 2114–2120, 2014.

[222] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras, "Star: Ultrafast universal rna-seq aligner," *Bioinformatics*, vol. 29, pp. 15–21, 2013.

[223] S. Anders, P. T. Pyl, and W. Huber, "Htseq-a python framework to work with high-throughput sequencing data," *Bioinformatics*, vol. 31, pp. 166–169, 2015.

[224] M. Smid, R. R. J. C. V. D. Braak, H. J. G. V. D. Werken, J. V. Riet, A. V. Galen, V. D. Weerd, M. V. D. Vlugt-daane, S. I. Bril, Z. S. Lalmahomed, W. P. Kloosterman, S. M. Wilting, J. A. Foekens, and J. N. M. Ijzermans, "Gene length corrected trimmed mean of m-values (getmm) processing of rna-seq data performs similarly in intersample analyses while improving intrasample comparisons," *BMC bioinformatics*, vol. 19, pp. 1–13, 2018.

[225] B. S. Carvalho and R. A. Irizarry, "A framework for oligonucleotide microarray preprocessing," *Bioinformatics*, vol. 26, pp. 2363–2367, 2010.

[226] R. CoreTeam, "R : A language and environment for statistical computing.," 2020.

[227] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth, "Limma powers differential expression analyses for rna-sequencing and microarray studies," *Nucleic Acids Res*, vol. 43, p. e47, 2015.

[228] M. J. Dunning, M. L. Smith, M. E. Ritchie, and S. Tavaré, "Beadarray: R classes and methods for illumina bead-based data," *Bioinformatics*, vol. 23, pp. 2183–2184, 2007.

[229] A. Kauffmann, T. F. Rayner, H. Parkinson, M. Kapushesky, M. Lukk, A. Brazma, and W. Huber, "Importing arrayexpress datasets into r/bioconductor," *Bioinformatics*, vol. 25, pp. 2092–2094, 2009.

[230] D. Sean and P. S. Meltzer, "Geoquery: A bridge between the gene expression omnibus (geo) and bioconductor," *Bioinformatics*, vol. 23, pp. 1846–1847, 2007.

[231] F. S. Foundation., "Bash 4.4.20," *[Unix shell program]. Retrieved from: http://tiswww.case.edu/php/chet/bash/bashtop.html*, 2007.

[232] J. Ferreira, V. Vieira, J. Gomes, S. Correia, and M. Rocha, "Troppo - a python framework for the reconstruction of context-specific metabolic models," *Advances in Intelligent Systems and Computing*, vol. 1005, pp. 146–153, 2020.

[233] V. Vieira and M. Rocha, "Cobamp: A python framework for metabolic pathway analysis in constraint-based models," *Bioinformatics*, vol. 35, pp. 5361–5362, 2019.

[234] A. Ebrahim, J. A. Lerman, B. O. Palsson, and D. R. Hyduke, "Cobrapy: Constraints-based reconstruction and analysis for python," *BMC Systems Biology*, vol. 7, pp. 1–6, 2013.

[235] W. McKinney, "Data structures for statistical computing in python," *Proceedings of the 9th Python in Science Conference*, vol. 455, pp. 56–61, 2010.

[236] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences*, vol. 102, pp. 15545–15550, 10 2005.

[237] C. Feng, C. Song, Y. Liu, F. Qian, Y. Gao, Z. Ning, Q. Wang, Y. Jiang, Y. Li, M. Li, J. Chen, J. Zhang, and C. Li, "Knocktf: a comprehensive human gene expression profile database with knockdown/knockout of transcription factors," *Nucleic Acids Research*, vol. 48, pp. D93–D100, 1 2020.

[238] H.-Y. Huang, Y.-C.-D. Lin, J. Li, K.-Y. Huang, S. Shrestha, H.-C. Hong, Y. Tang, Y.-G. Chen, C.-N. Jin, Y. Yu, J.-T. Xu, Y.-M. Li, X.-X. Cai, Z.-Y. Zhou, X.-H. Chen, Y.-Y. Pei, L. Hu, J.-J. Su, S.-D. Cui, F. Wang, Y.-Y. Xie, S.-Y. Ding, M.-F. Luo, C.-H. Chou, N.-W. Chang, K.-W. Chen, Y.-H. Cheng, X.-H. Wan, W.-L. Hsu, T.-Y. Lee, F.-X. Wei, and H.-D. Huang, "mirtarbase 2020: updates to the experimentally validated microrna–target interaction database," *Nucleic Acids Research*, vol. 48, pp. D148–D154, 10 2019.

[239] T. Ugai, N. Sasamoto, H.-Y. Lee, M. Ando, M. Song, R. M. Tamimi, I. Kawachi, P. T. Campbell, E. L. Giovannucci, E. Weiderpass, T. R.

Rebbeck, and S. Ogino, "Is early-onset cancer an emerging global epidemic? current evidence and future implications," *Nature Reviews Clinical Oncology*, vol. 19, pp. 656–673, 10 2022.

[240] T. Barata, V. Vieira, R. Rodrigues, R. P. das Neves, and M. Rocha, "Reconstruction of tissue-specific genome-scale metabolic models for human cancer stem cells," *Computers in Biology and Medicine*, vol. 142, pp. 1–12, 3 2022.

[241] A. Salehzadeh-Yazdi, Y. Asgari, A. A. Saboury, and A. Masoudi-Nejad, "Computational analysis of reciprocal association of metabolism and epigenetics in the budding yeast: A genome-scale metabolic model (gsmm) approach," *PLoS ONE*, vol. 9, p. e111686, 11 2014.

[242] M. P. Pacheco, E. John, T. Kaoma, M. Heinäniemi, N. Nicot, L. Vallar, J.-L. Bueb, L. Sinkkonen, and T. Sauter, "Integrated metabolic modelling reveals cell-type specific epigenetic control points of the macrophage metabolic network," *BMC Genomics*, vol. 16, p. 809, 12 2015.

[243] S. Chandrasekaran, J. Zhang, Z. Sun, L. Zhang, C. A. Ross, Y.-C. Huang, J. M. Asara, H. Li, G. Q. Daley, and J. J. Collins, "Comprehensive mapping of pluripotent stem cell metabolism using dynamic genome-scale network modeling," *Cell Reports*, vol. 21, pp. 2965–2977, 12 2017.

[244] F. Shen, L. Boccuto, R. Pauly, S. Srikanth, and S. Chandrasekaran, "Genome-scale network model of metabolism and histone acetylation reveals metabolic dependencies of histone deacetylase inhibitors," *Genome Biology*, vol. 20, pp. 1–15, 2019.

[245] M. Turpin and G. Salbert, "5-methylcytosine turnover: Mechanisms and therapeutic implications in cancer," *Frontiers in Molecular Biosciences*, vol. 9, 8 2022.

[246] D. Cheishvili, L. Boureau, and M. Szyf, "Dna demethylation and invasive cancer: implications for therapeutics," *British Journal of Pharmacology*, vol. 172, pp. 2705–2715, 6 2015.

[247] M. Ehrlich, "Dna methylation in cancer: too much, but also too little," *Oncogene*, vol. 21, pp. 5400–5413, 8 2002.

[248] M. Ghandi, F. W. Huang, J. Jané-Valbuena, G. V. Kryukov, C. C. Lo, E. R. McDonald, J. Barretina, E. T. Gelfand, C. M. Bielski, H. Li, K. Hu, A. Y. Andreev-Drakhlin, J. Kim, J. M. Hess, B. J. Haas, F. Aguet,

B. A. Weir, M. V. Rothberg, B. R. Paolella, M. S. Lawrence, R. Akbani, Y. Lu, H. L. Tiv, P. C. Gokhale, A. de Weck, A. A. Mansour, C. Oh, J. Shih, K. Hadi, Y. Rosen, J. Bistline, K. Venkatesan, A. Reddy, D. Sonkin, M. Liu, J. Lehar, J. M. Korn, D. A. Porter, M. D. Jones, J. Golji, G. Caponigro, J. E. Taylor, C. M. Dunning, A. L. Creech, A. C. Warren, J. M. McFarland, M. Zamanighomi, A. Kauffmann, N. Stransky, M. Imielinski, Y. E. Maruvka, A. D. Cherniack, A. Tsherniak, F. Vazquez, J. D. Jaffe, A. A. Lane, D. M. Weinstock, C. M. Johannessen, M. P. Morrissey, F. Stegmeier, R. Schlegel, W. C. Hahn, G. Getz, G. B. Mills, J. S. Boehm, T. R. Golub, L. A. Garraway, and W. R. Sellers, "Next-generation characterization of the cancer cell line encyclopedia," *Nature*, vol. 569, pp. 503–508, 5 2019.

[249] W. C. Reinhold, S. Varma, M. Sunshine, V. Rajapakse, A. Luna, K. W. Kohn, H. Stevenson, Y. Wang, H. Heyn, V. Nogales, S. Moran, D. J. Goldstein, J. H. Doroshow, P. S. Meltzer, M. Esteller, and Y. Pommier, "The nci-60 methylome and its integration into cellminer," *Cancer Research*, vol. 77, pp. 601–612, 2 2017.

[250] T. Nakagawa, M. A. Lanaspa, I. S. Millan, M. Fini, C. J. Rivard, L. G. Sanchez-Lozada, A. Andres-Hernando, D. R. Tolan, and R. J. Johnson, "Fructose contributes to the warburg effect for cancer growth," *Cancer & Metabolism*, vol. 8, p. 16, 12 2020.

[251] X. Ding, W. Zhang, S. Li, and H. Yang, "The role of cholesterol metabolism in cancer.," *American journal of cancer research*, vol. 9, pp. 219–227, 2019.

[252] C. Carson and H. A. Lawson, "Epigenetics of metabolic syndrome," *Physiological Genomics*, vol. 50, pp. 947–955, 11 2018.

[253] M. Jain, R. Nilsson, S. Sharma, N. Madhusudhan, T. Kitami, A. L. Souza, R. Kafri, M. W. Kirschner, C. B. Clish, and V. K. Mootha, "Metabolite profiling identifies a key role for glycine in rapid cancer cell proliferation," *Science*, vol. 336, pp. 1040–1044, 5 2012.

[254] D. C. Zielinski, N. Jamshidi, A. J. Corbett, A. Bordbar, A. Thomas, and B. O. Palsson, "Systems biology analysis of drivers underlying hallmarks of cancer cell metabolism," *Scientific Reports*, vol. 7, pp. 1–14, 2017.

[255] V. Pereira, F. Cruz, and M. Rocha, "Mewpy: a computational strain optimization workbench in python," *Bioinformatics*, vol. 37, pp. 2494–2496, 8 2021.

[256] B. He, C. Zhang, X. Zhang, Y. Fan, H. Zeng, J. Liu, H. Meng, D. Bai, J. Peng, Q. Zhang, W. Tao, and C. Yi, "Tissue-specific 5-hydroxymethylcytosine landscape of the human genome," *Nature Communications*, vol. 12, pp. 1–12, 2021.

[257] Y. Kawasaki, Y. Kuroda, I. Suetake, S. Tajima, F. Ishino, and T. Kohda, "A novel method for the simultaneous identification of methylcytosine and hydroxymethylcytosine at a single base resolution," *Nucleic acids research*, vol. 45, p. e24, 2 2017.

[258] M. J. Booth and S. Balasubramanian, "Reduced bisulfite sequencing: Quantitative base-resolution sequencing of 5-formylcytosine," *Methods in molecular biology (Clifton, N.J.)*, vol. 2272, pp. 3–12, 2021.

[259] U. T. Shankavaram, S. Varma, D. Kane, M. Sunshine, K. K. Chary, W. C. Reinhold, Y. Pommier, and J. N. Weinstein, "Cellminer: A relational database and query tool for the nci-60 cancer cell lines," *BMC Genomics*, vol. 10, pp. 1–10, 6 2009.

[260] R. M. Kohli and Y. Zhang, "Tet enzymes, tdg and the dynamics of dna demethylation," *Nature*, vol. 502, pp. 472–479, 10 2013.

[261] C. Schulz, T. Kumelj, E. Karlsen, and E. Almaas, "Genome-scale metabolic modelling when changes in environmental conditions affect biomass composition," *PLoS Computational Biology*, vol. 17, pp. 1–22, 2021.

[262] D. Dikicioglu, B. Kırdar, and S. G. Oliver, "Biomass composition: the "elephant in the room" of metabolic modelling," *Metabolomics*, vol. 11, pp. 1690–1701, 2015.

[263] H. T. Lee, S. Oh, D. H. Ro, H. Yoo, and Y. W. Kwon, "The key role of dna methylation and histone acetylation in epigenetics of atherosclerosis," *Journal of Lipid and Atherosclerosis*, vol. 9, p. 419, 2020.

[264] Z. Yu, T. G. Pestell, M. P. Lisanti, and R. G. Pestell, "Cancer stem cells," *The International Journal of Biochemistry & Cell Biology*, vol. 44, pp. 2144–2151, 12 2012.

[265] W. Zhao, Y. Li, and X. Zhang, "Stemness-related markers in cancer," *Cancer Translational Medicine*, vol. 3, pp. 87–95, 2017.

[266] M. Todaro, M. G. Francipane, J. P. Medema, and G. Stassi, "Colon cancer stem cells: Promise of targeted therapy," *Gastroenterology*, vol. 138, pp. 2151–2162, 5 2010.

[267] W. Zeijlemaker, T. Grob, R. Meijer, D. Hanekamp, A. Kelder, J. C. Carbaat-Ham, Y. J. M. Oussoren-Brockhoff, A. N. Snel, D. Veldhuizen, W. J. Scholten, J. Maertens, D. A. Breems, T. Pabst, M. G. Manz, V. H. J. van der Velden, J. Slomp, F. Preijers, J. Cloos, A. A. van de Loosdrecht, B. Löwenberg, P. J. M. Valk, M. Jongen-Lavrencic, G. J. Ossenkoppele, and G. J. Schuurhuis, "Cd34+cd38- leukemic stem cell frequency to predict outcome in acute myeloid leukemia," *Leukemia*, vol. 33, pp. 1102–1112, 5 2018.

[268] A. Fendler, D. Bauer, J. Busch, K. Jung, A. Wulf-Goldenberg, S. Kunz, K. Song, A. Myszczyszyn, S. Elezkurtaj, B. Erguen, S. Jung, W. Chen, and W. Birchmeier, "Inhibiting wnt and notch in renal cancer stem cells and the implications for human patients," *Nature Communications*, vol. 11, p. 929, 12 2020.

[269] I. A. Silva, S. Bai, K. McLean, K. Yang, K. Griffith, D. Thomas, C. Ginestier, C. Johnston, A. Kueck, R. K. Reynolds, M. S. Wicha, and R. J. Buckanovich, "Aldehyde dehydrogenase in combination with cd133 defines angiogenic ovarian cancer stem cells that portend poor patient survival," *Cancer Research*, vol. 71, pp. 3991–4001, 6 2011.

[270] B. A. Smith, A. Sokolov, V. Uzunangelov, R. Baertsch, Y. Newton, K. Graim, C. Mathis, D. Cheng, J. M. Stuart, and O. N. Witte, "A basal stem cell signature identifies aggressive prostate cancer phenotypes," *Proceedings of the National Academy of Sciences*, vol. 112, pp. E6544–E6552, 11 2015.

[271] Y. Wang, L. He, Y. Du, P. Zhu, G. Huang, J. Luo, X. Yan, B. Ye, C. Li, P. Xia, G. Zhang, Y. Tian, R. Chen, and Z. Fan, "The long noncoding rna lnctcf7 promotes self-renewal of human liver cancer stem cells through activation of wnt signaling," *Cell Stem Cell*, vol. 16, pp. 413–425, 4 2015.

[272] L. B. Hoang-Minh, F. A. Siebzehnrubl, C. Yang, S. Suzuki-Hatano, K. Dajac, T. Loche, N. Andrews, M. S. Massari, J. Patel, K. Amin, A. Vuong, A. Jimenez-Pascual, P. Kubilis, T. J. Garrett, C. Moneypenny, C. A. Pacak, J. Huang, E. J. Sayour, D. A. Mitchell, M. R. Sarkisian, B. A. Reynolds, and L. P. Deleyrolle, "Infiltrative and drug-resistant slow-cycling cells support metabolic heterogeneity in glioblastoma," *The EMBO Journal*, vol. 37, pp. 1–21, 12 2018.

[273] X. Dai, S. Zhang, H. Cheng, D. Cai, X. Chen, and Z. Huang, "Fa2h exhibits tumor suppressive roles on breast cancers via cancer stemness control," *Frontiers in Oncology*, vol. 9, pp. 1–14, 10 2019.

[274] T. Ishiguro, A. Sato, H. Ohata, Y. Ikarashi, R. u Takahashi, T. Ochiya, M. Yoshida, H. Tsuda, T. Onda, T. Kato, T. Kasamatsu, T. Enomoto, K. Tanaka, H. Nakagama, and K. Okamoto, "Establishment and characterization of an in vitro model of ovarian cancer stem-like cells with an enhanced proliferative capacity," *Cancer Research*, vol. 76, pp. 150–160, 1 2016.

[275] H. Gal, N. Amariglio, L. Trakhtenbrot, J. Jacob-Hirsh, O. Margalit, A. Avigdor, A. Nagler, S. Tavor, L. Ein-Dor, T. Lapidot, E. Domany, G. Rechavi, and D. Givol, "Gene expression profiles of aml derived stem cells; similarity to hematopoietic stem cells," *Leukemia*, vol. 20, pp. 2147–2154, 12 2006.

[276] V. Justilien, M. P. Walsh, S. A. Ali, E. A. Thompson, N. R. Murray, and A. P. Fields, "The prkci and sox2 oncogenes are coamplified and cooperate to activate hedgehog signaling in lung squamous cell carcinoma," *Cancer Cell*, vol. 25, pp. 139–151, 2014.

[277] P. Sancho, E. Burgos-Ramos, A. Tavera, T. B. Kheir, P. Jagust, M. Schoenhals, D. Barneda, K. Sellers, R. Campos-Olivas, O. Graña, C. R. Viera, M. Yuneva, B. Sainz, and C. Heeschen, "Myc/pgc-1α balance determines the metabolic phenotype and plasticity of pancreatic cancer stem cells," *Cell Metabolism*, vol. 22, pp. 590–605, 10 2015.

[278] G. D. Wilson, B. Marples, S. Galoforo, T. J. Geddes, B. J. Thibodeau, R. Grénman, and J. Akervall, "Isolation and genomic characterization of stem cells in head and neck cancer," *Head & Neck*, vol. 35, pp. 1573–1582, 11 2013.

[279] M. Kanehisa, "Kegg: Kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, vol. 28, pp. 27–30, 1 2000.

[280] M. Bochtler, A. Kolano, and G. L. Xu, "Dna demethylation pathways: Additional players and regulators," *BioEssays*, vol. 39, pp. 1–13, 2017.

[281] L. Hu, J. Lu, J. Cheng, Q. Rao, Z. Li, H. Hou, Z. Lou, L. Zhang, W. Li, W. Gong, M. Liu, C. Sun, X. Yin, J. Li, X. Tan, P. Wang, Y. Wang,

D. Fang, Q. Cui, P. Yang, C. He, H. Jiang, C. Luo, and Y. Xu, "Structural insight into substrate preference for tet-mediated oxidation," *Nature*, vol. 527, pp. 118–122, 11 2015.

[282] X. Yin and Y. Xu, "Structure and function of tet enzymes," *Advances in Experimental Medicine and Biology*, vol. 945, pp. 275–302, 11 2016.

[283] P. Bansal, A. Morgat, K. B. Axelsen, V. Muthukrishnan, E. Coudert, L. Aimo, N. Hyka-Nouspikel, E. Gasteiger, A. Kerhornou, T. B. Neto, M. Pozzato, M.-C. Blatter, A. Ignatchenko, N. Redaschi, and A. Bridge, "Rhea, the reaction knowledgebase in 2022," *Nucleic Acids Research*, vol. 50, pp. D693–D700, 1 2022.

[284] A. V. Popov, I. R. Grin, A. P. Dvornikova, B. T. Matkarimov, R. Groisman, M. Saparbaev, and D. O. Zharkov, "Reading targeted dna damage in the active demethylation pathway: Role of accessory domains of eukaryotic ap endonucleases and thymine-dna glycosylases," *Journal of Molecular Biology*, vol. 432, pp. 1747–1768, 3 2020.

[285] A. Schön, E. Kaminska, F. Schelter, E. Ponkkonen, E. Korytiaková, S. Schiffers, and T. Carell, "Analysis of an active deformylation mechanism of 5-formyl-deoxycytidine (fdc) in stem cells," *Angewandte Chemie - International Edition*, vol. 59, pp. 5591–5594, 2020.

[286] K. Iwan, R. Rahimoff, A. Kirchner, F. Spada, A. S. Schröder, O. Kosmatchev, S. Ferizaj, J. Steinbacher, E. Parsa, M. Müller, and T. Carell, "5-formylcytosine to cytosine conversion by c-c bond cleavage in vivo," *Nature Chemical Biology*, vol. 14, pp. 72–78, 2018.

[287] Y. Feng, N. B. Xie, W. B. Tao, J. H. Ding, X. J. You, C. J. Ma, X. Zhang, C. Yi, X. Zhou, B. F. Yuan, and Y. Q. Feng, "Transformation of 5-carboxylcytosine to cytosine through c-c bond cleavage in human cells constitutes a novel pathway for dna demethylation," *CCS Chemistry*, vol. 3, pp. 994–1008, 4 2021.

[288] N. Bhutani, D. Burns, and H. Blau, "Dna demethylation dynamics," *Cell*, vol. 146, pp. 866–872, 9 2011.

[289] Y. Tsukada, "Hydroxylation mediates chromatin demethylation," *Journal of Biochemistry*, vol. 151, pp. 229–246, 3 2012.

[290] I. R. Grin, S. N. Khodyreva, G. A. Nevinsky, and D. O. Zharkov, "Deoxyribophosphate lyase activity of mammalian endonuclease viii-like proteins," *FEBS Letters*, vol. 580, pp. 4916–4922, 9 2006.

[291] K. D. Rasmussen and K. Helin, "Role of tet enzymes in dna methylation, development, and cancer," *Genes and Development*, vol. 30, pp. 733–750, 2016.

[292] T. Mollick and S. Laín, "Modulating pyrimidine ribonucleotide levels for the treatment of cancer," *Cancer & Metabolism*, vol. 8, 12 2020.

[293] A. Bird, "Dna methylation patterns and epigenetic memory," *Genes & Development*, vol. 16, pp. 6–21, 1 2002.

[294] A. Bellacosa and A. C. Drohat, "Role of base excision repair in maintaining the genetic and epigenetic integrity of cpg sites," *DNA Repair*, vol. 32, pp. 33–42, 8 2015.

[295] W. Zhang and J. Xu, "Dna methyltransferases and their roles in tumorigenesis," *Biomarker Research*, vol. 5, p. 1, 12 2017.

[296] G. Tsitsiridis, R. Steinkamp, M. Giurgiu, B. Brauner, G. Fobo, G. Frishman, C. Montrone, and A. Ruepp, "Corum: the comprehensive resource of mammalian protein complexes–2022," *Nucleic Acids Research*, vol. 1, pp. 13–14, 2013.

[297] Y. Zhou and I. Grummt, "The phd finger/bromodomain of norc interacts with acetylated histone h4k16 and is sufficient for rdna silencing," *Current Biology*, vol. 15, pp. 1434–1438, 8 2005.

[298] T. M. Geiman, U. T. Sankpal, A. K. Robertson, Y. Chen, M. Mazumdar, J. T. Heale, J. A. Schmiesing, W. Kim, K. Yokomori, Y. Zhao, and K. D. Robertson, "Isolation and characterization of a novel dna methyltransferase complex linking dnmt3b with components of the mitotic chromosome condensation machinery," *Nucleic Acids Research*, vol. 32, pp. 2716–2729, 2004.

[299] I. Suetake, F. Shinozaki, J. Miyagawa, H. Takeshima, and S. Tajima, "Dnmt3l stimulates the dna methylation activity of dnmt3a and dnmt3b through a direct interaction," *Journal of Biological Chemistry*, vol. 279, pp. 27816–27823, 6 2004.

[300] T. Paysan-Lafosse, M. Blum, S. Chuguransky, T. Grego, B. L. Pinto, G. Salazar, M. Bileschi, P. Bork, A. Bridge, L. Colwell, J. Gough, D. Haft,

I. Letunić, A. Marchler-Bauer, H. Mi, D. Natale, C. Orengo, A. Pandurangan, C. Rivoire, C. J. A. Sigrist, I. Sillitoe, N. Thanki, P. D. Thomas, S. C. E. Tosatto, C. Wu, and A. Bateman, "Interpro in 2022.," *Nucleic Acids Research*, pp. gkac993–gkac993, 11 2022.

[301] K. Cervantes-Gracia, A. Gramalla-Schmitz, J. Weischedel, and R. Chahwan, "Apobecs orchestrate genomic and epigenomic editing across health and disease," *Trends in Genetics*, vol. 37, pp. 1028–1043, 11 2021.

[302] A. Bateman, M.-J. Martin, S. Orchard, M. Magrane, R. Agivetova, S. Ahmad, E. Alpi, E. H. Bowler-Barnett, R. Britto, B. Bursteinas, H. Bye-A-Jee, R. Coetzee, A. Cukura, A. D. Silva, P. Denny, T. Dogan, T. Ebenezer, J. Fan, L. G. Castro, P. Garmiri, G. Georghiou, L. Gonzales, E. Hatton-Ellis, A. Hussein, A. Ignatchenko, G. Insana, R. Ishtiaq, P. Jokinen, V. Joshi, D. Jyothi, A. Lock, R. Lopez, A. Luciani, J. Luo, Y. Lussi, A. MacDougall, F. Madeira, M. Mahmoudy, M. Menchi, A. Mishra, K. Moulang, A. Nightingale, C. S. Oliveira, S. Pundir, G. Qi, S. Raj, D. Rice, M. R. Lopez, R. Saidi, J. Sampson, T. Sawford, E. Speretta, E. Turner, N. Tyagi, P. Vasudev, V. Volynkin, K. Warner, X. Watkins, R. Zaru, H. Zellner, A. Bridge, S. Poux, N. Redaschi, L. Aimo, G. Argoud-Puy, A. Auchincloss, K. Axelsen, P. Bansal, D. Baratin, M.-C. Blatter, J. Bolleman, E. Boutet, L. Breuza, C. Casals-Casas, E. de Castro, K. C. Echioukh, E. Coudert, B. Cuche, M. Doche, D. Dornevil, A. Estreicher, M. L. Famiglietti, M. Feuermann, E. Gasteiger, S. Gehant, V. Gerritsen, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, N. Hyka-Nouspikel, F. Jungo, G. Keller, A. Kerhornou, V. Lara, P. L. Mercier, D. Lieberherr, T. Lombardot, X. Martin, P. Masson, A. Morgat, T. B. Neto, S. Paesano, I. Pedruzzi, S. Pilbout, L. Pourcel, M. Pozzato, M. Pruess, C. Rivoire, C. Sigrist, K. Sonesson, A. Stutz, S. Sundaram, M. Tognolli, L. Verbregue, C. H. Wu, C. N. Arighi, L. Arminski, C. Chen, Y. Chen, J. S. Garavelli, H. Huang, K. Laiho, P. McGarvey, D. A. Natale, K. Ross, C. R. Vinayaka, Q. Wang, Y. Wang, L.-S. Yeh, J. Zhang, P. Ruch, and D. Teodoro, "Uniprot: the universal protein knowledgebase in 2021," *Nucleic Acids Research*, vol. 49, pp. D480–D489, 1 2021.

Appendix A

Supplementary Figures - Chapter 2

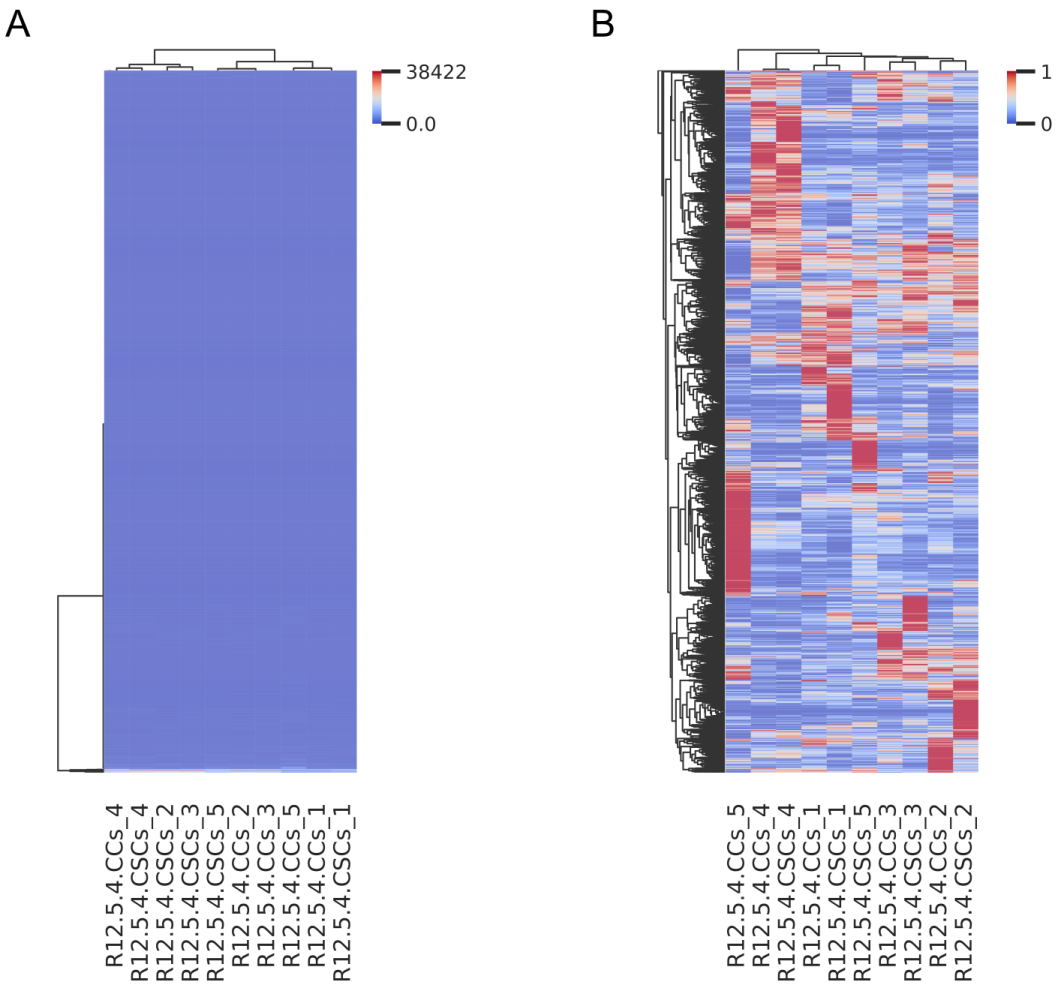

Figure A.1: Influence of min-max transformation on normalized gene expression data. Heatmaps show normalized gene expression data for the pancreas dataset without (**A**) and with (**B**) min-max transformation. There is a clear separation between samples with *min-max* normalization. Samples are overall grouped by donor, except CSC and CC samples of donor 5, which are respectively closer to the corresponding cell types of other donors than to each other.

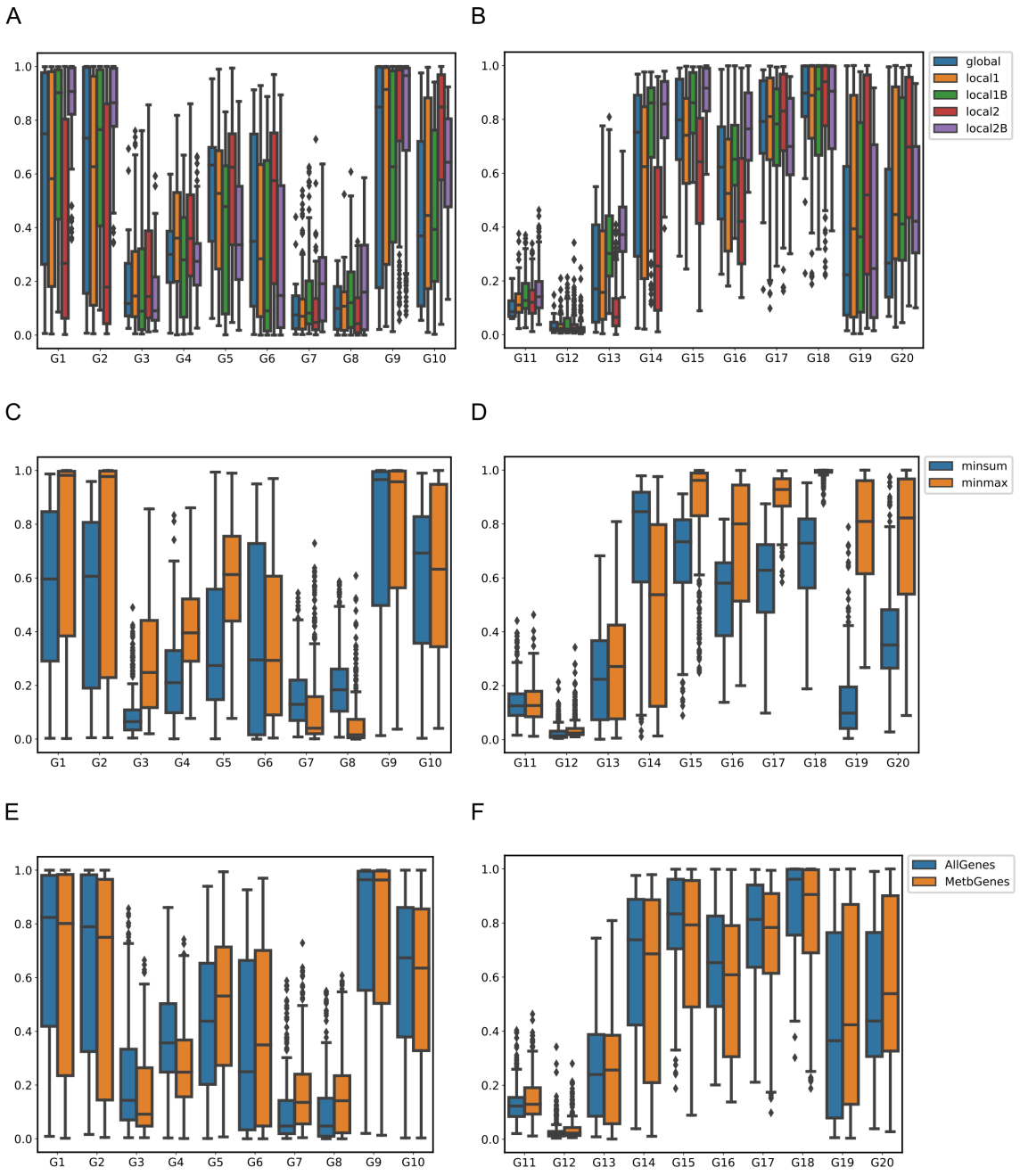

Figure A.2: Influence of transcriptomics data integration strategies in the similarity of reaction scores. The average Euclidean distance among reaction scores of samples of the same group (same cell type, same study, and donor/cell line) was determined for each gene threshold strategy (*global*, *local1*, *local2*, *local1B*, *local2B*), gene threshold (10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} percentile), gene to reaction scores conversion strategy (*min-max* and *min-sum*), and either including all genes or just the metabolic genes in the gene threshold calculation. An identical estimation was then made for simulated groups of random samples of the same size and the process was repeated 1000 times *(continues)*.

Figure A.2 *(continued from previous page)*: Boxplots show the proportion of simulations where the average Euclidean distance was smaller than the distance observed in the corresponding real sample groups (vertical axis) for each group (horizontal axis), and splits results by gene threshold strategy (**A**, **B**), conversion strategy (**C**, **D**) and decision to include all genes or just metabolic genes in gene threshold calculation (**E**, **F**). **A**, **C**, and **E** are sample groups corresponding to microarray studies. **B**, **D**, and **F** are sample groups of RNA-seq studies.

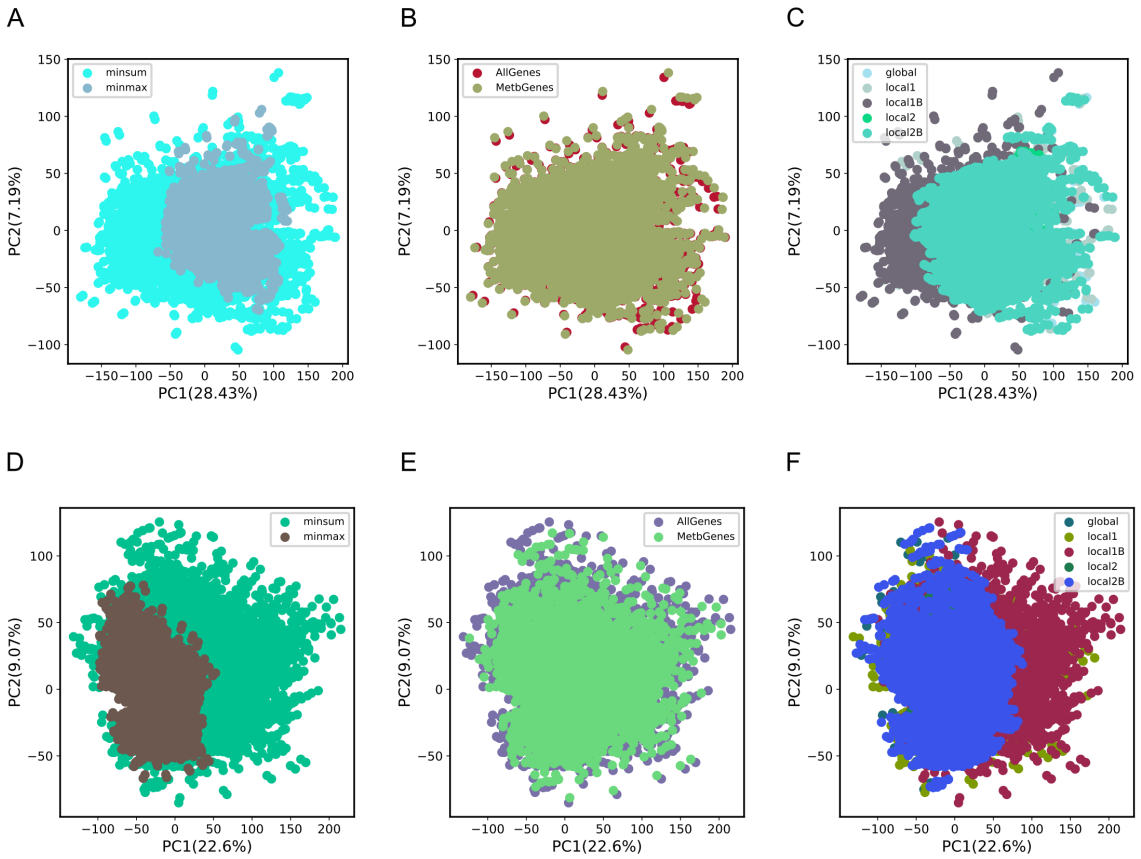

Figure A.3: Influence of transcriptomics data integration strategies in the variability of reaction scores. Principal Component Analysis (PCA) of reaction scores for different combinations of data integration parameters for microarray (**A**, **B**, **C**) and RNA-seq (**D**, **E**, **F**) datasets. None of the integration strategies *per se* are responsible for the variability in the first two components. **A** and **D**: gene to reaction scores conversion strategy. **B** and **E**: decision to include all genes or just metabolic genes in gene threshold calculation. **C** and **F**: gene threshold strategy.

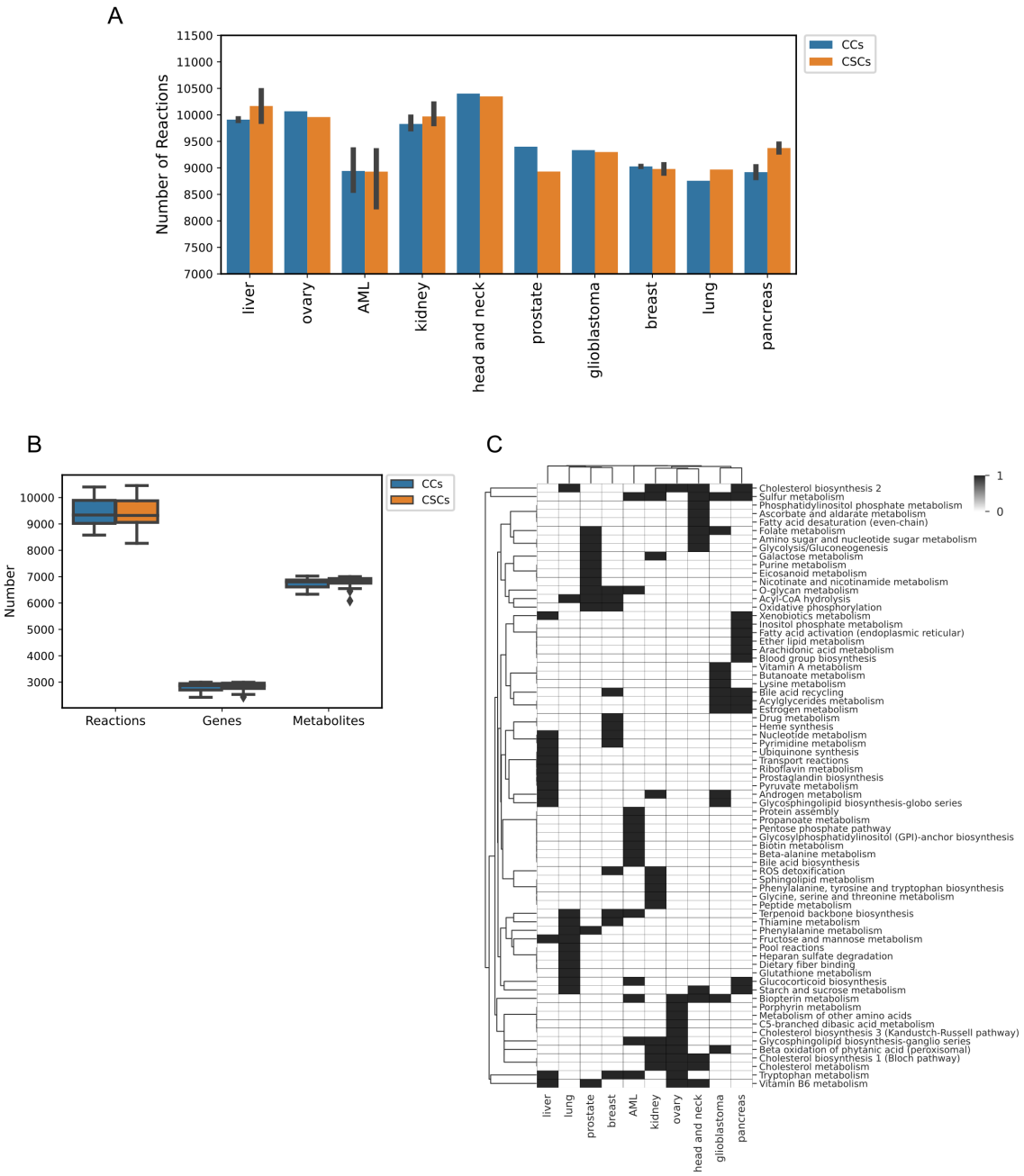

Figure A.4: Composition of reconstructed models. **A**: number of active reactions in reconstructed metabolic models of the different tissues. Tissues without error bars are those with just one donor/cell line. The error bars are 95% CI. **B**: Boxplot with the overall number of active reactions, genes, and metabolites of the different cell types. **C**: The median percentage of active reactions in each metabolic subsystem/pathway was calculated across models of different donors/cell lines of the same cell type and tissue. Then, the difference between the median values for CSCs and CCs of the same tissue was assessed. Metabolic subsystems were ranked from those with more percentage of active reactions in CSCs to those with more percentage of active reactions in CCs *(continues)*.

Figure A.4 *(continued from previous page)*: The top 10% subsystems in each tissue are shown. Subsystems with three or fewer reactions, or with no difference between CSCs and CCs across any tissue were excluded from the analysis. Note: percentage of active reactions was not directly compared among different tissues, because each tissue corresponds to an independent experiment/dataset.

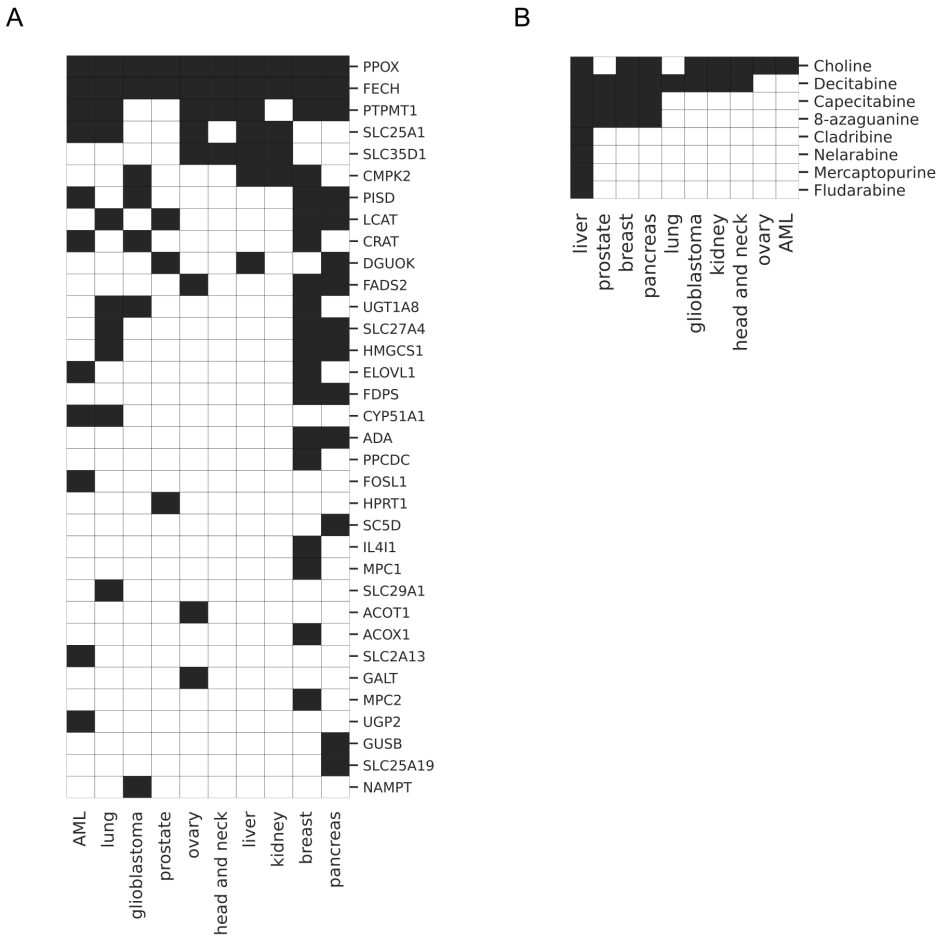

Figure A.5: Essential genes and antimetabolites predicted in both CSCs and CCs. **A**: essential genes predicted for models of CSCs and models of CCs. **B**: antimetabolites predicted to block the effect of the EMs that are common to both models of CSCs and models of CCs.

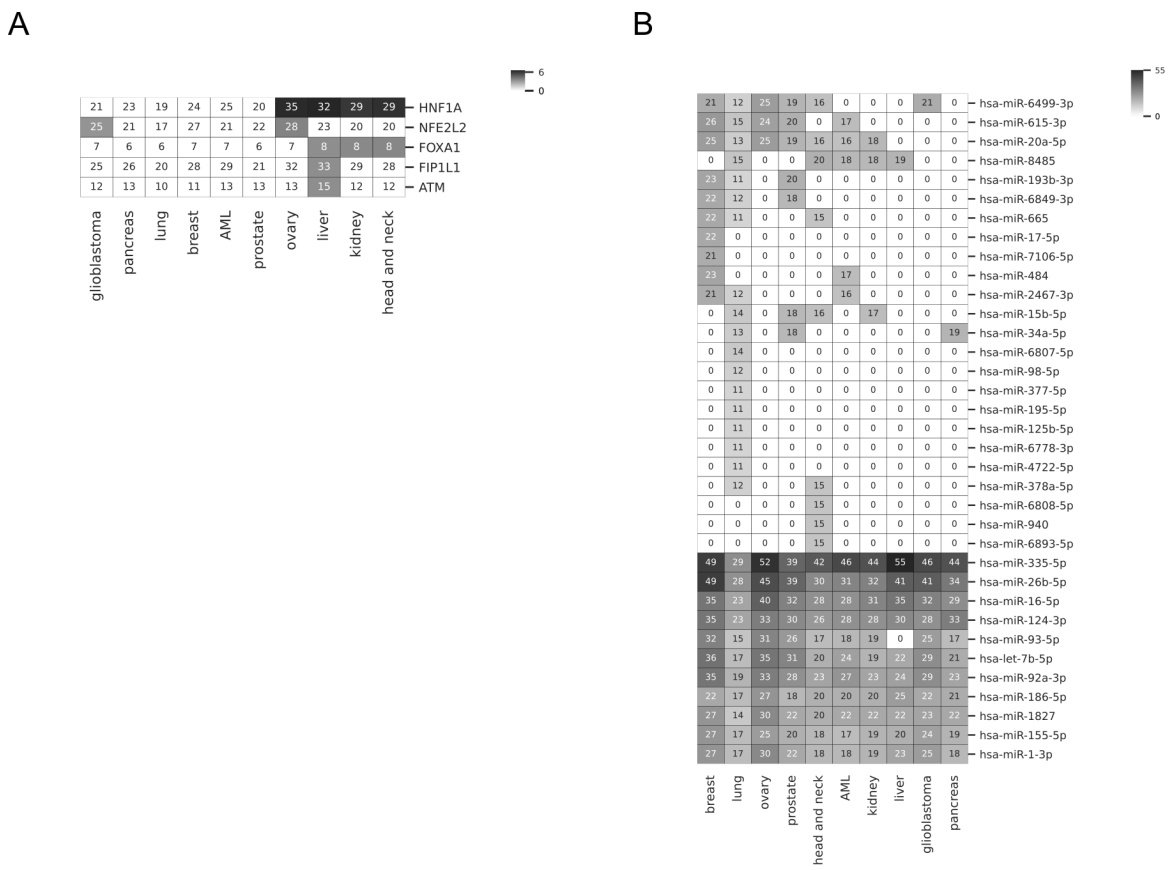

Figure A.6: TFs and miRs with potential to target genes that are correlated with biomass in both CSCs and CCs. **A**: TFs that when knocked-out decrease expression of genes that are directly correlated with biomass in both models of CSCs and of CCs. Only TFs significantly targeting those genes have color ($adjusted\ p\text{-}value < 0.05$) and color intensity is $-log(adjusted\ p\text{-}value)$. Each number counts the genes associated with biomass targeted by the TF. **B**: miRs that target genes that are directly correlated with biomass in both models of CSCs and of CCs. miRs targeting the top 10 numbers of targets in each tissue (maybe more than 10 miRs per tissue if different miRs have the same number of targets) are shown. Color intensity and numbers reflect the number of genes associated with biomass targeted by the miR.

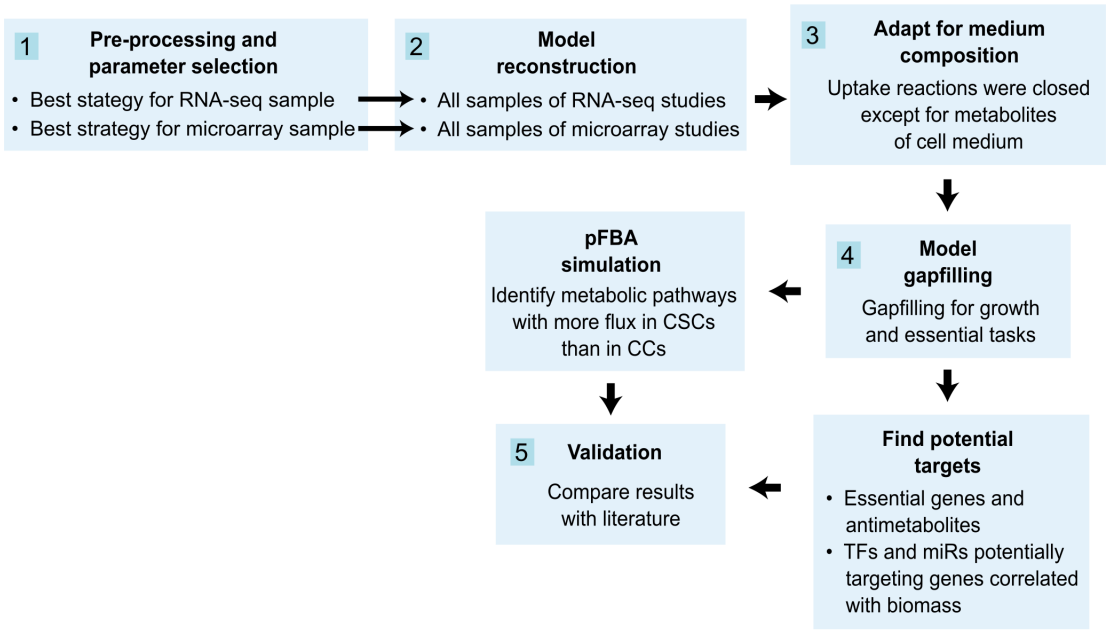

Figure A.7: Flow diagram of the overall study methodology. Full description of the overall methodology can be found in Materials and Methods. Details of pre-processing and parameter selection are represented in an independent flow diagram in Figure A.8.

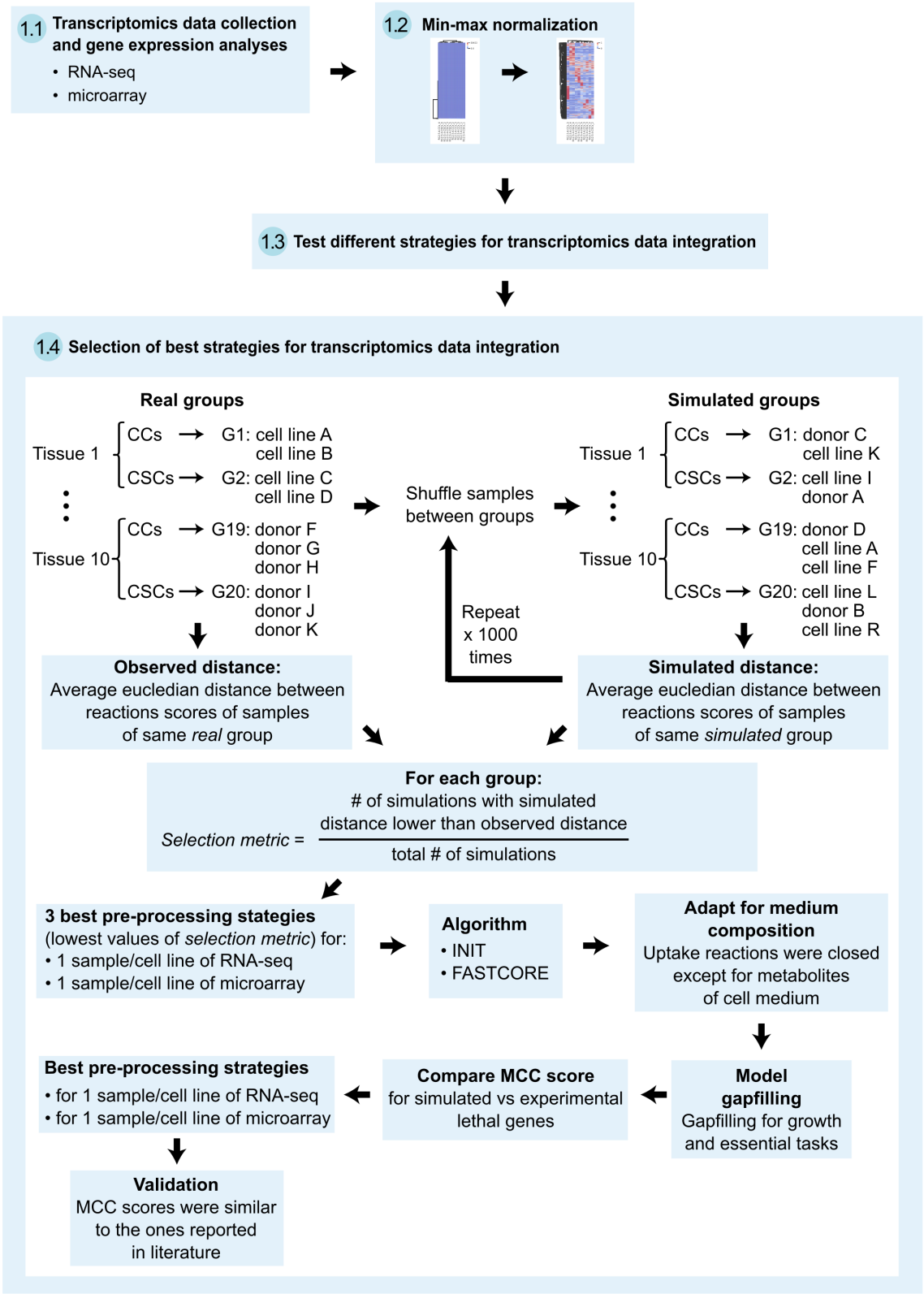

Figure A.8: Flow diagram of the pre-processing and parameter selection methodology. Full description of the pre-processing and parameter selection methodology can be found in Materials and Methods. Details of step 1.3: "Test different strategies for transcriptomics data integration" are represented in an independent flow diagram in Figure A.9.

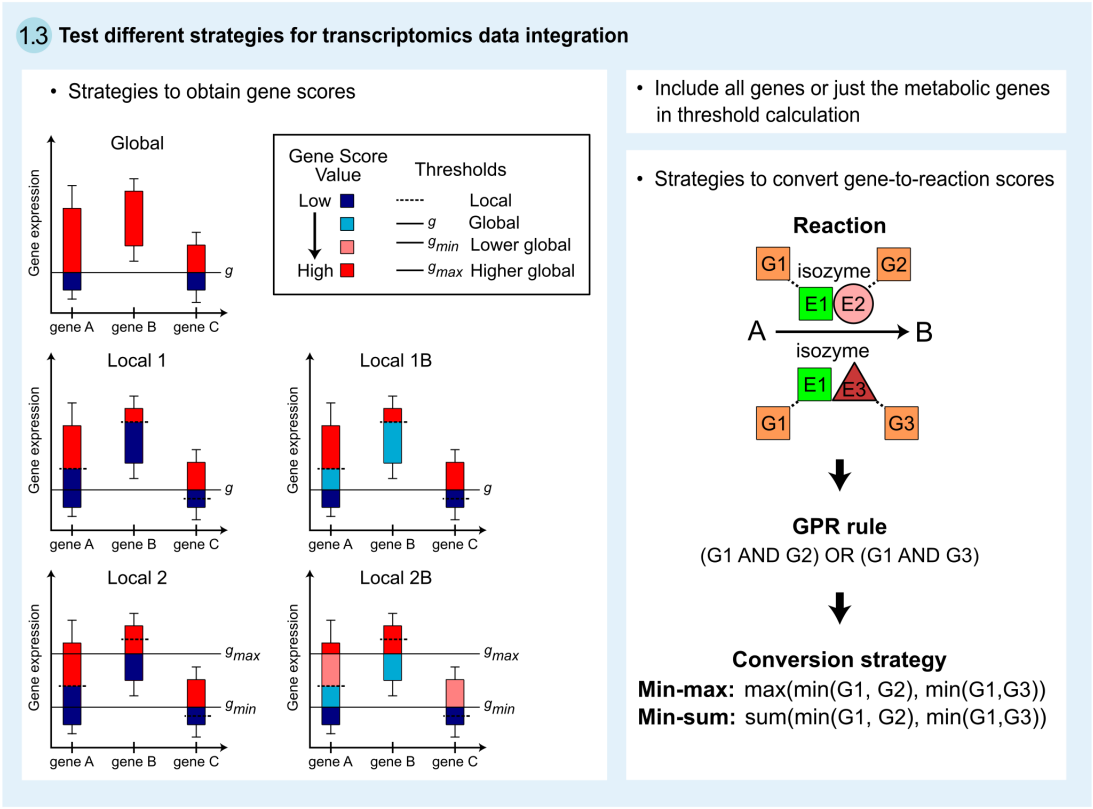

Figure A.9: Flow diagram of the pre-processing step – testing different strategies for transcriptomics data integration. Detailed representation of the different strategies applied for transcriptomics data integration.

Appendix B

Supplementary Figures - Chapter 3

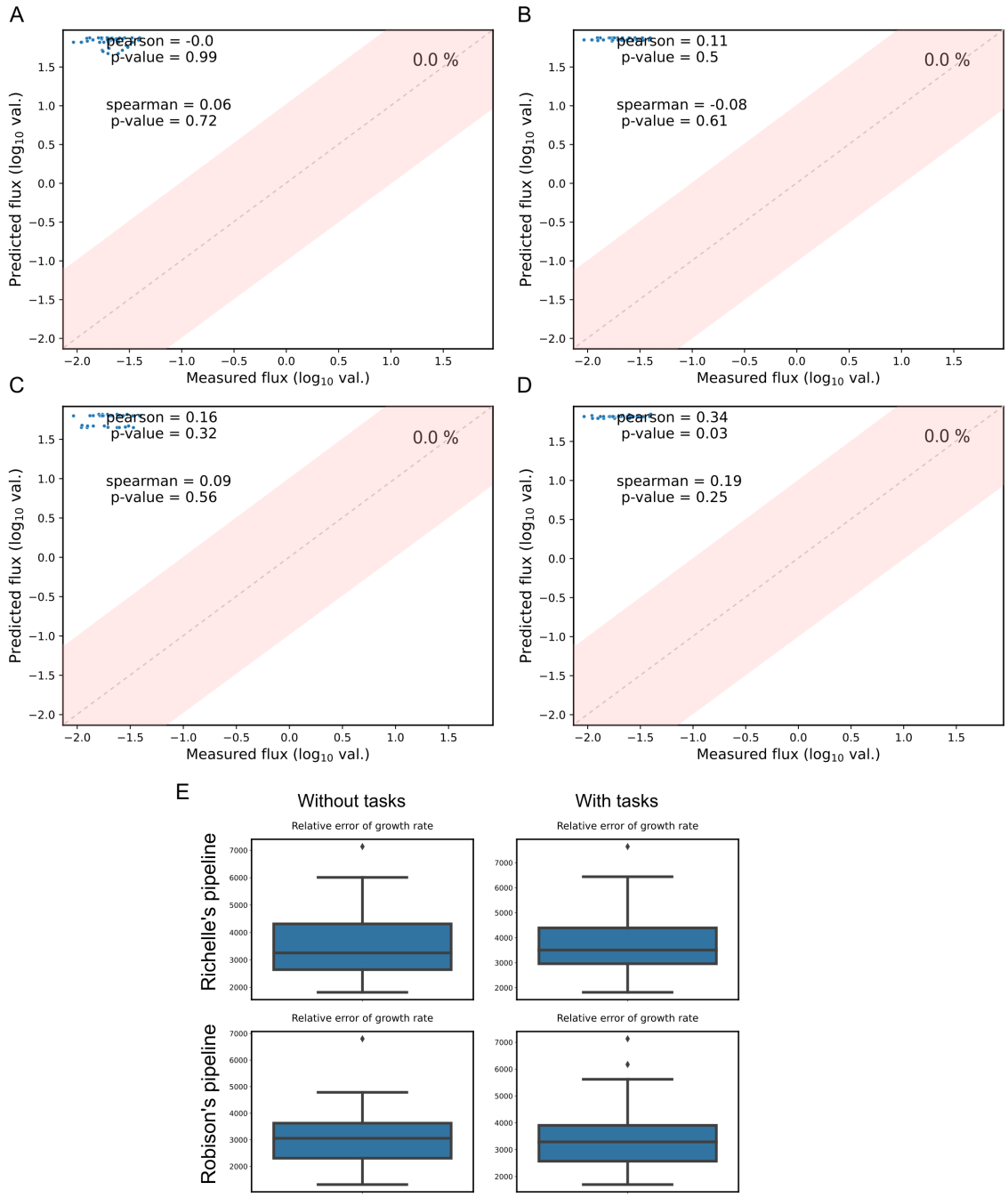

Figure B.1: Comparison of measured and simulated growth rates produced by traditional GSMMs where uptake/secretion rates of metabolites in Ham's media were loosely constrained (from -1000 to 1000). **A-D** are scatter plots with $\log_{10}$ of values of simulated and measured growth rates. Value at top right of each graph is the percentage of data points that are inside the pink area (where $\log_{10}(|\text{predicted value}|)$ is within $\log_{10}(|\text{measured value}|) \pm 1$). Either Richelle's pipeline using FASTCORE (**A**, **B**) and Robinson's pipeline using tINIT (**C**, **D**) were applied to reconstruct the models employed in the simulation. The effect of the integration (**B**, **D**) or not (**A**, **C**) of all tissue-specific metabolic tasks in those models was also assessed. **E**: Relative errors of predicted growth rates.

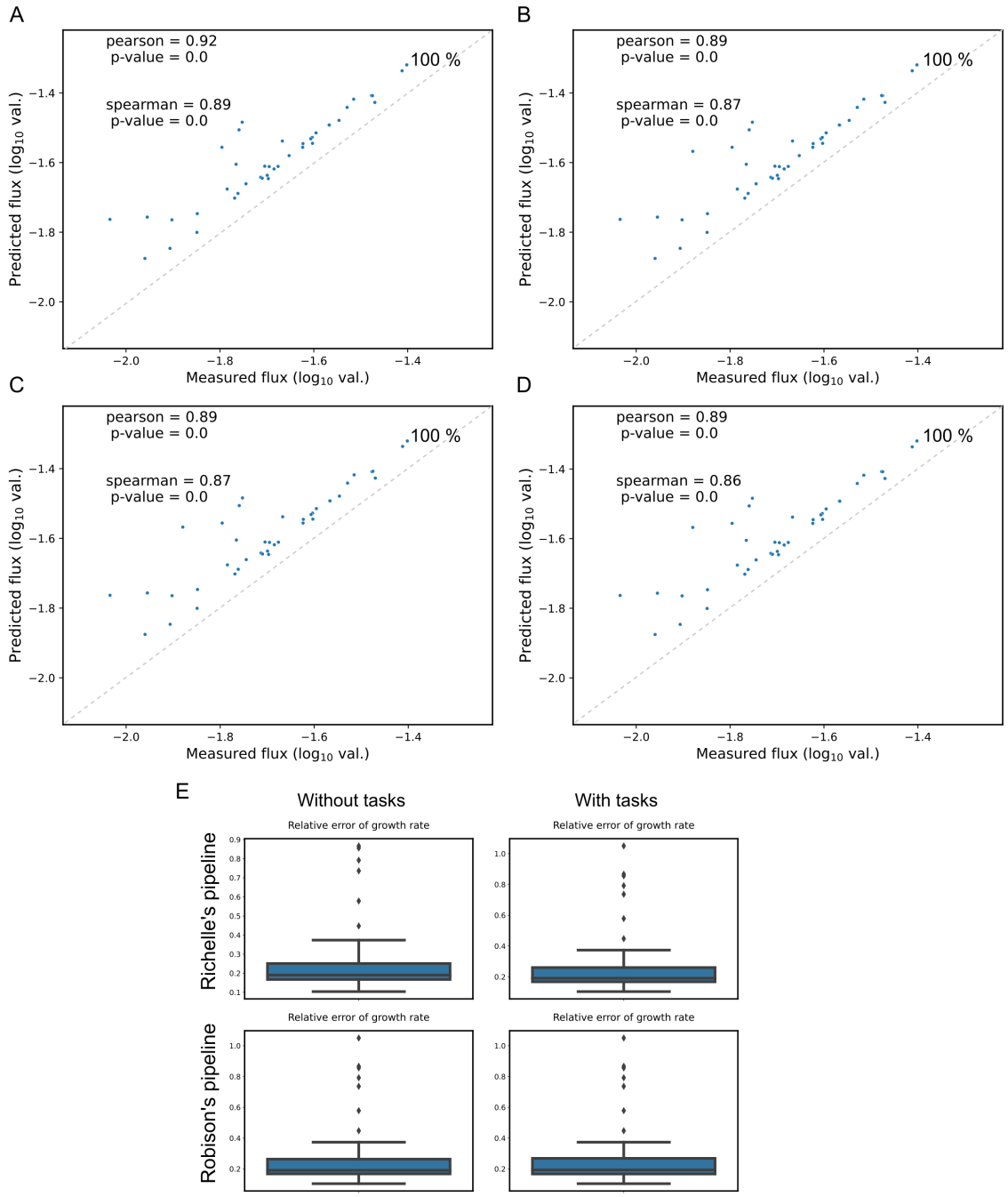

Figure B.2: Comparison of measured and simulated growth rates produced by traditional GSMMs where uptake/secretion rates of three metabolites (glucose, lactate and threonine) were constrained with measured fluxes. **A-D** are scatterplots with $\log_{10}$ of values of simulated and measured growth rates. Value at top right of each graph is the percentage of data points where $\log_{10}(|\text{predicted value}|)$ is within $\log_{10}(|\text{measured value}|) \pm 1$. Either Richelle's pipeline using FASTCORE (**A**, **B**) and Robinson's pipeline using tINIT (**C**, **D**) were applied to reconstruct the models employed in the simulation. The effect of the integration (**B**, **D**) or not (**A**, **C**) of all tissue-specific metabolic tasks in those models was also assessed. **E**: Relative errors of predicted growth rates.

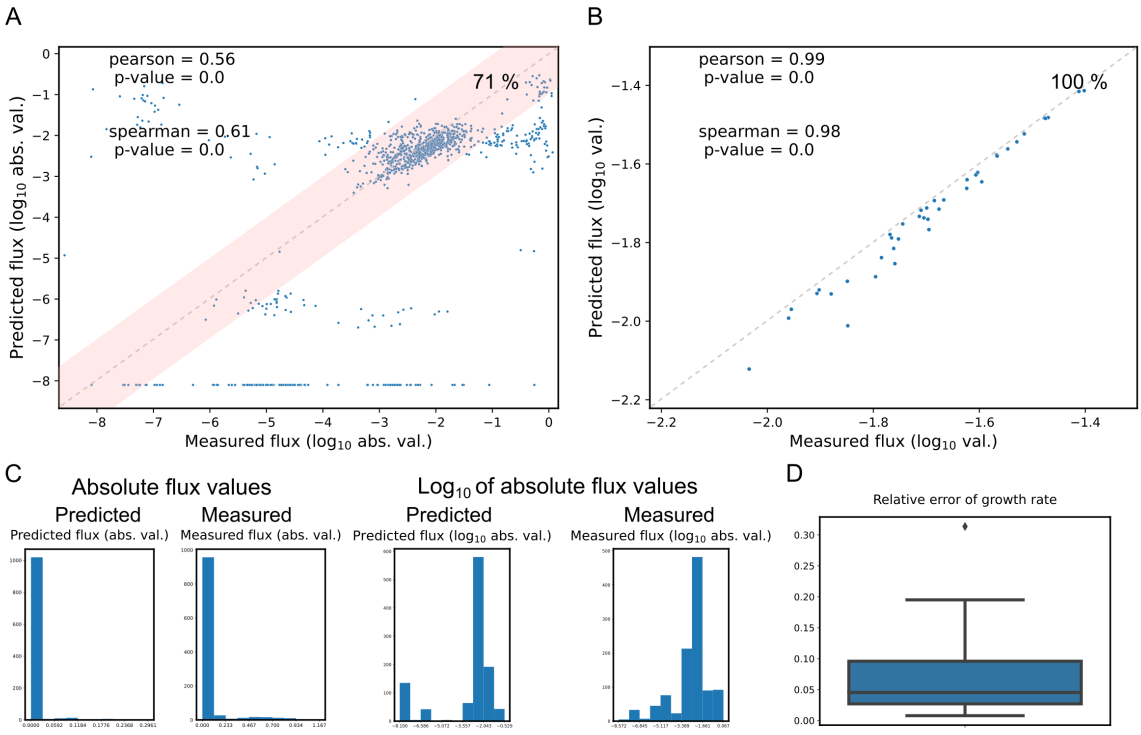

Figure B.3: Comparison of measured and simulated fluxes produced by traditional GSMMs constrained with measured growth rates. **A**: a scatter plot with $\log_{10}$ of values of simulated and measured fluxes of exchange reactions of 26 metabolites. Value at top right of **A** and **B** is the percentage of data points where $\log_{10}(|\text{predicted value}|)$ is within $\log_{10}(|\text{measured value}|) \pm 1$ (in **A**, it corresponds to the pink area). **B**: a scatter plot with $\log_{10}$ of values of simulated and measured growth rates. The correlation coefficients are not exactly one, because experimentally determined upper and lower bounds were used to constraint simulated biomass fluxes, while the value of the measured biomass in the graph is the average of those bounds. **C**: histograms with the distribution of absolute values of measured and simulated fluxes before and after logarithmization. **D**: Relative errors of predicted growth rates. Data points forming a line at the bottom of **A** correspond to metabolites with a predicted flux of zero, which are shown in the graphs as holding the lowest absolute measured value (besides zero), as the logarithm of zero is undefined. In **C**, the values of those metabolites fall in the lowest bin, creating an oddly tall bin at the beginning. Models used were reconstructed with Robinson's pipeline (using tINIT) and without tissue-specific tasks.

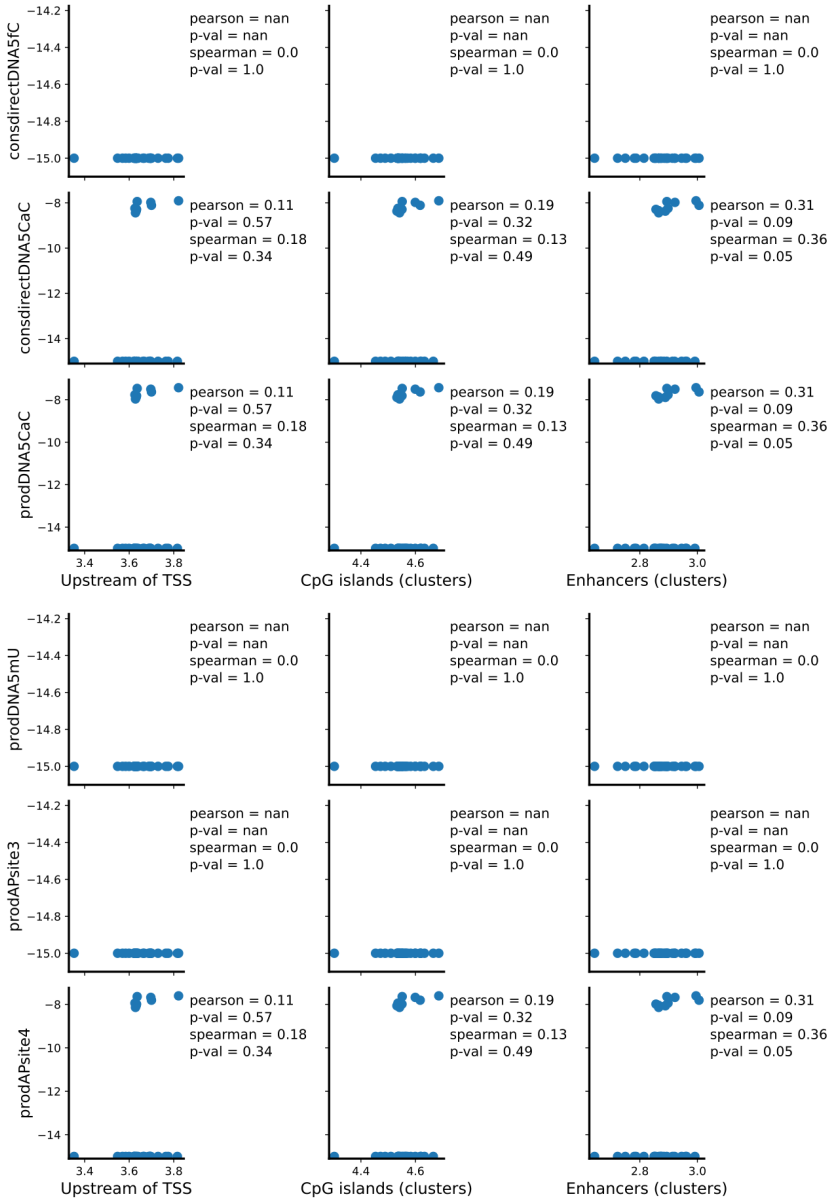

Figure B.4: Comparison of simulated fluxes of reactions involved in DNA demethylation and the estimated degree of DNA methylation for models with *methylation flux rules* and cell-specific methylation ratios - *Upstream of TSS*, *CpG islands*, *Enhancers*. **A**: scatter plots with $\log_{10}$ values of simulated fluxes versus the experimentally estimated degree of methylation across all genome or in close proximity to different genomic features. The procedure and the description of datasets applied to estimate the level of DNA methylation is detailed in *Comparison of fluxes of reactions involved in DNA (de)/methylation and the degree of DNA methylation* section of *Materials and methods*. The vertical labels are identifiers of reactions shown in Figure 3.1. Zero flux values were replaced with a very small value (1×10^{-15}) because $log_{10}(0)$ is undetermined. Only the 30 cell lines for which there was experimental data across all types of genomic intervals were here used.

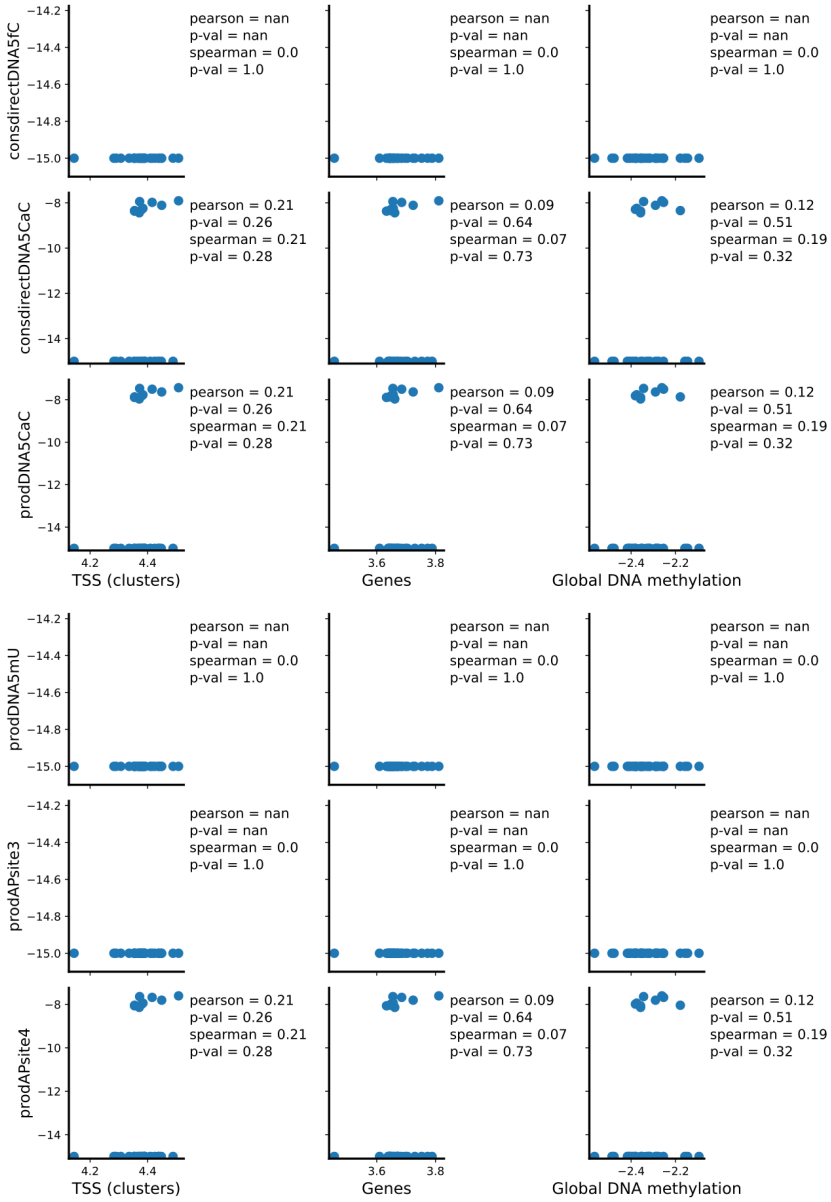

Figure B.5: Comparison of simulated fluxes of reactions involved in DNA demethylation and the estimated degree of DNA methylation for models with *methylation flux rules* and cell-specific methylation ratios - *TSS (clusters)*, *Genes*, *Global DNA methylation*. **A**: scatter plots with $\log_{10}$ values of simulated fluxes versus the experimentally estimated degree of methylation across all genome or in close proximity to different genomic features. The procedure and the description of datasets applied to estimate the level of DNA methylation is detailed in *Comparison of fluxes of reactions involved in DNA (de)/methylation and the degree of DNA methylation* section of *Materials and methods*.

Figure B.5 *(continued from previous page)*: The vertical labels are identifiers of reactions shown in Figure 3.1. Zero flux values were replaced with a very small value (1×10^{-15}) because $log_{10}(0)$ is undetermined. Only the 30 cell lines for which there was experimental data across all types of genomic intervals were here used.

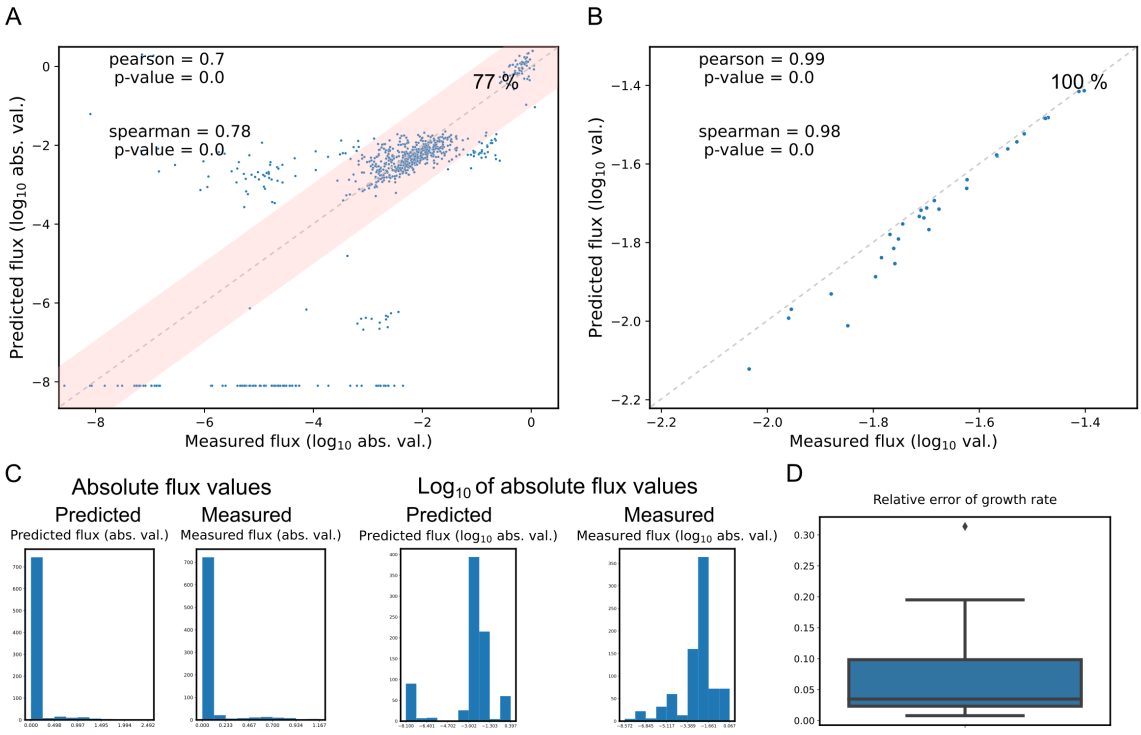

Figure B.6: Comparison of measured and simulated exchange fluxes produced by GECKO models constrained with measured growth rates, and containing *methylation flux rules* and cell-specific methylation ratios. **A**: a scatter plot with $\log_{10}$ of values of simulated and measured fluxes of exchange reactions of 26 metabolites. Value at top right of **A** and **B** is the percentage of data points where $\log_{10}(|\text{predicted value}|)$ is within $\log_{10}(|\text{measured value}|) \pm 1$ (in **A**, it corresponds to the pink area). **B**: a scatter plot with $\log_{10}$ of values of simulated and measured growth rates. The correlation coefficients are not exactly one, because experimentally determined upper and lower bounds were used to constraint simulated biomass fluxes, while the value of the measured biomass in the graph is the average of those bounds. **C**: histograms with the distribution of absolute values of measured and simulated fluxes before and after logarithmization. **D**: Relative errors of predicted growth rates. Data points forming a line at the bottom of **A** correspond to metabolites with a predicted flux of zero, which are shown in the graphs as holding the lowest absolute measured value (besides zero), as the logarithm of zero is undefined. In **C**, the values of those metabolites fall in the lowest bin, creating an oddly tall bin at the beginning. Models used were reconstructed with Robinson's pipeline (using tINIT) and without tissue-specific tasks.

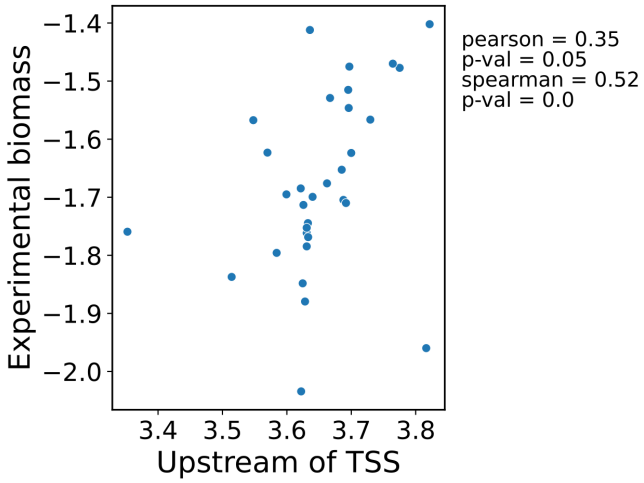

Figure B.7: Comparison of experimentally measured growth rates with degree of DNA methylation in gene promoters. Scatter plot with $\log_{10}$ values of measured growth rates versus the experimentally degree of methylation in the region Upstream of TSSs.

Appendix C

Supplementary Tables - Chapter 2

Table C.1: Markers of CSCs reported in literature.

Marker	Leukemia	Bladder	Breast	Colon	Gastric	Glioma/ Meduloblastoma	Head and Neck	Liver	Lung	Melanoma	Myeloma	Osteo-sarcoma	Ovarian	Pancreatic	Prostate	Kidney
ALCAM/CD166				1(c)											1(b)	
ALDH1A1		1(b)	1(a,b)	1(b)	1(b)		1(b)		1(b)				1(f)	1(b)	1(b)	
Aminopeptidase N/CD13								1(b)								
BMI-1	1(b)		1(b)	1(b)			1(b)							1(b)	1(b)	
CD117/c-kit	1(b)								1(b)				1(b)			
CD123/IL-3 R alpha	1(b)															
CD13								1(b)								
CD133			1(b)	1(a,b)		1(a,b)		1(b)	1(a,b)	1(b)			1(b)	1(b)	1(b)	
CD138/Syndecan-1											1(b)					
CD15/Lewis X					1(b)	1(b)										
CD151															1(b)	
CD166				1(b)											1(b)	
CD166/ALCAM										1(b)						
CD19											1(b)					
CD20/MS4A1											1(b)					
CD24				1(b)				1(b)					1(b)	1(a,b)		
CD26				1(b)												
CD27/TNFRSF7											1(b)					
CD29				1(b)												
CD34	1(b)															
CD34+CD38-	1(d)															
CD38											1(b)					
CD44	1(b)	1(b)	1(b)	1(b)	1(b)	1(b)	1(a,b)					1(b)	1(b)	1(a,b)	1(b)	1(e)
CD44+CD24-/low			1(a)													
CD45								1(b)								

Table C.1: Markers of CSCs reported in literature – continued from previous page.

Marker	Leukemia	Bladder	Breast	Colon	Gastric	Glioma/ Meduloblastoma	Head and Neck	Liver	Lung	Melanoma	Myeloma	Osteo-sarcoma	Ovarian	Pancreatic	Prostate	Kidney
CD47	1(b)	1(b)														
CD49f/Integrin alpha 6			1(b)			1(b)									1(b)	
CD90			1(b)					1(a)								
CD90/Thy1						1(b)		1(b)	1(b)							
CD96	1(b)															
CEACAM-6/CD66c		1(b)														
c-Myc	1(b)		1(b)	1(b)		1(b)								1(b)	1(b)	
CX3CR1						1(b)										
CXCR1/IL-8 RA			1(b)													
CXCR4			1(b)			1(b)								1(b)		1(e)
Endoglin/CD105													1(b)			
EpCAM/TROP1/ESA			1(b)	1(b)					1(b)					1(a,b)		
KLF4	1(b)														1(b)	
KLF4			1(b)													
Lgr5/GPR49				1(b)	1(b)		1(b)									
Musashi-1				1(b)		1(b)										
NANOG			1(b)	1(b)				1(b)					1(b)		1(b)	
Nestin						1(b)				1(b)		1(b)		1(b)		
OCT4/POU5F1	1(b)	1(b)											1(b)	1(b)	1(b)	
Rex1															1(b)	
SALL4	1(b)		1(b)	1(b)				1(b)					1(b)		1(b)	
SOX2			1(b)			1(b)		1(b)	1(b)						1(b)	
TIM3	1(b)															
TNFRSF16										1(b)						
TRA-1-60(R)															1(b)	

Table C.1: Markers of CSCs reported in literature – continued from previous page.

Marker	Leukemia	Bladder	Breast	Colon	Gastric	Glioma/ Meduloblastoma	Head and Neck	Liver	Lung	Melanoma	Myeloma	Osteo-sarcoma	Ovarian	Pancreatic	Prostate	Kidney
MET/RTK																1(e)

a) [264], b) [265], c) [266], d) [267], e) [268], f) [269]

Table C.2: Studies from which gene expression data sets were retrieved.

Study	**Cancer**	**Cell Subtype or Treatment**	**Patients**	**Cell Line**	**Technique**	**RNAseq Library**	**Sequencing Platform**	**Microarray**	**Array Platform**	**Dataset Ids**	**Paper**	**Comments**	**Criteria for study inclusion**
R1.1.1	prostate	luminal, basal	yes	no	RNAseq	NuGEN kit	Illumina HiSEq 2000			GSE82071 (Gene Expression Omnibus), E-GEOD-82071 (Array Express)	[270]	Here stem cells are the basal cells (or CD49f high) and non-stem cells are the luminal cells (or CD49f low). There are paired samples (dif. Cells from same patient).	**CSC marker CD49f**.
M2.2.13	liver	Hep3B $CD13^{+}CD133^{+}$, Huh7 $CD13^{+}CD133^{+}$, PLC/PRF/5 $CD13^{+}CD133^{+}$, versus Hep3B $CD13^{-}CD133^{-}$, Huh7 $CD13^{-}CD133^{-}$, PLC/PRF/5 $CD13^{-}CD133^{-}$	no	Hep3B, Huh7, PLC/ PRF/5	Microarray			Affymetrix	[HuGene-2_0-st] Affymetrix Human Gene 2.0 ST Array [transcript (gene) version	GSE66529 (Gene Expression Omnibus), E-GEOD-66529 (Array Express)	[271]	Although authors focus on lnRNAs, microarray was done for all mRNAs. Platform is the same as the one used for mRNAs and no specific isolation protocol for lnRNAs is done in study (total RNA was used)	**$CD13^{+}CD133^{+}$ markers**. Chemoresistance. Form **oncospheres**. **Higher tumorigenicity** .

Table C.2: Studies from which gene expression data sets were retrieved – continued from previous page.

Study	**Cancer**	**Cell Subtype or Treatment**	**Patients**	**Cell Line**	**Technique**	**RNAseq Library**	**Sequencing Platform**	**Microarray Array Platform**	**Dataset Ids**	**Paper**	**Comments**	**Criteria for study inclusion**
R4.10.2	glioblastoma	FCC, SCC	yes	no	RNAseq	NEBNext UltraTM RNA Library Prep Kit	Illumina HiSeq 2500		EGAD00001 004380 (European Genome-Phenome Archive)	[272]	SCC: slow-cycling cells are CSCs (they maintain their proliferation potential but just divide when needed - to keep homeostasis). FCC: fast-cycling cells are CCs. There is also a metabolomics dataset in same article.	Higher cell migration ability and **tumorigenicity**. Previous study PMID: 21515906 shown that SCCs are enriched in stem cell markers *in vitro*: **CD133$^+$/CD15$^+$** /ABCG2$^+$. Higher proliferation potential (but lower proliferation, as they are slow-cycling cells).
R5.8.2.TN	breast cancer - triple negative	nonCD44$^+$ CD24$^-$/low, CD44$^+$ CD24$^-$/low	no	SUM149PT, HCC1937, SUM159PT	RNAseq	unknown	Illumina HiSeq 2500		GSE132083 (Gene Expression Omnibus)	[273]		**CD44$^+$CD24$^-$ /low cells**

Table C.2: Studies from which gene expression data sets were retrieved – continued from previous page.

Study	**Cancer**	**Cell Subtype or Treatment**	**Patients**	**Cell Line**	**Technique**	**RNAseq Library**	**Sequencing Platform**	**Microarray**	**Array Platform**	**Dataset Ids**	**Paper**	**Comments**	**Criteria for study inclusion**
M6.3.14	ovary	SOC _differentiated, SOC _undifferentiated	yes	no	Microarray			Agilent	Agilent-039494 SurePrint G3 Human GE v2 8x60K Microarray 039381 (Probe Name version)	GSE64999 (Gene Expression Omnibus), E-GEOD-64999 (Array Express)	[274]	Treatment with Rock inhibitor produced CSCs. Treatment with FBS allowed to get CCs. soc stands for serum ovarian cancer	**Form spheroids**. **Tumorigenicity**. **ALDH1A1**, Nanog and Sox2 markers. Differentiated cells (CCs) showed reduced expression of ALDH1A1, Nanog and Sox2, epithelial -like morphology and differentiation marker CDK7.
M8.2.3	acute myeloid leukemia (AML)	AML ($CD34^+CD38^+$), AML ($CD34^+CD38^-$)	yes	no	Microarray			Affymetrix	[HG-U133A_2] Affymetrix Human Genome U133A 2.0 Array	GSE34044 (Gene Expression Omnibus), E-GEOD-34044 (Array Express)	[275]	CSCs are AML ($CD34^+CD38^-$) and CCs are AML ($CD34^+CD38^+$)	**$CD34^+CD38^-$** marker. **Over-representation of Notch pathway**.

Table C.2: Studies from which gene expression data sets were retrieved – continued from previous page.

Study	Cancer	Cell Subtype or Treatment	Patients	Cell Line	Technique	RNAseq Library	Sequencing Platform	Microarray	Array Platform	Dataset Ids	Paper	Comments	Criteria for study inclusion
M9.1.4	kidney	non-cultured cells, $CXCR4^{+}MET^{+}$ $CD44^{+}$, spheres	yes	no	Microarray			Illumina	Illumina HumanHT-12 V4.0 expression beadchip	GSE89461 (Gene Expression Omnibus)	[268]	Non-cultured cells are a mixture of CCs and some CSCs. $CXCR4^{+}MET^{+}$ $CD44^{+}$ are CSCs. As not all 'spheres' are CSCs, (most but not all are enriched for CSC markers), 'spheres' were excluded.	**$CXCR4^{+}MET^{+}$ $CD44^{+}$** markers. **Tumorigenesis**. **Activation of Wnt and Notch pathways.** Correlation with tumor aggressiveness and metastasis
R11.1.4	non-small cell lung cancer	ntp, nts	no	NCIH1703, NCIH1299, ChaGoK1	RNAseq	NuGEN kit	Illumina Genome Analyzer II			GSE48599 (Gene Expression Omnibus)	[276]	Samples ending in 'kds' mean knock down with shRNA, they are not going to be used in analysis. Samples ending in 'nts' are CSCs. Samples ending in 'ntp' are CCs.	**Oncosphere formation**. **Expression of Sox2**, Oct3/4, Nanog, **ALDH1 and CD133**. **Tumorigenicity**.

Table C.2: Studies from which gene expression data sets were retrieved – continued from previous page.

Study	Cancer	Cell Subtype or Treatment	Patients	Cell Line	Technique	RNAseq Library	Sequencing Platform	Microarray Platform	Dataset Ids	Paper	Comments	Criteria for study inclusion
R12.5.4	pancreas - pancreatic ductal adenocarcinoma	PDAC_Adh, PDAC_Sph	yes	no	RNAseq	Illumina TruSeq RNA Sample Prep Kit v2	Illumina Genome Analyzer IIx		E-MTAB-3808 (ArrayExpress)	[277]	CSCs are the spheroid cultured samples. Adherent culture samples are CCs. The same sample is split through several files (instead of being different runs of same sample), because all files of same sample had exackly 4M reads except for the last file which had variable number of reads. fastq.gz. Files of same sample were first merged with 'cat' command.	**Form spheres. Previous studies (PMID: 24204632) shown those spheres express CD133$^+$ and other markers associated with CSC**.

Table C.2: Studies from which gene expression data sets were retrieved – continued from previous page.

Study	Cancer	Cell Subtype or Treatment	Patients	Cell Line	Technique	RNAseq Library	Sequencing Platform	Microarray	Array Platform	Dataset Ids	Paper	Comments	Criteria for study inclusion
M13.2.11	head and neck squamous cell carcinoma (HNSCC) of the oral tongue/cavity	UT14, UT16A, UT24A, UT30, UT33, UT14CD44$^+$, UT16ACD44$^+$, UT24ACD44$^+$, UT30CD44$^+$, UT33CD44$^+$	no	UT14, UT16A, UT24A, UT30, UT33	Microarray			Affymetrix	[HuEx-1_0-st] Affymetrix Human Exon 1.0 ST Array [transcript (gene) version]	GSE55487 (Gene Expression Omnibus), E-GEOD-55487 (Array Express)	[278]	Cells were isolated by different methods. The method described in the article as the most effective (CD44 isolation) was the one chosen here.	**CD44**$^+$ marker. Chemoresistance

Table C.3: Essential metabolites predicted only in CSCs.

Metabolite	**liver**	**ovary**	**AML**	**kidney**	**head and neck**	**prostate**	**glioblastoma**	**breast**	**lung**	**pancreas**
malonyl-carnitin	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE
eicosa-(2E,8Z,11Z,14Z,17Z)-pentaenoyl-CoA	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE
2-methoxy-6-all trans-decaprenyl-2-methoxy-1,4-benzoquinol	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE
eicosenoylcarnitine(9)	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE
3-oxo-dihomo-gamma-linolenoyl-CoA	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE
3-oxo-tetracosa-12,15,18,21-all-cis-tetraenoyl-CoA	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE
2-trans-7,10,13,16,19-all-cis-docosahexaenoyl-CoA	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE
trans-2-cis,cis,cis-8,11,14-eicosatetraenoyl-CoA	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE
N-pantothenoylcysteine	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE
sphingosine	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE
deoxyguanosine	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE
adenine	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE
3-demethylubiquinol-10	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE
deoxyadenosine	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE
3-keto-eicosa-8,11,14,17-all-cis-tetraenoyl-CoA	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE
3(S)-hydroxy-docosa-7,10,13,16,19-all-cis-pentaenoyl-CoA	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE
arachidonyl-carnitine	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE
guanine	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE
cytidine	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE
3-oxo-docosa-7,10,13,16,19-all-cis-pentaenoyl-CoA	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE
2-deoxy-D-ribose-1-phosphate	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
3(S)-hydroxy-all-cis-8,11,14,17-eicosatetraenoyl-CoA	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE
dGMP	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE
3(S)-hydroxy-dihomo-gamma-linolenoyl-CoA	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE
1-acylglycerol-3P-gamma-lin	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
1-acylglycerol-3P-palm	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE

Table C.3: Essential metabolites predicted only in CSCs – continued from previous page.

Metabolite	liver	ovary	AML	kidney	head and neck	prostate	glioblastoma	breast	lung	pancreas
1-acylglycerol-3P-7,10,13,16,19-docosa	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
1-acylglycerol-3P-5-tetrade	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
1-acylglycerol-3P-tetraco	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
1-acylglycerol-3P-7-hexade	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
(9E)-octadecenoylcarnitine	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
1-acylglycerol-3P-palmn	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
1-acylglycerol-3P-6,9,12,15,18-tetraco	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
1-acylglycerol-3P-cis-vac	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
decanoyl-CoA	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE
1-acylglycerol-3P-8,11-eico	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
1-acylglycerol-3P-15-tetra	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
1-acylglycerol-3P-9,12,15,18,21-tetra	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
1-acylglycerol-3P-ol	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
1-acylglycerol-3P-myrist	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
1-acylglycerol-3P-8,11,14,17-eico	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
1-acylglycerol-3P-13,16,19-doco	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
1-acylglycerol-3P-trico	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
dGDP	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
1-acylglycerol-3P-eico	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
1-acylglycerol-3P-pentade	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
cholesterol-ester-ol	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
1-acylglycerol-3P-12,15,18,21-tetra	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
(9E)-octadecenoyl-CoA	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
1-acylglycerol-3P-7,10,13,16-docosa	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
1-acylglycerol-3P-10,13,16,19-doco	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
1-acylglycerol-3P-laur	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE

Table C.3: Essential metabolites predicted only in CSCs – continued from previous page.

Metabolite	liver	ovary	AML	kidney	head and neck	prostate	glioblastoma	breast	lung	pancreas
cholesterol-ester-arach	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
1-acylglycerol-3P-heneico	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
octadecenoylcarnitine(5)	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
1-acylglycerol-3P-11,14,17-eico	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
1-acylglycerol-3P-nanode	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
palmitoleoyl-CoA	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
1-acylglycerol-3P-tridec	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
1-acylglycerol-3P-13,16-docosa	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
1-acylglycerol-3P-6,9,12,15,18,21-tetra	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
1-acylglycerol-3P-10-heptade	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
1-acylglycerol-3P-hexacosa	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
bilirubin-monoglucuronoside	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE
1-acylglycerol-3P-4,7,10,13,16-docosa	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
palmitoleoyl-carnitine	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
1-acylglycerol-3P-docosa	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
sphingosine-1-phosphate	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
fatty acid pool	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE
1-acylglycerol-3P-linolen	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
10,13,16,19-docosatetraenoylcarnitine	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE
1-acylglycerol-3P-11,14-eicosa	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
1-acylglycerol-3P-hexecose	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
cholesterol-ester-linolen	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
1-acylglycerol-3P-13-docose	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
1-acylglycerol-3P-11-eico	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
1-acylglycerol-3P-9-octade	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
1-acylglycerol-3P-dihomo-gamma	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE

Table C.3: Essential metabolites predicted only in CSCs – continued from previous page.

Metabolite	liver	ovary	AML	kidney	head and neck	prostate	glioblastoma	breast	lung	pancreas
1-acylglycerol-3P-6,9,12,15-octa	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
1-acylglycerol-3P-stea	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
cholesterol-ester-5,8,11,14,17-eico	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
(2E)-pentadecenoyl-CoA	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE
1-acylglycerol-3P-7-octade	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
L-oleoylcarnitine	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE
1-acylglycerol-3P-9-tetrade	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
cholesterol-ester-lin	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
1-acylglycerol-3P-6,9-octa	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
bilirubin	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE
ocosatetraenoylcarnitine	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
3-hydroxypentadecanoyl-CoA	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE
1-acylglycerol-3P-9,12,15,18-tetraco	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
(15Z)-tetracosenoylcarnitine	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE
1-acylglycerol-3P-arach	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
1-acylglycerol-3P-9-eicose	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
cholesterol-ester-gamma-lin	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
cholesterol-ester-6,9,12,15-octa	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
1-acylglycerol-3P-lin	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
3-oxopentadecanoyl-CoA	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE
(5Z,8Z,11Z,14Z,17Z)-eicosapentaenoylcarnitine	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
pyridoxine-phosphate	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE
linoleic-carnitine	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
1-acylglycerol-3P-5,8,11-eico	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
ribulose-5-phosphate	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE
1-acylglycerol-3P-5,8,11,14,17-eico	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE

Table C.3: Essential metabolites predicted only in CSCs – continued from previous page.

Metabolite	liver	ovary	AML	kidney	head and neck	prostate	glioblastoma	breast	lung	pancreas
cholestenol	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE
1-acylglycerol-3P-11-docose	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
1-acylglycerol-3P-13-octade	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
1-acylglycerol-3P-4,7,10,13,16,19-doco	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
tricosanoylcarnitine	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
L-gulonate	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
heneicosanoylcarnitine	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
phosphatidate-LD-PC pool	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
[protein]-N6-methyl-L-lysine	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
6-lactoyl-5,6,7,8-tetrahydropterin	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
(6Z,9Z)-octadecadienoylcarnitine	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
(2E)-decenoyl-[ACP]	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
octanoyl-CoA	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
Alpha-Linolenyl Carnitine	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
cholesterol-ester-13,16,19-doco	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
hexanoyl-CoA	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
(11Z,14Z,17Z)-eicosatrienoylcarnitine	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
3-oxododecanoyl-[ACP]	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
(2E)-tetradecenoyl-[ACP]	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
butyryl-[ACP]	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
cholesterol-ester-12,15,18,21-tetracosa	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
trans,cis-hexadeca-2,7-dienoyl-CoA	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
L-palmitoylcarnitine	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
docosenoylcarnitine	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
cholesterol-ester-9,12,15,18-tetraco	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
cholesterol-ester-10,13,16,19-docosa	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

Table C.3: Essential metabolites predicted only in CSCs – continued from previous page.

Metabolite	liver	ovary	AML	kidney	head and neck	prostate	glioblastoma	breast	lung	pancreas
uridine	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
phosphatidate-LD-TAG pool	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
simvastatin dihydroxy acid form	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
hexacosenoylcarnitine	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
cholesterol-ester-palm	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
(2E)-dodecenoyl-[ACP]	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
cholesterol-ester-5-tetra	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
dodecanoylcarnitine	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
cholesterol-ester-13,16-docosa	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
CDP-ethanolamine	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
simvastatin-acyl-glucuronide	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
eicosenoylcarnitine(7)	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
cholesterol-ester-9-heptade	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
heptadecanoylcarnitine	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
D-xylulose	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
tetracosanoylcarnitine	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
cholesterol-ester-9-octa	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
hexacosanoylcarnitine	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
CDP-choline	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
(R)-3-hydroxydecanoyl-[ACP]	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
cholesterol-ester-tridec	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
cholesterol-ester-4,7,10,13,16-docosa	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
cholesterol-ester-9-eicose	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
hexanoyl-[ACP]	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
cholesterol-ester-9,12,15,18,21-tetra	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
eicosenoylcarnitine(11)	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

Table C.3: Essential metabolites predicted only in CSCs – continued from previous page.

Metabolite	liver	ovary	AML	kidney	head and neck	prostate	glioblastoma	breast	lung	pancreas
xylitol	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
octadecenoylcarnitine(11)	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
D-3-hydroxyhexanoyl-[ACP]	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
cholesterol-ester-dihomo-gamma	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
(4Z,7Z,10Z,13Z,16Z)-docosapentaenoylcarnitine	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
(2E)-hexenoyl-[ACP]	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
O-propanoylcarnitine	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
phosphatidylinositol-4,5-bisphosphate	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
HMA	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
(8Z,11Z,14Z,17Z)-eicosatetraenoylcarnitine	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
cholesterol-ester-docosa	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
thymidine	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
cholesterol-ester-9-tetrade	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
cholesterol-ester-hepta	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
(2E)-octadecenoyl-CoA	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
UDP-galactose	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
acetoacetyl-[ACP]	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
13-cis-retinol	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
docosanoylcarnitine	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
pyridoxal	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
(S)-3-hydroxy-7-hexadecenoyl-CoA	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
cholesterol-ester-7-octade	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
cholesterol-ester-trico	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
dodecanoyl-[ACP]	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
Tetradecenoyl Coenzyme A (N-C14:1)	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
cholesterol-ester-11-eico	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

Table C.3: Essential metabolites predicted only in CSCs – continued from previous page.

Metabolite	liver	ovary	AML	kidney	head and neck	prostate	glioblastoma	breast	lung	pancreas
(2E)-octenoyl-[ACP]	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
cholesterol-ester-tetraco	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
dUTP	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
3(S)-hydroxy-docosa-7,10,13,16-all-cis-tetraenoyl-CoA	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
tetradecanoyl-[ACP]	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
cholesterol-ester-eico	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
nicotinamide ribonucleoside	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
cholesterol-ester-stea	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
1,2-diacylglycerol-LD-TAG pool	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
cholesterol-ester-13-eicose	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
deamido-NAD	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
cholesterol-ester-7-tetrade	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
1-acylglycerol-LD-PE pool	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
cholesterol-ester-cis-vac	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
1-acylglycerol-LD-PC pool	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
3-oxooctanoyl-[ACP]	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
(S)-3-hydroxybutyryl-CoA	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
cholesterol-ester-11-docose	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
(2E,7Z,10Z,13Z,16Z)-docosapentaenoyl-CoA	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
alpha-D-galactose-1-phosphate	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
phosphocholine	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
cholesterol-ester-10,13,16-docosa	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
cholesterol-ester-11,14,17-eico	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
ribose-1-phosphate	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
1-phosphatidyl-1D-myo-inositol-5-phosphate	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
12,15,18,21-tetracosatetraenoylcarnitine	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE

Table C.3: Essential metabolites predicted only in CSCs – continued from previous page.

Metabolite	liver	ovary	AML	kidney	head and neck	prostate	glioblastoma	breast	lung	pancreas
glucono-1,5-lactone-6-phosphate	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
cholesterol-ester-8,11,14,17-eico	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
ascorbate	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
acetyl-[ACP]	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
cholesterol-ester-8,11-eico	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
cholesterol-ester-7-hexa	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
cholesterol-ester-6,9,12,15,18,21-tetra	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
trans,cis-hexadeca-2,9-dienoyl-CoA	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
(8Z,11Z)-eicosadienoylcarnitine	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
cholesterol-ester-7,10,13,16-docosa	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
eicosadienoylcarnitine	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
cholesterol-ester-6,9-octa	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
cholesterol-ester-15-tetra	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
cholesterol-ester-6,9,12,15,18-tetraco	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
11-Octadecenoyl Coenzyme A	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
eicosanoylcarnitine	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
ethanolamine-phosphate	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
cholesterol-ester-13-docose	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
3-oxohexanoyl-[ACP]	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
(9Z,12Z,15Z,18Z)-tetracosatetraenoylcarnitine	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
cholesterol-ester-penta	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
3-oxodecanoyl-[ACP]	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
(R)-3-hydroxybutanoyl-[ACP]	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
dehydroascorbic acid	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
3-oxopalmitoleoyl-CoA	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
cholesterol-ester-hexecose	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

Table C.3: Essential metabolites predicted only in CSCs – continued from previous page.

Metabolite	liver	ovary	AML	kidney	head and neck	prostate	glioblastoma	breast	lung	pancreas
cholesterol-ester-nanode	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
octanoyl-[ACP]	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
1,2-diacylglycerol-LD-PC pool	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
cholesterol-ester-11,14-eicosa	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
(9Z,12Z,15Z,18Z,21Z)-tetracosapentaenoylcarnitine	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
cholesterol-ester-hexacosa	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
cholesterol-ester-13-octade	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
cholesterol-ester-myrist	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
decanoyl-[ACP]	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
tetradecanoylcarnitine	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
crotonyl-CoA	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
dADP	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

Table C.4: Essential metabolites predicted in both CSCs and CCs.

Metabolite	**liver**	**ovary**	**AML**	**kidney**	**head and neck**	**prostate**	**glioblastoma**	**breast**	**lung**	**pancreas**
nicotinamide	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
nicotinamide D-ribonucleotide	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
dADP	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
1,2-diacylglycerol-LD-PC pool	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE
fatty acid pool	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	FALSE	TRUE
(9E)-octadecenoyl-CoA	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
dGDP	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE
1-acylglycerol-3P-13,16-docosa	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-hexecose	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-7,10,13,16-docosa	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-laur	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-nanode	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-tridec	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-tetraco	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-6,9,12,15,18,21-tetra	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-hexacosa	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-12,15,18,21-tetra	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-4,7,10,13,16-docosa	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-docosa	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-linolen	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-pentade	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-13,16,19-doco	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
2-deoxy-D-ribose-1-phosphate	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE
1-acylglycerol-3P-5-tetrade	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-ol	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-5,8,11-eico	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE

Table C.4: Essential metabolites predicted in both CSCs and CCs – continued from previous page.

Metabolite	liver	ovary	AML	kidney	head and neck	prostate	glioblastoma	breast	lung	pancreas
1-acylglycerol-3P-6,9,12,15-octa	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-9,12,15,18-tetraco	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-6,9-octa	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-5,8,11,14,17-eico	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-11,14-eicosa	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-8,11,14,17-eico	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-6,9,12,15,18-tetraco	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-10,13,16,19-doco	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-cis-vac	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-palmn	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-palm	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-trico	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
2-lysolecithin pool	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE
1-acylglycerol-3P-11,14,17-eico	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-heneico	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-eico	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-10-heptade	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-myrist	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-9,12,15,18,21-tetra	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-gamma-lin	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-4,7,10,13,16,19-doco	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-lin	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-9-tetrade	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-11-docose	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
sphingosine	TRUE	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
1-acylglycerol-3P-9-eicose	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE

Table C.4: Essential metabolites predicted in both CSCs and CCs – continued from previous page.

Metabolite	liver	ovary	AML	kidney	head and neck	prostate	glioblastoma	breast	lung	pancreas
1-acylglycerol-3P-9-octade	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-11-eico	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-dihomo-gamma	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-stea	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-13-octade	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-8,11-eico	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-7,10,13,16,19-docosa	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-7-hexade	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-7-octade	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-arach	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-13-docose	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
1-acylglycerol-3P-15-tetra	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
guanine	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE
bilirubin-monoglucuronoside	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE
CDP-choline	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE
dGMP	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE
bilirubin	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE
phosphatidate-LD-PC pool	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE
phosphocholine	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE
deoxyguanosine	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE
(2E)-octadecenoyl-CoA	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE
histone-N6-methyl-L-lysine	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE
N-pantothenoylcysteine	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
3-oxo-tetracosa-12,15,18,21-all-cis-tetraenoyl-CoA	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
deoxyadenosine	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE
O-propanoylcarnitine	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

Table C.4: Essential metabolites predicted in both CSCs and CCs – continued from previous page.

Metabolite	liver	ovary	AML	kidney	head and neck	prostate	glioblastoma	breast	lung	pancreas
pyridoxine-phosphate	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
eicosa-(2E,8Z,11Z,14Z,17Z)-pentaenoyl-CoA	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE
sphingosine-1-phosphate	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
3(S)-hydroxy-all-cis-8,11,14,17-eicosatetraenoyl-CoA	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE
palmitoleoyl-CoA	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
adenine	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE
3-keto-eicosa-8,11,14,17-all-cis-tetraenoyl-CoA	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE
pyridoxal	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE
malonyl-carnitin	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
decanoyl-CoA	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
(9E)-octadecenoylcarnitine	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
linoleic-carnitine	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
trans-2-cis,cis,cis-8,11,14-eicosatetraenoyl-CoA	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
2-Decaprenyl-6-Methoxy-1,4-Benzoquinone	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
12,15,18,21-tetracosatetraenoylcarnitine	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
(2E)-hexadecenoyl-CoA	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
Tetradecenoyl Coenzyme A (N-C14:1)	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
pantetheine	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
dodecanoylcarnitine	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
3-oxoheptadecanoyl-CoA	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
stearoylcarnitine	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
3-oxopalmitoleoyl-CoA	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
UDP-galactose	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
trans,cis-hexadeca-2,7-dienoyl-CoA	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
(15Z)-tetracosenoylcarnitine	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
2-amino-3-oxoadipate	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE

Table C.4: Essential metabolites predicted in both CSCs and CCs – continued from previous page.

Metabolite	liver	ovary	AML	kidney	head and neck	prostate	glioblastoma	breast	lung	pancreas
(2E)-heptadecenoyl-CoA	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
previtamin D3	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
docosenoylcarnitine	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
alpha-D-galactose-1-phosphate	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
trans,cis-hexadeca-2,9-dienoyl-CoA	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
eicosadienoylcarnitine	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
10,13,16,19-docosatetraenoylcarnitine	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
(S)-3-hydroxy-7-hexadecenoyl-CoA	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
3(S)-hydroxy-dihomo-gamma-linolenoyl-CoA	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
2-Decaprenyl-5-Hydroxy-6-Methoxy-3-Methyl-1,4-Benzoquinone	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
2-Decaprenyl-6-Methoxy-3-Methyl-1,4-Benzoquinone	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
3-oxo-dihomo-gamma-linolenoyl-CoA	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
3-hydroxyheptadecanoyl-CoA	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE

Table C.5: Composition of Ham's medium.

Reaction ID	**External metabolite**	**Name metabolite**	**Description**
HMR_9066	m01365s	arginine[s]	amino acid
HMR_9038	m02125s	histidine[s]	amino acid
HMR_9041	m02426s	lysine[s]	amino acid
HMR_9042	m02471s	methionine[s]	amino acid
HMR_9043	m02724s	phenylalanine[s]	amino acid
HMR_9045	m03089s	tryptophan[s]	amino acid
HMR_9064	m03101s	tyrosine[s]	amino acid
HMR_9061	m01307s	alanine[s]	amino acid
HMR_9067	m01986s	glycine[s]	amino acid
HMR_9069	m02896s	serine[s]	amino acid
HMR_9044	m02993s	threonine[s]	amino acid
HMR_9070	m01370s	aspartate[s]	amino acid
HMR_9071	m01974s	glutamate[s]	amino acid
HMR_9062	m01369s	asparagine[s]	amino acid
HMR_9063	m01975s	glutamine[s]	amino acid
HMR_9039	m02184s	isoleucine[s]	amino acid
HMR_9040	m02360s	leucine[s]	amino acid
HMR_9068	m02770s	proline[s]	amino acid
HMR_9046	m03135s	valine[s]	amino acid
HMR_9065	m01628s	cysteine[s]	amino acid
HMR_9159	m02982s	thiamin[s]	vitamin B1
HMR_9358	m02159s	hypoxanthine[s]	purine derivative fom in Ham's media
HMR_9146	m01830s	folate[s]	vitamin B9
HMR_9109	m01401s	biotin[s]	vitamin B7
HMR_9145	m02680s	pantothenate[s]	vitamin B5
HMR_9083	m01513s	choline[s]	essential nutrient with AA like metabolism: produced in cells, but insufficiently

HMR_9361	m02171s	inositol[s]	a sugar in media formulations
HMR_9378	m02583s	nicotinamide[s]	form of vitamin B3
HMR_9144	m02817s	pyridoxine[s]	vitamin B6
HMR_9143	m02842s	riboflavin[s]	vitamin B2
HMR_9423	m02996s	thymidine[s]	nucleotide in ham's media
HMR_9269	m01361s	aquacob(III)alamin[s]	vitamin B12
HMR_9167	m02394s	lipoic_acid[s]	derivative of fatty acid in ham's media
HMR_9034	m01965s	glucose[s]	
HMR_9074	m02946s	sulfate[s]	
HMR_9035	m02387s	linoleate[s]	essential fatty acid
HMR_9036	m02389s	linolenate[s]	essential fatty acid
HMR_9048	m02630s	O_2[s]	
HMR_9047	m02040s	H_2O[s]	
HMR_9404	m02833s	retinoate[s]	vitaminA
HMR_9076	m01821s	Fe^{2+}[s]	
HMR_9072	m02751s	Pi[s]	
HMR_9151	m01327s	alpha-tocopherol[s]	vitamin E
HMR_9153	m01935s	gamma-tocopherol[s]	vitamin E

Appendix D

Supplementary Tables - Chapter 3

Table D.1: Reactions involved in DNA methylation and demethylation that were added to the generic model *Human1*.

Reaction ID	Formula	Reasoning	Notes	Metabolites	Gene Rule	Reversible	Subsystem
prodDNAtot	0.009 DNA-5-methylcytosine[n] + 0.00031665646 DNA-5-hydroxymethylcytosine[n] + 0.0000027 DNA-5-formylcytosine + 0.99068064354 DNA[n] => **DNAtotal**[n]	To represent the presence/accumulation of modified cytosines in the DNA. Stoichiometric coefficients are based in the calculated percentages of genome with those modified cytosine species (see **Table D.2**).	Like guaninemethylcytosine/ hydroxymethylcytosine/ formylcytosine base pairs, guaninemethyluracil (a.k.a. guanine-thymine or GT) is eliminated by base-excision repair (BER). However, unlike the first guaninecytosine modified pairs, GT pairs do not accumulate in the genome. So, DNA-5-methyluracil, together with DNA-carboxylcytosine and other DNA modifications that are transient (like those involved in BER - DNA-APsite, DNA-dRPsite, DNA-hang-drPsite, nick-inDNA) were not included in this total DNA composition reaction.	MAM01722n:-0.009, DNA5hmCn:-0.00031665646, DNA5fCn:-0.0000027, MAM01721n:-0.99068064354, DNAtotn:1.0		False	Artificial reactions
adaptbiomass	45 ATP [c] + 0.0267 **DNA-total** [n] + 45 H2O [c] + 0.1124 RNA [c] + 0.4062 glycogen [c] + 0.0012 cofactor_pool_biomass [c] + 5.3375 protein_pool_biomass [c] + 0.2212 lipid_pool_biomass [c] + 0.4835 metabolite_pool_biomass [c] => 45 ADP [c] + 45 H+ [c] + 45 Pi [c] + biomass [c]	To include methylated nuclear DNA in biomass, to guarantee that the DNA methylation reaction MAR08641 is not blocked and that there is methylation when optimizing for cell growth.	DNA was just replaced by DNAtotal in the biomass reaction (MAR13082)	MAM01371c:-45.0, DNAtotn:-0.0267, MAM02040c:-45.0, MAM02847c:-0.1124, MAM03161c:-0.4062, MAM10012c:-0.0012, MAM10013c:-5.3375, MAM10014c:-0.2212, MAM10015c:-0.4835, MAM01285c:45.0, MAM02039c:45.0, MAM02751c:45.0, MAM03970c:1.0		False	Artificial reactions

Table D.1: Reactions involved in DNA methylation and demethylation that were added to the generic model *Human1* – continued from previous page.

Reaction ID	Formula	Reasoning or references	Notes	Metabolites	Gene Rule	Reversible	Subsystem
prodDNA5hmC	DNA-5-methylcytosine [n] + O2[n] + AKG[n] => Succinate[n] + CO2[n] + DNA-5-hydroxymethylcytosine[n]	Reaction R11030 from KEGG [279]. Article Bochtler et al. [280].	Although Fe(II) is used as co-factor and it is converted to Fe(III) and Fe(IV), it is also reconverted to Fe(II) in the same reaction [281, 282], so iron was not included in the reaction.	MAM01722n:-1.0, MAM02630n:-1.0, MAM01306n:-1.0, MAM02943n:1.0, MAM01596n:1.0, DNA5hmCn:1.0	ENSG00000138336 or ENSG00000168769 or ENSG00000187605	False	Dna (de)/ methylation
prodDNA5fC	DNA-5-hydroxymethylcytosine[n] + O2[n] + AKG[n] => Succinate[n] + CO2[n] + DNA-5-formylcytosine[n] + H2O[n]	Reaction 53828 from Rhea [283]. Articles Bochtler et al. [280] and Popov et al. [284].		DNA5hmCn:-1.0, MAM02630n:-1.0, MAM01306n:-1.0, MAM02943n:1.0, MAM01596n:1.0, DNA5fCn:1.0, MAM02040n:1.0	ENSG00000138336 or ENSG00000168769 or ENSG00000187605	False	Dna (de)/ methylation
prodDNA5CaC	DNA-5-formylcytosine[n] + O2[n] + AKG[n] => Succinate[n] + CO2[n] + DNA-5-carboxylcytosine[n] + H[n]	Reaction 53832 from Rhea [283]. Articles Bochtler et al. [280] and Popov et al. [284].		DNA5fCn:-1.0, MAM02630n:-1.0, MAM01306n:-1.0, MAM02943n:1.0, MAM01596n:1.0, DNA5CaCn:1.0, MAM02039n:1.0	ENSG00000138336 or ENSG00000168769 or ENSG00000187605	False	Dna (de)/ methylation
prodDNA5mU	DNA-5-methylcytosine[n] + H2O[n] => DNA-5-methyluracil[n] + NH3[n]	Reaction R01411 from KEGG [279]. Article Popov et al. [284].		MAM01722n:-1.0, MAM02040n:-1.0, DNA5mUn:1.0, MAM02578n:1.0	ENSG00000111732 or ENSG00000111701 or ENSG00000128383 or ENSG00000179750 or ENSG00000244509 or ENSG00000128394 or ENSG00000239713 or ENSG00000100298	False	Dna (de)/ methylation

Table D.1: Reactions involved in DNA methylation and demethylation that were added to the generic model *Human1* – continued from previous page.

Reaction ID	Formula	Reasoning or references	Notes	Metabolites	Gene Rule	Reversible	Subsystem
consdirectDNA5fC	DNA-5-formylcytosine[n] + H2O[n] => DNA[n] + formate[n] + H[n]	Articles Schon et al. [285] and Iwan et al. [286].	Formate has chemical formula HCOOH, but since in *Human1* it has less 1 H atom, we added H^+ to keep the mass and charge balance. There is no report of an enzyme that catalyzes this reaction. It is going to be considered a uncatalyzed reaction and protected during reconstruction of cell-specific models.	DNA5fCn:-1.0, MAM02040n:-1.0, MAM01721n:1.0, MAM01833n:1.0, MAM02039n:1.0	Uncatalyzed	False	Dna (de)/ methylation
consdirectDNA5CaC	DNA-5-carboxylcytosine[n] + H[n] => DNA[n] + CO2[n]	Articles Iwan et al. [286] and Feng et al. [287].	There is no report of an enzyme that catalyzes this reaction in humans. One study suggests the involvement of DNMTs, but it was done with bacterial and mouse enzymes [287]. It is going to be considered an uncatalyzed reaction and protected during reconstruction of tissue-specific models.	DNA5CaCn:-1.0, MAM02039n:-1.0, MAM01721n:1.0, MAM01596n:1.0	Uncatalyzed	False	Dna (de)/ methylation
prodAPsite1	DNA-5-methyluracil[n] + H2O[n] => DNA-APsite[n] + thymine[n]	Articles Bhutani et al. [288], Popov et al. [284], Tsukada et al. [289] and Grin et al. [290].	Grin et al. [290] explains mechanism of pyrimidine BER (Base Excision Repair).	DNA5mUn:-1.0, MAM02040n:-1.0, DNAapn:1.0, MAM02997n:1.0	ENSG00000139372 or ENSG00000129071 or ENSG00000140398	False	Dna (de)/ methylation
prodAPsite3	DNA-5-formylcytosine[n] + H2O[n] => DNA-APsite[n] + 5-formylcytosine[n]	Articles Rasmussen et al. [291], Tsukada et al. [289] and Grin et al. [290].	Grin et al. [290] explains mechanism of pyrimidine BER (Base Excision Repair).	DNA5fCn:-1.0, MAM02040n:-1.0, DNAapn:1.0, M5fCn:1.0	ENSG00000139372	False	Dna (de)/ methylation
prodAPsite4	DNA-5-carboxylcytosine[n] + H2O[n] => DNA-APsite[n] + 5-carboxylcytosine[n]	Articles Popov et al. [284], Tsukada et al. [289] and Grin et al. [290].	Grin et al. [290] explains mechanism of pyrimidine BER (Base Excision Repair).	DNA5CaCn:-1.0, MAM02040n:-1.0, DNAapn:1.0, M5CaCn:1.0	ENSG00000139372	False	Dna (de)/ methylation

Table D.1: Reactions involved in DNA methylation and demethylation that were added to the generic model *Human1* – continued from previous page.

Reaction ID	Formula	Reasoning or references	Notes	Metabolites	Gene Rule	Reversible	Subsystem
proddRPsite	DNA-APsite[n] + H2O[n] => DNA-dRPsite[n]	Articles Popov et al. [284], Grin et al. [290].	Grin et al. [290] explains mechanism of pyrimidine BER (Base Excision Repair).	DNAapn:-1.0, MAM02040n:-1.0, DNAdrpn:1.0	ENSG00000100823 or (ENSG00000172613 and ENSG00000113456 and ENSG00000136273 and ENSG00000100823) or ENSG00000169188 or ENSG00000154328	False	Dna (de)/ methylation
prodhangdRPsite	DNA-dRPsite[n] + dTTP[n] => DNA-hang-dRPsite[n] + PPi[n]	Articles Popov et al. [284], Grin et al. [290].	Grin et al. [290] explains mechanism of pyrimidine BER (Base Excision Repair).	DNAdrpn:-1.0, MAM01753n:-1.0, DNAhgdrpn:1.0, MAM02759n:1.0	ENSG00000070501 or (ENSG00000172613 and ENSG00000113456 and ENSG00000136273 and ENSG00000070501)	False	Dna (de)/ methylation
prodDNAnick	DNA-hang-dRPsite[n] + H2O[n] => nick-in-DNA[n] + 2-deoxy-D-ribose-5-phosphate[n] + H+[n]	Articles Popov et al. [284], Grin et al. [290].	Grin et al. [290] explains mechanism of pyrimidine BER (Base Excision Repair).	DNAhgdrpn:-1.0, MAM02040n:-1.0, DNAnick:1.0, MAM00640n:1.0, MAM02039n:1.0	ENSG00000070501 or ENSG00000166169	False	Dna (de)/ methylation

Table D.1: Reactions involved in DNA methylation and demethylation that were added to the generic model *Human1* – continued from previous page.

Reaction ID	Formula	Reasoning or references	Notes	Metabolites	Gene Rule	Reversible	Subsystem
ligateDNA	nick-in-DNA[n] => DNA[n] + H2O[n]	Articles Popov et al. [284], Grin et al. [290].	Grin et al. [290] explains mechanism of pyrimidine BER (Base Excision Repair).	DNAnick:-1.0, MAM01721n:1.0, MAM02040n:1.0	ENSG00000005156 or (ENSG00000005156 and ENSG00000143799) or (ENSG00000039650 and ENSG00000005156 and ENSG00000073050 and ENSG00000042088) or (ENSG00000005156 and ENSG00000073050)	False	Dna (de)/ methylation
consdirect5fC *	5-formylcytosine[n] +H2O[n] => cytosine[n] + formate[n] + H[n]	To allow flux through the reaction producing 5fC and the other reactions leading to that one (e.g. prodAPsite3, prodDNA5fC). Although there is no clear evidence that 5fC is converted to cytosine after being excised from DNA by BER, there is evidence that it occurs when the base is part of DNA (see reaction consdirectDNA5fC), so an assumption will be made that it can also occur after the 5fC is released from DNA backbone.		M5fCn:-1.0, MAM02040n:-1.0, MAM01632n:1.0, MAM01833n:1.0, MAM02039n:1.0	Uncatalyzed	False	Dna (de)/ methylation

Table D.1: Reactions involved in DNA methylation and demethylation that were added to the generic model *Human1* – continued from previous page.

Reaction ID	Formula	Reasoning or references	Notes	Metabolites	Gene Rule	Reversible	Subsystem
consdirect5CaC *	5-carboxylcytosine[n] + H[n] => cytosine[n] + CO2[n]	To allow flux through the reaction producing 5CaC and the other reactions leading to that one (e.g. prodAPsite4, prodDNA5CaC). Although there is no clear evidence that 5CaC is converted to cytosine after being excised from DNA by BER, there is evidence that it occurs when the base is part of DNA (see reaction consdirectDNA5CaC), so an assumption will be made that it can also occur after the 5CaC is released from DNA backbone.		M5CaCn:-1.0, MAM02039n:-1.0, MAM01632n:1.0, MAM01596n:1.0	Uncatalyzed	False	Dna (de)/methylation
transp2deox5ribP *	2-deoxy-D-ribose-5-phosphate[n] => 2-deoxy-D-ribose-5-phosphate[c]	To allow flux through the reaction that produces 2-deoxy-D-ribose-5-phosphate in the nucleus (prodDNAnick) and other reactions leading to that one. There are probably other reactions that use 2-deoxy-D-ribose-5-phosphate in the nucleus, just like it is used in the cytoplasm, but since there is no evidence for their occurrence in the nucleus, a direct transport reaction of 2-deoxy-D-ribose-5-phosphate to cytoplasm will be assumed.		MAM00640n:-1.0, MAM00640c:1.0	Uncatalyzed	False	Transport reactions

Table D.1: Reactions involved in DNA methylation and demethylation that were added to the generic model *Human1* – continued from previous page.

Reaction ID	Formula	Reasoning or references	Notes	Metabolites	Gene Rule	Reversible	Subsystem
transpthymine	thymine[n] <=> thymine[c]	To allow flux through the reaction that produces thymine in nucleus (prodAPsite1) and other reactions leading to that one. Human ENT2 (SLC29A2) transports pyrimidine nucleobases and nucleosides and exists in both cell and nuclear membranes [292]. In *human1*, the cytosolic to extracellular reaction is reversible (MAR04980), so we assume the same for transport from cytoplasm to nucleus.		MAM02997n:-1.0, MAM02997c:1.0	ENSG00000174669	True	Transport reactions
transpcytosine	cytosine[n] <=> cytosine[c]	To allow flux through the reaction that produces cytosine in nucleus (consdirect5fC, consdirect5CaC) and other reactions leading to that one. human ENT2 (SLC29A2) transports pyrimidine nucleobases and nucleosides and exists in both cell and nuclear membranes [292]. In *human1*, the cytosolic to extracellular reaction is reversible (MAR08636), so we assume the same for transport from cytoplasm to nucleus.		MAM01632n:-1.0, MAM01632c:1.0	ENSG00000174669	True	Transport reactions

Table D.1: Reactions involved in DNA methylation and demethylation that were added to the generic model *Human1* – continued from previous page.

Reaction ID	Formula	Reasoning or references	Notes	Metabolites	Gene Rule	Reversible	Subsystem
transpakg *	AKG[c] => AKG[n]	To allow flux through the reactions that use AKG in nucleus (prodDNA5hmC, prodDNA5fC, prodDNA5CaC). There is evidence for AKG to be used in these reactions in the nucleus, so it has to be transported to nucleus somehow.		MAM01306c:-1.0, MAM01306n:1.0		False	Transport reactions
transpsucc *	succinate[n] => succinate[c]	To allow flux through the reactions that produce succinate in nucleus (prodDNA5hmC, prodDNA5fC, prodDNA5CaC). There are probably other reactions that use succinate in the nucleus, just like it is used in the cytoplasm, but since there is no evidence for their occurrence in the nucleus, a direct transport reaction of succinate to cytoplasm will be assumed.		MAM02943n:-1.0, MAM02943c:1.0		False	Transport reactions

* Reaction included based on an assumption and to unblock other reactions

Table D.2: Calculation of generic composition of total DNA in terms of modified cytosines

Citations or sources	**Calculation**
"...5mC is found in all tissues, corresponding to ∼ **4%–5% of all cytosines**" from Rasmussen et al. [291]. "In human somatic cells, **m5C accounts for ∼ 1% of total DNA bases**" from Bird et al. [293].	• average of 4.5% of all cytosines in human genome is methylated. • proportion of cytosines in human genome is 20%. The stoichiometric coefficient of dCTP in the reaction of DNA formation MAR07160 is 0.2. So: 0.2*0.045 = **0.009** = 0.9% ∼ 1% **of human genome is methylated**"
Authors of He et al. [256] kindly provided us with the total number of 5hmC sites found in different normal human tissues. The calculated average percentage of the human genome with 5hmC sites across different tissues is ∼ **0.03%**.	• average ratio of human genome with 5hmC sites was calculated in **Table D.3**. The average number of 5hmC sites (at both DNA strands) across different tissues was divided by twice the total number of base pairs of the reference genome (hg38 used in the study He et al. [256]) to account for the two DNA strands. So, **0.00031665646** = 0.031665646% ∼ 0.03% **of human genome is hydroxymethylated**.
"The steady-state **levels** of 5fC and 5caC are much lower than those of 5hmC, **corresponding to approximately 0.03%** and 0.01% of 5mC levels, respectively" (note: in that context, levels of 5fC, 5caC, and 5mC refer to levels in the DNA) [294]. "**The stable levels of genomic 5CaC have not been experimentally determined, as they are often under the detection limit**" from Rasmussen et al. [291]. " (...) approximately **1/3 and 2/3 of the 5caC-DNA underwent direct decarboxylation and TDG-BER processing**, respectively." from Feng et al. [287]."	• 0.009 DNA5mC * 0.0003 = **0.0000027** = 0.00027 % of human genome has 5fC • although a study claims genomic 5CaC levels to be around 0.01% of hmC levels [294], a later study states its values are often under detection limit [291] (so reported levels might not be accurate) and a recent study reports that almost all genomic 5CaC is either directly or indirectly decarboxylated [287], suggesting that genomic 5mC levels are mostly transient. **Therefore, no accumulation of genomic CaC was assumed (no integration of CaC in DNAtot reaction)**.

Table D.3: Calculation of ratio of human genome with 5hmC sites.

Sample name	**Number of 5hmC sites** *	**Tissue ID**	**Tissue name**	**Number of 5hmC sites per tissue**	**Average of 5hmC sites**	**Number of genome (hg38) base pairs** **	**Number of bases in both strands of genome**	**Ratio of human genome with 5hmC**	**Average ratio of human genome with 5hmC**
AG_1	1776524	AG	Adrenal gland	1776524	2032482.331	3209286105	6418572210	0.000276779	0.00031665646
AO_1	1738469	AO	Aorta	1860046.5				0.000289791	
AO_4	1981624	BD	Bladder	1385939				0.000215926	
BD_4	1158842	BR	Brain	2702696				0.000421074	
BD_5	1613036	ES	Esophagus	2049301				0.000319277	
BR_5	2472751	HR	Heart	2434799.333				0.000379337	
BR_5_BS	2312037	KD	Kidney	2354625.667				0.000366846	
BR_5_CE	3076511	LE	Large Intestine	2102597.286				0.00032758	
BR_5_HT	2499648	LI	Liver	2281792.333				0.000355498	
BR_5_SC	2840152	LN	Lung	1953605.333				0.000304368	
BR_6_BF	2329263	OV	Ovary	2036697.5				0.000317313	
BR_7_BF	2595365	PA	Pancreas	1626069.5				0.000253338	
BR_8_BF	3495841	SE	Small Intestine	1918671.667				0.000298925	
ES_5	2049301	SK	Skin	1618916				0.000252224	
HR_1	2205456	ST	Stomatch	2006284.667				0.000312575	
HR_6_LV	2360358	SX	Spleen	2002641.333				0.000312007	
HR_8_LV	2738584	TC	Trachea	1902447.5				0.000296397	
KD_6	2368435	TR	Thyroid	2345359				0.000365402	
KD_7	2350514	UT	Uterus	2258150.667				0.000351815	
KD_8	2344928								
LE_4_AP	1504499								
LE_4_RE	2000723								
LE_5_AP	2154628								
LE_5_CL	2353520								

Table D.3: Calculation of ratio of human genome with 5hmC sites – continued from previous page.

Sample name	Number of 5hmC sites *	Tissue ID	Tissue name	Number of 5hmC sites per tissue	Average of 5hmC sites	Number of genome (hg38) base pairs **	Number of bases in both strands of genome	Ratio of human genome with 5hmC	Average ratio of human genome with 5hmC
LE_5_RE	2202961								
LE_6	2238781								
LE_8	2263069								
LI_1	2229338								
LI_4	2130782								
LI_8	2485257								
LN_1	1854015								
LN_6	1978612								
LN_7	2028189								
OV_7	1866030								
OV_8	2207365								
PA_5	1955406								
PA_6	1296733								
SE_4_IE	1998248								
SE_4_JE	1908925								
SE_5_JE	1601111								
SE_6	2131610								
SE_7	1518442								
SE_8	2353694								
SK_1	1610770								
SK_5	1571945								
SK_8	1674033								
ST_5	1971857								

Table D.3: Calculation of ratio of human genome with 5hmC sites – continued from previous page.

Sample name	Number of 5hmC sites *	Tissue ID	Tissue name	Number of 5hmC sites per tissue	Average of 5hmC sites	Number of genome (hg38) base pairs **	Number of bases in both strands of genome	Ratio of human genome with 5hmC	Average ratio of human genome with 5hmC
ST_6	2148742								
ST_7	1898255								
SX_1	1891294								
SX_6	2185887								
SX_7	1930743								
TC_4	1849401								
TC_5	1955494								
TR_4	2270133								
TR_5	2420585								
UT_6	2381044								
UT_7	2193241								
UT_8	2200167								

* The values were kindly provided by the authors of He et al. [256] and correspond to the number of 5hmC sites on both strands for each sample.
Samples of the same tissue start with the same two-letter-tissue identifier.
Some tissues of NCI-60 (breast, prostate, haemotopoietic and lymphoid tissue) were not tested in He et al. study.
** Value was taken from `genomewiki.ucsc.edu/index.php/Hg38_30-way-Genome.site_statistics`.

Table D.4: Updated gene rules of previously existing reactions and new rules associated with newly added reactions.

Reaction ID	Original gene rule	New gene rule	New gene rule (with gene IDs)	New protein rule (with Uniprot IDs)	Reasoning or references	EC number	Names of proteins inside complexes
MAR08641	(DNMT3B and DNMT1) or TRDMT1 or DNMT3A or DNMT3L	DNMT1 or DNMT3A or DNMT3B or (DNMT3B and DNMT1) or (DNMT3A and DNMT3L) or (DNMT3B and DNMT3L)	ENSG00000130816 or ENSG00000119772 or ENSG00000088305 or (ENSG00000088305 and ENSG00000130816) or (ENSG00000119772 and ENSG00000142182) or (ENSG00000088305 and ENSG00000142182)	P26358 or Q9Y6K1 or Q9UBC3 or (Q9UBC3 and P26358) or (Q9Y6K1 and Q9UJW3) or (Q9UBC3 and Q9UJW3)	• TRDMT1 a.k.a. DNMT2 methylates tRNA and there is still discussion whether it methylates DNA to a low degree or it just does not work as a DNA methylase [282, 295]. So, DNMT2 will not be inlude in the gene rule. • There is evidence that DNMT1 and DNMT3B are part of the same complex (retrieved from the CORUM database) [296, 297], but there is also evidence that DNMT3B is part of a complex of other proteins that do not include DNMT1 [298]. Also, each enzyme is mainly used in different situations: "DNMT1, DNMT3A and DNMT3B have different functions in the methylation process. DNMT1 is required for the maintenance of all methylation in the genome. During replication, DNMT1 restores the specific methylation pattern on the daughter strand in accordance with that of the parental DNA. DNMT3A and DNMT3B are referred to as de novo methyltransferases, which are responsible for establishing DNA methylation patterns during embryogenesis and setting up genomic imprints during germ cell development (...). Although they are highly expressed in early mammalian embryos, DNMT3A and DNMT3B decrease in expression over the course of cell differentiation" [295]. So, instead of only using "(DNMT3B and DNMT1)" is better to also include "DNMT3B or DNMT1", i.e. some redundancy will be applied. • DNMT3A and DNMT3L form a complex: "DNMT3L, an important regulator without catalytic activity, operates in the form of DNMT3L-DNMT3A heterotetramers" [295]. So, the gene rule should include "(DNMT3A and DNMT3L)", but some redundancy will be used, as DNMT3L seems to only increase the activity of DNMT3A (does not have catalytic activity) [295, 299]. Therefore, "or DNMT3A" will also be used. • DNMT3B and DNMT3L form a complex [299]. So, the gene rule should include (DNMT3B and DNMT3L), but some redundancy will also be assumed, as DNMT3L seems to only increase the activity of DNMT3B (does not have catalytic activity) [299]. Therefore, "or DNMT3B" alone will also be used.	2.1.1.37	DNA (cytosine-5)-methyltransferase 1, DNA (cytosine-5)-methyltransferase 3A, DNA (cytosine-5)-methyltransferase 3B, DNA (cytosine-5)-methyltransferase 3L

Table D.4: Updated gene rules of previously existing reactions and new rules associated with newly added reactions – continued from previous page.

Reaction ID	Original gene rule	New gene rule	New gene rule (with gene IDs)	New protein rule (with Uniprot IDs)	Reasoning or references	EC number	Names of proteins inside complexes
prodDNA5hmC		TET1 or TET2 or TET3	ENSG00000138336 or ENSG00000168769 or ENSG00000187605	Q8NFU7 or Q6N021 or O43151	• TET1 role at KEGG (entry: hsa:80312) [279]. • TET1,2,3 role at InterPro (entry: IPR040175) [300]. • TET1,2,3 role at Rasmussen et al. [291]. • There is no complex of these enzymes in CORUM database [296].	1.14.11	tet methylcytosine dioxygenase 1, tet methylcytosine dioxygenase 2, tet methylcytosine dioxygenase 3
prodDNA5fC		TET1 or TET2 or TET3	ENSG00000138336 or ENSG00000168769 or ENSG00000187605	Q8NFU7 or Q6N021 or O43151	• TET1,2,3 role at InterPro (entry: IPR040175) [300]. • TET1,2,3 role at Rasmussen et al. [291]. • There is no complex of these enzymes in CORUM database [296].	1.14.11	tet methylcytosine dioxygenase 1, tet methylcytosine dioxygenase 2, tet methylcytosine dioxygenase 3
prodDNA5CaC		TET1 or TET2 or TET3	ENSG00000138336 or ENSG00000168769 or ENSG00000187605	Q8NFU7 or Q6N021 or O43151	• TET1,2,3 role at InterPro (entry: IPR040175) [300]. • TET1,2,3 role at Rasmussen et al. [291]. • There is no complex of these enzymes in CORUM database [296].	1.14.11	tet methylcytosine dioxygenase 1, tet methylcytosine dioxygenase 2, tet methylcytosine dioxygenase 3
prodDNA5mU		AID/AICDA or APOBEC1 or APOBEC3A or APOBEC3B or APOBEC3C or APOBEC3F, APOBEC3G or APOBEC3H	ENSG00000111732 or ENSG00000111701 or ENSG00000128383 or ENSG00000179750 or ENSG00000244509 or ENSG00000128394 or ENSG00000239713 or ENSG00000100298	Q9GZX7 or P41238 or P31941 or Q9UH17 or Q9NRW3 or Q8IUX4 or Q9HC16 or Q6NTF7	• The function of AID, which is part of the family of APOBECs, is described in Bhutani et al. [288]. • The function of APOBECs family members is described in Popov et al. [284]. • All APOBECs except A3D, A2 and A4 can convert 5mC to thymine (same as 5-methyluracil) [301]. • There is no complex of the.se enzymes in CORUM database [296]. • There is no Ensembl gene id for APOBEC3E so it was excluded from the gene rule.	3.5.4.36, 3.5.4.38	Activation-induced cytidine deaminase (AID) a.k.a AICDA, apolipoprotein B mRNA-editing catalytic polypeptides (APOBECs) deaminases

Table D.4: Updated gene rules of previously existing reactions and new rules associated with newly added reactions – continued from previous page.

Reaction ID	Original gene rule	New gene rule	New gene rule (with gene IDs)	New protein rule (with Uniprot IDs)	Reasoning or references	EC number	Names of proteins inside complexes
prodAPsite1		TDG or MBD4 or NEIL1	ENSG00000139372 or ENSG00000129071 or ENSG00000140398	Q13569 or O95243 or Q96FI4	• TDG and MBD4 role at Popov et al. [284]. • TDG role at UniProt (entry: Q13569) [302]. • MBD4 role at UniProt (entry: Q95243) [302]. • NEIL1 has among other functions DNA glycosylase activity towards mismatched thymine, see UniProt (entry: Q96FI4) [302].	3.2.2.29	thymine DNA glycosylase, methyl-CpG-binding domain protein 4, Endonuclease 8-like 1
prodAPsite3		TDG	ENSG00000139372	Q13569	• TDG role at UniProt (entry: Q13569) [302]. • TDG role at Tsukada e. al. [289].	3.2.2.29	thymine DNA glycosylase
prodAPsite4		TDG	ENSG00000139372	Q13569	• TDG role at UniProt (entry: Q13569) [302]. • TDG role at Popov et al. [284].	3.2.2.29	thymine DNA glycosylase
proddRPsite		APE1 or (RAD9A and RAD1 and HUS1 and APE1) or APE2 or NEIL2	ENSG00000100823 or (ENSG00000172613 and ENSG00000113456 and ENSG00000136273 and ENSG00000100823) or ENSG00000169188 or ENSG00000154328	P27695 or (Q99638 and O60671 and O60921 and P27695) or Q9UBZ4 or Q969S2	• APE1 role at UniProt (entry: P27695) [302]. • APE1 role at Popov et al. [284], Tsukada et al. [289] and Grin et al. [290]. • CORUM database has a complex where APEX1 (a.k.a. APE1) is included that stimulates the activity of APEX1 for DNA repair [296]. • APE2 role at Uniprot (entry: Q9UBZ4) [302]. • NEIL2 has AP (apurinic/apyrimidinic) lyase activity, see Uniprot (entry: Q969S2) [302].	3.1.-.-, 3.1.11.2	apurinic/apyrimidinic endodeoxyribonuclease 1 (APE1 a.k.a. APEX1), RAD9 checkpoint clamp component A, RAD1 checkpoint DNA exonuclease, HUS1 checkpoint clamp component, endonuclease 8-like 2

Table D.4: Updated gene rules of previously existing reactions and new rules associated with newly added reactions – continued from previous page.

Reaction ID	Original gene rule	New gene rule	New gene rule (with gene IDs)	New protein rule (with Uniprot IDs)	Reasoning or references	EC number	Names of proteins inside complexes
prodhangdRPsite		POLB or (RAD9A and RAD1 and HUS1 and POLB)	ENSG00000070501 or (ENSG00000172613 and ENSG00000113456 and ENSG00000136273 and ENSG00000070501)	P06746 or (Q99638 and O60671 and O60921 and P06746)	• Beta-polymerase role at Popov et al. [284] and Grin et al. [290]. • Beta-polymerase role at UniProt (entry: P06746) [302]. • CORUM database has the complex 9-1-1 composed by RAD9A+RAD1+Hus1+POL1 that stimulates the activity of POLB thus recruiting POLB to DNA damage sites [296]. • CORUM database has complexes, where POLB interacts with enzymes that catalyze other steps of BER, since those steps are represented by other reactions that we include in the model (e.g. prodDNAnick, ligateDNA, etc.) such complexes, were excluded here (for e.g. complex PNKP+LIG3+ POLB+XRCC1 was excluded because besides POLB it has LIG3, which catalyzes DNA ligase shown below) [296].	2.7.7.7	DNA polymerase beta, RAD9 checkpoint clamp component A, RAD1 checkpoint DNA exonuclease, HUS1 checkpoint clamp component
prodDNAnick		POLB or POLL	ENSG00000070501 or ENSG00000166169	P06746 or P06746	• Beta-polymerase beta role at Grin et al. [290]. • Beta-polymerase beta role at UniProt (entry: P06746) [302]. • DNA-polymerase lambda has, among other functions, dRP-lyase activity, shown at UniProt (entry: Q9UGP5) [302].	4.2.99.-	DNA polymerase beta, DNA polymerase lambda

Table D.4: Updated gene rules of previously existing reactions and new rules associated with newly added reactions – continued from previous page.

Reaction ID	Original gene rule	New gene rule	New gene rule (with gene IDs)	New protein rule (with Uniprot IDs)	Reasoning or references	EC number	Names of proteins inside complexes
ligateDNA		LIG3 or (LIG3 and PARP1) or (PNKP and LIG3 and XRCC1 and TDP1) or (LIG3 and XRCC1)	ENSG00000005156 or (ENSG00000005156 and ENSG00000143799) or (ENSG00000039650 and ENSG00000005156 and ENSG00000073050 and ENSG00000042088) or (ENSG00000005156 and ENSG00000073050)	P49916 or (P49916 and P09874) or (Q96T60 and P49916 and P18887 and Q9NUW8) or (P49916 and P18887)	• DNA ligase III role at Popov et al. [284] and Grin et al. [290]. • DNA ligase III role at UniProt (entry: P49916) [302]. • CORUM database has complexes where LIG3 interacts with enzymes that catalyze other steps of BER, since those steps are represented by other reactions that we include in the model, such complexes were excluded here [296].	6.5.1.1	DNA ligase 3, poly(ADP-ribose) polymerase 1, polynucleotide kinase 3'-phosphatase, X-ray repair cross complementing 1, tyrosyl-DNA phosphodiesterase 1
transpthymine		ENT2	ENSG00000174669	Q14542	• human ENT2 (SLC29A2) transports pyrimidine nucleobases and nucleosides and exists in both cell and nuclear membranes. ENT1 (SLC29A1, which in Human1, like SLC29A2, transports deoxycytidine between cytoplasm and nucleus) transports nucleosides, but does not seem to transport nucleobases [292]. Therefore, ENT2 protein/gene was associated here with the reaction. Nonetheless, there may exist other transport proteins that can move pyrimidines across the nucleus membrane which may not have been discovered yet, as the transport reaction itself seems to not be described in databases(does not exist in KEGG [279] or Rhea [283]).		Equilibrative nucleoside transporter 2 (ENT2), solute carrier family 29 member 2

Table D.4: Updated gene rules of previously existing reactions and new rules associated with newly added reactions – continued from previous page.

Reaction ID	Original gene rule	New gene rule	New gene rule (with gene IDs)	New protein rule (with Uniprot IDs)	Reasoning or references	EC number	Names of proteins inside complexes
transpcytosine		ENT2	ENSG00000174669	Q14542	• human ENT2 (SLC29A2) transports pyrimidine nucleobases and nucleosides and exists in both cell and nuclear membranes. ENT1 (SLC29A1, which in Human1, like SLC29A2, transports deoxycytidine between cytoplasm and nucleus) transports nucleosides, but does not seem to transport nucleobases [292]. Therefore, ENT2 protein/gene was associated here with the reaction. Nonetheless, there may exist other transport proteins that can move pyrimidines across the nucleus membrane which may not have been discovered yet, as the transport reaction itself seems to not be described in databases(does not exist in KEGG [279] or Rhea [283]).		Equilibrative nucleoside transporter 2 (ENT2), solute carrier family 29 member 2

Note: Search for complexes was made in CORUM database looking for a core set only, i.e. a reduced set of complexes which is free from redundant entries.

Table D.5: New metabolites participating in the new reactions added to *Human1*.

Metabolite ID	Name	Formula	Charge	Compartment	Artificial	Notes
MAM01306n *	AKG	$C_5H_4O_5$	-2	nucleus	False	
MAM02943n *	succinate	$C_4H_4O_4$	-2	nucleus	False	
DNAtotn	DNA_total	$C_{10}H_{17}O_8PR_2$	0	nucleus	True	
DNA5hmCn	DNA-5-hydroxy-methylcytosine	$C_{11}H_{19}O_9PR_2$	0	nucleus	False	
DNA5fCn	DNA-5-formylcytosine	$C_{11}H_{17}O_9PR_2$	0	nucleus	False	
DNA5CaCn	DNA-5-carboxylcytosine	$C_{11}H_{16}O_10PR_2$	-1	nucleus	False	
DNA5mUn	DNA-5-methyluracil	$C_{11}H_{18}O_9PR_2 - N$	0	nucleus	False	Although methyluracil is the same as thymine, the model does not have the metabolite DNA containing thymine or containing methyluracil
DNAapn	DNA-APsite	$C_{10}H_{18}O_9PR$	0	nucleus	False	
MAM02997n *	5-methyl-uracil a.k.a thymine	$CH_2OR - N$	0	nucleus	False	
M5fCn	5-formylcytosine	$CHOR$	0	nucleus	False	
M5CaCn	5-carboxylcytosine	CO_2R	-1	nucleus	False	
DNAdrpn	DNA-dRPsite	$C_{10}H_{20}O_{10}PR$	0	nucleus	False	
DNAhgdrpn	DNA-hang-dRPsite	$C_{15}H_{27}O_{15}P_2R_2$	-1	nucleus	False	
MAM00640n *	2-deoxy-D-ribose-5-phosphate	$C_5H_9O_7P$	-2	nucleus	False	
DNAnick	nick-in-DNA	$C_{10}H_{19}O_9PR_2$	0	nucleus	False	
MAM01632n *	cytosine	RH	0	nucleus	False	

Metabolites with $*$ do not exist in the nucleus in the original *Human1* model, but exist in other compartments. All other metabolites do not exist in any compartment of the original model. Formulas of all metabolites are based on the formula of DNA in Human1 generic model (which includes two R groups and has a charge of 0): $C_{10}H_{17}O_8PR_2$. "R" represents the azotated base without an H atom (corresponding to the position where the base connects to DNA backbone). "$-N$" in the formulas means there is one less N atom in the R group (i.e. to cytosine).

Table D.6: Charge and mass balance calculations.

Table is in Excel spreadsheet format to let the readers view the mathematical formulas. The link to the spreadsheet is: https://docs.google.com/spreadsheets/d/1OE5QKwO6IAPHlwRAXRLciPJ9LPORPOqi/edit?usp=share_link&ouid=101146105014472512253&rtpof=true&sd=true

Table D.7: Estimation of the cell line-specific ratio of total DNA containing DNA5mC, DNA5hmC or DNA5fC.

Table is in Excel spreadsheet format to let the readers view the mathematical formulas. The link to the spreadsheet is: https://docs.google.com/spreadsheets/d/1Jk_jfhgBQFqyi-NY4WBfknZFB9Btj7dF/edit?usp=share_link&ouid=101146105014472512253&rtpof=true&sd=true

Table D.8: DNA methylation flux rules

Citation	**Reaction rate rules**
"... kinetic analyses of TET catalytic activity suggest that the rate of cytosine oxidation is significantly reduced for [. . .] and 5fC (**7.8-fold to 12.6-fold**) substrates compared with the initial oxidation reaction of 5mC." [291].	prodDNA5hmC rate / prodDNA5CaC rate = 7.8 to 12.6 fold. So, **10.2** fold average. Note that: TET catalyzed reaction where 5fC is used as substrate, is prodDNA5CaC.
"At 48 h following transfection, the levels of direct decarboxylation product, short-, and long-patch BER product were approximately 34%, 58%, and 8%, respectively [...]. Collectively, these results illustrate that approximately 1/3 and **2/3** of the 5caC-DNA underwent direct decarboxylation and **TDG-BER** processing, respectively." [287]	**2/3** of DNA5CaCn is converted to DNA by TDG-BER, so the flux of prodAPsite4 is 2/3 of the flux of prodDNA5CaC (which is the only reaction producing DNA5CaC).