Development of metabolic modelling methods to determine vulnerabilities associated to metabolic reprogramming in acute myeloid leukaemia

by

Effrosyni Karakitsou

A thesis submitted to the University of Birmingham for the degree of

DOCTOR OF PHILOSOPHY

Institute of Cancer and Genomic Sciences

Centre for Computational Biology

University of Birmingham

December 2020

Abstract

Metabolism refers to all the biochemical reactions that take place inside the cells of living organisms in order to sustain life. Metabolic deregulation has been observed in disease and has been established as a hallmark for cancer. The metabolic adaptation that occurs in cancer cells contributes in cancer progression, metastasis and the development of chemotherapy drug resistance. Thus, studying cancer metabolism can provide valuable insight into the mechanisms underlying pathogenesis and provide means to concur it.

The metabolic phenotype is a result of complex biological processes and regulatory mechanisms and therefore should be studied under the holistic approach of Systems Biology. Under this scope, different types of biological data representing the multiple layers of biological complexity should be integrated towards exploring and deciphering metabolism. Metabolic modelling provides the appropriate mathematical framework for the representation of the entirety of metabolic reactions and pathways. Additionally, the amount of genomic data and the functional annotation of entire genomes has made the reconstruction of metabolic networks at a genome scale and the representation of its full metabolic potential possible. Constraint-based modelling is broadly used to perform simulations on genome-scale metabolic models (GSMMs), since it can integrate previously established knowledge and experimentally generated -omic data (such as transcriptomics, metabolomics and proteomics) to build highly accurate condition-specific GSMMs for predictive studies.

As part of this Ph.D., we have developed computational methods to study the metabolic adaptations that emerge in Acute Myeloid Leukaemia (AML), aiming in the identification of new therapeutic and prognostic biomarkers. More specifically, a new computational platform able to integrate a high variety of –omic data into a genome-scale metabolic network reconstruction using constraint-based methods has been developed and employed in the study of different cell line models of AML (THP-1 and HL-60) under different stresses, such as drug treatment or specific gene inhibition. Ultimately, the AML GSMMs were subjected to a systematic simulation of gene knock-outs, which led to the identification of genes that would strongly compromise cell viability, predicting putative metabolic-related vulnerabilities that could be exploited in novel combination therapies.

Moreover, we reconstructed a consensus genome-scale model for AML, onto which patient-derived transcriptomic data from The Cancer Genome Atlas (TCGA) database were integrated, resulting in the reconstruction of AML patient-specific GSMMs. We designed a workflow combining constraint-based modelling and machine learning dimensionality reduction and classification to perform risk-stratification of AML patients and prognostic biomarker discovery. As a result, we have introduced a bioinformatics

approach focusing on personalised AML patient care and identified putative novel metabolic indicators that could expand the existing panel of prognostic biomarkers.

In memory of my father

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my supervisors, Prof. Marta Cascante, Prof. Jean-Baptiste Cazier and Prof. Pedro de Atauri, for their invaluable help and their continuous support through every step of my Ph.D. studies. The trust they bestowed on me and their invaluable advice and guidance truly helped me grow.

My appreciation goes to Dr. Vitaly Selivanov for the interesting discussions and continuous consideration.

I would also like to thank Dr. Carles Foguet for his tutelage and encouragement. My colleagues Claudia Hernandez Carro, Inês Baptista and Miriam Contreras Mostazo for their experimental work supporting my study. It is them, together with Dr. Míriam Tarrado, Dr. Erika Zodda, Dr. Josep Tarragó, Dr. Silvia Marín, Dr. Ibrahim Polat and Dr. Roldan Cortez that warmly welcomed me and unconditionally supported me through some of the most challenging times; and for that I will always be grateful.

I would like to acknowledge George Papageorgakopoulos, Natalia Kotsikou and Dr. Stelios Karozis, my closest friends, for their patience, encouragement and undiminished belief in me.

I would like to extent my gratitude to my mother, who has been a true inspiration and my greatest supporter throughout all those years.

Last, I would like to acknowledge that this project has received funding from the European Union's Horizon 2020 programme under the Marie Skłodowska-Curie grant agreement HaemMetabolome.

TABLE OF CONTENTS

1.

INTRODUCTION

[2]

1 Introduction

1.1 Cell and Metabolism

Cells are the basic functional and structural units of all living organisms, as first proposed by Schleiden and Schwann in the late 1830s [1]. They consist of several different biomolecules, such as nucleic acids, proteins and metabolites [2].

In greater detail, nucleic acids are long polymeric chains made from monomers called nucleotides. Each nucleotide consists of a five-carbon sugar (deoxyribose or ribose), a phosphate group and a nitrogenous base [3]. Five nitrogenous bases form nucleotides and these are adenine (A), guanine (G), cytosine (C), thymine (T) and uracil (U). Deoxyribonucleic acid (DNA) is an homopolimer containing deoxyribose and A, G, C and T, whereas ribonucleic acid (RNA) is also an homopolimer but contains ribose and the A, G, C and U bases. DNA is the macromolecule that stores the genetic information of the cell in the form of nucleotide sequences, the genes. In eukaryotic cells, DNA is kept in the nucleus in a double stranded helical arrangement [4]. The genetic instructions imprinted in the genome need to be translated into proteins. They need to leave the nucleus and reach the ribosomes, the dedicated cellular machinery that decodes amino acid sequences and translates them into proteins. For this reason, genes are transcribed into ribonucleic acid (RNA) molecules, called messenger RNA (mRNA).

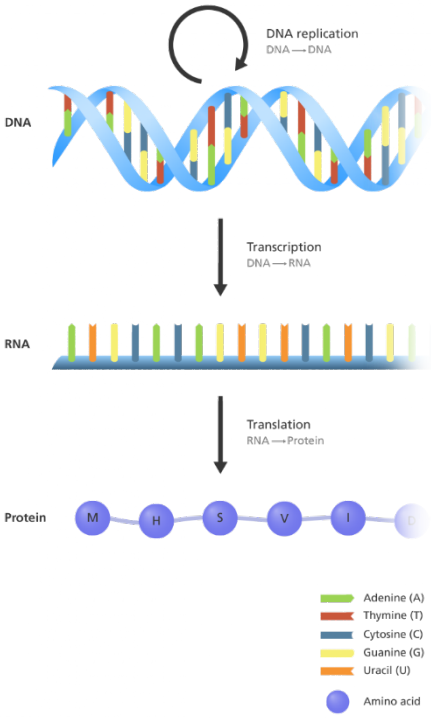


Figure 1.1: The central dogma of molecular biology was first formulated from Francis Crick in order to describe the flow of information from genetic information in the DNA to RNA and finally to protein.

[3]

These are single-stranded molecules, shorter that DNA and can carry one (or in some cases only a few) genes to the ribosomes for their translation into proteins. Francis Crick first formulated this flow of information from DNA to RNA to protein into the central dogma of molecular biology (Figure 1.1), which “deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred back from protein to either protein or nucleic acid” [5].

During translation from RNA to protein, the bases in the mRNA are read in groups of three, called the codons. All the possible combinations of the four bases that can be found in the sequence of an mRNA result in 64 codons. From these, 61 represent amino acids and three stand for stop codons, which terminate the translation process. Amino acids are the monomers that constitute proteins. Since only 20 amino acids make up the proteins in the human body, more than one of the 61 codons correspond to each amino acid [2]. Different types of proteins address the different needs of the cell. According to their function, proteins can broadly be categorised as structural proteins, enzymes, regulatory and transport proteins [6]. Briefly, structural proteins are involved in the shape and movement of a cell. Transport proteins are usually located in membranes and carry oxygen and other substances inside and out the cells and subcellular organelles or throughout the body, in the case of multicell organisms. Regulatory proteins regulate different activities, including the expression of genes. Finally, enzymes are biological catalysts and catalyse most of the biochemical reactions that constitute metabolism [6].

Metabolism has its origins in the Greek word “*μεταβολή*” (metavolí), which means change, reflecting on the interchange between proliferation and adaptation that constitutes life. It refers to all the biochemical reactions that take place inside the cells in order to sustain life. In principle, metabolic reactions permit energy to be collected, released, used for the biosynthesis of cellular components and then catabolized and recycled [7].

Cellular metabolism is carried out in a highly controlled sequential manner. Metabolic reactions are organised into pathways, which can be catabolic or anabolic. Complex nutrient molecules get broken down through catabolic pathways. Polysaccharides, proteins and nucleic acids are dissected to their composing elements, which are then taken up by anabolic pathways and used for the synthesis of new complex molecules. Catabolism is accompanied by the release of energy or the collection of it in in high energy phosphate bonds of phosphate nucleotides, mainly by phosphorylating adenosine diphosphate (ADP) into adenosine triphosphate (ATP), whereas anabolic pathways require the consumption of energy (e.g. ATP) [7].

Most of the biochemical reactions that make up the metabolic pathways occur with the help of enzymes, which lower the activation energy of the reaction. These molecules maintain an active site that can bind to their substrates, stabilise the transient state of the reaction and convert them to the end products of the particular reaction regulated by them [6]. Once the products are released from the enzyme, its active site is free to perform more catalytic cycles.

A distinctive trait of metabolism across organisms is the very wide range of time scales that different metabolic events require and therefore one approach for studying metabolic regulatory processes is grouping them according to the amount of time these require to cause the change [8]. The mechanisms that evolve during larger time scales (i.e. hours or days) affect the available amount of proteins through the rate of gene expression and the rate of protein degradation. Gene expression is regulated primarily at the first step of transcription. Proteins, called transcription factors (TFs), bind to specific DNA sequences adjacent to DNA gene coding regions. In this way, they activate or suppress the transcription of genes by helping or blocking RNA-polymerase, the dedicated transcription machinery of the cell, attach to DNA and initiate the production of an mRNA molecule [9]. On the other hand, protein degradation is mediated through the ubiquitin-proteasome pathway or lysosomal proteolysis. Proteins have varying half-lives to accommodate the specific needs of a cell depending on it function or external

stimuli [10]. Both gene expression and protein degradation can regulate metabolism by regulating the concentration of an enzyme at any given moment.

Metabolism-regulating mechanisms that unfold over medium time scales (i.e. seconds to minutes) involve the post-translational modification (PTM) of enzymes. PTM events, such as phosphorylation, acetylation and methylation, and their counterparts, dephosphorylation, deacetylation and demethylation, refer to the addition or removal of a functional group to a protein molecule that results in some alteration of its function. PTMs may affect the structural conformation and catalytic activity of enzymes or may serve as “labels” for the transportation of an enzyme to specific subcellular compartment [11], thus directly regulating enzyme catalysed reactions.

Finally, mechanisms that regulate metabolism at short time scales of seconds or even less refer to the non-covalent and reversible binding of regulatory molecules to the active site or an allosteric site (a site different than the active site) of enzymes [8]. Enzyme activators are small molecules that positively regulate the activity of an enzyme. Allosteric activators bind to allosteric sites of an enzyme and increase the activity of the active site. In some cases, the substrate of an enzyme can act as an activator molecule, in which case it binds to an allosteric site and enhances the activity of the active site. This process is called cooperativity [12]. On the contrary, enzyme inhibitors act by supressing or decreasing the enzymatic activity of a protein. Inhibitors can bind to the active site of an enzyme and therefore block the access to the substrates, in a process called competitive inhibition. In uncompetitive and non-competitive inhibitions, an inhibitor binds allosterically to an enzyme, in which case it changes the conformation of the enzyme to still allow binding of substrates to the active site, while losing the ability to stabilise the transient state of the reaction [13]. Finally, feedback inhibition refers to the inhibition of the activity of an enzyme by a product of the same pathway the enzyme-catalysed reaction belongs to. In this energy-saving process, high concentrations of a metabolite block its formation, when the cell has synthesised the necessary amount [10].

1.2 CANCER

The term cancer is used to describe a large ensemble of multifactorial diseases, characterised by the loss of physiological function and the uncontrollable proliferation of cells. Cancer can manifest in any part of the body and in some cases malignant cancer cells have the ability to invade other tissues and metastasize. According to the World Health Organisation (WHO), cancer is the second leading cause of death world-wide and the latest data revealed that more than 18 million of new cancer cases were diagnosed in 2018 and 9.6 million of

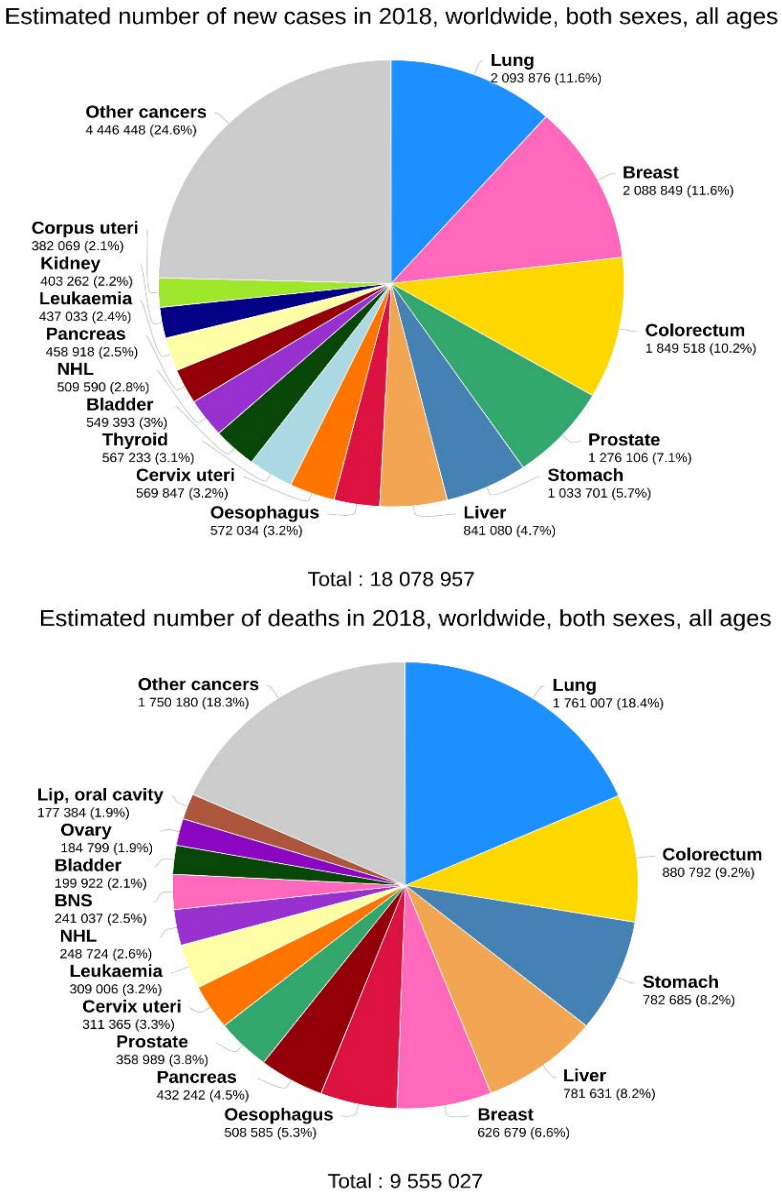


Figure 1.2: Estimated number of new cancer diagnoses and estimated number of cancer-caused deaths in 2018 worldwide, according to the World Health Organisation.

deaths were attributed to cancer (Figure 1.2). The most common types of cancer in men are lung, prostate and colon cancer, whereas in women breast, colorectal and lung cancer. Despite the great efforts towards understanding and combating cancer, the exact underlying mechanisms of this disease remain elusive and the development of more effective therapeutics is crucial.

1.2.1 HALLMARKS OF CANCER

The vast developments in cancer research have brought to light several distinctive features shared across different types of cancer cells, acquired upon their malignant transformation. In the beginning of the 21st century, Douglas Hanahan and Robert A. Weinberg produced an extensive review of such common traits, which resulted in the establishment of the hallmarks of cancer [14, 15]. Originally, six acquired capabilities of cancer cells were established as hallmarks, namely self-sufficiency in growth signals, insensitivity to anti-growth signals, evasion of apoptosis, limitless replicative potential, sustained angiogenesis and tissue invasion and metastasis [14]. A decade later, three more hallmarks of cancer cells were determined: tumor-promoting inflammation, evasion of immune destruction and metabolic reprogramming, together with “enabling characteristics” that facilitate the acquisition of such characteristics by cancer cells [15].

1.2.1.1 GENOMIC INSTABILITY

Genomic instability refers to the increased accumulation of genome alterations in cancer cells that contribute to cancer initiation and progression. These alterations may vary from gene mutations (e.g. insertions or deletions) to chromosomal alterations (e.g. translocations, copy number variations, inversions and others) [16]. Ultimately, genomic instability supports the activation or overexpression of oncogenes and the inactivation of tumour suppressor genes and genes connected to or responsible for DNA integrity and repair.

In more detail, several genes that code for proteins involved in the regulation of cell growth, division, proliferation or cell death (apoptosis) are called proto-oncogenes [17]. Mutations occurring in these genes can turn proto-oncogenes into oncogenes, that promote cancer development and progression. In principle, oncogenes over produce different proteins that stimulate cell division, increase cell proliferation, suppress cell differentiation and stop cell death. These traits are some of the hallmarks of cancer, as already described [14, 15, 18]. For example, the *ras* family gene members are activated as oncogenes due to point mutations of single substitutions that force the resulting oncogenic RAS proteins to be hyperactive and lead to uncontrollable proliferation [19]. Other proto-oncogenes are converted into oncogenes because of chromosomal translocations. One of the most well studied examples of such a case is the translocation of the *abl* gene from chromosome 9 to one end of chromosome 22, where the *bcr* gene is located, leading to a fused chromosome known as the Philadelphia chromosome [20]. The result of this translocation is a translatable fusion gene, whose product, the BCR-ABL protein, has high protein tyrosine kinase activity. The aberrant expression of BCR-ABL triggers other regulators of apoptosis, cell-division and the cell cycle towards enhancing proliferation and survival, most commonly found in myeloid and lymphocytic leukaemias [21].

Opposite to oncogenes, tumour suppressor genes normally control cell growth, replication, apoptosis and DNA repair. The inactivation or loss of tumour suppressor genes is another aspect of genomic instability contributing to sustaining cancer progression. The first characterisation of a tumour suppression gene was done in retinoblastoma studies, where chromosome 13 deletions containing the retinoblastoma (*rb*) gene led to the conclusion that the loss of it resulted in carcinogenesis [10]. Since then, the role of the *rb* gene as a master regulator of the cell cycle has been established, together with its crucial implications as a tumour suppressor gene affecting DNA repair and telomere maintenance, among others, in bladder, breast and lung cancers [22]. Another fundamentally important tumour suppressor gene, whose loss of function is caused by mutations or deletion (in most cases a combination of both in the two alleles) is

that of tumour protein 53 (*tp53*) gene across 50% of cancer types [10, 23]. *tp53* is referred to as the cell's gatekeeper. It can act as a transcription factor and it is a major regulator of cell cycle arrest, DNA repair, apoptosis induction in response to stress signals and oxidative stress and cell metabolism [24].

Apart from gene mutations and chromosomal alterations, another source of genomic instability in cancer comes from epigenetic alterations, most importantly DNA methylation and histone modifications. Hyper- and hypomethylation of specific regulatory regions of the target genes leads to the silencing of tumour suppressor genes and the upregulation of oncogenes expression, respectively [25]. In addition, the epigenetic histone modification results in alterations in the chromatin structure of DNA, thus allowing chromosomal instability and transcription regulation [26].

1.2.1.2 SIGNALLING PATHWAYS IN CANCER

The systematic study of genetic alterations in cancer, made possible by the rapid advances in high-throughput sequencing technologies, has revealed the most commonly affected cellular processes that sustain the acquisition of a cancer phenotype. A definite link between cancer progression and the deregulation of signalling pathways has been well established [27], while targeting these pathways is considered to hold great therapeutic promise [28].

1.2.1.2.1 PI3K/AKT PATHWAY

The overactivation of phosphoinositide-3-kinase (PI3K) signalling pathway is commonly observed across different cancers. PI3Ks are a family of plasma membrane-associated lipid kinases that respond to signals from various proteins and oncogenes, for example growth factors activate receptor tyrosine kinases (RTKs), cytokines, hormones and *ras [29]*. The activation of PI3K in turn activates AKT, a serine/threonine kinase, that has a great number of downstream effectors implicated in cell cycle regulation, cell proliferation, cell survival and metabolism, to name a few [30].

More specifically, Forkhead box O (FOXO) transcription factors are downstream effectors of AKT. The AKT mediated phosphorylation of FOXOs generates recognition motifs for the 14-3-3 family of phosphor-binding proteins, which transport FOXOs to the cytosol and prevents them from binding to DNA and initiating transcription. Therefore, the expression of genes regulated by FOXOs family members is supressed, genes implicated in apoptosis (e.g. FASL and TRAIL), cell cycle arrest (e.g. CDKN1a and CDKN1b), growth inhibition (e.g. Sestrin3, ATG12L and BNIP3L) and metabolism (e.g. G6PC and LPL) [30, 31].

Another downstream effector of AKT is the glycogen kinase 3 (GSK3). GSK3 binds to a variety of targets and “tags” them for proteasomal degradation. Phosphorylated forms of the induced myeloid leukaemia cell differentiation protein MCL-1, a member of the B-cell lymphoma 2 (BCL-2) apoptosis regulator family, and the oncogene c-MYC are targeted by GSK3 [30]. Additionally, GSK3 targets cellular metabolism by inhibiting metabolic enzymes, such as glycogen synthase (GS), and transcription factors, such as hypoxia inducible factor 1a (HIF1a) and nuclear factor erythroid 2–related factor 2 (NRF2) [32, 33]. Therefore, the overactivation of PI3K/AKT leads to the inactivation of GSK3 though AKT mediated phosphorylation and the subsequent stabilisation of antiapoptotic proteins, tumour promoting oncogenes and cell cycle regulators, helping in cancer cell proliferation and survival.

1.2.1.2.2 mTOR

The mechanistic target of rapamycin (mTOR) is a serine/threonine kinase. It constitutes of two structurally and functionally distinct protein complexes, the mTOR complex 1 (mTORC1) and the mTOR complex 2 (mTORC2). mTOR senses environmental and intracellular signals and regulates crucial cellular processes, such as cell survival, growth, autophagy, metabolism and cell homeostasis [34]. Forbes et al., 2011, reported that an overactivation of mTOR is observed in 70% of cancers and this can result from mutations in genes coding some components of the two complexes or by mutations affecting upstream regulators (promoters or suppressors) of mTOR [35].

Upstream regulators of mTORC1 are signalling pathways of mitogen activated protein kinase (MAPK) and PI3K/AKT, among others. The phosphorylation of AKT can activate mTORC1. However, the complete activation of AKT requires its phosphorylation at two residue sites, at the Thr308 by Phosphoinositide-dependent Kinase 1 (PDK1) and at Ser473 my mTORC2, creating a regulatory interdependency placing mTOR both upstream and downstream of AKT signalling.

Reported targets of mTORC1 are translation initiation factor 4E (eIF4E)-binding protein 1 (4EBP1) and S6 kinase 1 (S6K1), both of which regulate translation of proteins implicated in cancer progression, invasion and metastasis [36]. Moreover, mTORC1 plays a crucial role in metabolism through the activation of transcription factors sterol regulatory element-binding protein 1/2 (SREBP1/2), which regulate the expression of genes involved in lipid homeostasis [37, 38].

1.2.1.3 METABOLIC REPROGRAMMING IN CANCER

Metabolism plays a crucial role in most cellular processes and therefore the malignant transformation of cancer cells is accompanied by a rewiring of their metabolism, in order to support their increasing needs. Cancer cells are rapid proliferating cells and a shift in their metabolism is imperative to accommodate the expanded demand both in building blocks (for example lipids, amino acids and nucleotides) and required energy for growth signalling and DNA synthesis [39]. For this reason, overactivation of signalling pathways in cancer primarily promotes cell growth and proliferation. In addition, a one-carbon pool is highly required by cancer cells to be able to sustain epigenetic alterations, such as methylation and histone modification. Further contributions of metabolism in the cancer phenotype include the development of drug resistance, metastasis and immune escape [40-42]. Figure 1.3 illustrates the most commonly affected metabolic pathways in cancer (figure adapted from [43]).

1.2.1.3.1 GLYCOLYSIS AND THE WARBURG EFFECT

The glycolysis pathway metabolises glucose and other sugars into pyruvate in an oxygen independent manner. Through this process, energy is released in the form of ATP, together with other by-products that are precursors for macromolecule biosynthetic pathways [44]. Under aerobic conditions, pyruvate can enter the mitochondria and is oxidised through the oxygen dependent oxidative phosphorylation (OXPHOS), accompanied with a higher production of ATP and also reactive oxygen species (ROS). In the absence of oxygen, pyruvate remains in the cytoplasm, where it is reduced to lactate by lactate dehydrogenase (LDH-A) [45].

A first observation made by Otto Warburg and his colleagues 95 years ago revealed that cancer cells shift their bioenergetic metabolism from oxidative phosphorylation to glucose fermentation, even when oxygen is available, thus the term aerobic glycolysis [46, 47]. This phenomenon is known as the Warburg effect. Notably, the fermentation of one molecule of glucose to lactate can yield 2 molecules of ATP, whereas the oxidative phosphorylation of glucose in the

mitochondria can produce 31 molecules of ATP per molecule of glucose [48]. Warburg initially believed that defective mitochondria function was the reason for aerobic glycolysis in cancer cells, a theory that was later contradicted by evidence of intact mitochondria function in most types of cancer [49].

Provided that the Warburg effect can be observed in non-cancerous high proliferative cells (e.g. endothelial cells, macrophages and T-Helper lymphocytes) as well [50], it can be concluded that it offers significant advantage to cancer cells. A hypothesis as to why this is true focuses on the fact that the increased aerobic glycolysis can sustain the biosynthesis of cellular building blocks through glycolysis intermediates and the production of reductive power with the diversion of glycolytic flux through the oxidative branch of the pentose phosphate pathway [49, 51]. Another potential benefit of the Warburg effect to tumour cells is the acidification of the tumour microenvironment by the decrease of pH caused by the secretion of lactate. Evidence shows that this can contribute to the metastatic potential of cancer cells [52], as well as deprivation of glucose to the tumour infiltrating lymphocytes and subsequent inhibition of their immune response functions [53]. Finally, another hypothesized advantage of aerobic glycolysis over OXPHOS lies in the decreased production of ROS, which can in turn induce apoptotic signals

[54]. In this way, the Warburg effect supports cancer cells evade ROS mediated apoptosis [55].

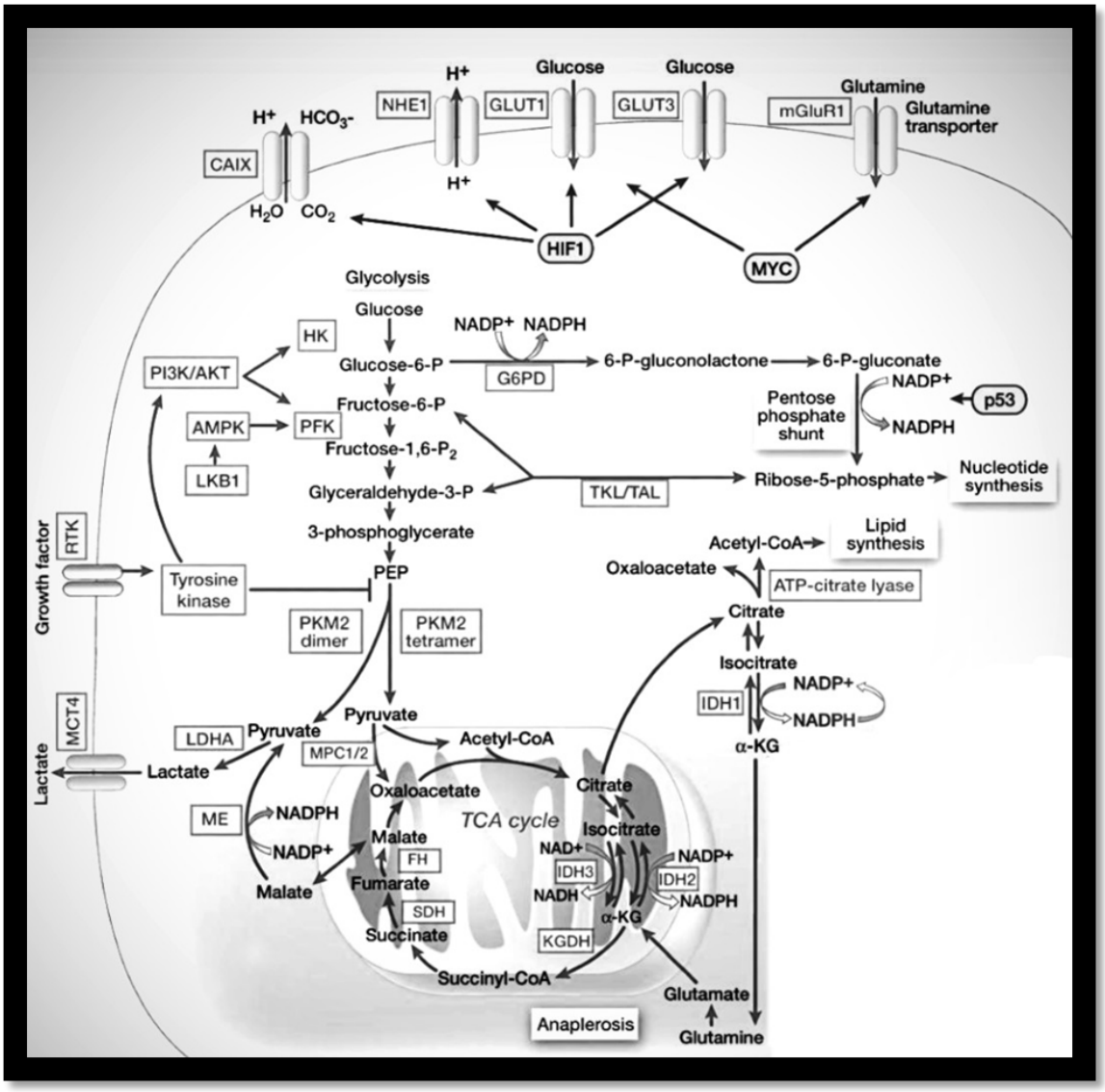


Figure 1.3: Metabolic reprogramming in cancer. HK, hexokinase; PFK, phosphofructokinase; G6PD, glucose 6-phosphate dehydrogenase; TKL, transketolase; TAL, transaldolase; PKM2, pyruvate kinase; LDHA, lactate dehydrogenase; ME, malic enzyme, MPC, pyruvate carboxylase, IDH2/3, isocitrate dhydrogenase, KGDH, α-ketoglutarate dehydrogenase, SDH; succinate dehydrogenase; FH, Fumarate hydratase

[15]

1.2.1.3.2 PENTOSE PHOSPHATE PATHWAY

The pentose phosphate pathway (PPP) plays a crucial role in the metabolism of cancer, since it contributes in the biosynthesis of ribose phosphate backbones for nucleotides and in the production of NADPH required for anabolism and redox homeostasis [56]. Hence, PPP is often upregulated in cancer, in order to support the increased anabolic needs of cancer cells and help counteract oxidative stress [57, 58].

The PPP consists of the oxidative branch and the non-oxidative branch. The irreversible oxidative branch converts glucose-6-phosphate, the intermediate product of the first step of glycolysis, into ribose-5-phosphate (R5P) and produces NADPH. The flux through this branch is highly dependent on the available concentrations of glucose-6-phosphate and the NADPH demands and is regulated in the first step, which is catalysed by glucose-6-phosphate dehydrogenase (G6PD) [59]. The non-oxidative branch of the PPP is reversible, and it interconverts

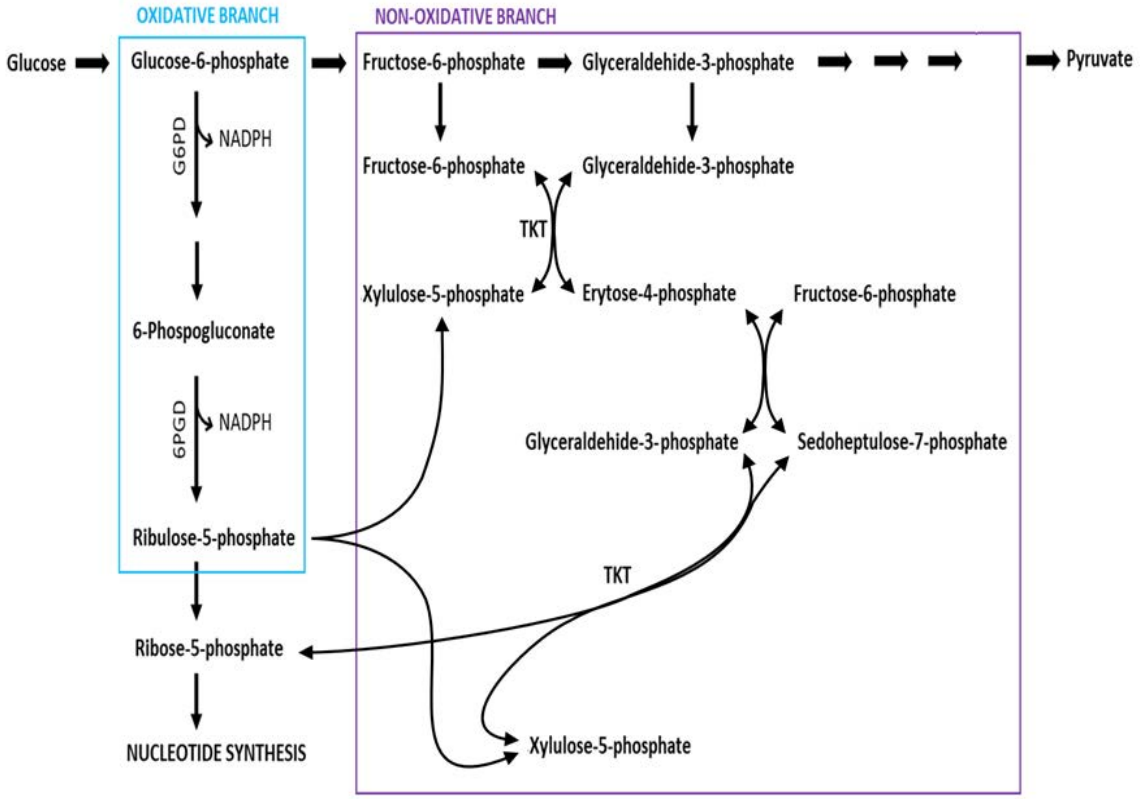


Figure 1.4: The Pentose Phosphate Pathway consists of two branches: the irreversible oxidative branch and the reversible non-oxidative branch.

R5P and intermediaries of the glycolysis pathway. The flux through the non-oxidative branch depends on the activity of the oxidative branch and the concentrations of the glycolytic intermediates and is regulated through transketolase (TKT) [60].

PI3K/AKT signalling can upregulate several members of both PPP branches (e.g. G6PD, TKT, Transaldolase) through NRF2 [61]. Additionally, G6PD has been shown to be regulated by epigenetic modifications , while tumour suppressor p53 (product of the *tp53* oncogene) has been shown to inhibit G6PD and consequently the PPP [62].

1.2.1.3.3 FOLATE METABOLISM

Folate metabolism and his role in one-carbon metabolism holds particular significance in *de novo* biosynthesis of nucleotides, epigenetic modification events and reductive metabolism. Hence, folate metabolism has a crucial role in the metabolic rewiring of cancer cells [63].

Folate, or vitamin B9, is transformed to tetrahydrofolate (THF) through two consecutive reactions catalysed by dihydrofolate reductase (DHFR) and can then serve as a carrier for one-carbon units (or methyl groups) [63]. The main donor for one-carbon units in cancer is serine, through its transformation into glycine by the cytosolic and mitochondrial serine hydroxymethyltransferases (SHMT1 and 2) [64]. It is suspected that two one-carbon pools, one devoted to the biosynthesis of thymidine monophosphate (dTMP) and the other to the synthesis of purines, are located in the cytosol and the mitochondria, respectively [65]. Hence, folate metabolism consists of two parallel branches unfolding in the cytosol and the mitochondria.

More specifically, one-carbon units intended for dTMP are products of the folate cycle in the cytosol, which involves SHMT1, thymidylate synthase (TYMS) and dihydrofolate (DHF) and THF. On the other hand, methyl groups intended for purine synthesis are first generated in the mitochondrial folate cycle involving SHMT2, 5,10-methenyl-THF-dehydrogenerase 2 (MTHFD2) and 10-formyl-THF

synthase (MTHFD1L). Through this process, a one-carbon unit is released in the form of formate, which is then transported to the cytosol where it forms the cytosolic 10-formyl-THF and gets directed to purine biosynthesis [66].

1.2.2 Acute myeloid leukaemia (AML)

In adults, haematopoiesis, the production of all cellular components that constitute the blood and the blood plasma, takes place in the bone marrow and starts with the production of Hematopoietic Stem Cells (HSCs). HSCs are multipotent precursors that produce progenitor cells for two lineages, the lymphoid and the myeloid [67]. Lymphoid lineage cells mature into T, B and natural killer cells, whereas myeloid lineage cells mature into erythrocytes, megakaryocytes,

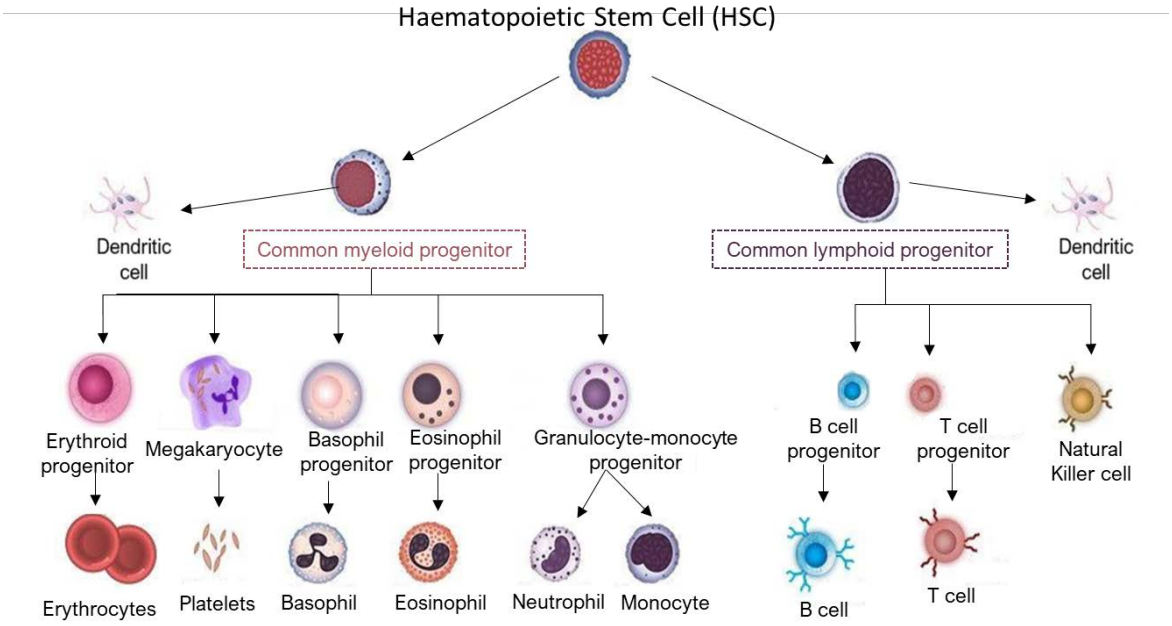


Figure 1.5: Hematopoietic stem cells are pluripotent precursor cells that produce the myeloid lineage and the lymphoid lineage blood cells.

granulocytes and macrophages. Dendritic cells can originate from both lymphoid and myeloid progenitor cells [68].

Haematological malignancies refer to cancers of blood cells. In most of these cancers, normal haematopoiesis is interrupted by the uncontrolled proliferation of a malignant type of blood cell. Leukaemia is the blood cancer characterised by the abnormal proliferation of white blood cells. Even though leukaemia was originally considered as a single disease, nowadays different subtypes have been identified. First, depending on the cell lineage affected, leukaemias can be

distinguished between lymphoblastic or lymphocytic and myelogenous or myeloid. Then, considering the rate of the disease progression, leukaemias are further categorised as chronic (slow progression) and acute (rapid progression).

Acute myeloid leukaemia (AML) is a result of various genetic abnormalities that disrupt normal growth and differentiation of the immature myeloid precursors or myeloblasts to mature white blood cells. This, in turn, leads to the accumulation of the immature myeloblasts in peripheral blood and the bone marrow, which are no longer able to produce healthy blood cells [69]. The clonal expansion of abnormal immature leukemic blasts which accumulate in the bone marrow leads to the clinical entity of AML.

AML is the most common type of leukaemia in adults and accounts for 80% of leukaemia cases diagnosed in adults worldwide [70]. Even though AML can occur as a result of treatments (e.g. radiation or chemotherapy treatment), it is primarily diagnosed as a *de novo* malignancy in previously healthy patients [70]. According to the American Society of Clinical Oncology, the 5-year survival rate for AML patients over 20 years old is estimated around 25%. Overall, AML is considered a rapidly progressing and aggressive malignancy with poor prognosis. Still, many aspects of the pathophysiology of the disease are unknown and there is a great need for the design of new and efficient therapeutic strategies.

1.3 Omic data in Systems Biology

Systems biology has its roots in the 1920s when Ludwig von Bertalanffy formulated his “systems theory of life”, according to which living organisms should be viewed as a whole, rather than broken down to their individual constituents and studied through the reductionist approach [71]. According to von Bertalanffy’s general systems theory, two principles characterize systems, hierarchical organization and complex links that hold their structural components together [71, 72]. It wasn’t until the beginning of the 21st century that this strategy was embraced in molecular biology. Systems biology has become recognized as the multidisciplinary field that combines biology, biochemistry, mathematics and computer science to study organisms as integrated systems of interconnected genetic, protein and metabolic components [73, 74]. Great advancements in

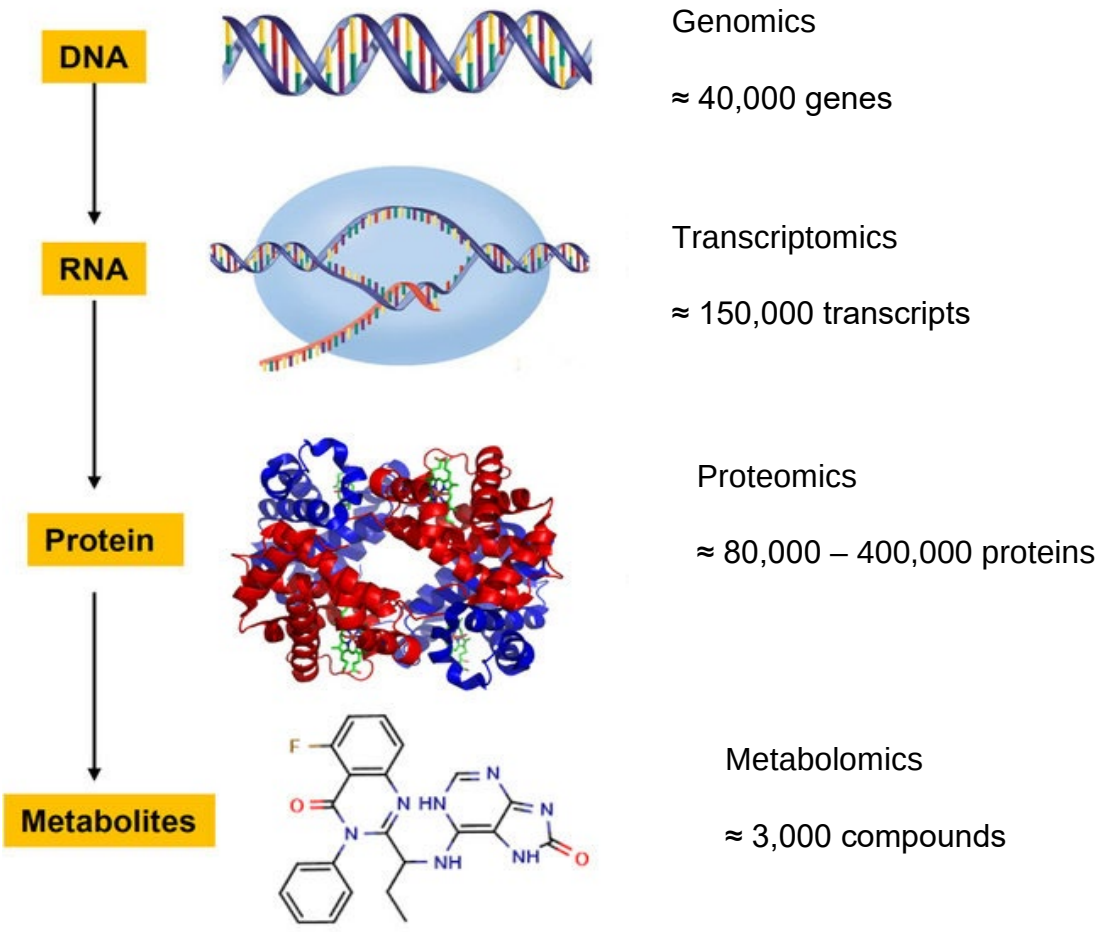


Figure 1.6: High-throughput -omic data provide massively parallel sequencing and quantification of all the levels of biological hierarchical organisation.

systems biology have coincided and were driven by the emergence of high-throughput technologies that permit the massively parallel sequencing and quantification of biomolecules, for example RNA transcripts, proteins and metabolites [75]. These high-throughput data, known as the -omic data, correspond to the different layers of biological organization (Figure 1.6).

1.3.1 TRANSCRIPTOMICS

The central dogma of Biology describes how genetic information stored in DNA is transcribed into messenger RNA (mRNA) and then translated into protein [76]. Transcriptomics refer to the study of the first step of gene expression, during which specific DNA sequences are transcribed into RNA molecules with the help of enzymes called RNA polymerases. Transcriptomic technologies aim at capturing the transcriptome, the entire set of RNA transcripts produced genome-wide in a cell under specific conditions, using high-throughput methods, such as microarray analysis and more recently RNA sequencing (RNA-seq). Nowadays, the addition of transcriptomic assays as the first step in system biology approaches is considered particularly informative, as it can provide a well-defined overview of the levels of gene expression, as well as information on the different transcript isoforms that occur in different cell conditions [77]. In cancer studies for example, transcriptomics help elucidate the functional role of genes, while also identifying common gene expression patterns between different types of malignancies[78]. Furthermore, transcriptomics can play a major role not only in the prediction of novel therapeutic strategies, but also in the investigation of the mechanisms underlying drug resistance.

1.3.1.1 DNA MICROARRAYS

After the first publication of the human genome by the Human Genome Project in 2001 [79], DNA microarrays covering the whole genome started to emerge in an effort to simultaneously monitor gene expression at a systems levels. The main idea behind microarrays is the use of the quantitative analysis of mRNA molecules present in a sample to estimate gene expression levels [80].

In greater detail, a DNA microarray constitutes of ordered DNA spots positioned on a glass slide. The spots contain short DNA sequences that could be fragments of a specific gene or of any other region of the genome of potential interest. The spots on the arrays are commonly referred to as probes and the molecules they contain serve as complementary sequences for the appropriately transformed mRNA molecules to bind. So, total RNA is extracted from cells and the mRNA molecules are isolated and reverse transcribed into complementary DNA (cDNA) molecules, referred to as the target. The nucleotides used to form the cDNA molecules are fluorescently labelled. Once this process is complete, the labelled cDNA is left to hybridize on to a DNA microarray and therefore the target molecules are allowed to interact and bind to the DNA fragments in the probes. Any cDNA molecules that do not get attached to a complementary sequence get washed away. Finally, the hybridized spots are excited with a laser and the array is scanned, so that image analysis software can measure the amount of fluorescent dye emitted due to the excitation [80].

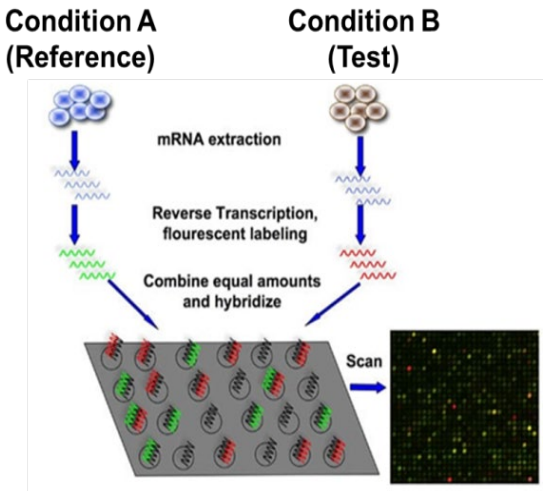


Figure 1.7: The workflow of a DNA microarray experiment for a differential expression analysis study includes the isolation of mRNA in the samples, the reverse transcription of the molecules to cDNA, the labelling with fluorescent dyes and finally their hybridization on the same array.

In the case of differential expression studies, the samples are grown under different conditions or stresses and then the mRNA is collected and processed as previously described. In order to compare gene expression between two different conditions, the condition of interest (condition A) and the reference state (condition B), two different fluorescent dyes are used, for example red dye for the test condition A and green dye for the reference condition B (Figure 1.7). The differentially labelled samples are then left to hybridize on to the same DNA microarray, so that the cDNA fragments present in the samples will bind to the spots on the array that contain

their complementary sequence. The amount of cDNA bound to a spot will be directly proportional to the initial number of RNA molecules present for that gene in both samples. Following the hybridization step, the spots in the hybridized microarray are excited by a laser and scanned at suitable wavelengths to detect the red and green dyes. The total signal intensity emitted upon excitation corresponds to the amount of cDNA bound to each spot of the array. If we consider that the expression of a gene was greater in condition A than in condition B, then the amount of nucleic acid from condition A bound to the corresponding probe will be greater than that from condition B and therefore the spot will be green. In the opposite case, the spot would appear red. If a gene was equally expressed in both conditions, the spot would be yellow, and if there was no gene expression in any of the two conditions, then the corresponding spot would be black.

The analysis of DNA microarray data can result in the characterization of the functional roles of genes in different cells, processes or diseases. The expression levels of large numbers of genes can be measured simultaneously and therefore their differential expression in various cell types, developmental stages and in response to some disease or treatment can be studied. Moreover, DNA microarrays can be used to identify the interaction between genes and gene products. Other applications of DNA microarrays include the detection of fusion transcripts, single nucleotide polymorphisms (SNPs) among alleles within or between populations and alternative splicing. Finally, the study of the evolution of gene regulation in different species has become possible with the use of microarrays.

1.3.1.2 RNA SEQUENCING

A more recently developed method to perform transcriptomic studies in a high-throughput manner is RNA Sequencing (RNA-seq). While microarrays are limited in detecting mRNA molecules whose sequence is already known, RNA-seq performs direct sequencing of the transcripts and therefore can be employed in the detection not only of known but also of novel sequences, allowing the realisation of discovery-based studies. In addition, it can be used in the *de novo* assembly of genomes for organisms and subsequently allowing for differential expression analysis even in cases where no genomic information was previously available [81, 82].

The first step in RNA-seq is the isolation of the total RNA present in a sample. High quality of isolated RNA is of crucial importance for the upcoming steps of the analysis. Next, a sequencing library is constructed. Towards that end, the total RNA isolated from the sample needs to be cleared from contaminating DNA and ribosomal RNA (rRNA) molecules. Next, the remaining RNA is fragmented to an appropriate size for sequencing. The RNA molecules get reverse-transcribed to single strands of cDNA, which then become double stranded molecules with the use of DNA polymerase. Subsequently, adaptors consisted of sequences necessary for the hybridization of the molecules to a flow cell are ligated to the 5' and/or 3' ends of the cDNA and the prepared library gets sequenced. Finally, the great number of short reads generated need to be aligned to a reference genome or transcriptome, so that they can be identified as exonic reads, junction reads and poly(A) end-reads [83]. Overall, the levels of gene expression are directly correlated with the number of reads that get mapped to a specific genomic region.

RNA-seq shows significant advantages compared to DNA microarrays [82]. Besides the analysis and quantification of the RNA molecules present in a sample, RNA-seq can also be employed to detect alternative splicing, post-transcriptional modifications, gene mutations and single nucleotide polymorphisms (SNPs), gene fusion and finally differentially expressed genes across time or between different

conditions. Moreover, RNA-seq can be used in the study of microRNA (miRNA), transfer RNA (tRNA) and long non-coding RNA (lncRNA) molecules.

1.3.1.3 COMPUTATIONAL ANALYSIS METHODS AND CHALLENGES

The ever-increasing volume of –omic data becoming available at extremely fast paces needs to be appropriately analysed so that an unbiased and accurate quantification of changes in gene expression can be achieved. The most important step in any transcriptomic data analysis workflow is the pre-processing and the normalization of the raw data.

In general, normalization is a term that is used to describe the process of eliminating systematic variations to allow appropriate comparison of data obtained from different samples. Due to the complexity and the several different steps taken towards the completion of transcriptomic experiments, a great number of sources of systematic variation could affect the measurements of gene expression levels. However, it is crucial to preserve the biological variance that occurs naturally due to changes at the transcriptional level. Microarrays and RNA-seq datasets are generated by fundamentally different methods and therefore the requirements in computational processing are different.

Overall, the most commonly used computational tools that perform normalization of raw transcriptomic data make two common assumptions, in regards to differential expression analysis. Firstly, only a small number of genes will be differentially expressed and secondly, it is assumed that the number of up-regulated genes will be approximately the same as the number of down-regulated genes.

1.3.1.3.1 DNA MICROARRAYS

In microarrays, the most common metric to translate the light intensities measured after the excitation of the red and green fluorescent dyes into gene expression levels would be the expression ratio. More specifically, the expression

ratio T_i is equal to the spot intensity metric for the test sample R_i divided by the spot intensity metric for the reference sample G_i:

$$T_i = \frac{R_i}{G_i},$$

where *i* is the *i*th probe on the array.

The use of the expression ratio is a rather direct method to illustrate the differences in the abundance of the mRNA levels in the samples. However, it lacks the ability to efficiently express up- and down-regulation, because down-regulation gets restricted between 0 and 1, whereas up-regulation can take any value from 1 to infinity. In order to resolve this issue, a logarithm transformation is applied on the values of the expression ratios and more specifically the logarithm with base 2 is used. In this way, the expression ratios are mapped in a continuous space and up- and down-regulation are mapped uniformly around zero.

A method to determine the quality of the raw data and therefore decide upon normalization is to visualize the intensities of red and green dies through a plot of the distribution of the intensity ratio (M) against the average intensity (A). However, in order to detect possible outliers or artifacts in the data, each array is compared to a pseudo-array that is composed by the median intensity of each probe across all arrays. So now, this more robust MA-plot shows the difference between the intensity of a probe on an array and the median intensity of that probe over all the arrays (M) plotted against the average of the median intensity of the probe on that array and of the median intensity of that probe across all the arrays (A). M and A follow the equations:

$$\mathrm{M} = \log_2(R/G) = \log_2(R) - \log_2(G)$$
$$A = \frac{1}{2}\log_2(R \cdot G) = \frac{1}{2}\left[\log_2(R) + \log_2(G)\right]$$

We assume that the majority of the genes are not differentially expressed and that the number of up and down regulated genes is equal. Therefore, one would expect the data to be equally distributed around M=0 (Figure 1.8).

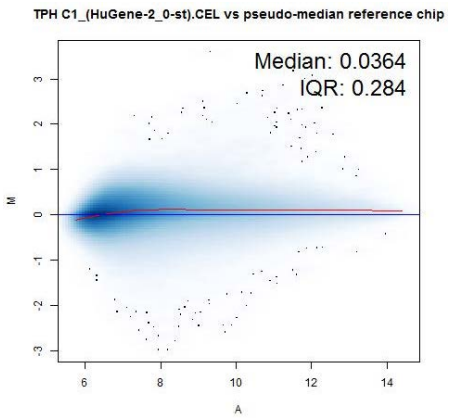


Figure 1.8: An MA plot is a visual representation of the dispersion of the raw data. It can help detect outliers in the dataset and assess the need for normalization of the data. The median value shows the central tendency, whereas the interquartile range (IQR) is a measure of the dispersion of the data.

Next, in the normalization process of transcriptomic microarray data, it is of vital importance to perform within array normalization in order to account for systematic biases present due to inadequate labelling between the two fluorescent dyes, unequal amount of mRNA originally extracted from the samples, as well as technical biases related to the quality of the DNA probes and the printing of the spots on the array. In this manner, true differences in gene expression can be detected.

The most commonly used approaches are the global normalisation methods that correct for location or scale bias within each array, thus adjusting around zero the center and spread of the distribution of log-ratios, respectively. The basic assumption of the location normalization methods is that the red and green intensities are correlated. If:

$R = k \cdot G$, then

$$\boldsymbol{log_2}(\boldsymbol{R}/\boldsymbol{G}) \rightarrow \boldsymbol{log_2}(\boldsymbol{R}/\boldsymbol{G}) - \mathbf{c} = \boldsymbol{log_2}\left(\boldsymbol{R}/\boldsymbol{kG}\right),$$

where $\mathbf{c} = \boldsymbol{log_2}(\boldsymbol{k})$, is the location parameter that most commonly is equal to the mean or the median of the log-intensity ratios[84].

In addition, the location normalization equation could be adjusted in order to account for spatial bias caused by the different print-tip groups in the array and for possible systematic differences and/or disfigurement among the print-tips [85]:

[28]

$$\boldsymbol{log_2}(\boldsymbol{R}/\boldsymbol{G}) \rightarrow \boldsymbol{log_2}(\boldsymbol{R}/\boldsymbol{G}) - \boldsymbol{c_i}(\mathbf{A}) = \boldsymbol{log_2}\left(\boldsymbol{R}/\boldsymbol{k_i(A)G}\right),$$

where $\boldsymbol{c_i}(\mathbf{A})$, represents the normalization parameter which is a function of the overall spot intensity A for the *i*th group of print-tips for *i*= 1,…,I and I is the total number of print-tip groups on the array.

Finally, the scale normalisation methods assume that the log ratios follow a normal distribution centred around zero and a variance equal to $\boldsymbol{a_i}^2\ \boldsymbol{\sigma}^2$, where $\boldsymbol{\sigma}^2$ is the variance of the true log-ratios and $\boldsymbol{a_i}^2$ is the scale factor of the *i*th element on the array. The scaling factors can be calculated using the median absolute deviation (MAD) [85]:

$$\boldsymbol{MAD_i} = \boldsymbol{median_j}\{|\boldsymbol{M_{ij}} - \boldsymbol{median_j}(\boldsymbol{M_{ij}})|\},$$

where *Mij* denotes the *j*th log ratio in the *i*th print-tip group and

$$\widehat{\boldsymbol{a}}_{\boldsymbol{i}} = \frac{\boldsymbol{MAD_i}}{\sqrt[I]{\prod_{i=1}^{I} \boldsymbol{MAD_i}}}.$$

After all the log ratios have been normalised within each array, the log-ratios of the different slides need to be adjusted appropriately, so that comparisons among them can be performed. The aforementioned scaling normalisation method MAD can be applied for between-array normalisation.

1.3.1.3.1.1 RNA Sequencing

Since the RNA-seq technology is completely different than that of microarrays, the sources of bias that could be introduced to the data and therefore the normalisation process that needs to be applied differs substantially. In detail, once the raw reads are generated by the sequencing platform, they need to be mapped to a reference genome or transcriptome database using an appropriate alignment software tool like STAR and TopHat [86, 87]. The mRNA abundance is considered to be directly proportional the total number of short reads that gets mapped to a specific gene. However, the sequencing depth and coverage can significantly affect the accuracy of the results. The sequence (or read) depth reflects how many times a specific nucleotide has been read in an experiment, whereas coverage

refers to the relative length of the assembled sequence over the length of the reference [88]. The fragmented nucleic molecules that get sequenced individually need to be then reassembled computationally to reproduce the reference sequence. During this process, some parts of the reads will be overlapping, thus the read count of bases of the same gene might differ (Figure 1.9). Therefore, the calculated expression levels of one gene might be affected by the expression of the rest of the genes in the experiment. In addition, the is a need to account for the library size bias introduced by the fact that a higher count will be assigned to the longer transcripts compared to the shorter fragments and consequently longer transcripts tend to be overrepresented in the dataset [89, 90].

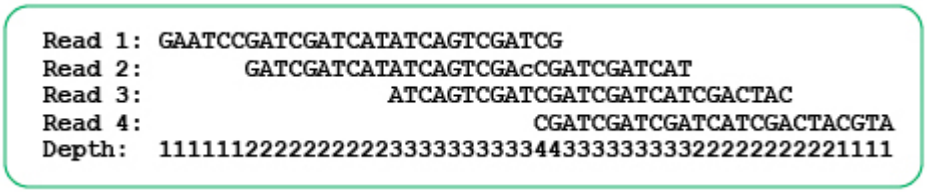


Figure 1.9: The overlapping of the reads during the alignment step results in nucleotides belonging to the same gene having different sequence depths.

Several tools have been developed in order to correct biases in RNA-seq data. First, the library size-based methods make use of the sequence library length to perform normalisation on the samples. The Total Count (TC) method uses the total library count to calculate a normalisation factor:

$$\boldsymbol{TC = \mu(total\ counts\ across\ the\ dataset) \times \frac{gene\ counts}{total\ counts\ in\ a\ lane}}.$$

On the other hand, the Upper Quartile (UQ) method uses the upper quartile of non-zero counts to perform the normalisation:

$$\boldsymbol{UQ = \mu(UQ\ of\ counts\ different\ from\ 0) \times \frac{gene\ counts}{total\ counts\ in\ a\ lane}}.$$

Finally, the Median Method (Med) accounts for outliers in the datasets by using the median of counts for each replicate:

$$\boldsymbol{Med = \mu(median\ counts\ different\ from\ 0) \times \frac{gene\ counts}{total\ counts\ in\ a\ lane}}.$$

Although these methods perform a rather direct correction of the raw data, they only account for the library size and make the assumption that the counts from different samples follow a similar distribution. Most frequently, this approach results in an inclination towards down-regulation [91]. Consequently, scaling factor-based methods were developed.

One of the most widely used tools in the processing of RNA-seq data, the DESeq Bioconductor package [92, 93], has implemented a scaling factor based correction. An mRNA ratio is calculated for each gene as the read count of the gene divided by its geometric mean across all samples. Then, it is considered that most genes will not be differentially expressed and therefore their expression ratio will be equal to 1. When this assumption is not true, then a scaling factor is applied to the data:

$$\widehat{s_j} = \boldsymbol{median}_i \frac{k_{ij}}{\left(\prod_{v=1}^{m} k_{iv}\right)^{1/m}},$$

where *i=1,...,n* indexes the genes, *j=1,…m* the samples, and k_{ij} is the *nxm* table of counts.

Moreover, a different scaling based method is implemented in the equally popular edgeR package [94]. The Trimmed Mean of M-values (TMM) estimates relative abundances of groups of genes across samples and scales the datasets accordingly.

Finally, methods to also account for bias introduced by the experimental design are widely implemented in the normalisation process of RNA-seq data. On one hand, the Reads Per Kilobase per Million reads (RPKM) and the Fragments Per Kilobase of transcript per Million fragments mapped (FPKM), try to calculate a score relative to the size of each transcript and in order to achieve a comparison at a transcript level. In principle, in both approaches, a scaling factor is calculated as the total reads in a sample divided by 10^{-6}. Then, the read counts in a sample are divided by the scaling factor and the result is the length-normalised reads or fragments per million (RPM or FPM) value. Finally, each RPM is divided by the

length of the gene in kilobases, which is equal to the normalised RPKM or FPKM value. The only difference between the RPKM and FPKM methods is that the latter was built for paired-end sequencing experiments, in which case two reads can map to the same fragment of the reference genome or transcriptome. On the other hand, Transcripts per Million (TPM) is an approach that corrects for the transcript length bias, but also for the length and the different number of transcripts aiming in estimating the relative molar concentration of RNA present in the samples. More specifically, the read counts are divided by the length of each gene in kilobases so that the reads per kilobase (RPK) values are determined. Next, the scaling factor is calculated by dividing the sum of all RPKs by 10^{-6} . Finally, the TPM normalised values are calculated by dividing RPK value by the calculated scaling factor [94].

1.3.1.4 Need for Biological Interpretation

Nowadays, there is a greater struggle and need to develop analyses to fully exploit, interpret and assign biological meaning to the data, than actually obtaining them. The most common approach to handling the results of the differential expression analysis would be a marginal examination of the gene list and the selection of a number of genes of interest based on the calculation of their statistical significance and fold change of expression between two or more conditions. Then, several different methods are being applied based on gene ontology and pathway enrichment, like DAVID and Reactome [95-98].This process could be very straightforward and effective in providing a small number of putative genes that play an important role in a specific study. However, there can be some drawbacks in selectively considering only the top up- or down-regulated genes. Firstly, there might be no biological correlation between the genes making biological interpretation of the results extremely challenging and laborious. Every gene needs to be manually examined, which implies limitations in the number of potential candidates and makes the accuracy and correctness of the outcome reliant on the ambiguously selected threshold and the expertise of the researcher

conducting the analysis. Finally, by considering only major changes in the expression of individual genes, one could fail to detect changes in entire pathways.

1.3.1.4.1 Gene Set Enrichment Analysis (GSEA)

A different approach developed to deal with the limitations presented in the marginal consideration of the differential expression analysis results is called aggregate score approach. It is based on the use of groups of genes that represent a specific biological motif in order to calculate an aggregate score for each gene that was expressed in the conducted experiment [99]. The most widely used tool implementing the aforementioned technique is the Gene Set Enrichment Analysis (GSEA) developed by Subramanian et al. [100].

GSEA analyses a ranked list of genes provided by a genome-wide transcriptomics experiment and aims in examining the distribution of the members of a predefined gene set across this experimental list. More specifically, the genes in the list provided by a differential expression analysis get ranked based on their fold change in expression so that the up-regulated genes appear for example at the top of the list and the down-regulated genes at the bottom. Since there is no specific selection process, all the genes are included in the ranked list including those that exhibited small changes between the compared conditions. Next, appropriate sets comprising genes that share a common biological theme are constructed. The sets could include all the protein coding genes of a metabolic pathway, genes that share the same molecular function or are involved in the same biological process. In this way, well established biological knowledge can be implemented in the interpretation of the results. Finally, the distribution of the members of the predefined gene set among the experimental list and calculates whether they tend to cluster more towards the top or the bottom of the ranked list and therefore correlates the observed changes in expression with specific phenotypic profiles.

1.3.2 METABOLOMICS

Metabolomics refer to the characterisation and quantification of all or a large number of all the metabolites present in a cell, tissue or organism and the study of the chemical reactions that occur among them [101]. It is the analytical examination of the unique chemical fingerprint of cellular processes under a specific physiological condition and can provide an evaluation of the phenotypic state of a system [102]. Metabolomic studies find a plethora of applications, in toxicology [103], monitoring adverse effects of drugs or other environmental stresses [104], cancer research [105], nutrition [106], to name a few. Depending on the scope of scientific inquest, two metabolomic approaches can be followed: the untargeted and the targeted.

Untargeted metabolomics, or discovery metabolomics, aim in the exhaustive and unbiased identification and analysis of all compounds present in a sample, including unknowns. Thus, this approach holds particular advantage in exploratory and hypothesis generating studies. Untargeted studies provide a relative quantification of analytes that exhibit a statistically significant change in abundance across control and tests samples [107]. Large datasets are generated that first need to be filtered using advanced chemometrics and then annotated using existing metabolic repositories or analytical chemistry techniques [108, 109].

On the other hand, targeted metabolomics are employed for the quantitative measurement of a predefined group of known metabolites in a sample [102]. The steps followed in protocols for targeted metabolomic assays can be tightly controlled, in order to improve the detection efficiency of the target metabolites in the samples [110]. Also, measuring the abundance of well characterised analytes results in the elimination of artifacts and biases in the subsequent analysis. Targeted metabolomic assays can be applied for the quantification of metabolites at a large scale, therefore satisfying the needs of systems biology studies.

The two main analytical techniques used in metabolomics for the qualitative and/or quantitative measurement of metabolites in samples are mass spectroscopy (MS) and nuclear magnetic resonance (NMR). In MS analyses,

multiple ions are generated from the individual compounds present in a sample using electron ionisation. Product ions undergo fragmentation and then get separated according to their mass-to-charge ratio (m/z). After the ions are converted into electrical signals, their abundance can be measured. MS provides the mass spectrum of a molecule and the ion abundance is plotted against the mass-to-charge ratio [111]. Analytical approaches based on MS have been developed and applied in metabolomics, such as gas chromatography-mass spectrometry (GC-MS) [112], gas chromatography-time-of-flight mass spectroscopy (GC-TOF) [113], capillary electrophoresis-mass spectrometry (CE-MS) and liquid chromatography-mass spectrometry (LC-MS) [114]. On the other hand, NMR relies on the fact that the nuclei of the molecules in a sample have an electric charge and most of them have spin. The sample is placed in a magnetic field and the nuclei are excited by radio waves of specific frequencies. The frequencies of the energy emitted while the spin of the nuclei return to their base levels are measured and appropriately processed to provide the NMR spectra of the analytes present in a sample [115, 116]. Finally, bearing in mind the high heterogeneity of chemical and physical properties of the large number of metabolites that exist in a cell, a combination of analytical techniques can provide better detection and identification coverage [117, 118].

1.4 Metabolic modelling at genome scale

The greatest challenge in our -omics era is to achieve the optimal correlation and integration of the different -omic data representing the various layers of biological organisation and to define and interpret the “genotype-phenotype” relationship [119, 120]. Contrary to what was originally thought, connecting gene expression to translation to protein and ultimately to a specific phenotypic outcome is a complex and non-linear process [120]. The role of metabolism in the growth, survival and proliferation of living organisms, as well as in the pathogenesis of several different diseases, such as cancer, has been firmly established. What is more, metabolomics provide a snapshot of the physiology of the system that is the closest possible to its phenotype and therefore should be a crucial component in systems data integration studies [121-124].

Mathematical representations of the metabolism, called metabolic models, are gaining increasing use in Systems Biology, since they provide an appropriate mathematical platform that allows the integration of different sets of data, such as metabolomics, transcriptomics, proteomics and others. They can be applied for the simulation of different aspects of the metabolic state of a system, including the simulation of its metabolic potential or its response to a perturbation, its dynamic behaviour over time of a limited number of pathways by including enzyme kinetics or the steady state of the whole set of biochemical reactions in all pathways by applying various constrains [125, 126]. Hence, metabolic models can be broadly categorised into kinetic and constraint-based models.

1.4.1 The stoichiometric matrix

If we perceive metabolic models as networks of interacting components, then immediately the principles of graph theory can be applied in their formulation and interpretation [127]. The biochemical reactions that take place among the various metabolites in the system can be represented as stoichiometric equations. The stoichiometry of these equations holds the information about the quantitative relationship between the various substrates and products of the biochemical

reactions. In this way, the mathematical interpretation of the entire network of interconnected elements in metabolic networks is the stoichiometric matrix, which is the combination of the whole system of stoichiometric equations in the network [74].

The stoichiometric matrix *S* will have *m* number of rows, equal to the number of metabolites in the network, and *r* number of columns, equal to the number of biochemical reactions. Now, let us consider a vector v_r of fluxes, i.e. the rate of production or consumption of metabolites per unit of area per unit of time, and a vector x_m of the derivatives of metabolite concentrations over time. Then, the stoichiometric matrix is the linear transformation of the flux vector to the vector of concentration derivatives and can be concisely written as follows:

$$\frac{dx}{dt} = Sv \quad (1)$$

Equation (1) is also known as the mass balance equation. The five metabolites A, B, C, D and E in Figure 1.10 participate in four reactions, two of which are reversible and therefore six fluxes can be determined in this network. Note that in the case where a metabolite does not participate in a reaction, its stoichiometric coefficient is equal to 0.

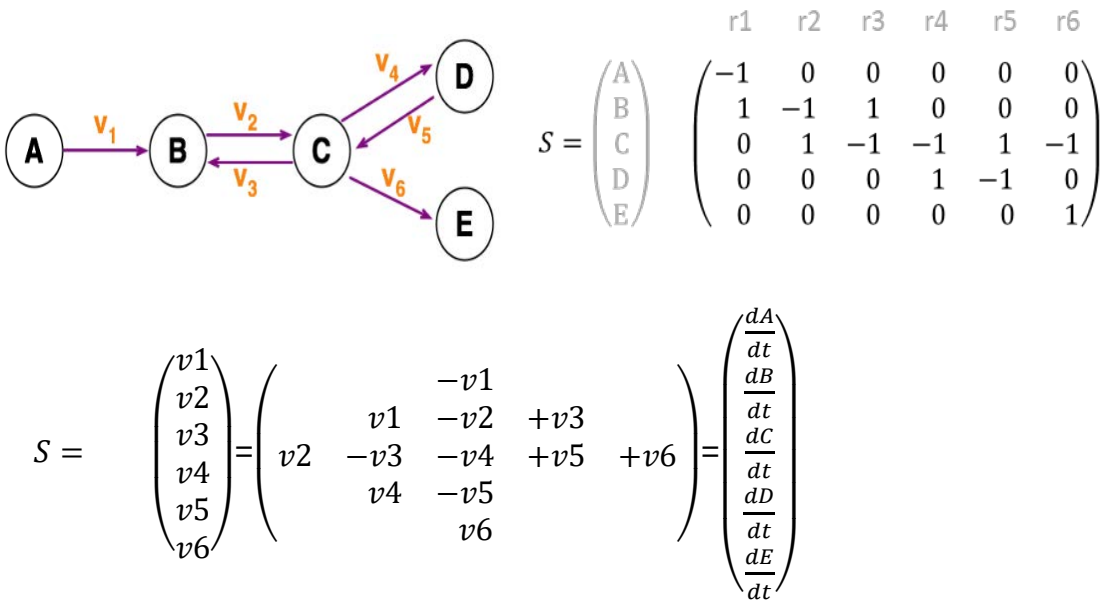


Figure 1.10: Example of the stochiometric matrix S and the balance equations for a toy metabolic network of five metabolites and the reactions among them.

1.4.2 KINETIC MODELS

Kinetic models are dynamic models of the metabolism. Therefore, reaction fluxes are represented as functions of the change in concentrations of the metabolites and the enzymes catalysing the different biochemical reactions that take place. As a result, non-linear systems of ordinary differential equations (ODEs) are formed in order to dynamically represent a metabolic network and they incorporate the initial metabolite concentrations, reaction rate equations and kinetic parameters [128].

The complexity and the high degree of parameterisation that is inherent in kinetic models were initially dealt with by simplifying the mathematical expressions used to describe the problem. According to the Law of Mass Action (LMA), the rate of any chemical reaction is proportional to the concentrations of the reactants, with the concentration of each of the reactants raised to a power equal to the corresponding stoichiometric coefficient in that chemical reaction. Based on this principle, several methods have been introduced in an attempt to address the issues of the fitting of various parameters in the system, as well as the non-linear mathematical representation of the rate laws. The application of the Generalised Mass Action (GMA) and the S-system models resulted in several significant studies [129], the description of the sphingolipid metabolism in yeast [130] and of genetic circuit networks in E. Coli [131] being among others. In addition, the log-lin and lin-log kinetic models implemented linear logarithmic approximations, enhancing the computational power of the models [132-134]. However, these approaches have been limited to specific small-scale networks and in some cases sacrificed the thermodynamics of the reactions.

On the other hand, mechanistic models that focused on the mechanism of action for each enzyme and were built on the mass conservation and thermodynamic laws of each reaction were introduced to dynamic modelling. Kinetic models of human red blood cells [135, 136] and of the central carbon metabolism of model organisms *Saccharomyces cerevisiae* [137, 138] and E. Coli [139] were built but met certain limitations. The number of parameters that need to be estimated was immense, the modelling was again restricted to specific parts [39]

of the metabolic network, while at the same time enzyme kinetic data calculated in vitro were used to predict enzymatic activity in vivo.

The expansion of kinetic models from pathway specific to large, even genome, scale requires both stoichiometry and thermodynamics of the system to be satisfied and the integration of kinetic and regulatory information needs to be integrated to the reconstructions [128]. However, specific obstacles still hinder the completion of larger scale kinetic models: extensive lack of regulatory data, especially for human models, and the complexity of fitting the large number of parameters in the reconstruction.

1.4.3 Constraint-based models

As we established in the previous section, kinetic models can be widely applied to study the dynamic behaviour of the metabolism of organisms, however they are limited to modelling a certain part of metabolism.

Constraint-based models adopt a characteristic feature of biological systems: their tendency towards a metabolic steady state. Living organisms consume nutrients from the environment and excrete waste products while maintaining homeostasis, which is greatly dependent on the flux of matter and energy through metabolic pathways. This process is referred to as dynamic equilibrium and gives rise to the steady state assumption applied to constraint-based models. The concentrations of internal metabolites remain constant, since the rate of production has been perfectly balanced to the rate of their degradation. The supply of input fluxes from the environment is constant, as is the excretion (or output) fluxes and no metabolites can accumulate or be depleted [140]. By applying a metabolic pseudo-steady state, constraint-based models rely on the stoichiometry of the system to produce steady-state flux distributions [141]. The models can be further constrained by imposing upper and lower bounds to fluxes, in order to account for thermodynamic rules, availability or lack of external nutrients, specific physiological conditions [142, 143].

If we apply a steady-state approximation to equation (1), then the vector of concentration derivatives over time is equal to 0 and the formulation of the problem now is:

$$Sv = 0, \qquad lb \leq v \leq ub \quad (2)$$

where S is the stoichiometric matrix of the metabolic network, v is the vector of metabolic fluxes and lb and ub are the lower and upper bounds imposed to fluxes, respectively.

Now, the originally formulated system of differential equations has been reduced to a system of linear equations that depends strictly on the stoichiometry of the system and its flux values and requires no information of initial metabolite concentrations, enzyme kinetics or reaction rates. The system, with the appropriate constraints, can be well defined and solved using linear programming methods. Therefore, constraint-based models have been employed in the simulation and studying of large-scale metabolic networks, even at the genome scale.

1.4.3.1 Genome Scale Metabolic Models (GSMMs)

The wealth of available genomic data and the functional annotation of entire genomes for many organisms has provided the necessary information for the reconstruction of Genome Scale Metabolic Models (GSMMs). All the instructions for the formation of all the necessary components of metabolism, e.g. the enzymes that catalyse metabolic reactions, are imprinted in the genome of an organism in the form of genes. Therefore, the sequencing and annotation of a genome uncovers the entire metabolic potential for an organism, meaning all the metabolic reactions and consequently all the metabolic pathways that can occur in that organism [144]. Therefore, a GSMM is a network representation of the entire collection of metabolic reactions, together with the gene-reaction associations, for an organism [142].

More specifically, the core of a GSMM is the whole collection of metabolic reactions and transport processes and their stoichiometry, which is represented with the stoichiometric matrix. Information on subcellular compartments and the localisation of reactions and metabolites appropriately can also be included in the model. In addition, whether a reaction is reversible or irreversible, as well as the normal ranges for the participating metabolite concentrations are also implemented in a GSMM. Finally, and very importantly, GSMMs include gene-protein reaction (GPR) associations [145]. Such rules indicate whether a functional protein is part of a protein complex, if an enzyme catalyses one or multiple reactions and if one reaction can be catalysed be multiple enzymes. GPR associations are vital in the resulting accuracy of a GSMM, offer the basis for transcriptomic or proteomic data integration and greatly affect the outcome of gene knockout simulation studies [145, 146].

As additional information, not of structural importance, GSMMs may include various information on reaction names and their formulas, the pathways a reaction may participate in, names of enzymes and metabolites, and others.

1.4.3.1.1 Genome Scale Metabolic reconstruction

The first GSMM of *Haemophilus influenzae* Rd was constructed in 1999 [147]. Nowadays, 183 GSMMs have been manually curated to study the metabolism of various different systems [146], such as important model organisms of *Escherichia coli [148]* and *Saccharomyces cerevisiae* [149], bacteria [150] and plants [151]. The first human constraint-based GSMM reconstruction was of the erythrocyte [152]. Since then the accumulating data and technological progress have allowed the reconstruction of the entire human metabolic network and its constant refinement. Recon 1 [153] was first released and was the basis for Recon2 [154] and the Human Metabolic Reaction database (HMR) [155]. Then, Recon2.2 was released [156], as a refinement to Recon2. Finally, Recon3D [157] was built based on Recon2.2 and HMR.

As described very methodically by Thiele & Pallson, 2010, the reconstruction of GSMMs is an elaborate bottom-up process of 96 steps [145]. A prerequisite for starting the reconstruction of an organism's GSMM is a well annotated genome, which provides enough information on metabolic genes, enzymes and metabolic reactions catalysed by them. A draft reconstruction can be built based on these preliminary data, that needs to be curated using available biochemical databases, generated experimental data and literature sources.

Some key points in the curation of the draft reconstruction include the manual refinement of pathways to ensure that all reported reactions are included and to identify and add any missing reactions. In addition, the protonation state of the metabolites should be validated for each metabolite, since usually biochemical databases list metabolites with their uncharged formula. Then, the stoichiometry for the reactions can be determined and the addition of protons and water molecules where appropriate can be performed. In this way, the mass and charge of every reaction will be balanced and therefore the uncounted production of protons and ATP molecules will be prevented. Reaction directionality should also be determined, and this is possible by estimating the standard Gibbs free energy of formation ($\Delta_f G'°$) and of reaction ($\Delta_r G'°$) or consulting the literature for information on reaction directionality. Equally important is the curation of gene and reaction localisation information, which can be found in the literature or predicted based on the sequence of proteins using appropriate databases. Spontaneous reactions should be added, as well as intracellular transport reactions, in the case of networks with multiple compartments.

Finally, it is important to refer to the curation of the biomass reaction. This is a reaction created to describe the amount of energy and metabolites required by the target organism for normal growth and proliferation. In general, the formulation of the biomass reaction depends on the relative abundance of macromolecules that make up the system to be modelled and consequently by the fraction of metabolites that constitute those macromolecules and their reaction pathways [158]. Hence, the biomass reaction is the sum of all necessary metabolite

precursor fractions for the production of 1 gram dry weight of cells [145]. This task is particularly challenging and equally essential for the quality of the resulting GSMM reconstruction. It can differ among organisms, cell types or specific conditions and therefore should be appropriately adjusted to the characteristics and requirements of the modelled system. Unfortunately, it might not always be possible to accurately measure the exact biomass composition of interest and it might be necessary to rely on estimations derived from the genome and appropriate databases.

1.4.3.2 PREDICTING METABOLIC FLUX DISTRIBUTIONS AT A GENOME SCALE

In constraint-based metabolic modelling, the quality of a reconstructed GSMM and the validity of the defined biomass reaction play a crucial role in the outcome of the simulations. However, in most cases, even after the steady-state constraint has been imposed, there is a number of possible feasible solutions to the system and therefore additional steps need to be taken to identify a unique steady-state flux distribution [74]. For this reason, several approaches have been developed with the purpose to reduce the space of possible solutions and identify the one that best satisfies all the imposed requirements. Flux Balance Analysis (FBA) and Flux Variability Analysis (FVA) are two of these methods and will be described subsequently.

Moreover, not all metabolic reactions included in a GSMM reconstruction for an organism can occur under specific conditions or by a specific cell type. Therefore, the system can be further constrained and consequently the solution space further restricted by integrating experimental metabolomic and gene expression data [142]. Tools used for such integrations will also be presented.

1.4.3.2.1 FLUX BALANCE ANALYSIS (FBA)

The most widely used constraint-based modelling application is Flux Balance Analysis (FBA) [159]. The principle behind FBA is biological optimisation, the notion that living organisms tend towards efficiency, minimising the cost by optimising a specific necessary process [160]. Hence, FBA seeks a steady-state flux distribution, on the edge of the possible solution space, that maximises or minimises a specific objective function. Depending on the system being modelled or the nature of the scientific question at hand, different objective functions can be defined. In the case of high proliferative cancer cells, the goal is usually the maximisation of growth and proliferation, therefore the objective function is the biomass yield (Figure 1.11) [143].

If we define an objective function as follows:

$$f(v) = \sum_{i=1}^{r} c_i v_i \quad , \quad (3)$$

where c_i represent the weights for the individual flux rates v_i, then the optimisation problem can be formulated as follows:

$$FBA_{opt} = max(\boldsymbol{c^T v}) \ , \quad (4)$$

given that $Sv = 0, lb \leq v \leq ub$. FBA_{opt} is the optimal solution of the FBA objective and c^T is the transpose vector of maximisation or minimisation weights for the reaction fluxes.

One limitation of FBA is the appropriate selection of fluxes for optimisation, given the complexity of particular cellular processes or responses to perturbations. Most importantly, even though FBA seeks for a unique solution, it is possible that a number of alternative optimum solutions exists and this might be due to redundant pathways in the network or alternative pathways not accounted for in the objective function formation [74].

1.4.3.2.2 Flux Variability Analysis (FVA)

Flux Variability Analysis (FVA) works under the scope of FBA and seeks the minimum and maximum range of flux for the possible reactions in the network [161]. The double optimisation problem at hand can be formulated as follows:

$$vmax_i^{FVA} = maxv_i \text{ and } vmin_i^{FVA} = minv_i, \quad (5) , (6)$$

given that $Sv = 0, lb \leq v \leq ub \; and \; v \cdot c \geq FBA_{opt} \cdot (1 - T^{FBA})$, where $vmax_i^{FVA}$ and $vmin_i^{FVA}$ are the maximum and minimum feasible flux values for flux i, respectively, and T^{FBA} is the deviation threshold from the FBA_{opt} (see equation 4).

FVA has found applications in the investigation of alternative optimum solutions in GSMMs [162], challenging network flexibility and pathway redundancy [163], metabolic engineering [164-166], to mention a few.

1.4.3.2.3 Integration of gene expression data

The integration of gene expression data can further constrain a GSMM reconstruction to better simulate the system at a specific condition. As we already mentioned in section 1.4.3.1.1, GPR associations provide the appropriate

framework for the integration of such data and therefore the implementation of valuable information on the expression and availability of specific enzymes and consequently of the metabolic regulation that occurs under a specific condition [167].

1.4.3.2.3.1 Gene Inactivity Moderated by Metabolism and Expression (GIMME)

Gene Inactivity Moderated by Metabolism and Expression (GIMME) [168] (similarly the later introduced Gene Inactivation Moderated by Metabolism, Metabolomics and Expression (GIM^3E) [169]) in an algorithm developed for the integration of transcriptomic data to GSMM reconstructions and building condition-specific GSMMs. It imposes a penalty on reactions catalysed by lowly expressed enzymes, that can be formulated as follows:

$$w_i = max(ge^{th} - ge_i, 0), \qquad (7)$$

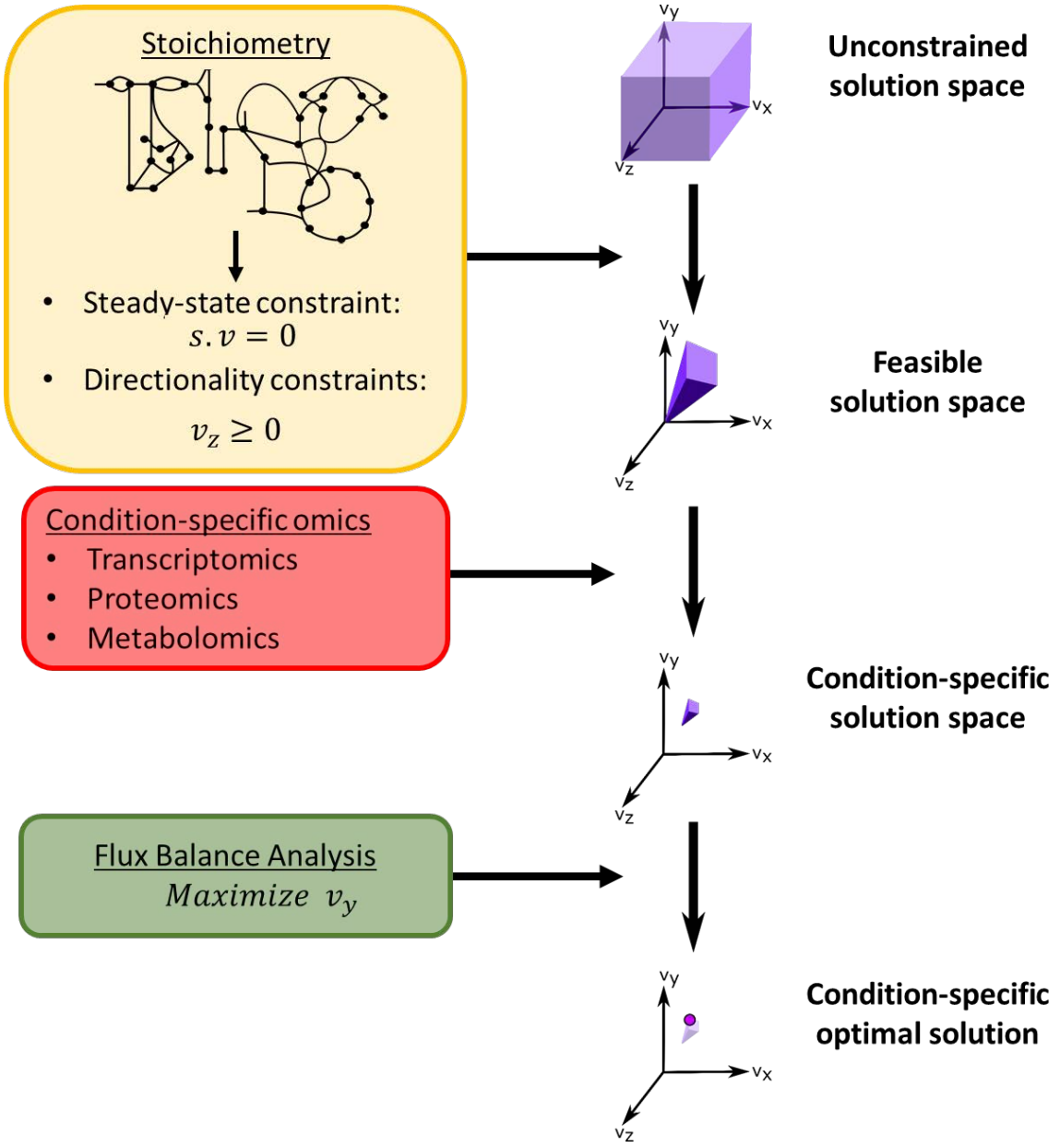


Figure 1.11: Constraint-based modelling and condition specific metabolic models. In constraint-based modelling, a steady-state approximation ($Sv = 0$) and directionality rules are imposed to the infinite space of possible solutions to obtain a constraint space of feasible flux distributions. Additional integration of -omic data, such as transcriptomics, proteomics and metabolomics, will lead to further constraining the feasible solution space to a condition-specific solution space. Finally, Flux Balance Analysis can be applied to identify optimum solution(s) that both satisfy imposed constrains and optimise a target phenotype.

where ge_i is the gene expression of gene i and ge^{th} is a threshold set to indicate the gene expression value below which the penalty would be imposed to the reaction. In addition, the algorithm converts the system into an irreversible, by splitting forward and reverse reactions [169], and now the formulation of the new linear problem to solve is as follows:

$$GIMME_{opt} = min \sum_i w_i v_i^{irr} \quad , \qquad (8)$$

given that $$S^{irr} v^{irr} = 0, \quad lb^{irr} \leq v^{irr} \leq ub^{irr} \; and \qquad (9)$$

$$v^{irr} \cdot c^{irr} \geq FBA_{opt} \cdot (1 - T^{FBA}) \qquad (10),$$

where $GIMME_{opt}$ is the optimal value of the GIMME objective, S^{irr} is transformed stoichiometric matrix of the irreversible system, v^{irr} the vector of fluxes of the irreversible reactions, c^{irr} their objective coefficients and lb^{irr} and ub^{irr} the lower and upper bounds of the irreversible reactions, respectively.

1.4.3.2.4 Genome Scale Metabolic Models in drug discovery

Constraint-based analysis of GSMMs provides the opportunity of reconstructing condition specific genome-scale models to simulate different types of cells and most importantly different types of pathological states, for example cancer. The condition specific GSMMs can be interrogated towards their metabolic and phenotypic response against different perturbations. Therefore, this approach in metabolic modelling has found a wide range of applications from exploration studies to discover a function of an enzyme, to bioengineering and to the identification of putative metabolic vulnerabilities in pathogens [146, 170].

In cancer research, GSMMs are gaining increasing interest as valuable tools in the investigation of potential druggable metabolic targets with the prospects of identifying novel therapeutic targets, combination therapies or appropriate repurposing of already approved and used (in cancer or other pathologies) drugs.

The identification of such vulnerabilities in cancer can be viewed as the identification of genes whose knock out would severely compromise the viability of the cells under investigation and under specific conditions. These genes are

referred to as essential genes and they can be further explored as therapeutic targets aiming to kill cancer cells or they can provide a wide characterisation of tumour metabolic dependencies . Towards that purpose, project DRIVE (deep RNAi interrogation of viability effects in cancer) the effects of the systematic knock down of 7,837 genes was assessed in 398 cancer cell lines [171].

In the framework of GSMMs and constraint-based modelling, the systematic interrogation of a biological system towards metabolic vulnerabilities can be achieved computationally and at a genome-scale. The GPR rules embedded in the reconstruction of GSMMs can be employed to identify which reactions will be rendered inactive due to the knock out of a gene and constraint-based modelling tools can be applied to assess the consequences to the biomass yield. Such efforts have already provided promising results in hepatocellular carcinoma and other cancer cell lines [172, 173].

1.4.3.2.4.1 Simulating gene knock outs with Flux Balance Analysis

FBA can be applied to systematically simulate gene knock outs in genome scale metabolic reconstructions. The flux through the reactions predicted to be inactive due to the loss of function of a gene is constraint to zero and the mathematical formulation of the problem can be as follows:

$$FBA_{opt}^{KO} = max(\boldsymbol{c^T v}) \ , \qquad (11)$$

given that: $v_i = 0$, $i \in KO^{reactions}$ and $Sv = 0, lb \leq v \leq ub$. FBA_{opt}^{KO} is the optimal value of the objective function after the effect of a gene knock out. A threshold can be set for FBA_{opt}^{KO} to determine the value below which a gene will be considered as essential.

When FBA is used to simulate gene knock outs in GSMMs, we must assume that the system under investigation will rewire its metabolism after each gene knock out and FBA will look for the optimal flux distribution after the adjustment.

1.4.3.2.4.2 Simulating gene knock outs with Minimisation of Metabolic Adjustment (MOMA)

Minimisation of metabolic adjustment (MOMA) is a method developed to simulate knock outs and predict the flux distribution of the system after the effect of the gene knock out that is optimal and at the same time closest to the optimal flux distribution before the perturbation [174]. In greater detail, MOMA is based on the assumption that the system will perform a readjustment of its metabolic fluxes but with the minimum possible alterations and will remain as close as possible to the original flux distribution of the system before the gene knock out. The problem that MOMA intends to solve is a quadratic function and can be formulated as follows:

$$min \sum_i (v_i^{wt} - v_i)^2 \ , \quad (12)$$

given that $v_i = 0$, $i \in KO^{reactions}$ and $Sv = 0, lb \leq v \leq ub$. Now, v_i^{wt} is the steady-state flux distribution of the unperturbed system. This flux distribution can be referred to as reference flux distribution and both FBA and GIMME can be used for its calculation. Overall, MOMA, as a minimum, will result in the identification of the same solutions as FBA and some additional essential genes, highly dependent on the reference distribution.

1.5 Machine Learning in Systems Biology

As highlighted in previous sections, the -omics era has brought a grand amount of biological data rich in insightful information on complex biological processes. With the continuous advancements in omic technologies the generation of large-scale datasets is becoming increasingly more efficient and both time and cost-effective [175]. For these reasons, biological data are now “big data” and they can be either structured or unstructured, characterised by great variability and the inability to be analysed by traditional and long standing statistical analysis methods [176-178]. Current needs in the analysis of -omic data include and are not limited to data mining, pattern recognition, data inference, classification and reconstruction of predictive models, all with the ultimate goal of extracting meaningful and coherent biological knowledge from such complex datasets [179].

Machine (and deep) learning (ML) methods are significantly gaining ground in omic data processing, analysis, exploration and interpretation over the last years [180, 181]. ML refers to a branch of artificial intelligence that focuses on the development of models that can learn from and be trained on an appropriate processable input so that they can increase their prediction accuracy [182]. ML algorithms have found numerous applications in systems biology and omic data processing, for example in pattern recognition in DNA or RNA sequences to predict protein binding sites [183], prediction of protein structure [184], reconstruction of phylogenetic trees [185], imputation of missing SNPs and DNA methylation states [186, 187]. The integration of constraint-based models with machine learning algorithms is fairly new and limited to only a few studies so far [182]. Focusing on studies of human metabolism, Shaked et al., 2016 coupled constraint-based models with machine learning to predict drug side effects and achieved a higher accuracy than baseline methods [188]. Li et al., 2010, reconstructed condition-specific GSMMs of NCI-60 cancer cell lines and applied ML to identify drug targets [189].

ML algorithms can be broadly categorised into supervised and unsupervised. In supervised learning, part of the data set, the training set, is used to train the

model towards formulating a function that correctly maps an input to a specific output (label). The ultimate goal is the generalisation of this function to correctly predict the label of newly introduced inputs (outside of the training set). Supervised learning problems can be further categorised into classification and regression problems. The former deals with problems where the output labels are categorical (e.g. malignant vs benign), whereas the later handles problems where the output are continuous values (e.g. pathogenicity risk score). Widely used supervised learning algorithms are Support Vector Machines (SVNs), Artificial Neural Networks (ANNs) and Random Forests (RFs) that can perform both classification and regression tasks, K-Nearest Neighbours and Naïve Bayes for classifications and Simple Linear Regression and LASSO regression more specific to regression problems [190].

Unsupervised ML, the algorithm deals with unlabelled inputs, input data for which the desired outcome is not known and tries to identify any underlying structure or patterns in the input data. Unsupervised learning problems can be further categorised into clustering and association. In clustering problems, the ML algorithm groups data based on inherent characteristics, whereas in association problems the algorithm seeks for latent rules or trends that apply to a large portion of the data. The most commonly used clustering ML algorithms are K-means and hierarchical clustering [191].

1.5.1 Feature selection in Machine Learning

A great challenge posed by -omic data, and big data in general, that ML algorithms are called to face is the so called “curse of dimensionality”. More specifically, the dimensionality of a dataset refers to the number of features that are attributed to each observation/sample. So, high dimensional datasets are those characterised by a number of features in the order of magnitude of thousands or even more. In the case of biological datasets, most commonly the number of biological markers quantified (e.g. single-nucleotide polymorphisms, gene expression measurements) greatly exceeds the number of samples (e.g.

patients) [192]. High dimensionality causes data sparsity, multicollinearity and distance concentration. As a consequence ML models are unable to accurately classify data or identify underlying patterns, suffer from overfitting (“overtraining”), so that ultimately they have a skewed performance when presented with new data [193]. Despite the rapid advancements in -omic technologies, the possibility of datasets with comparable numbers of samples and quantified features seems rather unlikely. Therefore, feature selection methods are often used to prepare the data for further analysis with ML algorithms.

1.5.1.1 SUPERVISED SPARSE CANONICAL CORRELATION ANALYSIS (SSCCA)

Canonical Correlation Analysis (CCA) is a multivariate statistical method, first introduced by Harold Hotelling in 1935 [194]. CCA tries to identify and measure linear correlations between variables of two datasets. Canonical variables or latent variables in statistics refer to variables that can not be directly observed and in latent mathematical models refer to the hidden underlying traits that exist in datasets. Witten et al., 2009, proposed a penalised CCA as an extension to the classic method introducing sparsity, in order to allow its application to high-throughput biological datasets, where the number of features is much higher than the number of available samples [195]. For two datasets X_1 and X_2 of dimensions $n \times m_1$ and $n \times m_2$ and consisting of measurements quantified for the same n samples, then the penalised CCA seeks to maximise:

$$max\ w_1^T X_1^T X_2 w_2, \quad (13)$$

subject to $\|w_1\|^2 \leq 1$, $\|w_2\|^2 \leq 1$, $P_1(w_1) \leq c_1$, $P_2(w_2) \leq c_2$, where P_1 and P_2 are convex penalty functions imposed on the vectors of the canonical weights w_1 and w_2, c_1 and c_2 depend on the size of the dataset and it is assumed that $1 \leq c_1 \leq \sqrt{m_1}$ and $1 \leq c_2 \leq \sqrt{m_2}$. CCA and sparse CCA can be characterised as unsupervised methods, since they search for maximal correlation patterns without any knowledge over the input data. In the case of existing data labels, then it might be interest to identify highly correlated features between two datasets that are

additionally associated with a specific outcome (or phenotype) [195]. If y is a vector of quantitative outcome labels, then the supervised sparse CCA can be formulated as follows:

$max\ w_1^T X_1^T X_2 w_2,$

subject to $\|w_1\|^2 \leq 1$, $\|w_2\|^2 \leq 1$, $P_1(w_1) \leq c_1$, $P_2(w_2) \leq c_2$ and

$w_{1j} = 0\ \forall\ j\ \notin\ Q_1\ and\ w_{2j} = 0\ \forall\ j\ \notin\ Q_2,$

where Q_1 is the set of features of X_1 that are maximally correlated with y and Q_2 is the set of features of X_2 that are maximally correlated with the outcome y.

1.5.2 Random Forests (RFs)

Random forests (RFs) is an ensemble supervised learning method that can be applied to classification and regression problems. RFs consist of a large number of uncorrelated decision trees that work together and "vote" the class, an input sample belongs to, or the mean/average outcome, in the case of regression applications. RFs benefit from the high accuracy of decision trees, while at the same time correct for their inherent disadvantage of overfitting, especially in the case of high dimensional data sets. The formulation of RFs started in 1996, with the independent works of Tin Kam Ho and Amit & German [196-198]. The basis of their proposals was the reconstruction of a classifier consisting of multiple trees and different implementations of reconstructing each of the trees based on a randomly selected subspace of the features. Finally, Leo Breiman introduced the notion of out-of-bag estimates for a randomised selection of features at each node of the individual decision trees, further improving the performance of RFs. Breiman also demonstrated how the error rate of the RFs depends on the accuracy of the individual trees in the forest and by the measure of correlation between any two trees in the forest [199]. The training of each individual tree of the forest on a random subset of the feature space, with the parallel randomisation of feature selection for each node of each tree to build decision rules reduces significantly

both the variance and overfitting of the RF model. Additionally, it allows RFs to handle efficiently any large number of features in high dimensional datasets.

1.5.2.1 DECISION TREES

Decision trees can be considered as binary tests organised hierarchically and therefore their visualisation resembles a tree, made from nodes n (Figure 1.12). The root node (input point to the tree) and the internal nodes enclose binary tests that split them into child nodes and decide the direction of the data. Finally, the leaf

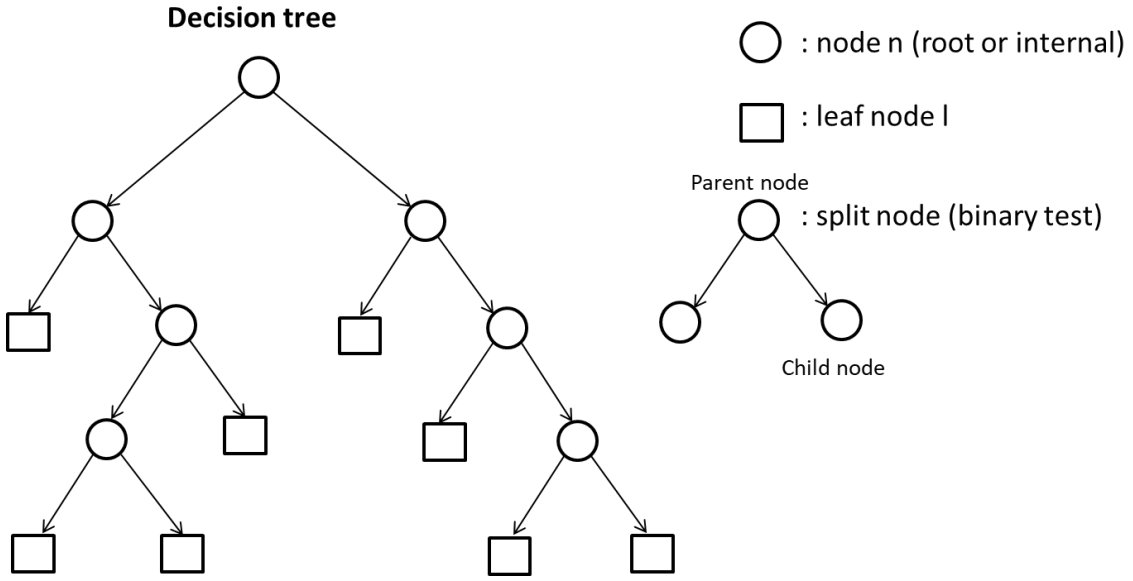


Figure 1.12: Example of a decision tree and its elements. Decission trees are hierarchically organised sets of binary rules (split nodes) that analyse the input variables to reach to a decision (leaf node).

node is the decision provided by the tree and therefore cannot be further split into children nodes.

If we consider a vector of samples **x** and its set of features F and its labels **l**, then the binary rules φ for the root and internal nodes n can be formulated as follows:

$$\varphi_n == \begin{cases} 0\,, F_{j_n} < t_n \\ 1\,, F_{j_n} \geq t_n \end{cases}, \quad (14)$$

Where F_{j_n} is the jth element in node n and t_n a threshold for node n. Based on this formulation, an input variable from **x** enters the tree at the root node (n=0) and it is directed towards the right child node (φ_n=0) or towards the left child node (φ_n=1), progressively, until it reaches the leaf nodes [200]. At the leaf node a prediction λ is made for the variable, based on the generalised prediction function formulated by the algorithm in the training phase.

1.5.2.2 MATHEMATICAL EXTENSION TO RANDOM FORESTS (RFS)

Bearing in mind that RFs are an ensemble of decision trees and that each tree in the forest is reconstructed independently and therefore will have a different size and split structure (Figure 1.13), the independent predictions provided by each tree for the same input variable can be integrated to a single prediction as follows:

$$p(\lambda_f) = \frac{1}{T}\sum_{t=1}^{T}\delta(\lambda_t)\,, \quad (15)$$

Where λ_f represents the prediction of the forest, whereas λ_t the prediction of a tree in a forest of T trees and δ is the Dirac delta function. If the output **I** is categorical then the most represented label in equation (15) will be the final prediction of the RF. If the output labels are continuous values (regression problems), then the medium or median of all tree predictions can be calculated.

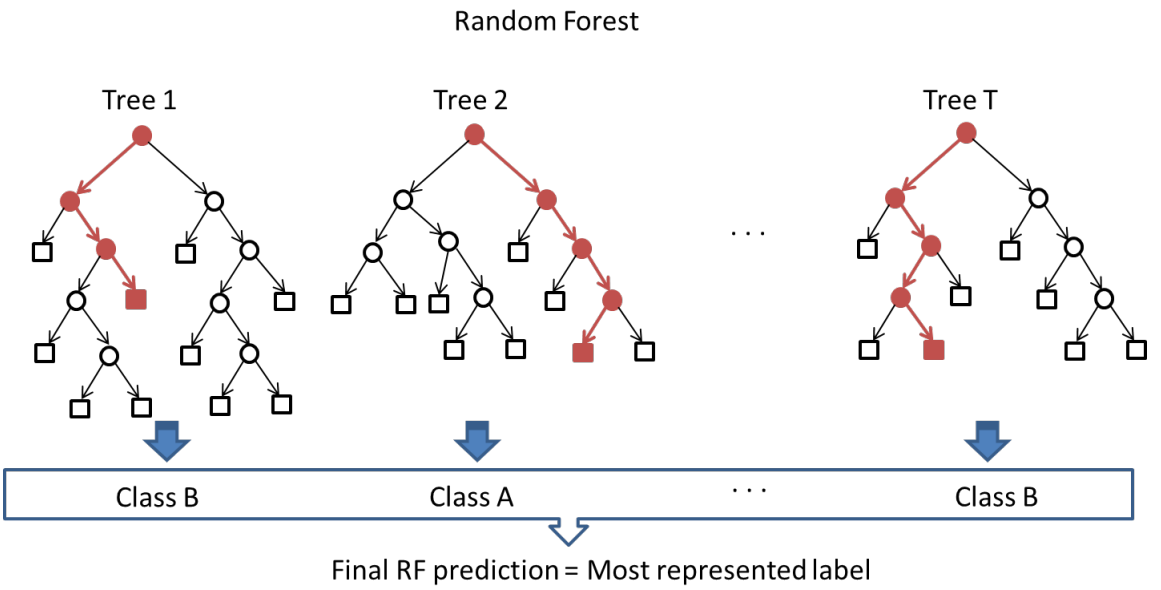


Figure 1.13: Example of a Random Forest (RF) performing a classification task. The forest consists of T number of independant trees, each one analysing each input data and providing a class prediction. The final prediction of the Random Forest for each input data point will be the most represented class in the predictions of the trees.

1.5.2.3 Hyperparameters in Random Forests

Hyperparameters in ML algorithms refer to parameters that can not be estimated from the data and concern the configuration of the learning process. They can have different effects on the algorithm: increase its performance or the computational time required to perform the task. Hyperparameter optimisation or tuning entails the selection of the optimal combination of hyperparameters for the learning algorithm to address a specific task. It is a commonly used and beneficial process in the reconstruction process of RF (and most ML) models.

In RFs, the number of estimators refers to the number of decision trees the model reconstructs to perform a classification or regression task. In general, increasing the number of estimators increases the overall performance of the model as well [201]. Given the two levels of “randomness” in the reconstruction process, RFs do not suffer from overfitting that a large number of estimators could potentially cause. However, increasing the number of estimators comes with a

computational cost. Additionally, the maximum depth refers to the depth of every tree in the forest. Increasing the depth of the trees involves the increase in the number of nodes/rules in every tree and consequently the complexity of the decision function. A tree can be allowed to "grow", without any constraints regarding the depth, or the parameter can be tuned in the case of noisy data. Again, computation time increases with deeper trees. Moreover, the maximum number of features is another hyperparameter in RFs and it refers to the maximum number of features that will be considered at each node split-rule. This hyperparameter highly depends on the type of RF model (i.e. regression or classification). Breiman, 2001, in his original suggestion for randomising the features selected for each node of each tree to decrease the correlation among trees in the forest, he determined the optimum performance in RFs with $\boldsymbol{p}$ number of features is achieved when the maximum number of features is equal to: $\frac{\boldsymbol{p}}{\boldsymbol{3}}$ for regression tasks and $\sqrt{\boldsymbol{p}}$ for classification tasks [199, 202].

1.5.2.4 Evaluation of a Classification model

As mentioned earlier, in the reconstruction of ML models the original data set is divided into a training and a test set. The former consists of the data that will be used in the training process of the model and the later consists of data that will be presented to the built ML model for the first time, after its reconstruction process has been completed. In this way, the model's performance on newly introduced "unknown" data and its predictive power can be evaluated. Several evaluation metrics can be applied for this purpose. We will focus on widely used metrics to evaluate the performance of supervised methods (including RFs), such as classification accuracy, Area Under the Curve (AUC) and precision-recall curves. The selection of the most appropriate ML algorithm for a specific task is dictated by the comparison of their evaluation metrics.

For simplicity reasons, we will consider a binary classification task where samples should be classified into classes TRUE and FALSE. All metrics presented

subsequently can be appropriately adjusted to evaluate multilabel tasks (i.e. classification problems with three or more different outcome labels). We can define true positives (TP) as the correctly predicted labels (samples of class TRUE correctly predicted as TRUE), false positives (FP) as samples predicted TRUE but actually belonged to class FALSE, true negatives (TN) as the samples correctly predicted as FALSE and, finally, false negatives (FN) as samples predicted as FALSE but belonged to the class TRUE.

1.5.2.4.1 Classification accuracy

Classification accuracy (or simply referred to as accuracy and precision) of a classifier refers to the ratio of correct predictions divided by the total number of input samples to the model:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ samples} = \frac{TP+TN}{TP+FP+TN+FN} \quad .$$

Classification accuracy might be a misleading metric in problems dealing with unbalanced data, where the majority of samples belongs to one class and therefore the second class is extremely underrepresented.

1.5.2.4.2 Area Under the Curve (AUC)

Area Under the Curve (AUC) is one of the most widely used evaluation metrics in ML models. It is the area enclosed under a Receiver Operating Characteristic (ROC) curve. Let us first define:

- True Positive Rate or Sensitivity or Recall as the rate of correct predictions over the total number of true samples in the dataset:

$$TruePositiveRate = \frac{TP}{TP + FN}$$

- True Negative Rate or Specificity as the rate of samples correctly predicted as FALSE over the total number of FALSE samples in the dataset:

$$TrueNegativeRate = \frac{TN}{TN + FP}$$

- False Positive Rate as the rate of wrongly classified FALSE samples over the total number of FALSE samples in the dataset:

$$FalsePositiveRate = \frac{FP}{FP + TN} = 1 - Specificity$$

Now, if we plot the False Positive Rate (1-Specificity) against the True Positive Rate (Sensitivity), the resulting curve is called ROC curve and the measured area under it (AUC) is the evaluation metric for the performance of the classifier regarding (Figure 1.14). The higher the AUC ROC curve, the higher the performance of the classifier.

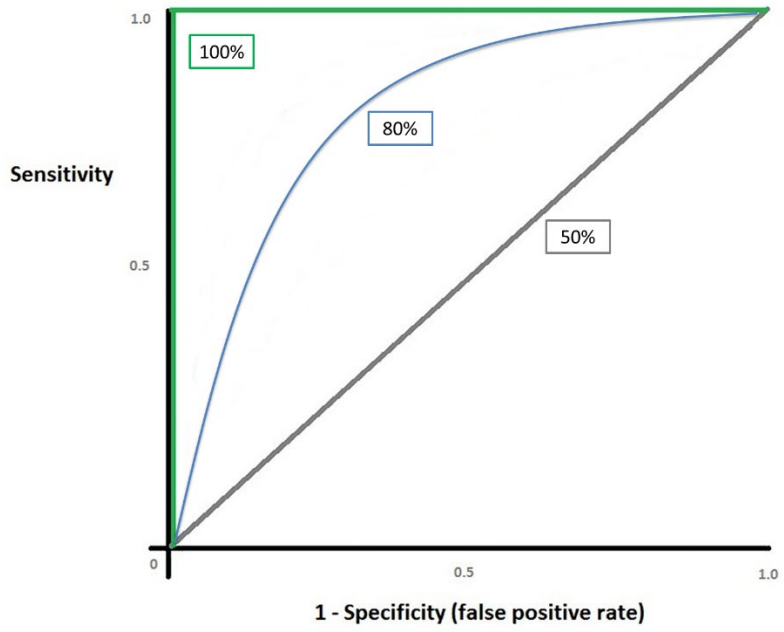


Figure 1.14: A Receiver Operating Characteristic (ROC) curve can be plotted as the False Positive Rate against the True Positive Rate of the predictions of a classifier and the Area Under the Curve (AUC) is a widely used evaluations metric of classification algorithms in machine learning. The perfect classifier would have an AUC of 100%, denoting both high sensitivity and specificity in predicting the correct label for a sample, whereas a weak classifier would have a 50% AUC score.

1.5.2.4.3 PRECISION – RECALL CURVE

Finally, we can define precision as the rate of correctly identified TRUE samples over all the TRUE samples in the dataset:

$$Precision = \frac{TP}{FP + TP}$$

And the F-score as the harmonic mean of precision and recall (or sensitivity as defined above):

$$F - score = \frac{2}{\frac{1}{precision} + \frac{1}{recall}}$$

Then, recall can be plotted against precision to denote the performance of a classifier in regards to its accuracy and robustness (Figure 1.15). In general, the higher the F-score and consequently the AUC of the precision – recall curve, the higher the performance of the classifier.

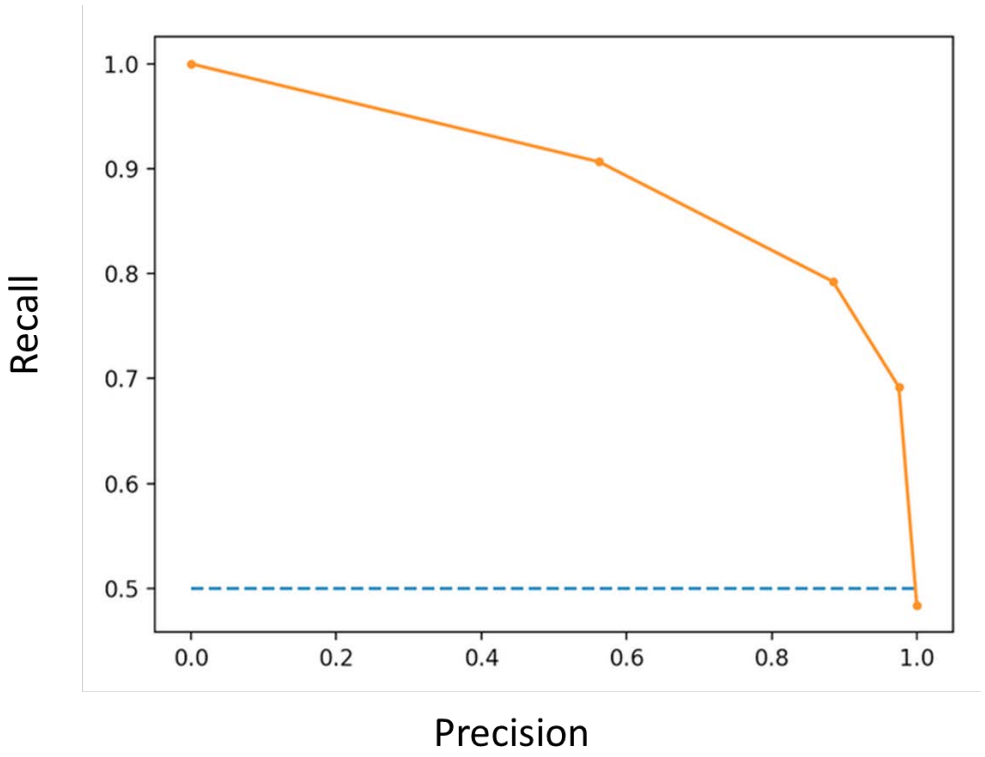


Figure 1.15: Precision and recall curves and the area under the curve are a powerful evaluation metric for classification algoritms, calculating the accuracy of the method in correctly classifying the samples (precision) and its robustness in predicting lowly represented samples in the dataset (sensitivity). The higher the area under the curve the higher the performance of the classifier.

Overall, the choice of evaluation metrics could depend on the nature of the classification problem, but a combination of evaluation metrics could provide a more integrated view on the performance of a ML classifier.

2.

OBJECTIVES

2 OBJECTIVES

The general objective of this Ph.D. thesis is to develop computational methods and workflows for metabolic flux analysis, to study the metabolic adaptations that occur in Acute Myeloid Leukaemia and identify putative metabolic biomarkers to use in therapy or in disease prognosis.

Hence, to meet the main objective of this thesis, the following specific objectives were set:

1. To develop a constraint-based modelling framework to functionally characterise the oncogene Transketolase-like 1 in Acute Myeloid Leukaemia and assess its therapeutic potential
2. To develop a constraint-based modelling framework able to integrate a diverse set of -omic data in order to explore the mechanisms and metabolic adaptations that occur in drug-resistant Acute Myeloid Leukaemia cells and identify potential metabolic vulnerabilities to use as combination therapy
3. To develop a computational workflow combining constraint-based modelling and machine learning algorithms supporting the reconstruction of patient-specific genome scale metabolic models for patient risk-stratification and prognostic biomarker discovery

Successful fulfilment of these objectives will provide valuable insight on the metabolic reprogramming of Acute Myeloid Leukaemia, as well as the key players in drug resistance and disease prognosis. The genome-scale metabolic reconstructions introduced will provide a new tool to study Acute Myeloid Leukaemia, both globally and at a patient-specific level.

3.

MATERIALS AND METHODS

3 Materials and Methods

3.1 Cell lines and cell culture

The human THP-1 acute monocytic leukemia cell line was obtained from the American Type Culture Collection (Manassas, VA, US). Cell work was performed in a standard incubator at 21% O2 simulating normoxic.

THP-1 WT and THP-1 TKTL-1 knock-down cells were generated by Sirion Biotech GmbH (Munich, Germany). The THP-1 cell lines were grown in RPMI 1640 culture medium supplemented with 10% heat-inactivated FCS and antibiotics: 100 U/ml penicillin, 100 ug/ml streptomycin and 50 ug/ml gentamicin. THP-1 and HL60 resistant cells to Cytarabine (Ara-C) and Doxorubicin (DOX) were generated by Prof. Jindrich Cinatl from Goethe University of Frankfurt (GUF).

3.2 DNA Microarrays

Total RNA was isolated and extracted using the RNeasy Mini Kit (Qiagen).

In total, six Affymetrix GeneChip® Human Gene 2.0 ST Arrays were hybridized with cDNA from THP-1 cells. More specifically, three arrays were hybridized with cDNA from cells transfected with an empty vector and were used as the control, whereas three arrays were hybridized with cDNA from the TKTL-1 knock-down cells.

3.3 RNA Sequencing

Cell pellets from the THP-1 and HL60 AML parental and resistant cells under normoxic conditions were collected and frozen. Total RNA was extracted using RNeasy Mini Kit (Qiagen, Hilden, Germany). RNA integrity was further tested using lab-on-a-chip technology on the BioAnalyzer 2100.

High quality RNA-seq (transcriptome) was performed in the CNIC genomic unit using the Illumina HiSeq 2500 sequencer. There, 200 ng of total RNA were used to generate barcoded RNA-seq libraries using the NEBNext Ultra RNA Library preparation kit (New England Biolabs). Briefly, poly A+ RNA was purified using poly-T oligo- attached magnetic beads followed by fragmentation and then first and second cDNA strand synthesis. Next, cDNA ends were repaired and adenylated. The NEBNext adaptor was then ligated, followed by uracile excision from the adaptor and PCR amplification. Finally, the size of the libraries was checked using the Agilent 2100 Bioanalyzer DNA 1000 chip and their concentration was determined using the Qubit® fluorometer (Life Technologies). Libraries were sequenced on a HiSeq2500 (Illumina) to generate 60 bases single-end reads. Finally, FastQ files were obtained for each sample using CASAVA v1.8 software (Illumina). In total, RNA from 18 samples were sequenced: three replicates from each of the THP-1 and the HL60 parental, the Cytarabine (Ara-C) resistant and the Doxorubicin (DOX) resistant cells.

3.4 TCGA AML PATIENT DATASET

The results presented in Chapter 4.3 ‘Application of machine learning techniques on GSMMs for AML patient stratification and metabolic biomarker identification’ are in part based upon data generated by the Cancer Genome Atlas (TCGA) Research Network: https://www.cancer.gov/tcga. The transcriptome profiling and clinical data from the TCGA-LAML dataset were used.

The TCGA LAML dataset consisted of transcriptomic data for 183 AML patients. RNA-seq analysis was performed using an Illumina HiSeq platform. The gene abundance estimation was done using the RNA-Seq by Expectation Maximization (RSEM) software and provided in the format of normalised binary logarithm RPKM values [203].

The cytogenetic risk classification of patients in the TCGA LAML dataset was based on the definition by the Cancer and Leukaemia Group B (CALGB) [204]. This information was missing in three of the 183 patients in the dataset, therefore

we decided not to include these patients in our analysis. The 180 patients were classified into Poor risk (23.3%), Normal/Intermediate risk (57.2%) and Favourable risk (19.5%) groups.

Finally, for the 180 patients we selected to include in our work, we extracted the information on their age at initial diagnosis (continuous values ranging from 18 to 88 years of age), days to death (continuous values ranging from 0 to 1706 days), ethnicity (i.e. Asian, Black or African American and White), gender (i.e. male 54.4% and female 45.5%) and cytogenetic abnormalities.

3.5 Bioinformatics analysis

All computational analyses were performed using CaStLeS resources (both compute and storage) which are an integrated element of Birmingham Environment for Academic Research (**BEAR**) Cloud and the wider BEAR infrastructure of the intranet of Birmingham University. The differential expression analysis and the Gene Set Enrichment Analysis for both transcriptomic experiments were performed using R programming language version 3.2.2 [205] in the integrated development environment (IDE) RStudio (2016) and packages from *Bioconductor* [206].

More specifically, the pre-processing of the raw data obtained from the DNA microarray experiment was performed using the *Oligo* package from *Bioconductor* [207]. The Robust Multichip Average (RMA) algorithm was used to perform background subtraction, quantile normalization and summarization through median-polish [208-210].

On the other hand, Quality Control (QC) of the RNA-seq data was performed using FastQC, a computational Quality control tool for high throughput sequence data in Java developed by Simon Andrews and the Babraham Bioinformatics group (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The alignment of the reads was performed using STAR version 2.5.2a open source

software [86] and the FASTA sequences were generated using the *Homo sapiens* high coverage GRCh37.75.dna.primary_assembly.

3.5.1 Differential Expression analysis

The differential expression (DE) analysis for the microarray data was performed using the *Limma* package from Bioconductor [211, 212]. The annotation of the probes was done using the *hugene20stprobeset.db* platform, which corresponds to the type of microarrays used in the experiments. The annotations were based on the exon probeset level.

Accordingly, in the case of the RNA-seq experiments, the differentially expressed genes were identified using the Differential gene expression analysis based on the negative binomial distribution 2 (DESeq2) package from Bioconductor [93].

3.5.2 Gene Set Enrichment Analysis

Gene Set Enrichment Analysis (GSEA) was applied to the lists of statistically significant genes resulting from the different experiments in order to identify enriched terms in the datasets. GSEA was performed using the Generally Applicable Gene-set Enrichment for Pathway Analysis (Gage) package from Bioconductor in R [213]. Genes were considered to be statistically significant if their adjusted p-value was $< 10^{-5}$ and GSEA was carried out considering gene sets from the Gene Ontology (GO) and KEGG databases. More specifically, signal transduction and metabolic (total of 2868 gene sets) and disease pathways (total of 85 gene sets) and terms related to biological processes (total of 7,573 gene sets) were used [214-216].

3.6 Metabolic modelling

Genome Scale Metabolic Model (GSMM) reconstructions were built for the THP-1 and the HL60 Acute Myeloid Leukaemia (AML) cell-lines under different treatments and/or experimental conditions. All computational analyses for the model reconstructions were performed in Python programming language, using the COBRApy-0.17.1 Constraints Based Reconstruction and Analysis for [74]

constraint-based modelling of biological networks [217]. The CPLEX 12.6.2 solver was used for the computational simulations. Recon 2.2, comprising of 5324 metabolites, 7785 reactions and 1675 associated genes, was used as a base model for the reconstructions [218]. Consumption and production rates of glucose, lactate, glutamine and glutamate, as well as essential amino acids in the media, were calculated as follows and added as constraints:

$$\boldsymbol{\mu} = \frac{\boldsymbol{ln}\left(\frac{\boldsymbol{N_f}}{\boldsymbol{N_0}}\right)}{\boldsymbol{t_f}} \quad \text{and} \quad \boldsymbol{k_{pc}} = \frac{\boldsymbol{\Delta M}}{\boldsymbol{\Delta N}} \times \boldsymbol{\mu},$$

where μ represents the growth rate, N_0 and N_f the initial and final number of cells respectively given in [millions of cells] for a time t_f given in [h], $\Delta M = M_f - M_0$ represents the increase or decrease in concentration for a metabolite for the time t_f, an initial concentration M_0, a final concentration M_f and it is given in [μmol]. Finally, the consumption or production rate of a metabolite k_{pc} is calculated in [μmol/ (million cells ·h)]. Where appropriate, flux rates were normalized by total protein to account for the difference in the size of the cells.

The rate of basal cellular respiration and the rate of ATP production from both glycolysis and mitochondria were integrated directly as constrains to the models. In addition, gene expression data obtained by the transcriptomic experiments were integrated to the reconstructions using GIM3E [219].

The Flux Variability Analysis (FVA) implementation in the COBRApy package was applied first as a preprocessing step to identify and remove inactive reactions from the reconstructions and ultimately to compute the optimum flux distributions for the GSMMs [162].

The models were further interrogated towards identifying essential genes/reactions whose inhibition would impair the survival of the cells. Gene essentiality analysis was performed using both COBRApy implementations for Flux Balance Analysis (FBA) and minimization of metabolic adjustment (MOMA) [159, 220]. Overall, a gene was considered essential for the viability of the cell if the simulated biomass production, after its knock-down, dropped at less than 10%.

3.6.1 Acute Myeloid Leukaemia consensus model

We decided to reconstruct a consensus AML model, to later use as a base model for the integration of AML patient-derived transcriptomic data. The reason behind this was the lack of metabolomic data from each patient in the TCGA LAML dataset. Therefore, we constrained Recon 2.2 using metabolomic data from the THP-1 and HL60 cell lines. We treated the three replicates from each of the two cell lines as six biological replicates. Then, we calculated the consumption and production rates of glucose, lactate, glutamine and glutamate, as well as essential amino acids in the media using the mean value and standard error of mean (SEM) over the six measured concentrations, to predict the upper or lower bounds within a 95% confidence interval.

For the consumption or production rate of a metabolite k_{pc} we calculated the standard deviation (SD) of the mean:

$$SD = \sqrt{\frac{\sum(k_{pc} - \overline{k_{pc}})}{N}}$$

And then the SEM:

$$SEM = \frac{SD}{\sqrt{N}}$$

where $\overline{k_{pc}}$ is the mean value of the consumption or production rate of a metabolite and N=6 the sample size. Since our sample size is N=6 (<30), we need to assume a t-distribution for our data and the t statistic for a 95% confidence interval is t=2.447 [221].

3.7 Machine learning algorithms

3.7.1 supervised sparse Canonical Correlation Analysis (ssCCA)

The supervised extension of the sparse Canonical Correlation Analysis (ssCCA) was performed using the PMA: Penalized Multivariate Analysis package version 1.2.1 in R [195, 222]. Two matrices were used as inputs to ssCCA: the

matrix of gene expression levels of AML patients from the LAML TCGA data base and the matrix containing the simulated metabolic flux distributions for AML patients, calculated from patient specific GSMM reconstructions. The third input to ssCCA was a vector of phenotypes, which was the Cancer and Leukemia Group B (CALGB) risk classification of the patients provided from the LAML TCGA database.

3.7.2 Random Forests (RFs)

The Random Forest (RF) classifier was built using the scikit-learn library version 0.23.2 in Python [223].

First, missing values were handled applying multivariate iterative imputation using the IterativeImputer function. Missing values in categorical features (i.e. cytogenetic anormalities and ethnicity) were imputed using the most frequent strategy, whereas missing values in numerical features were imputed based on the mean statistic.

Then, the dataset was divided into a train set and a test set with a 60/40 split. A grid search 5-fold cross validation was performed in order to identify the best hyperparameters for the model. This led to the use of the following parameters: max_depth=8, min_sample_leaf=1, min_samples_split=10, n_estimators=500.

We selected the Area Under the Curve (AUC) of Receiver Operating Curves (ROC) and Pecision – Recall (PR) curves as performance metrics of the RF. Since our classification problem was a multi-class problem, meaning that the data could be classified into three risk categories, the AUC ROC and PR metrics were calculated for each class individually. In addition, the micro and macro average scores for the two metrics were also calculated as follows:

For each class $c_i \, \epsilon \, C = \{1,2,3\}$, the micro-average scores:

- True Positive Rate or Sensitivity or Recall:

$$TruePositiveRate_{micro} = Recall_{micro} = \frac{\sum_{i=1}^{3} TP_i}{\sum_{i=1}^{3}(TP_i + FN_i)}$$

- True Negative Rate or Specificity:

$$TrueNegativeRate_{micro} = \frac{\sum_{i=1}^{3} TN_i}{\sum_{i=1}^{3}(TN_i + FP_i)}$$

- Precision:

$$Precision_{micro} = \frac{\sum_{i=1}^{3} TP_i}{\sum_{i=1}^{3}(TP_i + FP_i)}$$

- And the F-score as the harmonic mean of precision and recall:

$$F_{micro} = \frac{2}{\frac{1}{precision_{micro}} + \frac{1}{recall_{micro}}}$$

And the macro-average scores for the ROC AUC curve:

- True Positive Rate or Sensitivity or Recall:

$$TruePositiveRate_{macro} = Recall_{macro} = \frac{1}{3}\frac{\sum_{i=1}^{3} TP_i}{\sum_{i=1}^{3}(TP_i + FN_i)}$$

- True Negative Rate or Specificity:

$$TrueNegativeRate_{macro} = \frac{1}{3}\frac{\sum_{i=1}^{3} TN_i}{\sum_{i=1}^{3}(TN_i + FP_i)}$$

Finally, the attempt to specify the importance of features in our ML model refers to the identification and ranking of features based on their contribution and significance in building the model. As a result, feature importance as a metric, was employed to provide insights into the putative role of different features as putative biomarkers and help in the interpretation of the results. To extract the feature importance scores the impurity-based feature importance function ‘feature_importances_’ implemented in the scikit-learn toolbox was used.

4.

SUMMARY OF THE RESULTS

4 Summary of the Results

4.1 Functional characterization of Transketolase-like 1 in Acute Myeloid Leukaemia

4.1.1 Introduction

Three transketolase genes are found in the human genome: transketolase (TKT), transketolase-like 1 (TKTL1) and transketolase-like 2 (TKTL2). While it has been established that TKT plays a crucial role in the central metabolism, very little is known about the other two isoforms and great efforts are underway to explore their cellular and enzymatic functions.

More specifically, TKT is a thiamine-dependent enzyme that links glycolysis and the non-oxidative branch of the Pentose Phosphate Pathway (PPP). Glucose enters the cell and is transformed into glucose-6-phosphate (G6P), which then can either be converted to pyruvate through glycolysis or enter the PPP through the oxidative branch of PPP producing ribose-5-phosphate (R5P), CO2 and reduced nicotinamide adenine dinucleotide phosphate (NADPH). Alternatively intermediate metabolites of glycolysis (fructose-6-phosphate (F6P) and glyceraldheide-3-phosphate (G3P) can be converted into R5P, and *vice versa*, through the non-oxidative branch of PPP that comprises a series of reversible reactions that transfer two carbon or three carbon units, being TKT the main player in these pathway [224-226]. R5P is necessary for the synthesis of nucleic acids, complex sugar molecules and co-enzymes. The latter acts as an electron donor and therefore is required in reductive biosynthetic processes. NADPH is also involved in the detoxification of the cells, in the defence mechanisms against oxidative stress through the synthesis of glutathione and finally in the generation of reactive oxygen species (ROS) [227]. The fact that the non-oxidative PPP is a reversible pathway and regulated mainly by mass balance confer to this pathway a key role in ensuring the proper adaptation of cell metabolism to a high demand of NADPH (the two branches operated sequentially recycling the R5P into glycolytic intermediates and producing CO2 and NADPH) or to a high demand of R5P (the

two branches operate as parallel processes producing R5P either directly from G6P or indirectly through TKT reaction involving F6P and G3P). Thus, any dysregulation in TKT activity affects all the aforementioned biochemical processes and has been linked to diabetes, Alzheimer's disease and cancer [228-230].

Regarding the two other known isoforms of transketolase, TKTL1 and TKTL2, several studies have attempted to explore their biological function and whether they even have any enzymatic activity [231]. According to the most recent findings, TKTL1 protein lacks 38 amino acids in its active site compared to TKT, including two histidine residues and has a substitution mutation [232]. Throughout evolution, TKTL1 has been conserved by 81% compared to TKT [231, 232].

There has been increasing interest around TKTL1, since it has been closely linked to various aspects of different types of cancer, as well as diabetes. An overexpression of the gene has been reported in different cancers [9-13] and has been linked to poor disease prognosis, cancer progression and metastasis [10, 14-16]. It has been associated with an increased resistance to cisplatin treatment in nasopharyngeal carcinomas [17]. Moreover, it has been shown that TKTL1 promotes cancer cell proliferation, while preventing apoptosis and oxidative stress [18-20]. Li et al. demonstrated that TKTL1 protein levels change throughout the different cell-cycle phases and identified TKTL1 as a positive regulator of Ribose 5-phosphate (R5P) [21].

The aim of this study was to investigate the role of TKTL1 in Acute Myeloid Leukaemia (AML) and explore its potential as a biomarker or a putative drug target for this multifactorial disease. In order to do so, the human THP-1 monocytic leukaemia cell line was selected as the most suitable cell model to collect data and perform suitable validation experiments, as the THP-1 cells exhibit a high TKTL1 gene expression. Previously generated experimental transcriptomic and metabolomics data were collected from THP-1 control cells and from THP-1 TKTL1 knock-down (KD) cells (see Materials and Methods section 3.1). First, the transcriptomic data were analysed extensively, and differential expression analysis and gene set enrichment analysis were performed. Next, the transcriptomics were

integrated with the metabolomic data and genome-scale metabolic models (GSMMs) were reconstructed. This integrative approach paved the way to a deeper exploration and systemic identification of putative vulnerabilities that could be used as combination therapies with the knock-down of TKTL1 in AML. Finally, all the hypotheses formulated and based on the described computational analyses were validated experimentally.

4.1.2 Results

4.1.2.1 Analysis of transcriptomic data

First, a thorough computational analysis was performed on the transcriptomic data, including the appropriate processing of the microarray data, as well as the differential expression analysis to identify the most significant changes in gene expression after the knock-down of TKTL1 and the gene set enrichment analysis to provide greater biological insight on patterns of gene expression. To identify the differentially expressed genes between THP-1 control and THP-1 TKTL1 knock-down, six microarrays (three microarrays for each cell line, performed in the same cell culture conditions), have been analysed.

4.1.2.1.1 Pre-processing of microarray data

A method to determine the quality of the raw data, detect possible outliers or artefacts in the data and decide upon normalization, the intensities of red (R) and green (G) dies can be visualised through an MA-plot of the distribution of the intensity ratio (M) against the average intensity (A). We decided to make the analysis more robust by comparing each array to a pseudo-array, composed by the median intensity of each probe across all arrays. The resulting robust MA-plot depicts the difference between the intensity of a probe on an array and the median intensity of that probe over all the arrays (M) plotted against the average of the median intensity of the probe on that array and of the median intensity of that probe across all the arrays (A). M and A follow the equations:

$$\mathrm{M} = \log_2(R/G) = \log_2(R) - \log_2(G)$$
$$A = \frac{1}{2}\log_2(R \cdot G) = \frac{1}{2}\left[\log_2(R) + \log_2(G)\right],$$

[85]

where R and G represent the intensities of the red and green dies, respectively.

In addition, the metric of the median value calculated from the plot shows the central tendency, whereas the interquartile range (IQR) is a measure of the dispersion of the data.

The distribution of the data of the three microarrays from the THP-1 control (Figure 4.1A) and the microarrays of the THP-1 TKTL1 knock-down (Figure 4.1B) gets more and more dispersed as A increases and therefore the loess curve, shown in red, deviates from M=0. After normalization, the data is more evenly distributed, for both the control (Figure 4.2A) and the TKTL1 knock-down (Figure 4.2B) repicates, clearly indicating that the variability of the data has been eliminated to a great extent.

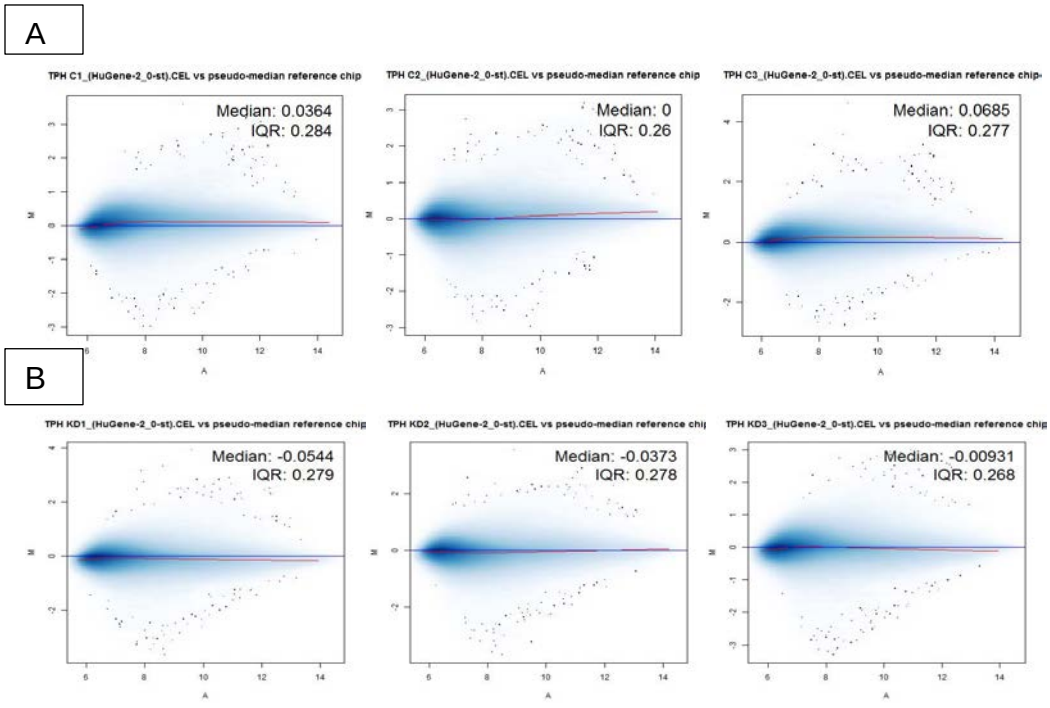


Figure 4.1: The raw probe intensities for each of the six arrays used were visualized in an MA-plot. The three arrays corresponding to the control cell line are on the top [A] and the arrays corresponding to the cells with the TKTL-1 knock-down are shown on the bottom [B].

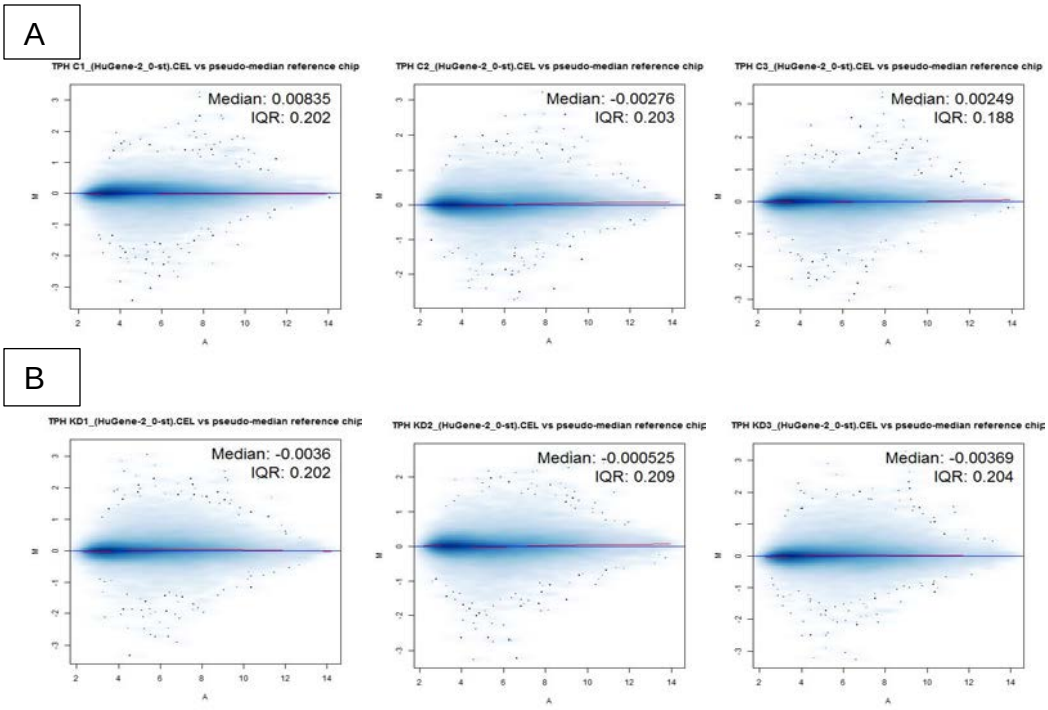


Figure 4.2: The normalized probe intensities in MA-plots. The three arrays corresponding to the control cell line are shown on the top [A], whereas the three arrays corresponding to the cells with the TKTL-1 knock-down are shown below [B].

4.1.2.1.2 Differential expression analysis

The differential expression (DE) analysis was performed comparing the TKTL-1 knock-down cells vs the Control cell-line. Differentially expressed genes were selected according to the following thresholds regarding the binary logarithm of the fold change (logFC) in gene expression and its statistical significance measured by a p-value corrected for multiple testing (adj. p-value): |logFC|≥1 and adj. p-value<5x10-4. Most commonly, a far more strict threshold for differential expression is used, for example a |logFC|≥2.5 or in some cases even ≥3.0, but since the objectives of this study included the identification of changes in the expression of metabolic genes due to the knock-down of TKTL1, as well as the exploration of the functional role of this gene in AML, including small alterations in gene expression was considered more appropriate.

Under this scope, 112 genes were differentially expressed. More specifically, 42 of these genes were down-regulated with a logarithmic fold change in expression ranging between -6.54 and -1.35, whereas 70 genes were up-regulated with a logarithmic fold change in expression ranging between 1.41 and 4.88.

4.1.2.1.3 Gene Set Enrichment Analysis

Gene Set Enrichment Analysis (GSEA) was performed on the entire list of statistically significant genes (adj. p-value<$5x10^{-4}$) provided by the differential expression analysis, as described in 4.1.2.1.2 [100]. The logFC in gene expression was used as the criteria to rank the genes for subsequent GSEA analysis. Significantly enriched pathways were identified both in down and up-regulation. The False Discovery Rate adjusted p-values (or q-values), which in principal depict the corrected for multiple testing p-values, were used to filter the results. Only gene sets/pathways with a calculated q-value<0.25 were considered significant.

Firstly, GSEA was performed using the disease gene sets from the KEGG database and significant enrichment was identified in 56 of the 80 annotated pathways, suggesting an important role of TKTL1 in various different pathologies.

Table 1 summarizes the results and includes the sets that describe the diseases that have been to TKTL1 by previous publications, validating the robustness of our analysis.

Table 1: Significantly enriched pathways in down-regulation in TKTL-1 knock-down vs control, using the disease gene sets from KEGG database. From a total of 56 pathways identified from *gage* algorithm, the 19 that are of specific interest to this study are shown. The pathways are ranked from lower to higher q-value.

Disease related pathway	p-value	q-value	Set size
Transcriptional misregulation in cancer	1.15E-07	8.19E-06	159
MicroRNAs in cancer	3.27E-06	3.87E-05	242
Pathways in cancer	4.67E-06	4.74E-05	373
Viral carcinogenesis	0.00025	0.0013	177
Melanoma	0.0012	0.0039	66
Central carbon metabolism in cancer	0.0016	0.0046	59
Glioma	0.0024	0.0067	58
Pancreatic cancer	0.0049	0.011	61
Prostate cancer	0.0051	0.011	80
Non-small cell lung cancer	0.0068	0.013	52
Choline metabolism in cancer	0.0080	0.014	92
Endometrial cancer	0.0091	0.016	47
Basal cell carcinoma	0.010	0.017	53
Chronic myeloid leukemia	0.015	0.024	66
Small cell lung cancer	0.019	0.030	83
Chemical carcinogenesis	0.020	0.030	55
Colorectal cancer	0.025	0.034	57
Bladder cancer	0.033	0.044	38
Acute myeloid leukemia	0.038	0.048	50

As part of the aforementioned disease gene sets, the KEGG database has proposed 22 gene sets related to carcinogenesis and different types of cancer. The GSEA analysis identified 19 out of the total of predefined 22 cancer gene sets as enriched after the knock-down of TKTL1, suggesting an important role of this gene in cancer and the existence of gene co-expression patterns among AML and other cancer types (Table 1). Therefore, we looked for a common gene signature between the AML gene set and the melanoma, prostate cancer, small cell lung carcinoma and colorectal cancer sets, based on publications that have proved an important implication of TKTL1 in these cancers (Figure 4.3). A text mining search among the genes that form these sets identified a list of six genes that are common

for all them, including the three AKT isoforms (AKT1, AKT2 and AKT3), Cyclin D1 (CCND1) and two members of the Phosphoinositide 3-kinase (PI3K) regulatory subunit, namely the PIK3R3 and PIK3R5.

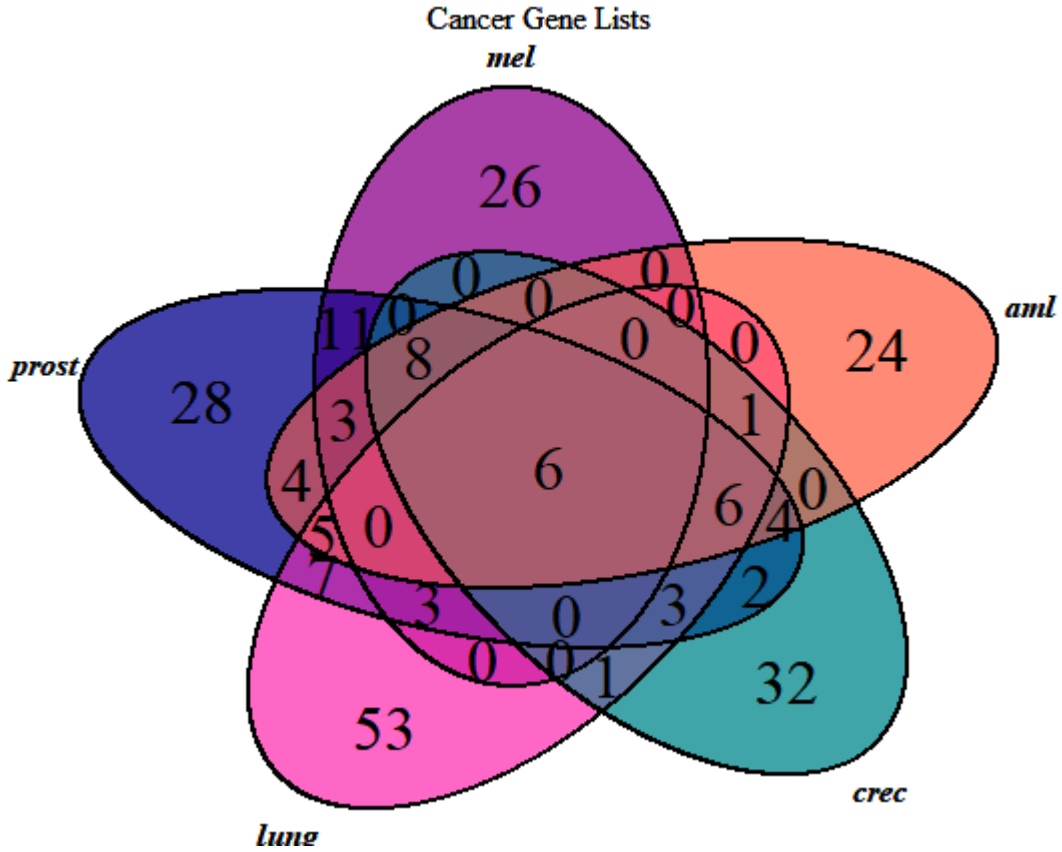


Figure 4.3: Venn diagram of the genes in five of the 19 cancer related pathways that were enriched after the knock-down of TKTL1. The abbreviations that appear in the figure correspond to mel: melanoma, aml: acute myeloid leukaemia, crec: colorectal cancer, lung: small cell lung carcinoma and prost: prostate cancer.

Next, the terms associated with biological processes from the GO database were used to perform a GSEA, in order to further explore which pathways and cellular activities could be affected by the knock-down of TKTL1. A great number of gene sets were highlighted as significantly perturbed. In particular, 3002 of 7530 gene sets were found to be enriched and with remarkably low p- and q-values. The fact that so many GO terms were found to be strongly regulated is of great interest

and could indicate that TKTL1 plays a vital role in several biological processes in AML cells.

In order to reduce the size of the list of significant GO gene sets and at the same time increase the robustness and the biological relevance of the results, the non-redundant gene sets were identified. In other words, gene sets that overlap greatly in their main list of perturbed genes with truly significant gene sets were eliminated. Again, the resulting list of enriched gene sets was extensive and consisted of 1682 sets. From this list, four themes were most prominent, and these were related to the cell cycle, signal transduction, immune response and DNA repair processes and, finally, response to oxidative stress and the regulation of ROS levels (Table 2).

Table 2: Amongst the most significantly enriched gene sets from the Gene Ontology consortium were terms related to the cell cycle, signal transduction, immune response and DNA repair and response to oxidative stress and regulation of ROS levels.

GO Essential Sets	p-value	q-value	Set size
Energy derivation by oxidation of organic compounds	3.11E-14	1.41E-11	341
Defence response to another organism	2.98E-13	8.58E-11	493
Wnt signalling pathway	1.30E-11	1.96E-09	382
Cell growth	1.33E-11	1.96E-09	426
Cellular cation homeostasis	1.61E-11	2.23E-09	481
Regulation of hormone levels	2.03E-11	2.58E-09	450
Methylation	1.13E-10	9.67E-09	307
Signal release	3.01E-11	3.50E-09	429
Positive regulation of protein transport	7.33E-11	6.78E-09	394
Activation of protein kinase activity	3.06E-10	1.94E-08	440
Regulation of cell cycle process	3.48E-10	2.11E-08	499

Next, GSEA was performed using a list of terms comprising of the canonical (signal transduction) and metabolic pathways from KEGG database. In total, 101 of 2232 non-redundant gene sets were identified as significantly enriched. Table 3 and Table 4 show the most enriched metabolic and signal transduction related pathways, respectively.

Table 3: GSEA analysis identified 22 metabolic pathways among the 101 KEGG gene sets that are affected by the knock-down of TKTL1.

KEGG Metabolic pathways	p-value	q-value	Set size
Carbon metabolism	2.28E-07	1.02E-05	110
Biosynthesis of amino acids	1.34E-05	0.000185	68
Purine metabolism	1.74E-05	0.000225	161
Glycolysis / Gluconeogenesis	9.57E-05	0.000944	65
Oxidative phosphorylation	0.000134	0.001259	112
Valine, leucine and isoleucine degradation	0.000269	0.002061	47
Starch and sucrose metabolism	0.000881	0.004319	32
Glycerophospholipid metabolism	0.000877	0.004319	90
Butanoate metabolism	0.001029	0.004842	28
Arginine and proline metabolism	0.002472	0.008123	44
Pyrimidine metabolism	0.003259	0.010068	89
Fatty acid metabolism	0.003464	0.010339	43
Cysteine and methionine metabolism	0.003865	0.010667	42
Inositol phosphate metabolism	0.003865	0.010667	67
Tyrosine metabolism	0.004465	0.0117	33
Pentose phosphate pathway	0.005909	0.014391	29
2-Oxocarboxylic acid metabolism	0.006449	0.015169	16
Fructose and mannose metabolism	0.007314	0.01679	31
Galactose metabolism	0.007789	0.016882	30
Mineral absorption	0.007662	0.016882	46
Metabolism of xenobiotics by cytochrome P450	0.008276	0.017481	53
Fatty acid degradation	0.009118	0.018874	40

Table 4: GSEA analysis identified 79 signal transduction pathways among the 101 KEGG gene sets that are affected by the knock-down of TKTL1. The top 30 of these pathways according to the calculated q-value are summarised below.

KEGG Signalling pathways	p-value	q-value	Set size
MAPK signalling pathway	2.47E-07	1.02E-05	244
PI3K-Akt signalling pathway	8.80E-07	3.04E-05	319

mTOR signalling pathway	1.09E-06	3.17E-05	142
Insulin signalling pathway	4.13E-06	8.55E-05	132
Wnt signalling pathway	1.10E-05	0.000162	132
Ras signalling pathway	2.15E-05	0.000262	211
Signalling pathways regulating pluripotency of stem cells	4.99E-05	0.000574	133
cAMP signalling pathway	8.47E-05	0.000876	186
Rap1 signalling pathway	0.000206	0.00164	196
AMPK signalling pathway	0.000284	0.002098	113
cGMP-PKG signalling pathway	0.000296	0.002116	151
TNF signalling pathway	0.000335	0.002313	102
Chemokine signalling pathway	0.000464	0.002822	168
HIF-1 signalling pathway	0.000859	0.004319	97
Sphingolipid signalling pathway	0.001217	0.005478	109
Jak-STAT signalling pathway	0.001399	0.005909	147
Oxytocin signalling pathway	0.001437	0.005948	143
Calcium signalling pathway	0.001549	0.005992	173
T cell receptor signalling pathway	0.002	0.007395	101
NF-kappa B signalling pathway	0.002644	0.008553	88
PPAR signalling pathway	0.003222	0.010068	67
Glucagon signalling pathway	0.003646	0.010339	96
Thyroid hormone signalling pathway	0.003519	0.010339	110
Neurotrophin signalling pathway	0.003415	0.010339	112
Phosphatidylinositol signalling system	0.004277	0.011498	90
Adipocytokine signalling pathway	0.004407	0.011697	65
Prolactin signalling pathway	0.004773	0.01235	64
Notch signalling pathway	0.006136	0.014599	46
Hedgehog signalling pathway	0.008913	0.018636	44
Toll-like receptor signalling pathway	0.009803	0.020089	95

Taking a closer look in the significantly enriched metabolic gene sets, the knock-down of TKTL1 caused a deregulation in the gene expression of members of the “carbon metabolism”, “biosynthesis of amino acids”, “purine metabolism” and “cysteine and methionine metabolism” (Table 10). These results give indications of a potential link between the argued enzymatic activity of TKTL1 in the PPP affecting the availability of R5P, a fundamental precursor in the *de novo* biosynthesis of purines. Additionally, a deregulation of genes involved in one-carbon and folate metabolism was also observed, whose role is crucial in epigenetic methylation events commonly deregulated in cancer and especially in AML [233]. Complementary evidence was also identified in the enrichment of gene

sets representing the PI3K/AKT and mTOR signalling pathways, since it has been reported in the literature that amino acids activate the PI3K/AKT and the mTOR Complex I, a mechanism employed by cancer cells to promote cell growth and proliferation [234].

Altogether, the GSEA analysis performed illustrated that TKTL1 may very well be a moonlight enzyme implicated in a great number of different cellular processes. Therefore, we decided to explore in greater depth the role of TKTL1 in AML cell cycle regulation, AKT and mTOR signalling and in epigenetic regulation.

4.1.2.2 TKTL1 AFFECTS CELL-CYCLE PROGRESSION IN AML

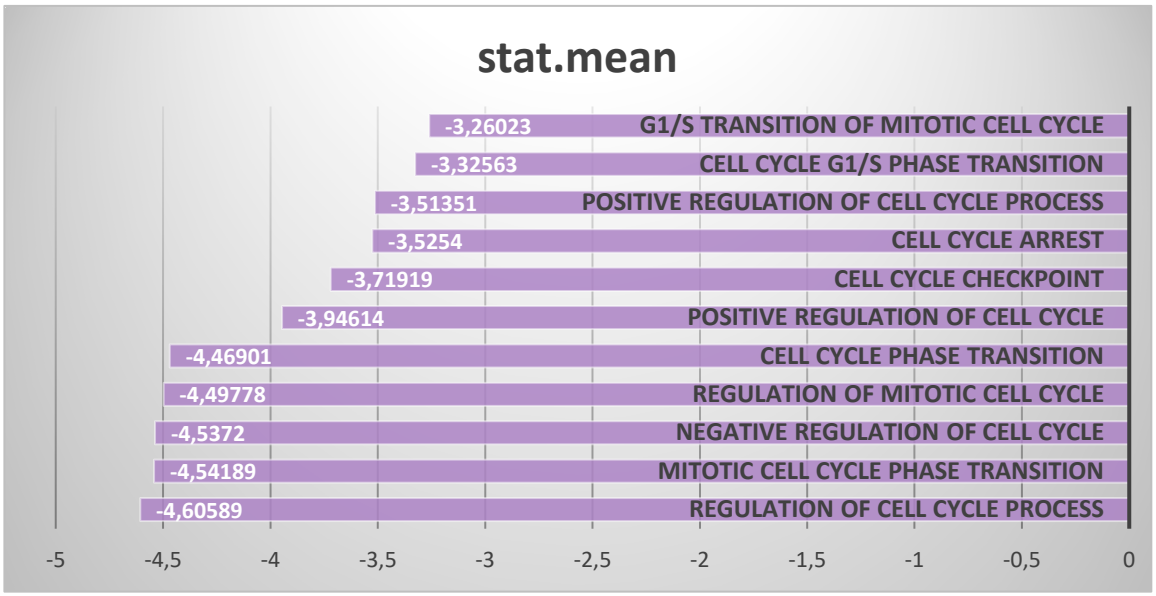


Figure 4.4: Statistically enriched gene sets representing several cell cycle related processes identified upon GSEA analysis of the differentially expressed genes comparing THP-1 TKTL1 knock-down vs parental cells.

As depicted in the results of the GSEA (Figure 4.4), TKTL1 might be implicated in cell-cycle regulation and phase transition, as the knock-down of TKTL1 resulted in the downregulation of several relevant Gene Ontology (GO) gene sets (summarised in Table 2). In addition, the lack of TKTL1 caused a decrease in the expression of ribose 5-phosphate isomerase A (RPIA) and Transketolase (TKT), as identified by the DE analysis. These last two enzymes belong to the PPP and are required for the synthesis of ribose-5-phosphate (R5P) through the non-oxidative branch, which is required for DNA and RNA synthesis and is known to

be regulated during cell-cycle. Li et al. have demonstrated how TKTL1 levels change throughout the cell-cycle and suggested that TKTL1 could be a positive regulator of R5P in both human embryonic kidney and human cervical cancer Hela cells [21]. To further investigate the implication of TKTL1 in the phase transition of the cell-cycle in AML we performed cell cycle analysis (Figure 4.5). The results showed less cells going through the G1 phase, almost the same number of cells in the S phase and increased number of cells in the G2 phase, after the knock-down of TKTL1.

In addition, the DE analysis shows an overexpression of the antiapoptotic gene BCL2 by a logFC=0.76, following the loss of TKTL1 and the GSEA identified a downregulation in gene sets related with apoptosis (Table 5Table 5: Statistically enriched down-regulated gene sets representing apoptosis events, as identified by the GSEA abalysis of the differentially enriched genes comparing THP-1 TKTL1 KD cells vs parental.), implying that the TKTL1 knockdown cells will be more resistant to apoptosis. In order to determine the effects of TKTL1 knockdown on apoptosis in AML cells, we quantified the cells going through the different stages of the programmed cell death and determined that overall less cells were in apoptosis (Figure 4.6). Therefore, evidence suggests that TKTL1 is implicated in the regulation of the cell cycle of AML cells by impairing phase transition and at the same time employing antiapoptotic mechanisms. We can also conclude that, in accordance with previous findings in different cancer types, the loss of TKTL1 inhibits cell proliferation. However, in our study this comes as a result of an induced arrest of the cells in the G2 phase of the cell cycle and not because TKTL1 knockdown impairs the viability of the cells.

Table 5: Statistically enriched down-regulated gene sets representing apoptosis events, as identified by the GSEA abalysis of the differentially enriched genes comparing THP-1 TKTL1 KD cells vs parental.

GO apoptosis gene sets	stat.mean	p-val	q-val
cell-type specific apoptotic process	-4.50584	5.77E-10	2.94E-08
regulation of apoptotic signaling pathway	-4.17991	9.71E-09	2.87E-07
negative regulation of apoptotic signaling pathway	-2.70764	0.000151	0.001003
extrinsic apoptotic signaling pathway	-2.63596	0.000234	0.00144

intrinsic apoptotic signaling pathway in response to DNA damage	-1.99011	0.003948	0.014404
apoptotic mitochondrial changes	-1.98196	0.004191	0.01502
leukocyte apoptotic process	-1.90726	0.005537	0.018575

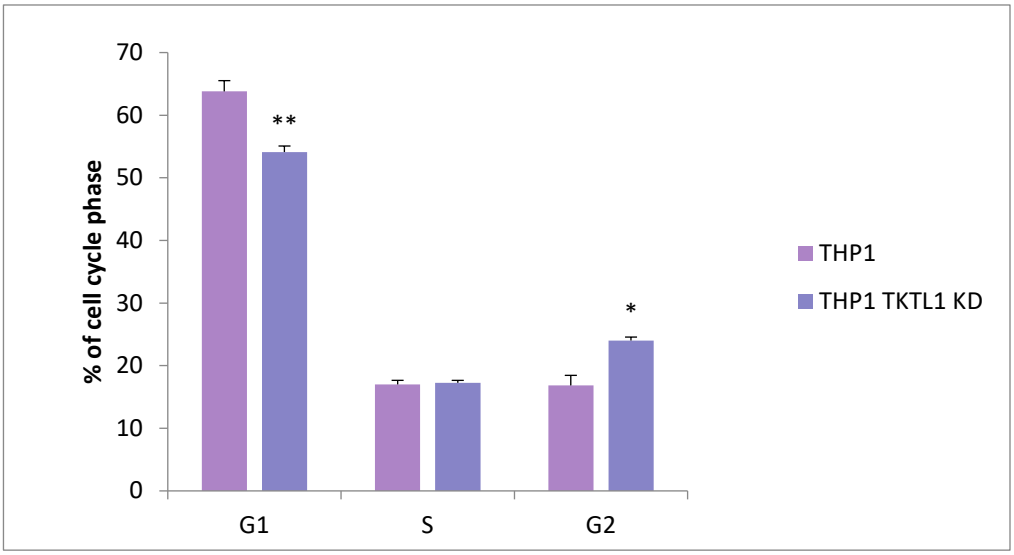


Figure 4.5: The results of the cell cycle analysis showed that the number of cells in G1 phase decreased, whereas more cells went through the G2 phase after the knock down of TKTL1, implying that TKTL1 inhibits the proliferation of AML cells.

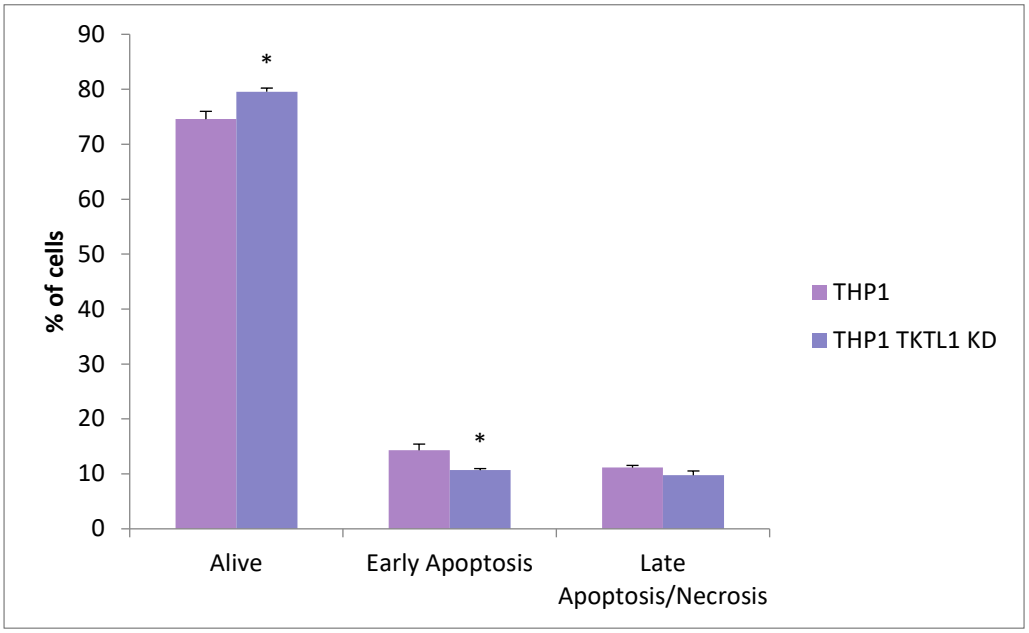


Figure 4.6: By performing an apoptosis assay we determined that the number of AML cells going through early and late apoptosis and necrosis decreased after the knock-down of TKTL1.

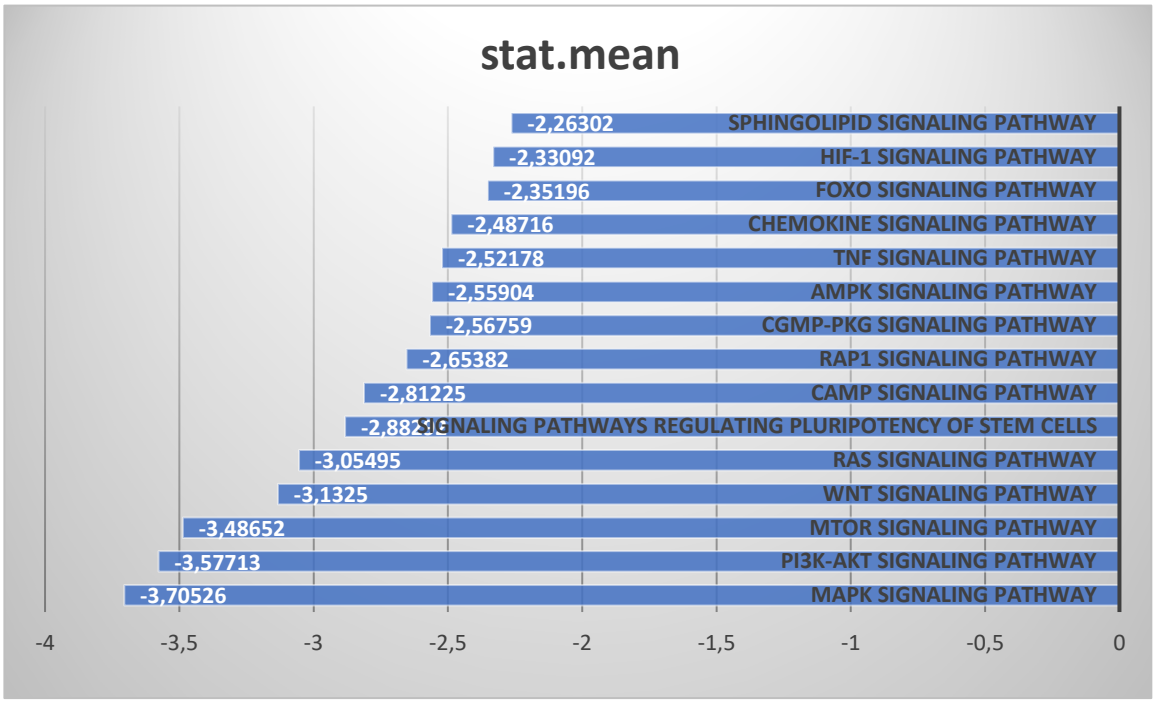


Figure 4.7: Statistically enriched gene sets representing signalling pathways identified upon GSEA analysis of the differentially expressed genes comparing THP-1 TKTL1 knock-down vs parental cells.

Following the DE analysis and the GSEA (Figure 4.7), a significant downregulation of several signalling pathways was evident, as a result of the knock-down of TKTL1. The most significantly enriched and affected pathways were the MAPK and the PI3K/AKT/mTOR signalling cascade, related with cell proliferation and cell cycle regulation, among others. Due to the crucial role of the PI3K/Akt/mTOR in cancer and specifically the increasing interest in the role of mTOR in AML and its association with a poor AML-

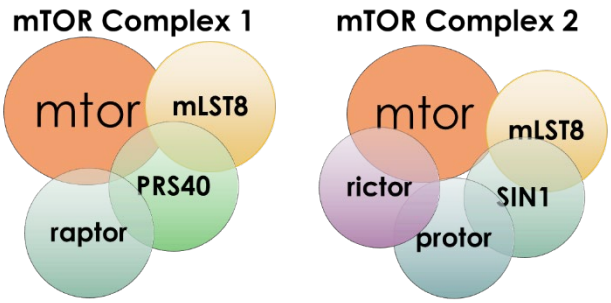


Figure 6: The mammalian target of rapamycin (mTOR) forms two distinct complexes, mTOR Complex 1 and mTOR Complex 2, that control different biological and cellular functions.

disease prognosis, we decided to investigate further the role of TKTL1 in these signalling events [23].

The mammalian target of rapamycin (mTOR) is a serine/threonine kinase that binds with other proteins to form two complexes mTOR Complex 1 (mTORC1) and mTOR Complex 2 (mTORC2), that perform different vital cellular processes (Figure 6).

Starting from mTOR Complex 1 (mTORC1), it was confirmed by Western Blot (WB) analysis that the protein levels of its major component, mTOR, were invariable between the KD cells and the Control, as predicted from the transcriptomics (Figure 4.9). To further validate the activity of mTORC1 the cells

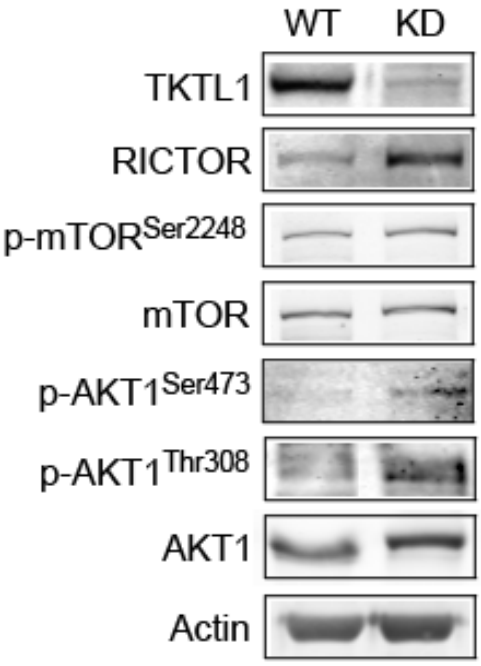


Figure 4.9: Western Blot analysis of important members of mTOR.

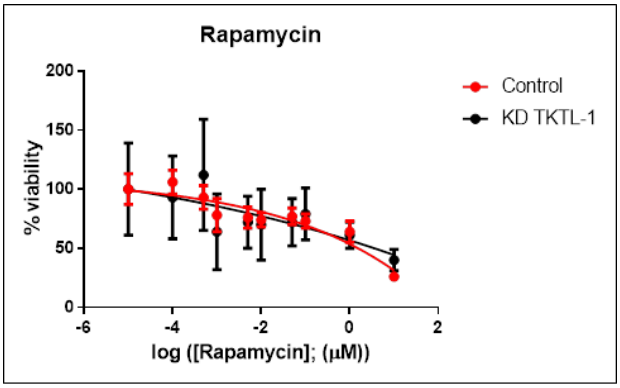


Figure 4.8: IC50 assay of the effect of Rapamycin on THP-1 parental and TKTL1 KD cells.

were treated with Rapamycin, an inhibitor of mTORC1. After short-term treatment, Rapamycin had no effect on the viability of the TKTL1 knockdown cells as expected (Figure 4.9), confirming that mTORC1 was inactive.

However, regarding mTROC2 and despite the downregulation in gene expression, it was determined by WB analysis that RICTOR protein, the main member of this complex, was more abundant in the KD cells (Figure 4.9). In addition, both key residues of Akt, T308 and S473, were phosphorylated; thus, Akt was activated after the knock down of TKTL1. One of the first reported and most vital substrates of Akt is the Glycogen Synthase Kinase 3 (GSK3), which in turn regulates various substrates related to cell survival, proliferation and metabolism and subsequently induces their inhibition or proteasomal degradation [24]. The phosphorylation of Akt leads to the phosphorylation of GSK3 that renders the binding pocket of this protein inaccessible to its substrates. It has been shown that RICTOR protein is regulated by GSK3-dependent degradation [25]. Therefore, the downregulation of GSK3 (logFC=-0.75) and its phosphorylation stabilised RICTOR protein, activated mTORC2 and facilitated the survival of AML cells.

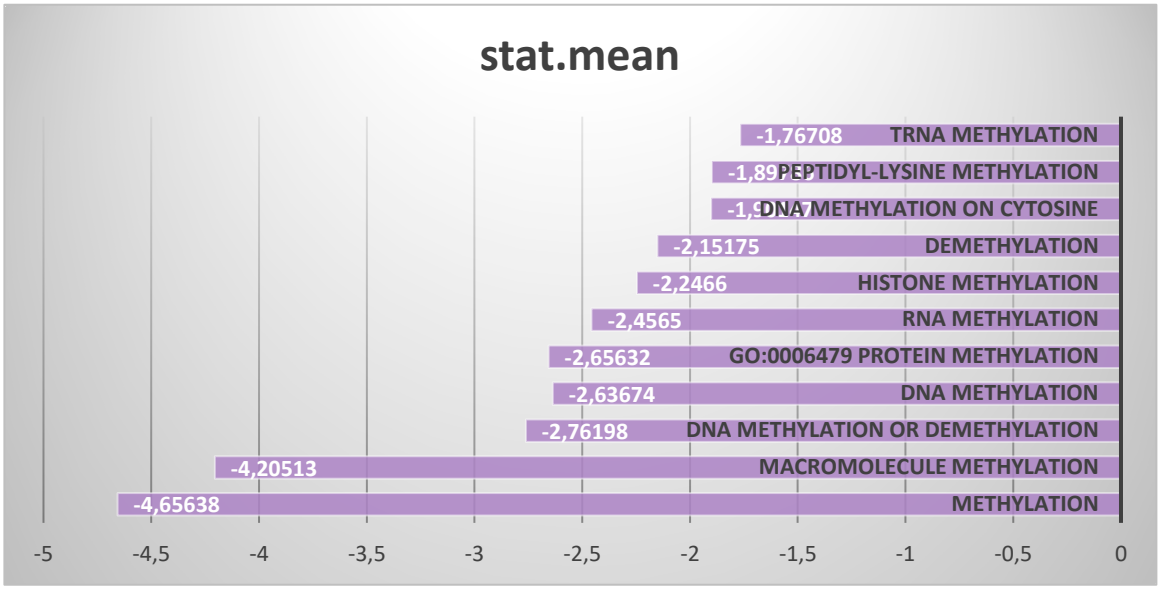


Figure 4.10: Statistically enriched gene sets representing epigenetic regulation events, identified upon GSEA analysis of the differentially expressed genes comparing THP-1 TKTL1 knock-down vs parental cells.

Based on the GSEA analysis performed on GO terms (Figure 4.10), we also identified significant enrichment methylation related gene sets, as a result of TKTL1 knock down. Epigenetic control of gene expression and specifically DNA methylation are vital regulatory events in the cell. More specifically, DNA methylation entails the addition of methyl groups to a DNA molecule with the purpose of controlling gene expression, without any change in the actual DNA sequence. The disruption of the normal DNA methylation patterns has been identified as a hallmark of cancer [26]. Over the last years, numerous studies have been exploring the methylation landscape in AML, unveiling not only the importance of epigenetics in the pathogenesis of this disease, but also prognostic biomarkers and promising therapeutic strategies based on methylating agents [27-30].

In addition, Jayachandran et al. , have demonstrated an association between TKTL1 and DNA hypomethylation in Chronic Myeloid Leukaemia (CML) [31]. Therefore, we hypothesized that TKTL1 may be involved in promoter methylation, suppressing the expression of several genes in AML. After a literature search, we

constructed the current list of all the genes whose expression has been reported to be epigenetically regulated by promoter methylation in AML and compared it with the results of our DE analysis[233, 235, 236]. Interestingly, genes controlled by promoter hypomethylation in AML were significantly upregulated after the loss of TKTL1 (Table 6), suggesting that the KD of this gene has the same effect on AML cells as the administration of hypomethylation agents.

Table 6: Several genes whose expression is known to be regulated by promoter hypomethylation in AML were found to be upregulated after the knock down of TKTL1 in THP-1 cells.

Gene name	log2fc	p-value	adj. p-value
NPAS2	0.54444	0.000207	0.00461
ELAVL2	0.664429	0.005276	0.043256
RNASE2	0.69321	0.0002	0.004525
BCL9L	0.753513	0.000182	0.00423
SCRN1	1.189047	3.78E-05	0.00156
CLEC11A	1.497442	4.07E-07	0.000118
PGR	2.539477	6.67E-07	0.000162
CNR1	3.465268	1.18E-08	1.70E-05
TIMP3	4.7578495	9.02E-10	5.38E-06

Mei et al. have demonstrated the synergetic effect of hypomethylating agents with Venetoclax in AML [32]. Based on the predictions of the combined computational DE and GSEA analysis that revealed a developed mechanism of TKTL1 knock-down cells to avoid apoptosis by overexpressing the antiapoptotic gene BCL2 in combination with the hypomethylating function of TKTL1, we hypothesized that TKTL1 KD cells would be more sensitive to treatment with Venetoclax. Indeed, the antiproliferative effect of administering Venetoclax to the THP-1 cells was distinctively more evident in the TKTL1 KD than in the Control cells (Figure 4.11).

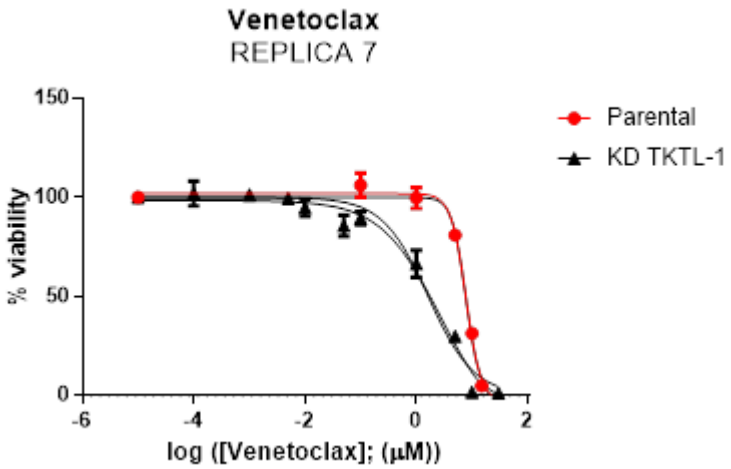


Figure 4.11: Venetoclax treatment seriously compromised the viability of AML cells carrying a TKTL1 knock down, making it a promising putative therapeutic strategy.

4.1.2.5 Genome Scale Metabolic Modelling and Gene Essentiality Analysis

The next step in our integrative approach was the reconstruction of Genome Scale Metabolic Models (GSMMs) of the THP-1 Control and TKTL1 KD cells, in order to explore in greater depth the metabolic reprogramming of the cells caused by the loss of TKTL1, as well as to perform high-throughput gene knock-out simulations and determine metabolic vulnerabilities of the AML cells. Recon 2.2 [22] was used as the base model for the reconstructions and the available experimental data were introduced as flux constraints to create condition specific reconstructions. More specifically, the measured consumption and production rates of glucose, lactate, glutamine and glutamate, and other metabolites present in the media were used in the first step of constraining the reconstructions. Then, the gene expression levels were integrated and THP-1 Control and THP-1 KD specific reconstructions were built (see Materials and Methods section 3.6).

A gene essentiality analysis was performed on both reconstructions using FBA. More specifically, simulations were run testing one by one the knock-down of all the genes in the reconstruction and how the loss of each gene affected the biomass production in the models. Finally, a gene was considered essential if the simulated biomass dropped at less than 10%, after the knock-out of that gene.

Our analysis identified 25 essential genes that could potentially constitute metabolic vulnerabilities for the viability of the THP-1 Control cells (Table 7). The results included enzymes catalysing important reactions for sphingolipid biosynthesis, pyrimidine biosynthesis, de novo purine synthesis and purine metabolism. In addition, thymidylate synthase and several enzymes catalysing reactions in the folate biosynthesis and the one-carbon folate metabolism pathways were also determined as crucial. Therefore, AML THP-1 cells rely prominently on nucleotide synthesis and methylation reactions for survival.

Table 7: Gene essentiality analysis from the systematic knock-out simulations performed based on the GSMM reconstructions identified 25 essential genes for the THP-1 parental cells.

gene_symbol	reaction_id	reaction_name
ATIC	AICART, IMPC	phosphoribosylaminoimidazolecarboxamide formyltransferase, IMP cyclohydrolase
BCAT1	ILETA, VALTA, LEUTA	isoleucine transaminase, valine transaminase, leucine transaminase
BCAT2	ILETAm, VALTAm, LEUTAm	isoleucine transaminase, mitochondrial, valine transaminase, mitochondiral, leucine transaminase, mitochondrial
CAD	ASPCTr, CBPS, DHORTS	aspartate carbamoyltransferase (reversible), carbamoyl-phosphate synthase (glutamine-hydrolysing), dihydroorotase
SLC25A10	r0819, MALSO4tm, r0836, r0829, FUMSO4tm, r0834, FUMtm, FUMTSULtm, r0821, MALtm, r0885, SUCCt2m, r2420, r0830, r0822, FUMSO3tm, MALTSULtm, MALSO3tm, r2419, r0835, AKGMALtm	fumarate transport, mitochondrial, succinate transport, mitochondrial
DHFR	r0514, r0512, DHFR, r0224, r0226, FOLR2	Dihydrofolate One carbon pool by folate, Folate biosynthesis
DHODH	DHORD9	dihydoorotic acid dehydrogenase (quinone10)
KDSR	3DSPHR	3-Dehydrosphinganine reductase
GART	PRAGSr, r0666, GARFT	phosphoribosylglycinamide synthase
LCAT	LCAT1e	Lecithin-cholesterol acyltransferase
LIPA	r1172, r1177, CHOLESTle, r1179	hydrolysis of cholesterol ester by cholesterol esterase
PFAS	PRFGS	phosphoribosylformylglycinamidine synthase
PPAT	GLUPRT	glutamine phosphoribosyldiphosphate amidotransferase
TYMS	TMDS	thymidylate synthase

UMPS	OMPDC, ORPT	orotidine-5-phosphate decarboxylase, orotate phosphoribosyltransferase
PGS1	PGPPT	glycerophosphate phosphatidyltransferase
SPTLC2	SERPT	serine C-palmitoyltransferase
SPTLC1	SERPT	serine C-palmitoyltransferase
PAICS	AIRCr, PRASCS	phosphoribosylaminoimidazole carboxylase, phosphoribosylaminoimidazolesuccinocarboxamide synthase
RPIA	RPI, r0249	ribose-5-phosphate isomerase, D-Ribose-5-phosphate ketol-isomerase Pentose phosphate pathway
CMPK1	CYTK7n, CYTK9, CYTK11, CYTK2n, CYTK13n, CYTK9n, UMPK6n, CYTK10, CYTK1n, UMPK2, CYTK5n, CYTK10n, UMPK4, UMPK7, CYTK3n, CYTK14n, CYTK12n, CYTK2, UMPK2n, CYTK4n, UMPK3, UMPK5, CYTK6, CYTK8n, CYTK6n, CYTK11n, CYTK7, CYTK8, UMPK4n, UMPK3n, CYTK12, UMPK, CYTK1, UMPK7n, UMPK6, UMPK5n, CYTK5, CYTK14,	cytidylate kinase (dCMP)

	CYTK13, UMPKn	
CRLS1	CLS_hs	cardiolipin synthase
SPTLC3	SERPT	serine C-palmitoyltransferase
PTPMT1	PGPP_hs	Phosphatidylglycerol phosphate phosphatase
SGMS1	SMS	Sphingomyelin synthase

Then, the same analysis was performed for THP-1 TKTL1 KD and 59 genes were identified as essential for these cells. Interestingly, the 25 essential genes identified in the THP-1 parental reconstruction were included in the results for the THP-1 TKTL1 KD. Thus, 34 specific metabolic vulnerabilities for the KD cells were proposed by our method (Table 8). In greater detail, the mitochondrial isoenzyme of fumarate hydratase, which catalyses a reaction in the Tricarboxylic Acid Cycle (TCA), was identified as a potential target for the TKTL1 KD cells. Moreover, some subunits of the mitochondrial ATP synthase, involved in oxidative phosphorylation, were also identified, together with all the subunits forming the ubiquinol-cytochrome c reductase (or Complex III) and the cytochrome c oxidase (or Complex IV) of the respiratory chain. In conclusion, a clear dependency of AML cells on TCA, respiration, and oxidative phosphorylation, all of which are coupled processes for production of energy in the form of ATP, was illustrated by the analysis.

Table 8: the gene essentiality analysis performed on the THP-1 TKTL1 KD GSMM resulted in the identification of 34 condition specific metabolic vulnerabilities, that could be targeted as a combination therapy in AML.

gene_symbol	reaction_id	reaction_name
ATP5F1B	ATPS4m	ATP synthase (four protons for one ATP)
ATP5ME	ATPS4m	ATP synthase (four protons for one ATP)
COX4I1	CYOOm3	cytochrome c oxidase, mitochondrial Complex IV
COX5B	CYOOm3	cytochrome c oxidase, mitochondrial Complex IV
COX6A1	CYOOm3	cytochrome c oxidase, mitochondrial Complex IV
COX6A2	CYOOm3	cytochrome c oxidase, mitochondrial Complex IV

COX6B1	CYOOm3	cytochrome c oxidase, mitochondrial Complex IV
COX6C	CYOOm3	cytochrome c oxidase, mitochondrial Complex IV
COX7A1	CYOOm3	cytochrome c oxidase, mitochondrial Complex IV
COX7A2	CYOOm3	cytochrome c oxidase, mitochondrial Complex IV
COX7B	CYOOm3	cytochrome c oxidase, mitochondrial Complex IV
COX7C	CYOOm3	cytochrome c oxidase, mitochondrial Complex IV
COX8A	CYOOm3	cytochrome c oxidase, mitochondrial Complex IV
CYC1	CYOR_u10m	ubiquinol-6 cytochrome c reductase, Complex III
FH	FUM, FUMm	fumarase, fumarase, mitochondrial
MT-CO1	CYOOm3	cytochrome c oxidase, mitochondrial Complex IV
MT-CO2	CYOOm3	cytochrome c oxidase, mitochondrial Complex IV
MT-CO3	CYOOm3	cytochrome c oxidase, mitochondrial Complex IV
MT-CYB	CYOR_u10m	ubiquinol-6 cytochrome c reductase, Complex III
NSF	ATPS4m	ATP synthase (four protons for one ATP)
UQCRB	CYOR_u10m	ubiquinol-6 cytochrome c reductase, Complex III
UQCRC1	CYOR_u10m	ubiquinol-6 cytochrome c reductase, Complex III
UQCRC2	CYOR_u10m	ubiquinol-6 cytochrome c reductase, Complex III
UQCRFS1	CYOR_u10m	ubiquinol-6 cytochrome c reductase, Complex III
UQCRH	CYOR_u10m	ubiquinol-6 cytochrome c reductase, Complex III
COX7A2L	CYOOm3	cytochrome c oxidase, mitochondrial Complex IV
COX5A	CYOOm3	cytochrome c oxidase, mitochondrial Complex IV
ATP5MG	ATPS4m	ATP synthase (four protons for one ATP)
UQCR11	CYOR_u10m	ubiquinol-6 cytochrome c reductase, Complex III
UQCRQ	CYOR_u10m	ubiquinol-6 cytochrome c reductase, Complex III

[107]

COX4I2	CYOOm3	cytochrome c oxidase, mitochondrial Complex IV
COX6B2	CYOOm3	cytochrome c oxidase, mitochondrial Complex IV
COX7B2	CYOOm3	cytochrome c oxidase, mitochondrial Complex IV
COX8C	CYOOm3	cytochrome c oxidase, mitochondrial Complex IV

4.1.3 Discussion

Transketolase-like 1 (TKTL1) is a key enzyme of the non-oxidative branch of glycolysis. Although still a lot is unknown about its biological and enzymatic function, TKTL1 has been linked to various different cancer types, as an important element in their pathogenesis or as a prognostic marker. The overexpression of this gene in Acute Myeloid Leukaemia (AML) patients suggested a putative role of TKTL1 in this highly aggressive and heterogeneous disease, that has not been investigated so far.

Our study provides a detailed bioinformatics analysis of transcriptomic data, as well as the reconstruction of a genome scale metabolic model (GSMM) of AML cells integrating gene expression and metabolomics as a tool to explore the function of TKTL1 in AML *in vitro* and to identify potential drug targets that could be used in combination with the inhibition of TKTL1 as a new therapeutic approach.

The Gene Set Enrichment Analysis (GSEA) revealed an implication of TKTL1 in cell cycle and phase transition in AML, which was experimentally validated. The loss of TKTL1 decreased the proliferation rate of AML cells, but did not compromise their viability. As predicted by the GSMM-based simulations, TKTL1 KD cells presented an increased flux through purine and pyrimidine *de novo* biosynthesis pathways, which facilitated their transition to the S phase and DNA replication.

Interestingly, the knock down of TKTL1 triggered a signalling cascade and a mechanism of survival in AML cells by activating AKT, which in turn stabilised RICTOR protein and activated mTOR Complex 2. This regulatory loop controls cell survival by blocking GSK3-mediated degradation of antiapoptotic Myeloid Cell Leukaemia 1 (Mcl1) protein, among others [30]. At the same time, TKTL1 knock down led to an overexpression of BCL2, a major antiapoptotic regulator in the cell.

Furthermore, the high-throughput simulations based on the TKTL1 knock down GSMM revealed a major dependency of the cells to components of the TCA, respiratory chain and oxidative phosphorylation. Targeting respiration would result in releasing cytochrome c oxidase outside the mitochondria which in turn would

trigger apoptosis. In fact, the use of oxidative phosphorylation inhibitors in cancer and specifically in leukaemia is considered as a very promising therapeutic strategy [33].

Finally, our integrative analysis of transcriptomic and metabolomic data revealed an implication of TKTL1 in methylation events. The importance of epigenetic regulation in AML has been well established and the methylation landscape of this malignancy is the focus of thorough research. The results from our analysis imply a direct role of TKTL1 in DNA methylation. The knock down of TKTL1 led to the upregulation of several genes regulated by promoter hypomethylation in AML and caused a significant flux increase through one Carbon metabolism and the folate cycle, which contribute to DNA methylation [34]. Therefore, we propose that TKTL1 can control hypomethylation, since its knock down had the same effect on AML cells as the administration of hypomethylating agents. Mei et al., had previously stated the synergetic effect of BCL2 inhibitor Venetoclax with hypomethylation agents in AML [32]. Indeed, the combination of TKTL1 inhibition and Venetoclax reduced dramatically the viability of AML cells, proving the methylation attributes of TKTL1. Moreover, this approach might have significant therapeutic implications not only in AML, but in other cancer types that exhibit high TKTL1 expression as well.

4.2 Tackling drug resistance in Acute Myeloid Leukaemia

4.2.1 Introduction

Numerous different therapeutic approaches are employed in the battle against cancer, including different types of surgery, radiation therapy, immunotherapy, hormone and gene therapy and finally chemotherapy. The latter still remains one of the most common treatments up until today [237]. Some chemotherapeutics derive from specific types of plants and others are synthetized [238-241] and can be broadly categorized into six major groups: alkalyting agents, antimetabolites, anthracyclins, topoisomerase inhibitors, plant alkaloids and corticosteroids. In the case of AML, the use of several drugs has been approved, but the two most prevalent are cytarabine (cytosine arabinoside or AraC) and doxorubicin (DOX) [242].

AraC is an antimetabolite drug that acts as a pyrimidine antagonist. It consists of the nucleobase cytosine attached to an arabinose sugar, making it analogous to cytidine. In this way, once inside the cell, AraC can take the place of cytidine and be incorporated into DNA during the S phase of the cell cycle and stop DNA replication because the arabinose sterically hinders the rotation of the molecule within DNA [243]. In addition, AraC further interferes with DNA repair and synthesis by inhibiting DNA and RNA polymerases [244].

On the other hand, DOX is an anthracycline drug that was first extracted from the bacterium *Streptomyces peucetius var. caesius* 50 years ago [245]. DOX has different mechanisms of action, that are still not fully defined and remain the interest of many clinical studies [246]. DOX can intercalate into DNA and interfere with transcription and replication by causing a non-repairable DNA double-strand break, leading the cells to apoptosis [247, 248]. It also inhibits topoisomerase II, which blocks DNA transcription, replication and repair processes and induces apoptosis [248]. Additionally, it has been reported that once it enters the cell and in the presence of several oxidoreductases like NADH dehydrogenase, DOX gets oxidised to its semiquinone radical form, which is then re-oxidised back to DOX through a reaction that results in the formation of ROS [249]. Elevated levels of

ROS can eventually cause DNA damage, oxidative stress, lipid peroxidation and membrane damage [250].

Although great advancement is being made in optimizing existing chemotherapeutics and developing new ones, drug resistance to practically all types of therapeutics still poses the greatest threat against the successful treatment of cancer. In fact, statistical reports show that 90% of cancer patients deaths are due to drug resistance [251]. Two types of resistance to chemotherapeutics can be observed: intrinsic, that is pre-existing resistance to an administered drug, or acquired drug resistance, which refers to the ability that the cancerous cells develop to adapt and survive the treatment. The exact mechanisms of drug resistance are reportedly complex and not well understood [251, 252]; thus, arising the imperative need to find a way to overcome resistance to anti-cancer drugs.

Metabolism has not only been established as a hallmark of carcinogenesis, but several studies have recognised its importance in the emergence and instatement of drug resistance in different types of cancer [253-255]. Our aim in this chapter was to explore the metabolic rewiring that occurs in AML cells after they become resistant to AraC and DOX and identify putative metabolic vulnerabilities that could serve as drug targets against the resistant cells.

Towards that end we designed a novel bioinformatics pipeline to integrate a variety of omic data under the framework of GSMMs. First, we performed an extensive transcriptomic characterisation of two AML cell lines, the THP-1 and the HL60, and identified specific gene expression patterns, revealing potential mechanisms that drive the establishment of resistance. Then, we integrated the extensive metabolomic data generated from these cell lines with the transcriptomics and introduced measurements of oxygen consumption rates of these cells. By doing so, we were able to build condition specific GSMM reconstructions of AML parental and drug resistant cells. Finally, we interrogated these reconstructions towards metabolic alterations taking place across the whole

of cellular metabolism and identified druggable metabolic vulnerabilities to tackle AML drug resistance.

4.2.2 Results

4.2.2.1 Analysis of transcriptomic data

First, a thorough computational analysis was performed on the transcriptomic data collected from the THP-1 and the HL60 AML cell-lines. The appropriate processing of the RNA seq data and the subsequent statistical analyses were performed in order to identify the most significant changes in gene expression comparing the AraC- and DOX-resistant vs the parental AML cell-lines . GSEA was applied to provide deeper insight on patterns of gene expression changes, before and after the AML cells become resistant to the drugs. To identify the differentially expressed genes between the parental untreated THP-1 and HL60 (control) and the drug resistant cell lines, RNA-seq analysis was performed. In total, 18 samples were sequenced and further analysed (three samples for each cell line grown in the same cell culture conditions, see Materials and Methods section 3.1).

4.2.2.1.1 Pre-processing of RNA-seq data

The quality control check was performed using FastQC on each of the 18 fastq files containing the sequenced samples, as generated by the Illumina sequencing platform (see Materials and Methods section3.5). The results of the analysis were very consistent across all the files. Therefore, to be concise, only the results generated for the quality of the sequenced sample for replicate 1 of the THP1 Control cell line will be presented here. This was a 61 base pair single-end read experiment. Very high-quality scores were calculated per base and per sequence, indicating that there was no need for trimming or filtering out any part of the sequences.

After the sequenced reads were aligned to the reference genome sequence (see Materials and Methods section 3.5), the counts table was generated, comprising of the number of times a specific gene was sequenced in each sample for both cell lines. To assess whether any of the sequenced samples were contaminated in any way, rendering them inappropriate for further inclusion in the

analysis, the distance matrix of the samples was generated. First, since the data in the RNA-seq counts table are not homoscedastic, a regularised-logarithm (rlog) transformation was applied, so that the subsequent clustering of the samples would be not affected by the variance introduced by the mean of the gene counts. Then, the Euclidian distance between the samples was calculated and a hierarchical clustering of the samples was performed (Figure 4.12) [256]. The samples clustered together as expected, according to cell line, THP-1 and HL60, and to the different treatment conditions, namely AraC-resistant, DOX-resistant and the Parental-untreated conditions.

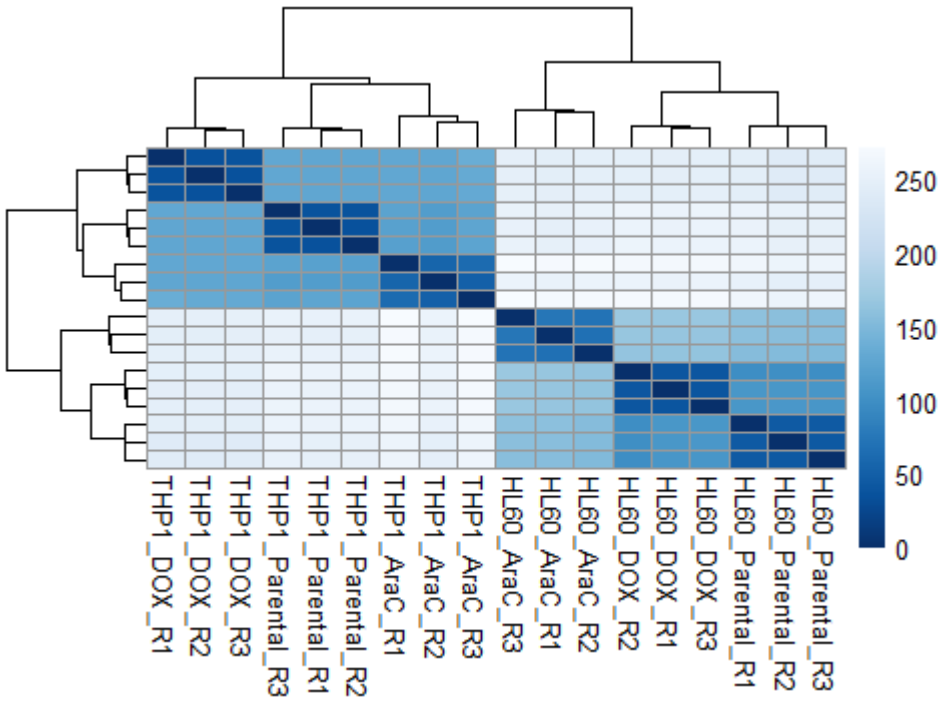


Figure 4.12: The hierarchical clustering based on the Euclidian distance of the sequenced samples confirmed the high quality of the transcriptomic data.

4.2.2.1.2 Unveiling gene expression patterns in drug-resistant AML cell lines

The differential expression (DE) analysis was performed comparing gene expression in the drug resistant cells vs the non-treated parental cells, in both cell lines. More specifically, differentially expressed genes were identified between AraC-resistant vs Parental THP-1 cells, DOX-resistant vs Parental THP-1 cells, Arac-resistant vs Parental HL60 cells and DOX-resistant vs Parental HL60 cells.

Differentially expressed genes were selected according to the following thresholds regarding the binary logarithm of the fold change (log2FC) in gene expression and its statistical significance measured by a p-value corrected for multiple testing (adj. p-value): $|log2FC| \geq 3$ and adj. p-value<10e-4. The results from the DE analysis were subjected to Gene Set Enrichment Analysis (GSEA), considering gene sets from the Gene Ontology (GO) and KEGG databases related to biological processes, metabolic and signalling pathways. The log2FC in gene expression was used as the criteria to rank the genes for the GSEA analysis. Significantly enriched pathways were identified both in down and up-regulation. The False Discovery Rate adjusted p-values (or q-values), which in principal depict the corrected for multiple testing p-values, were used to filter the results. Only gene sets/pathways with a calculated q-value<0.25 were considered significant.

The acquisition of resistance to AraC in both THP-1 and HL60 cells caused significant changes in gene expression. In the case of the THP-1 AraC resistant cells compared to the untreated parental cells, the DE analysis identified 455 differentially expressed genes, 378 of which were downregulated and 77 were upregulated (Figure 4.13).

THP-1 AraC-resistant cells exhibited a downregulation of gene terms related to DNA replication and gene expression processes (e.g. ‘GO: RNA processing’, ‘GO: posttranscriptional regulation of gene expression’, ‘GO: translation’), as well as terms related to mitosis (e.g. ‘GO: mitotic cell cycle’, ‘GO: mitotic cell cycle process’, ‘GO: chromosome segregation’), describing the anticipated mechanism of action of AraC on the cells (Figure 4.14).

On the other hand, the establishment of resistance to AraC triggered a significant upregulation of several signal transduction related events in THP-1 cells (i.e. ‘GO: G-protein coupled receptor signaling pathway’, ‘KEGG: Rap1 signaling pathway’, ‘KEGG: PI3K-Akt signaling pathway’, ‘KEGG: MAPK signaling pathway’ and ‘KEGG: HIF-1 signaling pathway’) (Figure 4.14). The upregulation of components of the G-protein coupled receptor and Rap1 signalling pathways leads to the activation of MAPK and the PI3K-Akt pathways, major regulators of cell

proliferation and survival and one of the documented mechanisms of drug resistance in AML [257, 258].

After analysing gene expression changes in the HL60 AraC-resistant cells, 632 differentially expressed genes were identified, 177 of which were downregulated and 455 were upregulated (Figure 4.15). In a similar manner, AraC resistance in HL60 cells seems to regulate gene expression, but in this case by affecting post-translational regulation, since ubiquitination was significantly downregulated (e.g. ‘GO: protein ubiquitination’, ‘GO: ubiquitin-dependant protein catabolic process’) (Figure 4.16).

HL60 cells resistant to AraC compared to the untreated Parental cells exhibited an upregulation in several metabolic pathways (Figure 4.16). Increased fatty acid and glycerolipid metabolism (‘KEGG: Fatty acid metabolism’,’ KEGG: Glycerolipid metabolism’) have been positively correlated with AML progression and disease prognosis [259]. In addition, oxidative phosphorylation (‘KEGG: Oxidative phosphorylation’) has already been observed in AraC resistant AML cells and

targeting oxidative phosphorylation (OXPHOS) has been reported to improve the efficacy of the drug [260].

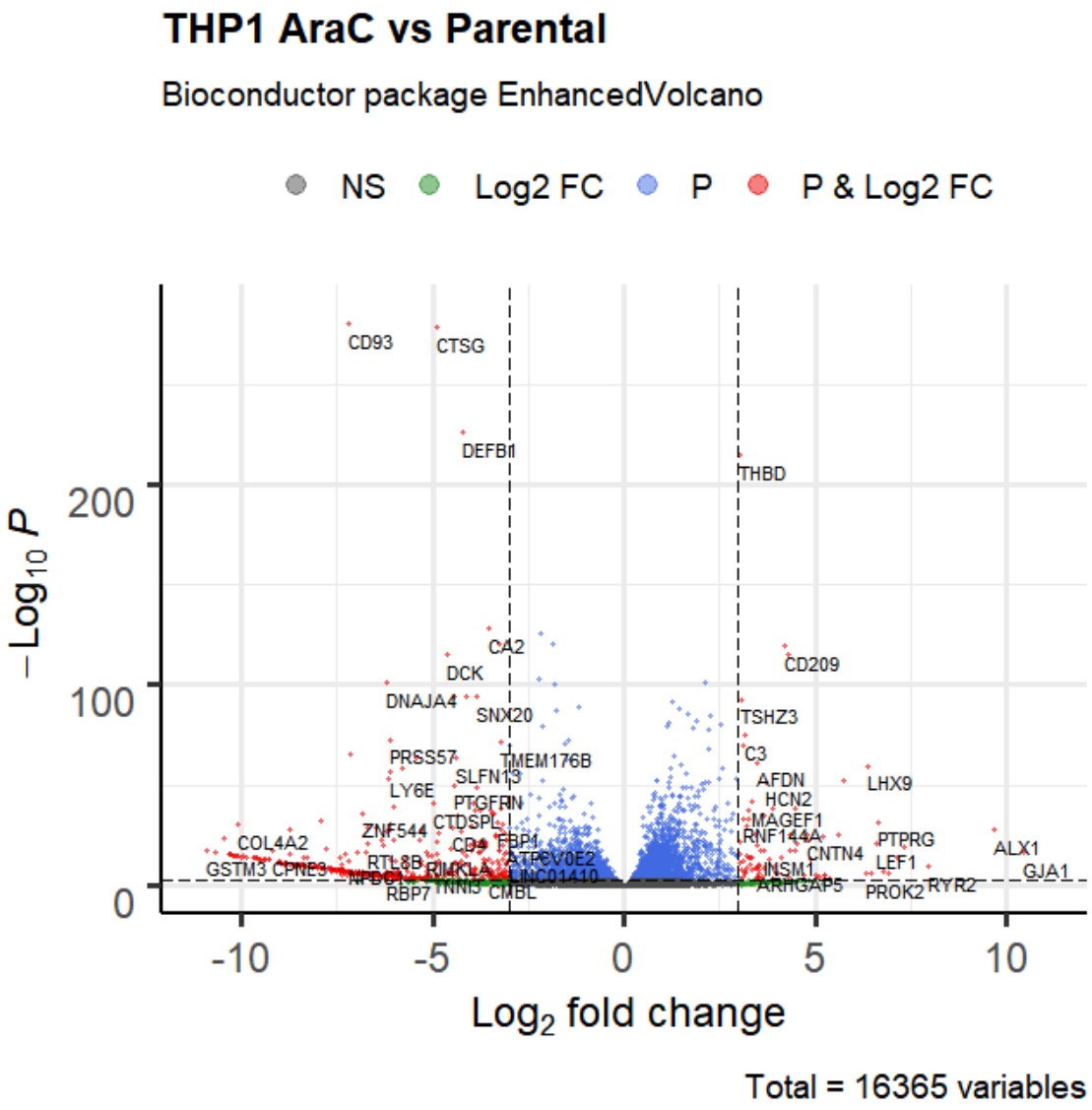


Figure 4.13: Volcano plot of the differentially expressed genes in the THP-1 cell line, comparing the AraC-resistant cells vs the untreated parental cells. The logarithm base 2 of the fold change in gene expression (x-axis) is plotted against the logarithm transformed adjusted p-value of each observation (y-axis). The grey coloured points represent genes that are not significant (NS) according to the applied thresholds for gene expression |log2 fold change|≥3 and statistical significance adjusted p-value<10e-4. The green points represent genes that satisfy only the gene expression threshold. Likewise, the blue points represent genes that satisfy only the statistical significance threshold. The genes that appear as red points are the statistically significant up- and down-regulated genes that are of interest in this study.

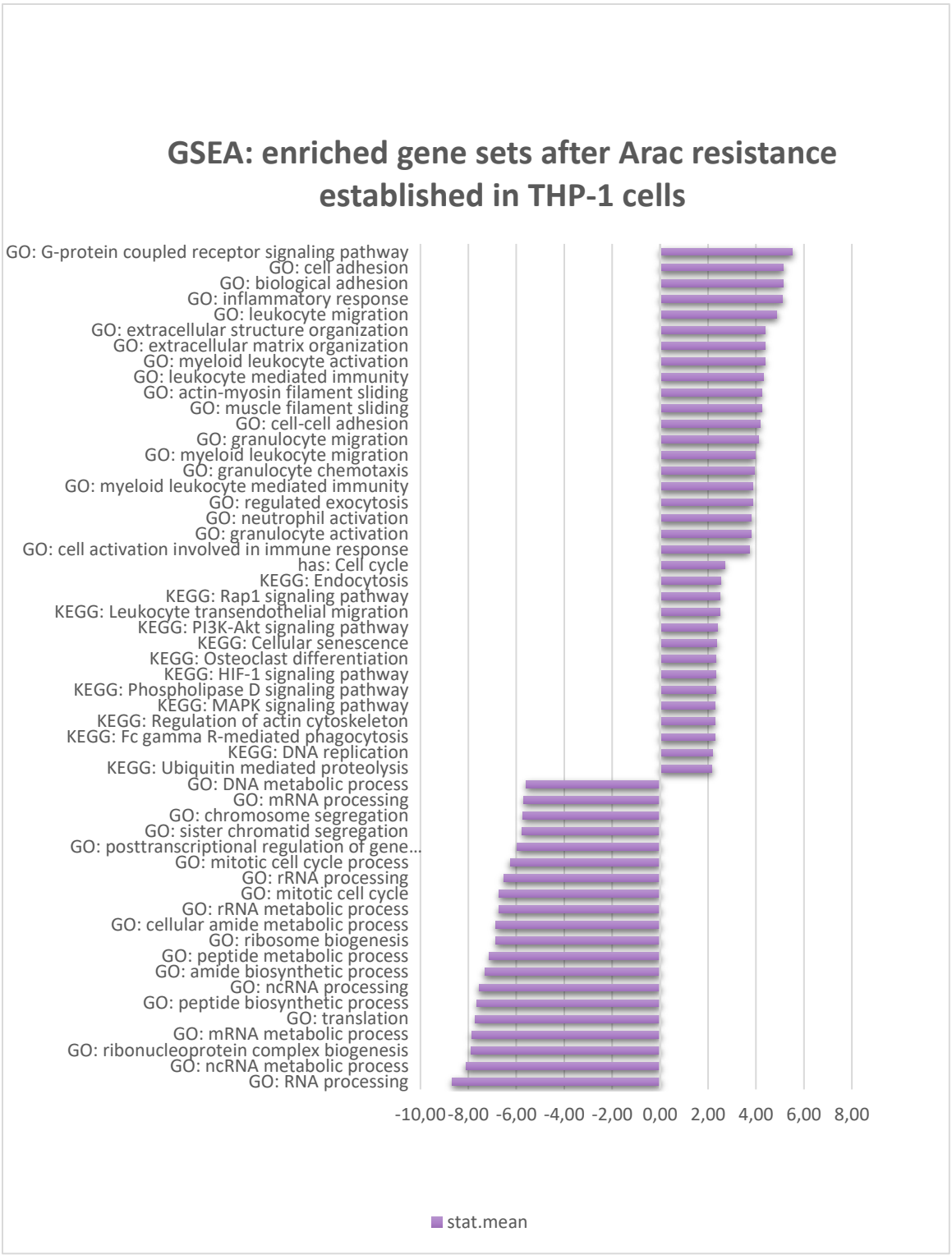


Figure 4.14: Figure summarises the results from the GSEA analysis performed on the differentially expressed genes comparing AraC-resistant vs Parental THP-1 cells. The analysis considered gene sets from the Gene Ontology (GO) and the KEGG databases, including sets describing biological processes, metabolic and signalling pathways.

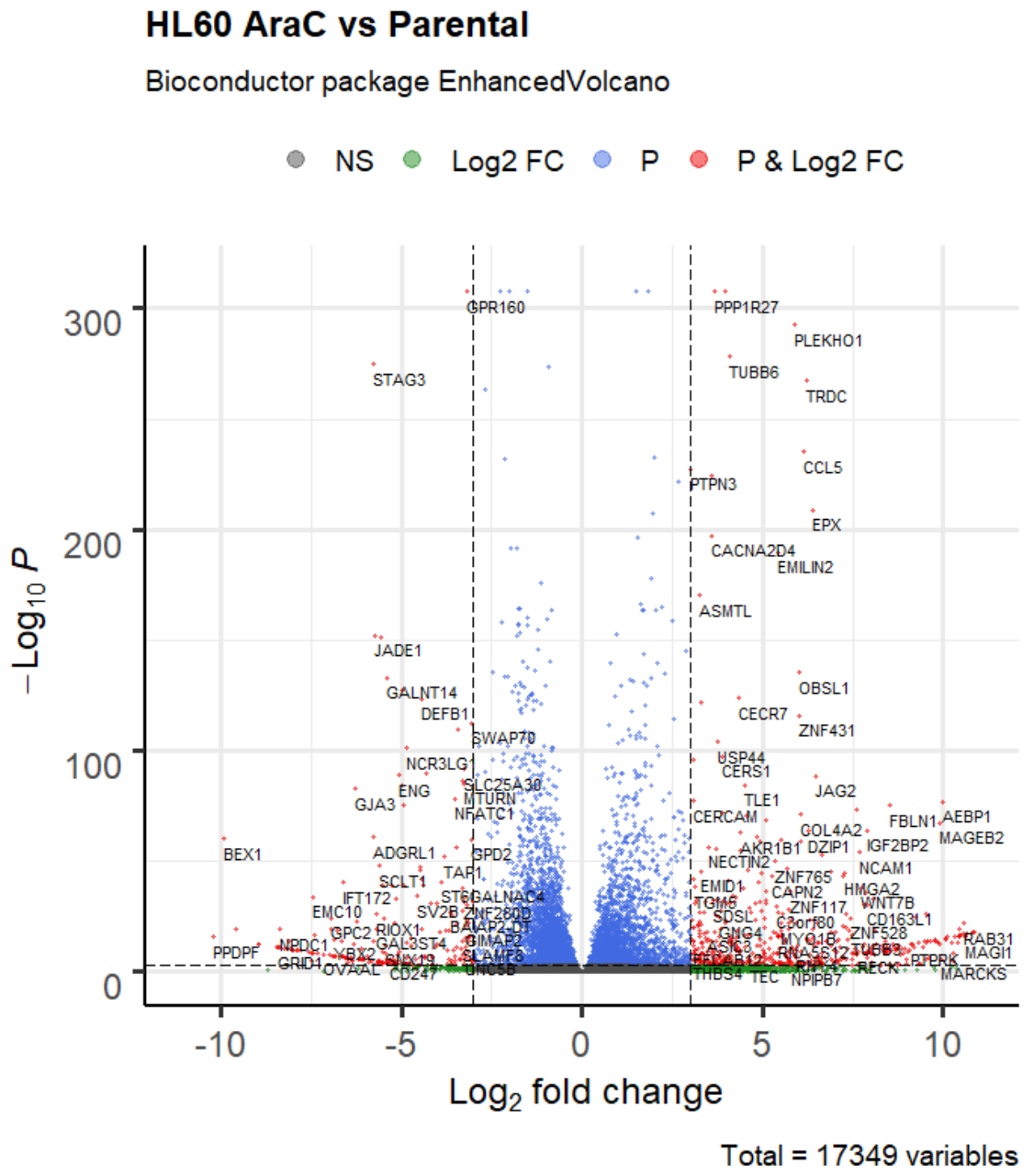


Figure 4.15: Volcano plot of the differentially expressed genes in the HL60 cell line, comparing the AraC-resistant cells vs the untreated parental cells. The logarithm base 2 of the fold change in gene expression (x-axis) is plotted against the logarithm transformed adjusted p-value of each observation (y-axis). The grey coloured points represent genes that are not significant (NS) according to the applied thresholds for gene expression |log2 fold change|≥3 and statistical significance adjusted p-value<10e-4. The green points represent genes that satisfy only the gene expression threshold. Likewise, the blue points represent genes that satisfy only the statistical significance threshold. The genes that appear as red points are the statistically significant up- and down-regulated genes that are of interest in this study.

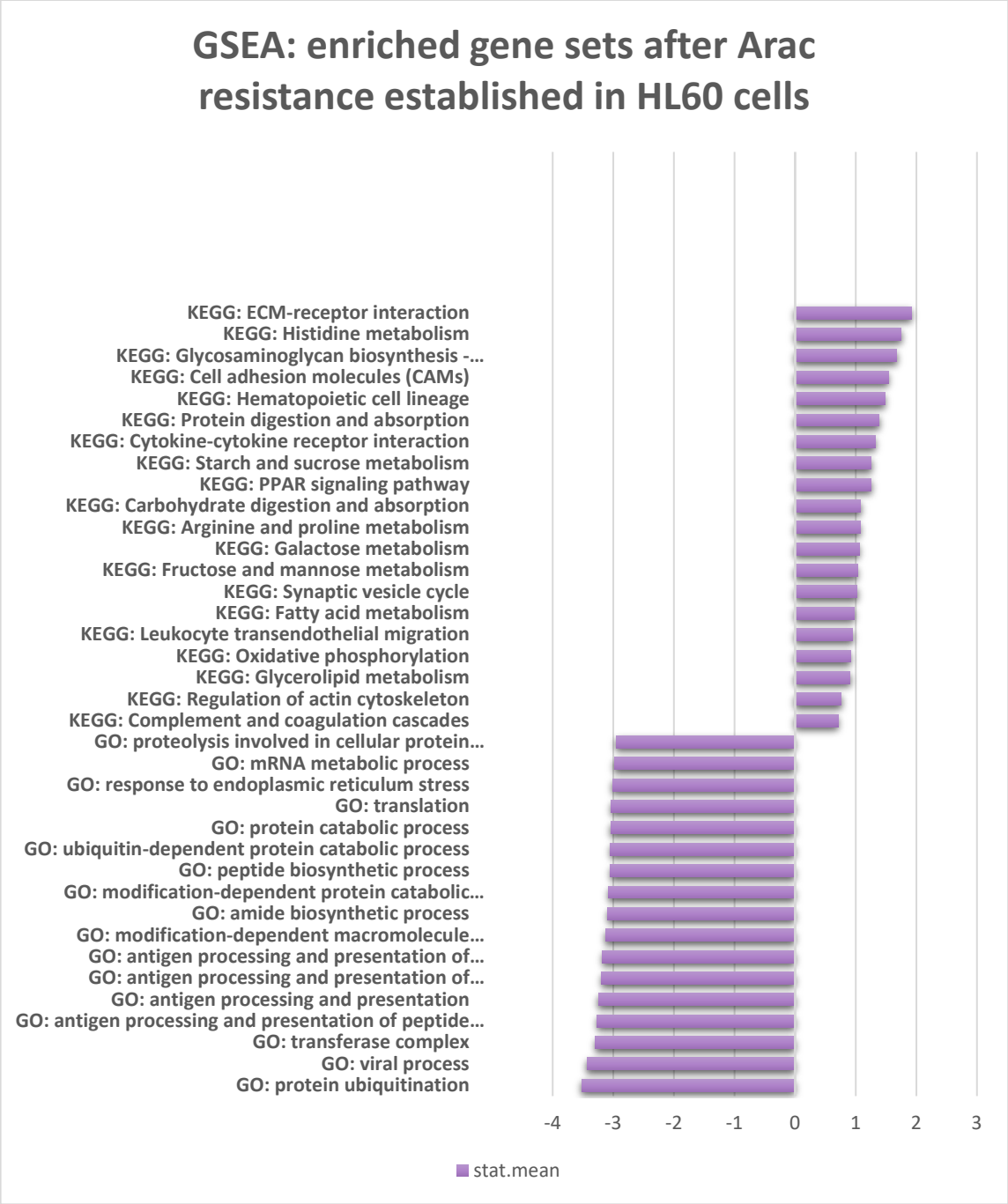


Figure 4.16: Figure summarises the results from the GSEA analysis performed on the differentially expressed genes comparing AraC-resistant vs Parental HL60 cells. The analysis considered gene sets from the Gene Ontology (GO) and the KEGG databases, including sets describing biological processes, metabolic and signalling pathways.

The same analysis for the THP-1 DOX resistance cells resulted in the identification of 596 differentially expressed genes, 454 of which were downregulated and 142 were upregulated (Figure 4.17).

The development of DOX resistance by the THP-1 cells triggered an upregulation of processes related to DNA packaging (i.e. ‘GO: DNA packaging complex’,’ GO: nucleosome’,’ GO: chromosomal part’) (Figure 4.18). As already mentioned earlier, the primary mechanism of action of DOX is its intercalation into DNA; thus, it has been reported that DOX destabilises both chromatin structure and the function of the nucleosome [261]. Another response to DOX mediated DNA damage is the extensive ubiquitination of ribosomal proteins, as reported by Halim et al., which ultimately leads to the inhibition of ribosome function [262]. In our results, the KEGG gene set representing the ribosome appears upregulated in the THP-1 DOX-resistant cells, potentially to counteract the effects of the drug.

Curiously, significant downregulation was observed on terms related with the adaptive immune system. More specifically, a decreased gene expression of genes related with ‘KEGG: Th1 and Th2 cell differentiation’, ‘KEGG: Th17 cell differentiation’, ‘KEGG: Haematopoietic cell lineage’ and ‘KEGG: Antigen processing and presentation’ was identified. Helper T cells are white blood cells of the lymphoid lineage that play a central role in the immune system by producing cytokines to stimulate immune response or by producing antibodies [3]. Potentially, the downregulation of genes that are involved in T helper cell gene sets in our results on THP-1 cells projects a more general trend of suppression of immune response mechanisms, due to the establishment of DOX resistance. However, it is worth mentioning that T helper type 1 (Th1) and Interleukin (IL)-17-producing helper T (Th17) have been reported as prognostic biomarkers and as new AML immunotherapy targets and are gaining increasing interest [263, 264]. A link between Th1 cells and reduced immune function in AML patients has been reported [265]. Interestingly, increased levels of Th17 cells have been found in

AML patients, which suppressed their immune system and were thought to facilitate further disease progression [266].

In the case of HL60 DOX resistant DE analysis, 261 genes were identified, 25 were downregulated and 236 were upregulated (Figure 4.19).

Regarding HL60 DOX-resistant compared to the untreated HL60 cells, the GSEA analysis revealed a downregulation of genes involved in metabolic pathways, namely ‘KEGG: Cysteine and methionine metabolism’, ‘KEGG: Purine metabolism’ and ‘KEGG: Oxidative phosphorylation’. In addition, DOX repressed DNA replication and gene expression processes, represented by the downregulation of terms such as ‘GO: translation’, ‘GO: mRNA metabolic process’ and ‘GO: RNA splicing’ (Figure 4.20).

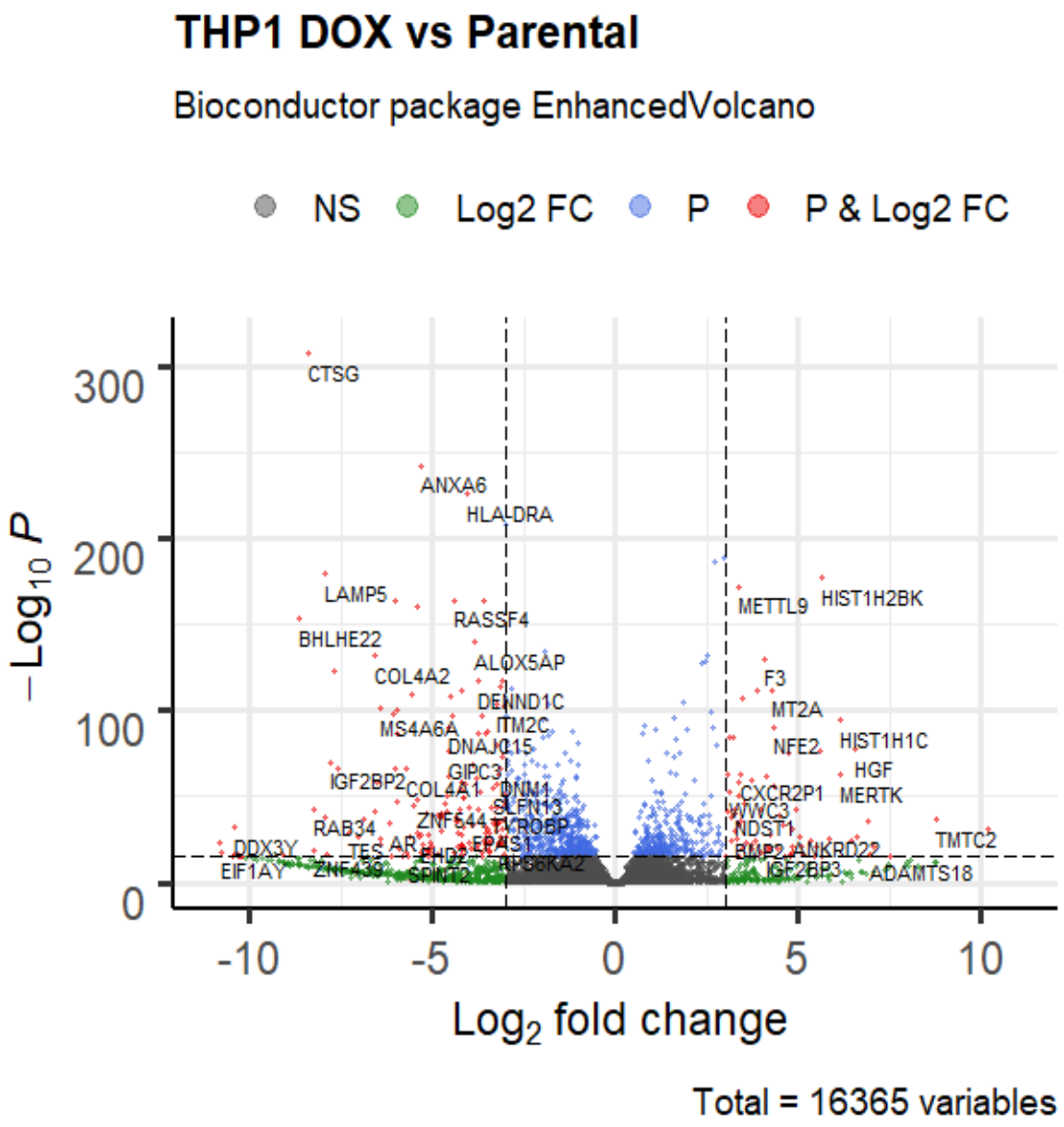


Figure 4.17: Volcano plot of the differentially expressed genes in the THP-1 cell line, comparing the DOX-resistant cells vs the untreated parental cells. The logarithm base 2 of the fold change in gene expression (x-axis) is plotted against the logarithm transformed adjusted p-value of each observation (y-axis). The grey coloured points represent genes that are not significant (NS) according to the applied thresholds for gene expression |log2 fold change|≥3 and statistical significance adjusted p-value<10e-4. The green points represent genes that satisfy only the gene expression threshold. Likewise, the blue points represent genes that satisfy only the statistical significance threshold. The genes that appear as red points are the statistically significant up- and down-regulated genes that are of interest in this study.

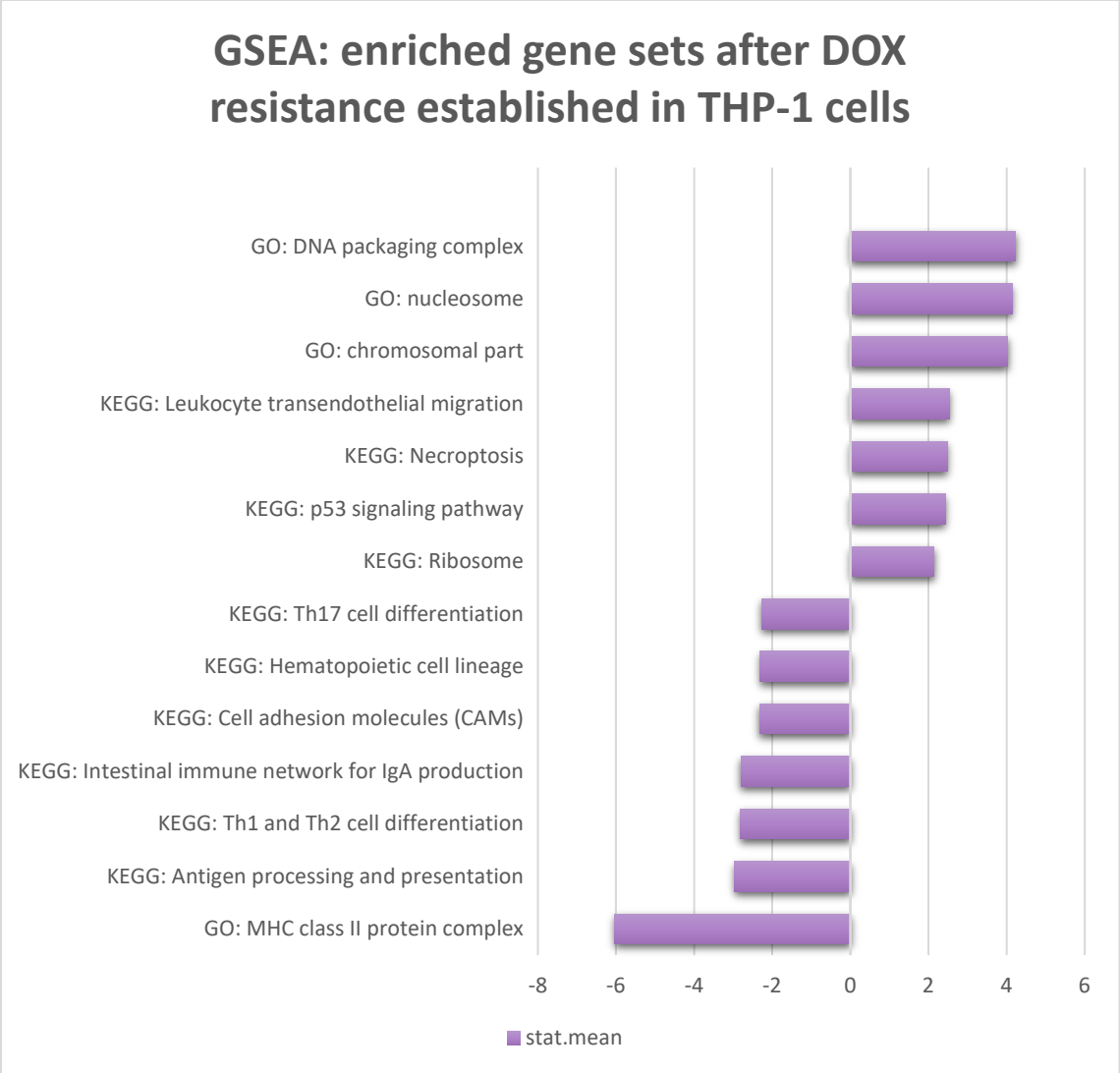


Figure 4.18: Figure summarises the results from the GSEA analysis performed on the differentially expressed genes comparing DOX-resistant vs Parental THP-1 cells. The analysis considered gene sets from the Gene Ontology (GO) and the KEGG databases, including sets describing biological processes, metabolic and signalling pathways.

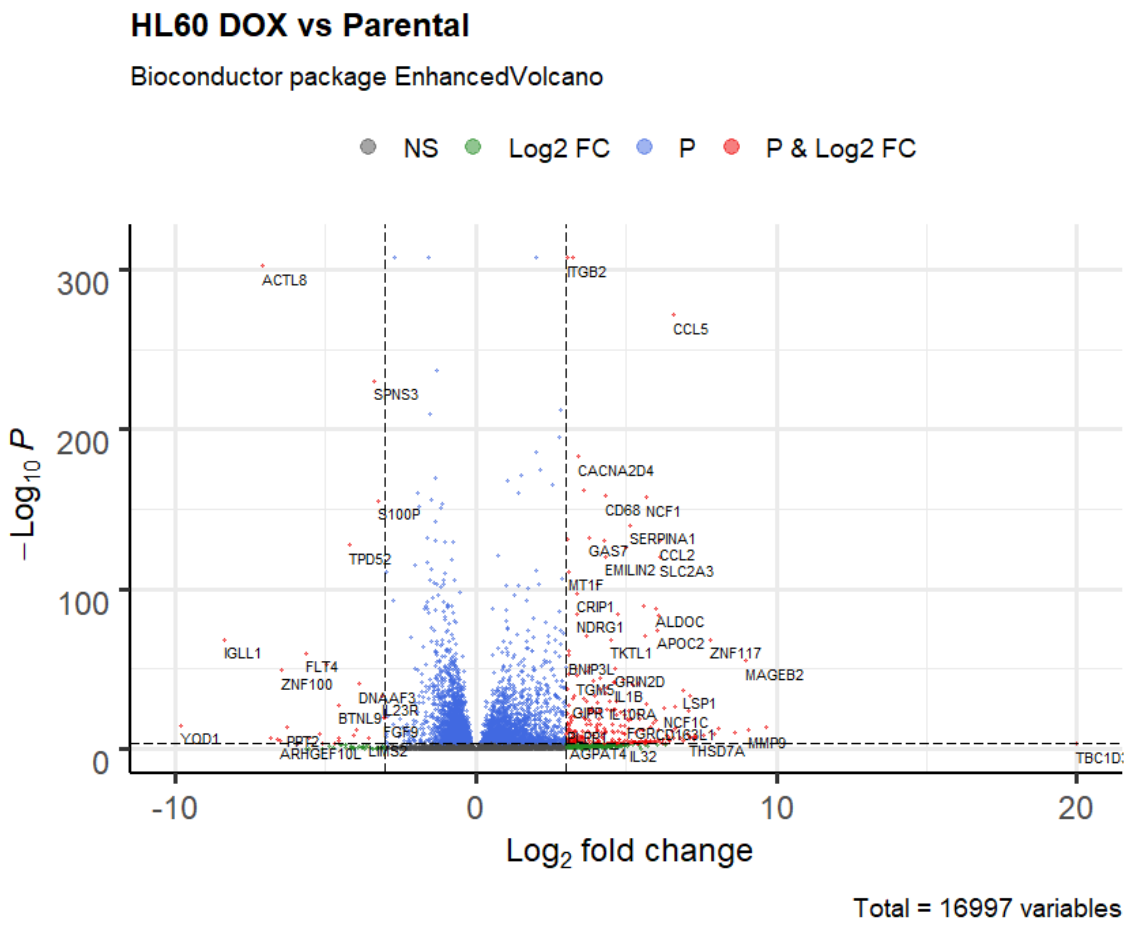


Figure 4.19: Volcano plot of the differentially expressed genes in the HL60 cell line, comparing the DOX-resistant cells vs the untreated parental cells. The logarithm base 2 of the fold change in gene expression (x-axis) is plotted against the logarithm transformed adjusted p-value of each observation (y-axis). The grey coloured points represent genes that are not significant (NS) according to the applied thresholds for gene expression |log2 fold change|≥3 and statistical significance adjusted p-value<10e-4. The green points represent genes that satisfy only the gene expression threshold. Likewise, the blue points represent genes that satisfy only the statistical significance threshold. The genes that appear as red points are the statistically significant up- and down-regulated genes that are of interest in this study.

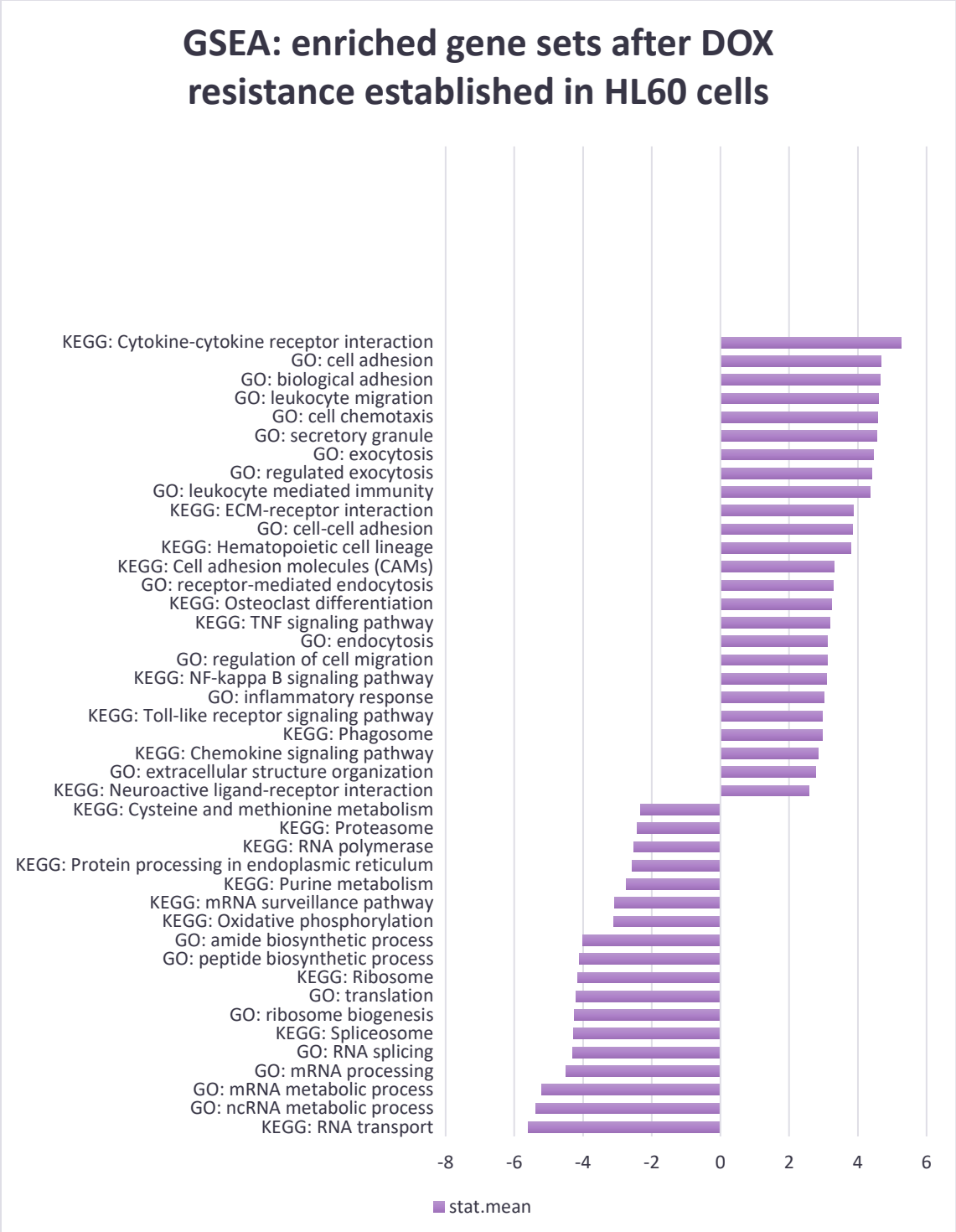


Figure 4.20: Figure summarises the results from the GSEA analysis performed on the differentially expressed genes comparing DOX-resistant vs Parental HL60 cells. The analysis considered gene sets from the Gene Ontology (GO) and the KEGG databases, including sets describing biological processes, metabolic and signalling pathways.

A common gene expression pattern observed in THP-1 and HL60 AraC-resistant cells and in DOX-resistant HL60 cells was the upregulation on gene sets related to cell adhesion (e.g. see ‘GO: cell-cell adhesion’, ‘GO: cell adhesion’ and ‘GO: biological adhesion’ in Figure 4.14 for the THP-1 AraC-resistant cell line and ‘KEGG: Cell adhesion molecules (CAMs)’ in Figure 4.16 for the HL60 AraC-resistant cell line and ‘GO: cell-cell adhesion’, ‘GO: cell adhesion’, ‘GO: biological adhesion’ and ‘KEGG: Cell adhesion molecules (CAMs)’ in Figure 4.20 for the HL60 DOX-resistant cell line). Adhesion molecules play an important role in the attachment of Leukemic Stem Cells (LSCs) and AML blasts to the microenvironment of the bone marrow and specifically to the niche. Until now, deregulation of cell-cell and cell-matrix adhesion was viewed as a distinguishing trait of solid tumours. However, it has been identified that the overexpression of adhesion molecules (AM) in AML contributes distinctly in the pathogenesis and progression of AML, since the attachment of LSCs to the bone marrow niches can transform them into self-sustainable leukemic niches [267-269]. Furthermore, the adhesion of AML cells to the niche provides them with protection from apoptosis and helps them acquire a quiescent phenotype. Consequently, LSCs are able to survive treatment and cause relapse of the disease [270-272]. This mechanism is known as environment-mediated drug resistance (EMDR) and it has already been reported for the AraC resistant HL60 cell line [273], together with the observation that drug resistant HL60 cells exhibit a stem cell-like phenotype [274]. Evidence from our analysis implies that the THP-1 AML cells employ the same mechanism against AraC treatment as the drug resistant HL60 cells and that THP-1 cells exhibit a stem cell-like phenotype when exposed to AraC treatment; facts which to our knowledge have not been reported thus far.

Another trait of solid tumours is the ability to metastasize and invade other tissues. Cancer Stem Cells (CSCs) are thought to be responsible for the metastasis of the different cancer types [275]. Apart from the implication of cell-cell and cell-matrix adhesion in drug resistance, these processes have an equally crucial contribution in metastasis, since the differential expression of cell adhesion molecules (CAMs) is responsible for the detachment of CSCs from the tumour and

their subsequent attachment at the metastatic site. Moreover, CAMs mediate leukocyte extravasation, a multi-step process by which leukocytes leave blood circulation and adhere to a site of inflammation [276]. It has been well established that CSCs mimic leukocytes and follow a similar process to the leukocyte extravasation cascade, taking advantage of the immune system-specific signalling, in order to reach and invade metastatic sites [277]. Finally, metastatic CSCs rely heavily on the interaction between the C-X-C chemokine receptor type 4 (CXCR4) and its ligand the matrix-derived factor-1 (SDF-1 or CXCL12) for their migration to other tissues and organs [278, 279]. In AML patients, CXCR4 has been positively correlated with poor disease prognosis, recurrence and progression and it has been found to be highly expressed in high migration drug-surviving (short-term) MOLT4 leukaemia cells [280, 281]. It has not yet been definitively proven that LSCs possess metastatic ability, but accumulating evidence points towards that direction [282]. Our results illustrate a putative mechanism employed by both THP-1 and HL60 cells in order to facilitate the migration and metastasis of LSCs. The AraC and DOX-resistant THP-1 cells exhibited an upregulation on gene sets related to inflammatory response and leukocyte activation, migration and chemotaxis (e.g. ‘GO: inflammatory response’, ‘GO: leukocyte migration’, ‘KEGG: Leukocyte mediated immunity’). HL60 cells with acquired resistance to AraC and DOX displayed very similar traits with gene upregulation in terms related to inflammatory response and leukocyte migration (e.g. ‘KEGG: Cytokine-cytokine receptor interaction’, ‘KEGG: Leukocyte transendothelial migration’). Uniquely, AraC-resistant THP-1 and HL60 cells showed an upregulation of the ‘KEGG: Regulation of actin cytoskeleton’ gene set, which has been found to be the case for aggressive and metastatic malignancies [283]. In addition, CXCR4 was upregulated by both THP1 and HL60 AraC and DOX-resistant cells. Overall, given the LSC-like phenotype demonstrated from both cell lines, we hypothesise that THP1 and HL60 drug-resistant cells differentially regulate CAMs in order for LSCs to detach from their site of homing, the leukemic niche, and later successfully perform the different steps of the extravasation process, facilitated by the distinctively increased immune response and activation of leukocytes. Most

probable metastatic sites would be those having high SDF-1 expression, which could explain the AML metastatic pattern to the spleen, since it has the second highest SDF-1 gene expression levels [284].

4.2.2.2 METABOLIC REPROGRAMMING OF DRUG-RESISTANT AML CELL LINES

Genome scale metabolic reconstructions of the THP-1 and HL60 AML cell lines were used to identify the metabolic rewiring that occurs after the establishment of resistance to AraC and DOX. Towards that end, comparisons were made between the AraC-resistant vs the parental GSMMs and the DOX-resistant vs the parental GSMMs, for both THP-1 and HL60 cells, in order to calculate any variations in flux distributions across all metabolic pathways.

For the most part, different metabolic pathways were altered due to the establishment of AraC resistance in the two cell lines (Figure 4.21). In THP-1, Cholesterol metabolism, the TCA cycle, Oxidative phosphorylation and Serine metabolism were downregulated, whereas the pathways of Pyruvate metabolism, Glycolysis/Gluconeogenesis and the Pentose phosphate pathway were upregulated. On the other hand, the HL60 AraC-resistant cells showed a downregulation of Bile acid synthesis and Fatty acid oxidation pathways, whereas isoleucine, phenylalanine and folate metabolism were upregulated. Contrary to THP-1 cells, AraC resistance caused an upregulation of the TCA cycle and serine metabolism in the HL60 cells. Finally, both cell lines had a flux increase through the Methionine and Cysteine pathway and Glycolysis/Gluconeogenesis. Therefore, putative metabolic targets to reverse AraC resistance in AML cells could derive from these pathways. Actually, administration of glycolysis inhibitors has been shown to improve the efficacy of AraC treatment in AML cell lines and primary samples [285, 286]. In addition, targeting methionine metabolism has been gaining increasing interest in cancer therapy and several ongoing clinical trials are considered promising [287].

The metabolic alterations identified in DOX resistant THP-1 and HL60 cells compared to the parental untreated cells were more extensive and for the most part DOX resistance caused a decrease in the flux through several metabolic

pathways (Figure 4.22). In both cell lines, Folate, Glutamate and Pyruvate metabolism, Glycolysis/Gluconeogenesis, the TCA cycle, Urea cycle, Glycerophospholipid metabolism, and Bile acid synthesis were downregulated. Interestingly, the Pentose phosphate pathway and Oxidative phosphorylation were upregulated in the THP-1 DOX-resistant cells and downregulated in the HL60 DOX-resistant cells. Finally, in the case of HL60 DOX-resistant cells, only the pathway of Histidine metabolism appeared to be upregulated, suggesting that targeting this pathway could provide a new approach to tackle DOX resistance in this cell line.

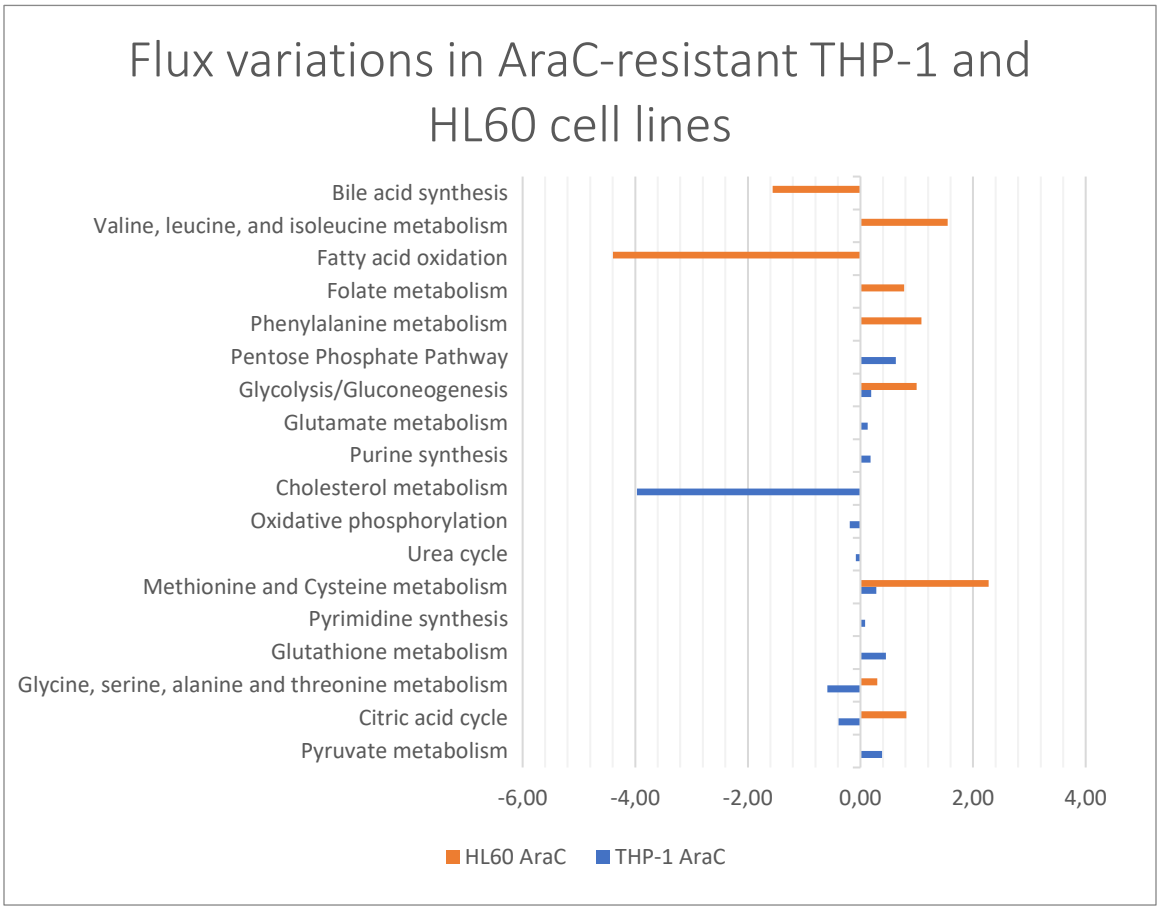


Figure 4.21: The comparison between flux distributions, as calculated using the reconstructed GSMMs of the AraC-resistant vs the parental THP-1 and HL60 cells, revealed metabolic alterations in the two AML cell lines caused by drug resistance. AraC mediated resistance had a completely different effect on the metabolic pathways of the two cell lines.

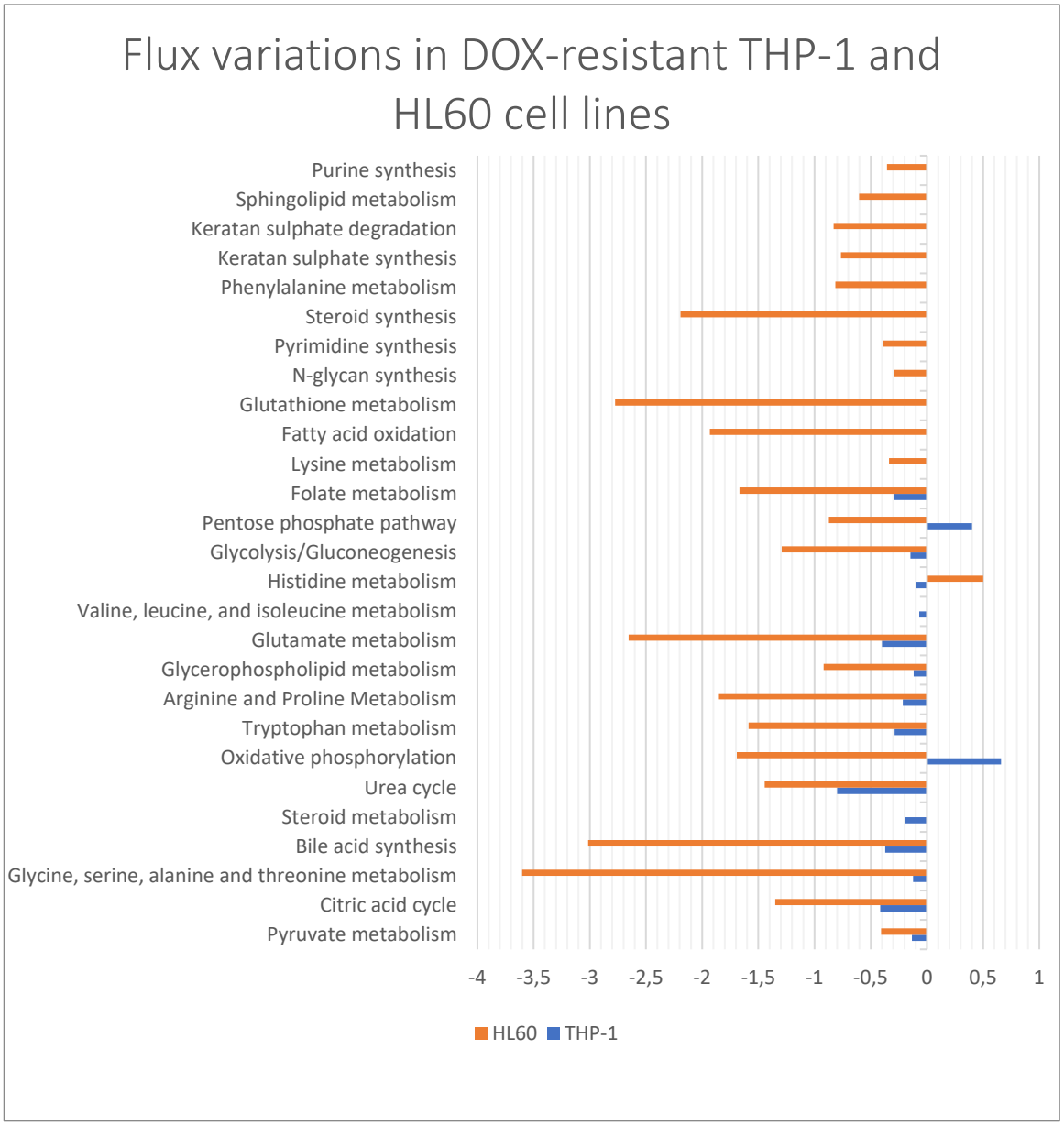


Figure 4.22: The comparison between flux distributions, as calculated using the reconstructed GSMMs of the DOX-resistant vs the parental THP-1 and HL60 cells, revealed metabolic alterations in the two AML cell lines caused by drug resistance. DOX mediated resistance had a completely different effect on the metabolic pathways of the two cell lines.

4.2.2.3 IDENTIFYING PUTATIVE METABOLIC TARGETS AGAINST ARAC AND DOX RESISTANCE IN AML

Using the reconstructed GSMMs of the AraC-, DOX-resistant and parental THP-1 and HL60 cells, we performed an exhaustive search simulating the knock-out (KO) of every gene in each model. Hence, we managed to identify the genes whose KO would severely compromise the viability of the resistant cells, forward referred to as essential genes (see Materials and Methods section 3.6). Several potential metabolic vulnerabilities for each condition were proposed by our pipeline, as summarized in Table 9.

Table 9: Using the GSMM reconstructions of the parental untreated and the AraC- and DOX-resistant THP-1 and DOX cell lines, we identified all the genes that were essential for the viability of the cells in each condition. To achieve this task, we simulated the knock down of every gene in each GSMM reconstruction and selected those genes whose knock down caused a decrease of the biomass production to less than 10%.

THP-1			
	Parental untreated	AraC-resistant	DOX-resistant
Number of essential genes	28	26	89
HL60			
	Parental untreated	AraC-resistant	DOX-resistant
Number of essential genes	44	54	25

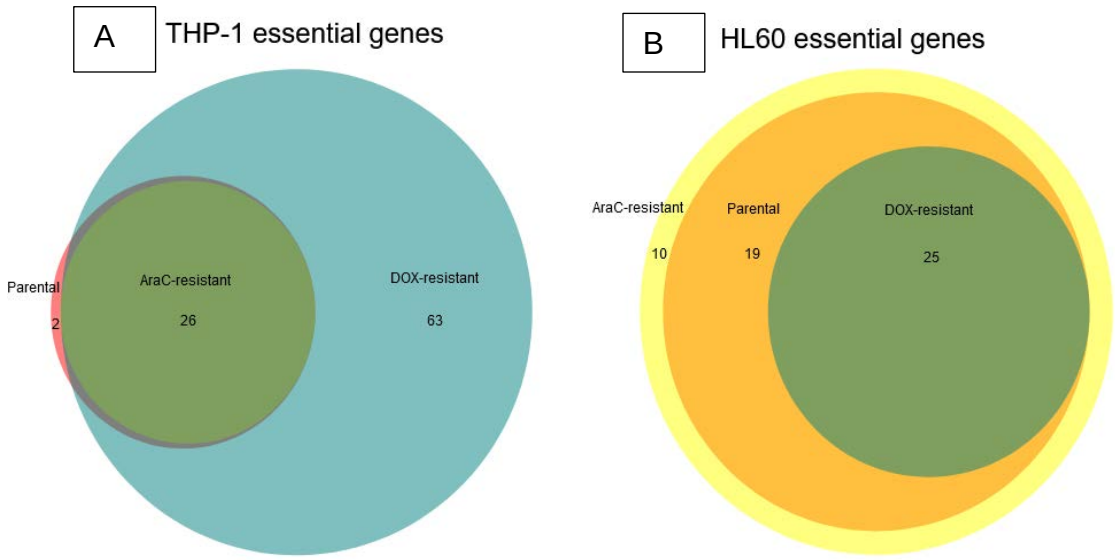


Figure 4.23: The Genome Scale Model reconstructions were interrogated towards identifying the essential genes, whose knock-down would compromise the viability of the AML parental and drug-resistant cells. [A] A Venn diagram of the common essential genes that were identified for the THP-1 parental, AraC- and DOX-resistant cells. [B] A Venn diagram of the common essential genes that were identified for the HL60 parental, AraC- and DOX-resistant cells.

First, we examined the results for each cell line separately. For the THP-1 cell line, 26 essential genes were identified for the AraC-resistant cells and 89 for the DOX-resistant cells. As illustrated with a Venn diagram in Figure 4.23A comparing the essential genes identified for the drug resistant cells vs the essential genes identified for the parental untreated cells, no specific metabolic vulnerabilities were found for the THP-1 AraC-resistant cells, whereas 63 specific metabolic vulnerabilities were found for the THP-1 DOX-resistant cells (Appendix I). Several enzymes that catalyse key reactions in the cholesterol biosynthesis pathway, namely mevalonate kinase (MVK), squalene synthase (SQLS), lanosterol synthase (LSS), were characterised as essential for the survival of the THP-1 DOX-resistant cells. We evaluated the effects of cholesterol biosynthesis pathway inhibition by using YM-53601, a SQLS inhibitor. Indeed, the proliferation of THP-1 DOX-resistant cells was severely compromised compared to the THP-1 parental untreated cells, suggesting that targeting cholesterol biosynthesis could be a promising therapeutic strategy for these cells (Figure 4.24). Important to mention,

that as anticipated by our results, SQLS inhibition had no effect on THP-1 AraC-resistant and HL60 AraC- and DOX-resistant cells.

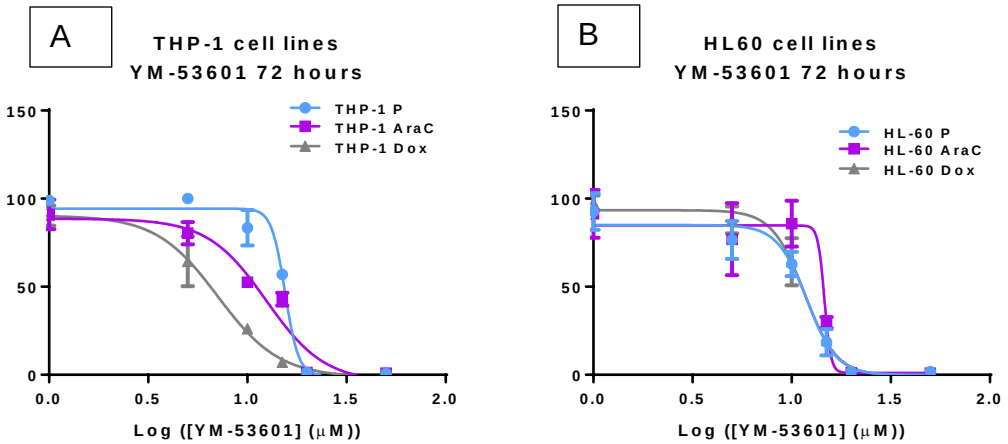


Figure 4.24: Effect of YM-53601 squalene synthase inhibitor on AML THP-1 parental, AraC-resistant and DOX-resistant cell lines. Cell viability was assessed after 72h incubation with either MTX or PTX (nM). Values represent mean ± SD of n=3.

Then, we identified the essential genes that would impair the growth of AraC- and DOX-resistant cells of both cell lines. The 25 common putative metabolic targets across all drug resistant cells are presented in Table 10, together with the metabolic reactions they participate in. It should be noted that these 25 common genes are in fact all the essential genes identified by our method for the HL60 DOX-resistant cells.

Table 10: In total, 25 metabolic genes that could serve as potential drug targets to compromise the viability of both THP-1 and HL60 AML cell lines resistant to Cytarabine and Doxorubicin were identified by our proposed workflow.

Gene symbol	Gene Name	Reaction ID
ADSL	adenylosuccinate lyase	‘ADSL1', 'ADSL2'
ASNS	asparagine synthetase (glutamine-hydrolyzing)	‘ASNS1'
ATIC	5-aminoimidazole-4-carboxamide ribonucleotide formyltransferase/IMP cyclohydrolase	‘IMPC', 'AICART'
CAD	carbamoyl-phosphate synthetase 2, aspartate transcarbamylase, and dihydroorotase	‘CBPS', 'ASPCTr', 'DHORTS'
DHFR	dihydrofolate reductase	‘r0224', 'r0512', 'DHFR', 'r0514', 'FOLR2', 'r0226'
DHODH	dihydroorotate dehydrogenase (quinone)	‘DHORD9'
KDSR	3-ketodihydrosphingosine reductase	‘3DSPHR'
GART	phosphoribosylglycinamide formyltransferase, phosphoribosylglycinamide synthetase, phosphoribosylaminoimidazole synthetase	‘GARFT', 'PRAGSr', 'r0666'
LCAT	lecithin-cholesterol acyltransferase	‘LCAT1e'
lipA	lipase A, lysosomal acid type	‘r1172', 'CHOLESTle', 'r1179', 'r1177'
PFAS	phosphoribosylformylglycinamidine synthase	‘PRFGS'
PPAT	phosphoribosyl pyrophosphate amidotransferase	‘GLUPRT'
TXNRD1	thioredoxin reductase 1	‘TRDR2', 'TRDR3', 'r0027', 'r1433', 'TRDR'
TYMS	thymidylate synthetase	‘TMDS'
UMPS	orotate phosphoribosyltransferase, orotidine-5-phosphate decarboxylase	‘ORPT', 'OMPDC'
PGS1	phosphatidylglycerophosphate synthase 1	‘PGPPT'
SPTLC2	serine palmitoyltransferase long chain base subunit 2	‘SERPT'
SPTLC1	serine palmitoyltransferase long chain base subunit 1	‘SERPT'
PAICS	phosphoribosylaminoimidazole carboxylase; phosphoribosylaminoimidazolesuccinocarboxamide synthase	‘AIRCr', 'PRASCS'
RPIA	ribose 5-phosphate isomerase A	‘RPI', 'r0249'

CMPK1	cytidine/uridine monophosphate kinase 1	‘UMPK4n', 'CYTK8', 'CYTK9'
CRLS1	cardiolipin synthase 1	‘CLS_hs'
SPTLC3	serine palmitoyltransferase long chain base subunit 3	‘SERPT'
PTPMT1	protein tyrosine phosphatase, mitochondrial 1	‘PGPP_hs'
SGMS1	sphingomyelin synthase 1	‘SMS'

Even though Folate metabolism appeared upregulated only in the THP-1 and HL60 AraC-resistant cells, two enzymes related with this pathway were identified as common metabolic vulnerabilities of both AraC- and DOX-resistant cells, namely thymidylate synthetase (TYMS) and dihydrofolate reductase (DHFR). Both enzymes have a crucial role in DNA biosynthesis [288]. TYMS catalyses the methylation of 2′-deoxyuridine-5′-monophosphate (dUMP) and 5,10-methylenetetrahydrofolate (CH2THF) to 2′-deoxythymidine-5′-monophosphate (dTMP, thymidylate) and 7,8-dihydrofolate (7,8-DHF) [289]. Then, DHFR catalyses the first of the two reactions needed to transform the reduced 7,8-DHF back to CH2THF [289]. Inhibitors for both enzymes have been used, usually in combination with other chemotherapeutics, as anticancer drugs [290]. We therefore decided to evaluate the efficacy of two commonly used drugs, Pemetrexed, which inhibits TYMS and DHFR, and Methotrexate, which inhibits DHFR. Both drugs seemed to have a strong effect on THP-1 and HL60 AraC-resistant cells, with the exception of HL60 cells and Methotrexate, where the effect of the drug was equally strong (Figure 4.25).

Additionally, protein tyrosine phosphatase mitochondrial 1 (PTPMT1) and cardiolipin synthase 1 (CRLS1) were also included in the list of common putative metabolic vulnerabilities for the drug resistant AML cell lines. Both enzymes have crucial roles in the cardiolipin biosynthesis pathway. Inside the mitochondria, PTPMT1 dephosphorylates phosphatidylglycerol phosphate (PGP) to phosphatidylglycerol (PG), which is then converted into cardiolipin by CRLS1 [291, 292]. Inhibition of PTPMT1 has been shown to compromise mitochondrial respiration, while Niemi et al. have demonstrated that PTPMT1 depletion leads to

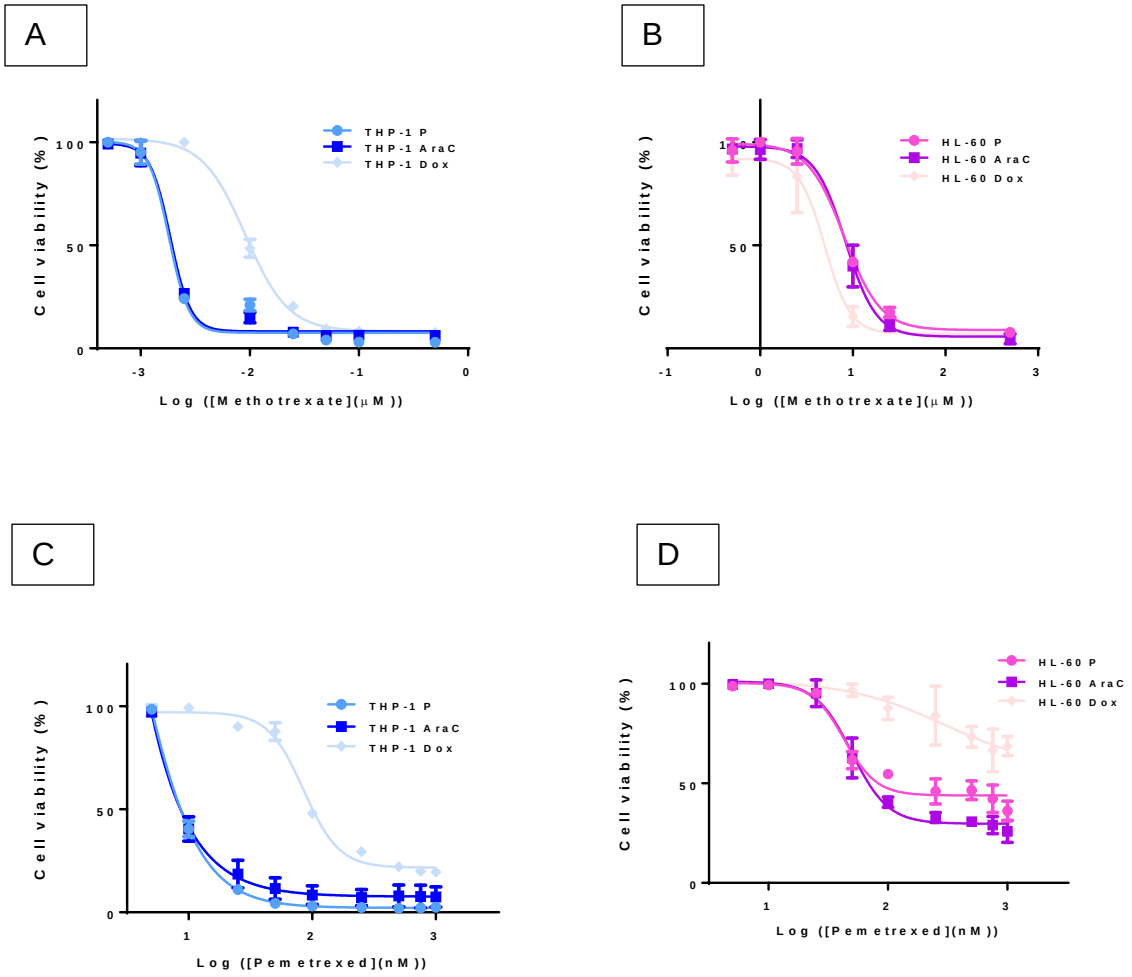


Figure 4.25: Effect of antifolate analogues Pemetrexed and Methotrexate on AML cell lines. Cell viability was assessed after 72h incubation with either MTX or PTX (nM). Values represent mean ± SD of n=3. [A] Effect of Pemetrexed on THP-1 parental, AraC-resistant and DOX-resistant cells. [B] Effect of Pemetrexed on HL60 parental, AraC-resistant and DOX-resistant cells. [C] Effect of Methotrexed on THP-1 parental, AraC-resistant and DOX-resistant cells. [D] Effect of Methotrexed on HL60 parental, AraC-resistant and DOX-resistant cells.

cancer cell death [291, 293]. Thus, we decided to evaluate the effect of U.S. Food and Drug Administration approved anticancer drug alexidine dihydrochloride on the parental, AraC-resistant and DOX-resistant AML cell lines. The drug seemed to have a stronger effect on HL60 DOX resistant cells compared to the untreated parental cells (Figure 4.26). However, Alexidine did not seem to have any effect on the THP-1 cells, as would be expected since THP-1 cells are more dependent on OXPHOS than HL60.

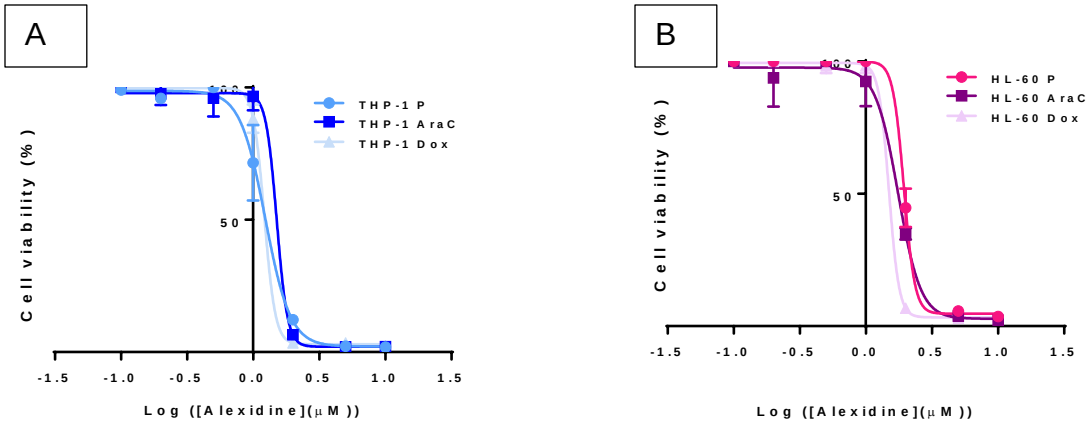


Figure 4.26: Effect of Alexidine dihydrochloride on AML cell lines. Cell viability was assessed after 72h incubation with Alexidine (nM). Values represent mean ± SD of n=3. [A] Effect of Alexidine dihydrochloride on THP-1 parental, AraC-resistant and DOX-resistant cells. [B] Effect of Alexidine dihydrochloride on HL60 parental, AraC-resistant and DOX-resistant cells.

4.2.3 Discussion

Despite the great efforts put towards designing more effective therapeutic strategies for AML, the fact still remains that drug resistance is the leading cause for treatment failure in this type of cancer. In an effort to address AML resistance to two commonly used chemotherapeutic drugs, we designed a multi-omic data integration pipeline under the framework of constraint-based modelling. We achieved the reconstruction of condition specific GSMMs of different AML cell lines under AraC and DOX resistance, which resulted in the identification of metabolic adaptation mechanisms employed by the resistant cells to survive treatment, as well as the identification of putative metabolic vulnerabilities.

Our analysis revealed that AraC and DOX resistant AML cells exhibit a stem-like phenotype and upregulate cell-cell and cell-matrix adhesion molecules, as well as inflammatory response and leukocyte activation and migration pathways. This mechanism employed by resistant cells, not only appears to aid them escape chemotherapeutic drugs, but may also provide evidence of the metastatic potential of LSCs.

The reconstruction of GSMMs of THP-1 and HL60 AML cells under AraC and DOX resistance and their interrogation towards the identification of metabolic alterations caused by the establishment of resistance led to the illustration of different metabolic signatures across cell lines and conditions. Our results depict the complexity of the metabolic rewiring that takes place in different cell lines of AML, highlighting the heterogeneity of this disease and the need for designing more personalised therapeutic regimes.

Taking full advantage of the potential of the reconstructed GSMMs, we managed to identify several common targets against AraC- and DOX-resistant THP-1 and HL60 cells. Inhibition of the cholesterol biosynthesis pathway using a squalene synthase inhibitor greatly compromised the viability of THP-1 DOX resistant cells. As predicted by our method, squalene synthase was a highly specific metabolic vulnerability and its validation highlighted both the great

phenotypic heterogeneity of AML, as well as the robustness of the workflow we designed to reconstruct and interrogate condition specific GSMMs.

Finally, our evaluation of two antifolate agents provided very promising preliminary results, since they had a strong antiproliferative effect on the resistant cells. Additionally, our analysis suggested that targeting these metabolic vulnerabilities, would also affect the parental untreated cells, as was actually confirmed upon experimental validation. We did not find this discouraging, since folate metabolism is an emerging target against different types of primary and chemo-resistant cancer cells [294, 295]. Thus, we believe that future work should focus on investigating in greater detail the potential therapeutic advantages of targeting folate metabolism in AML.

4.3 Application of machine learning techniques on GSMMs for AML patient stratification and metabolic biomarker identification

4.3.1 Introduction

Acute Myeloid Leukaemia (AML) is a highly heterogeneous disease regarding both its phenotypic and genetic traits, making the identification of accurate prognostic and predictive biomarkers and the specific definition of AML subgroups a great challenge up to this day.

The diagnostic procedure for AML and the classification of the different subgroups of the disease, proposed by the World Health Organisation (WHO), is based on morphologic, immunologic, cytogenetic and molecular classification methods (MICM) and is constantly being revised to include new findings provided by the rapidly evolving high-throughput -omic technologies. The first attempt in AML classification relied on the identification of specific cytogenetic abnormalities observed in patients. These non-random chromosomal rearrangements have been well established and are found in about 60% of *de novo* AML patients [296]. They can be unbalanced (e.g. deletions, monosomies or trisomies), balanced (e.g. reciprocal translocations, insertions, and inversions) or can correspond to a complex karyotype containing at least three chromosomal abnormalities. The first whole genome sequencing of an AML patient and the advancements in Next Generation Sequencing (NGS) methods that followed made the identification of additional mutational events and gene expression patterns possible. This still increasing pool of information has been particularly enlightening for the high percentage of AML patients with a normal karyotype [297]. It has also contributed in revealing functional groups of mutated genes; thus, providing biological insight into the causes of AML pathology, as well as new diagnostic and therapeutic approaches to the disease.

AML patients are classified into different subgroups depending on their acquired cytogenetic abnormalities, and the defined subgroups fall under three prognostic categories depicting the potential clinical outcome of the patient: the favourable, the intermediate and the poor-risk group. However, this risk stratification approach does not include the large subgroup of AML patients with

normal karyotypes, which is also considered as the most genetically heterogeneous group [298]. Great efforts have been made towards a more accurate characterisation of the distinctive molecular features of the different AML subgroups. Notably, Papaemmanouil et al. employed refined statistical models and managed to classify AML patients in 11 cytogenetically and molecularly distinct subgroups, proposing a more elaborate approach [299].

However, an accurate panel of prognostic biomarkers to enable risk stratification of AML and thus assist the selection of the most appropriate therapeutic strategy for each patient is still lacking. In this chapter, we propose a new machine learning based decision support tool aiming in the risk stratification of AML taking into consideration not only the molecular genetics and the demographic data (e.g. race, gender, age) of the patients, but also the alterations that occur in AML metabolism though the integration of metabolomics and simulated AML patient-specific metabolic reaction fluxes. Our goal was to introduce metabolic features in the process of classifying AML patients, in order to incorporate new and additional knowledge into the existing attempt towards prognostic models for the disease.

A dataset of 180 AML patients available from the Cancer Genome Atlas (TCGA) database, including transcriptomic and cytogenetic data, as well as demographic information and the cytogenetic risk category each patient has been assigned to, was used to perform this study. In order to build personalised GSMMs for the patients and due to missing metabolomic assays corresponding to each of them, we decided to construct an AML consensus model from the data collected from the THP-1 and the HL60 cell lines (see Materials and Methods section 3.1). Since these two cell lines are phenotypically different, we believe that the resulting GSMM is an adequate approximation of the AML metabolic landscape. The transcriptomic data of each patient were integrated into the consensus AML GSMM and patient-specific reconstructions were built. These were then used to simulate the maximum metabolic capability of each individual patient and simulate the metabolic flux through the reactions in the models (see Materials and Methods

3.6). In order to reduce the number of parameters to use in the subsequent classification analysis, a penalised matrix decomposition approach applied to sparse Canonical Correlation Analysis (sCCA) was performed in regards to the desired phenotype, which in this case was the risk category each patient was assigned to in the TCGA database (see Materials and Methods 3.7.1). Finally, a Random Forest (RF) classifier was built using the gene expression and cytogenetic data (see Materials and Methods 3.7.2), the metabolomics and the demographic information of the patients. We selected the specific ML algorithms, namely the ssCCA and RFs, because of their inherent capability to handle high-dimensional datasets. First, ssCCA, having the sparsity extension was an excellent choice to identify latent relationships between the transcriptomic and the simulated metabolic flux rates datasets. This allowed the reduction in the number of gene expression and metabolic features to use for building an ML classifier. Then, we used RFs to perform the classification task and biomarker identification, because of their resistance to overfitting, given the very low number of samples/patients compared to the very larger number of features in our dataset.

With our proposed workflow and resulting classification tool we managed to identify specific genes and metabolic pathways driving the risk stratification of AML patients and could also be potential biomarkers to explore as a personalised therapeutic approach for patients in the different risk groups.

4.3.2 Description of the dataset

The dataset we compiled for the purposes of this chapter was partly based on AML patient data provided by the TCGA LAML dataset and partly based on simulated metabolic reaction fluxes calculated using the reconstructed patient-specific GSMMs (see Materials and Methods section 3.4).

The total number of samples/ patients was 180. The total number of features (gene expression, metabolic reaction fluxes, cytogenetic abnormalities, age at initial diagnosis, days to death, ethnicity, gender and CALGB cytogenetic risk category) was 49,363. Missing values were found in the cytogenetic abnormalities, days to death and the ethnicity of the patients. These missing values were

randomly distributed across patients and inferred by the RF algorithm appropriately (see Materials and Methods section 3.7.2).

Table 11: In total 8 different types of features were used to build a Random Forest classifier for the risk stratification of Acute Myeloid Leukaemia patients.

TCGA-LAML project			
180 AML patient dataset			
Features	# features per sample	Data type	Missing values
Reaction fluxes	4,310	numerical continuous variable	0
Gene expression	45,047	numerical continuous variable	0
Cytogenetic abnormalities	1	categorical – 32 class labels	19
Age at initial diagnosis	1	numerical discrete variable	0
Days to death	1	numerical discrete	13
Ethnicity	1	categorical – 3 class labels	2
Gender	1	categorical – 2 class labels	0
AML CALGB cytogenetic risk category	1	categorical - 3 class labels	0

4.3.3 Results

The original dataset described in 4.3.2 was split into training (60%) and test (40%) sets, which were then used to train and evaluate the performance of an RF classifier, respectively. However, the classifier was not able to identify a comprehensive list of important features driving the classification of patients. For this reason, we decided to apply ssCCA to find the latent relationships between the gene expression and the metabolic reaction fluxes datasets, in regard to the AML CALGB cytogenetic risk category of each patient. Thus, we managed to reduce the number of gene expression and metabolic features and consequently the total number of features to use for the RF classification task.

4.3.3.1 SUPERVISED SPARSE CANONICAL CORRELATION ANALYSIS (SSCCA)

We applied sparse supervised Canonical Correlation Analysis (ssCCA) on the set of gene expression data (a data matrix with 180 patients as rows and 45,047 transcripts as columns) and the set of metabolic fluxes (a data matrix with 180 patients as rows and 4,310 reaction fluxes as columns). The ssCCA was performed towards the desired phenotype, which in our study was the risk category each patient was assigned to in the TCGA database (see section 3.7.1 of Materials and Methods) [195]. In this way, we managed to identify the most correlated genes and metabolic reactions regarding the classification of patients into risk categories and achieved a substantial reduction of 91.8% in the number of genes and 93.8% in the number of metabolic reactions to include in the RF classifier. More specifically, the ssCCA revealed that 3,685 genes and 267 metabolic reactions were highly correlated with the classification of AML patients into the three risk categories, the Poor, the Intermediate/Normal and the Favourable.

We performed an enrichment analysis using DAVID on the list of 3,685 genes identified by ssCCA to explore the functional group they belong to [300, 301]. This analysis was performed using UniProt Knowledgebase (UniProtKB)) keywords, which provide a comprehensive and extensive database for functional information on protein-coding genes [302]. The statistical significance of the enrichment was evaluated based on a modified Fisher's exact test (Appendix II) [303]. From the 1,200 UniProtKB keywords, 52 were highly enriched in our results with a modified Fisher's exact score < 0.1 (Figure 4.27). More specifically, terms related to posttranscriptional modification, such as 'Acetylation', 'Prenylation' and 'Methylation', were enriched and the most statistically significant term was 'Phosphoprotein', which represents proteins that can be posttranslationally modified. These findings suggest a putative association of posttranscriptional regulation and risk prognosis in AML patients. Interestingly, several keywords related to metabolic pathways and processes were also enriched ('Kinase', 'Serine/threonine-protein kinase', 'Oxidoreductase', 'Lipid metabolism', 'Fatty acid metabolism' and others), highlighting the implication of metabolism in AML risk classification. It is worth mentioning that the functional enrichment analysis

identified enrichment in the terms related to ‘immunity’ and ‘innate immunity’, processes which played a crucial role in the mechanism employed by AML cells in order to develop drug resistance, as described in the previous chapter (see section 4.2.2.1.2).

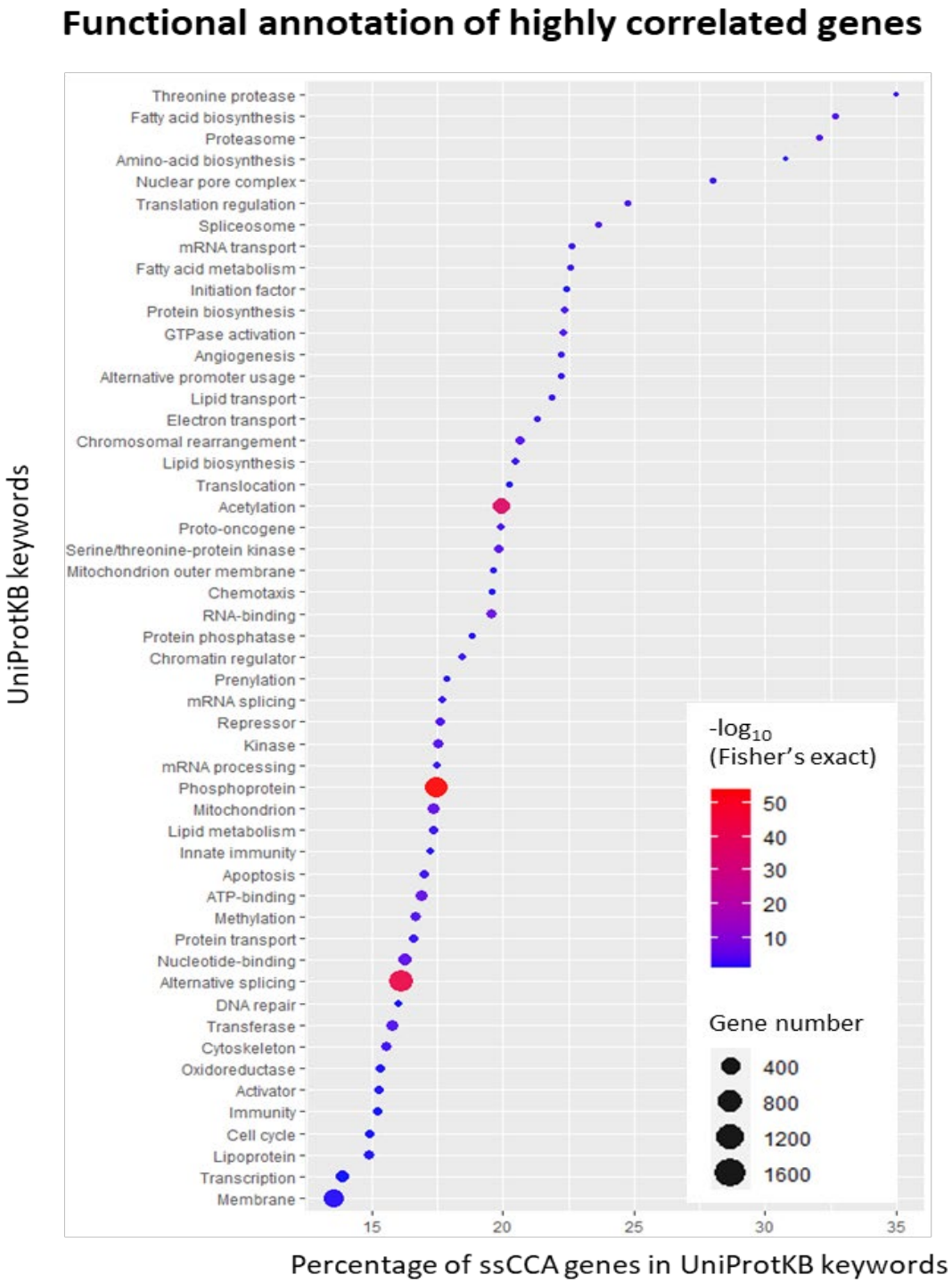


Figure 4.27: Visualisation of the functional annotation enrichment analysis performed on the genes identified by ssCCA as highly correlated with the risk classification of AML patients using UniProt Knowledgebase keywords. The results are presented in a bubble chart. The y-axis represents the significantly enriched keywords, the x-axis represents the percentage of ssCCA genes in the different keywords, the size of the bubbles is proportional to the number of genes assigned to a particular keyword and the colour of the bubbles represents the -log10 transformation of the modified Fisher's exact score. The larger the number of genes belonging to a specific keyword, the larger the size of the bubble. The higher the statistical significance of a keyword, the higher on the colour bar the bubble will be.

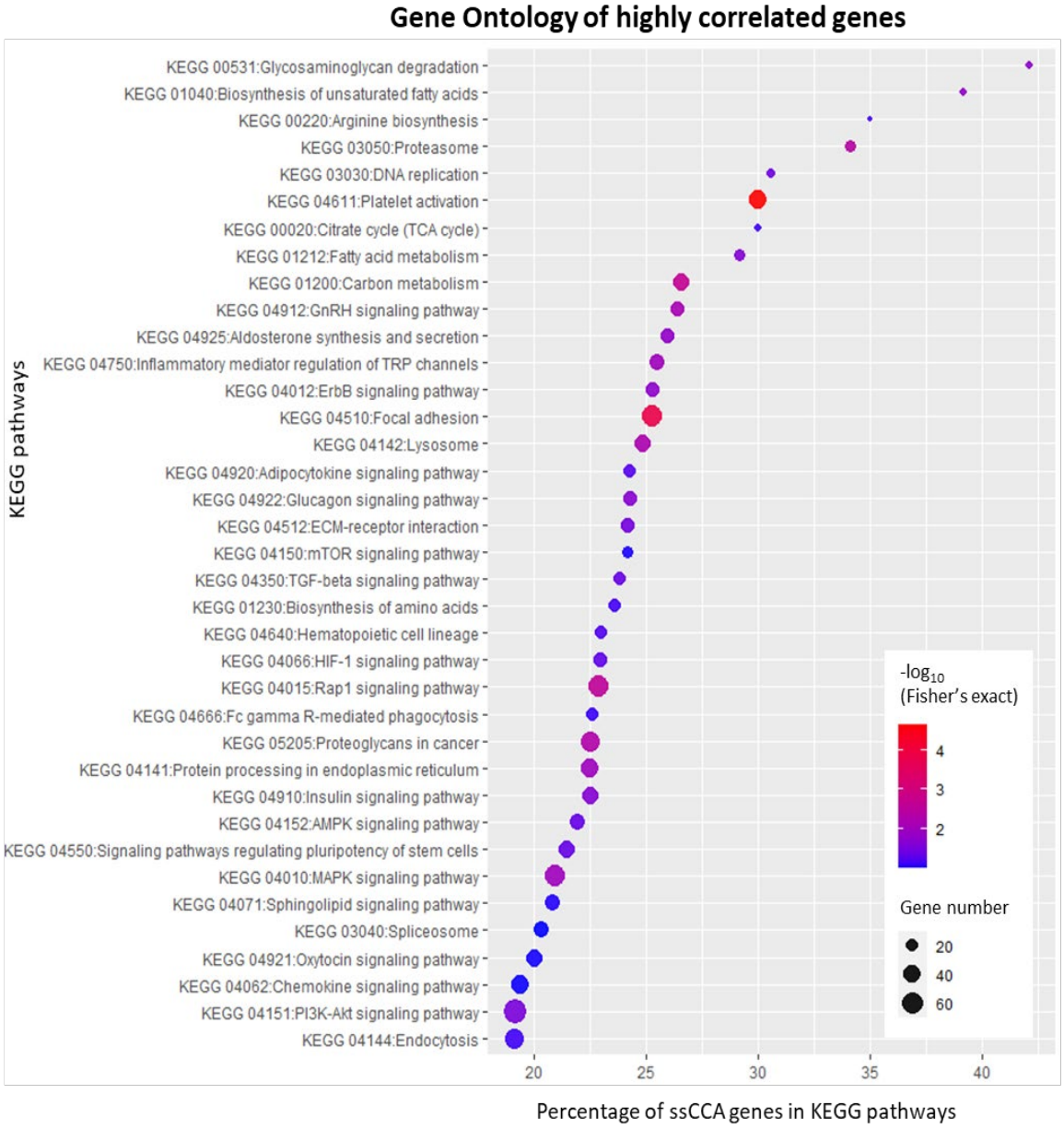


Figure 4.28: Visualisation of the gene ontology (GO) enrichment analysis performed on the genes identified by ssCCA as highly correlated with the risk classification of AML patients using KEGG signalling and metabolic pathways. The results are presented in a bubble chart. The y-axis represents the significantly enriched KEGG terms, the x-axis represents the percentage of ssCCA genes in the different pathways, the size of the bubbles is proportional to the number of genes assigned to a particular pathway and the colour of the bubbles represents the -log10(Fisher's exact). The larger the number of genes belonging to a specific keyword, the larger the size of the bubble. The higher the statistical significance of a keyword, the higher on the colour bar the bubble will be. KEGG: Kyoto Encyclopaedia of Genes and Genomes.

Similarly, we performed a Gene Ontology analysis using DAVID on the list of 3,685 genes identified by ssCCA to discover enriched Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathway terms, including signal transduction and metabolic pathways. The statistical significance of the enrichment was evaluated based on a modified Fisher's exact test (Appendix II) [303]. The analysis identified the 37 most significantly enriched KEGG terms in our results (Figure 4.28). Two of the most statistically significant KEGG pathways was the 'Chemokine signalling pathway' and 'Focal adhesion', both of which link back to the previous chapter on drug resistance. The implication of chemokines in both the development of drug resistance and the risk stratification of AML patients could indicate whether a subpopulation of patients would benefit from receiving the typical aggressive chemotherapy treatment or they have a predisposition in developing drug resistance and therefore would be in need of a different therapeutic regime. In fact, Kornblau et al., 2010, showed eight recurring cytokines and chemokines expression signatures in AML patients and managed to demonstrate their prognostic significance and relevance to remission, primary resistance, relapse rates, and overall survival of the patients [304]. More recently, Carter et al., 2017, investigated the role of focal adhesion kinase (FAK) in AML and established a correlation of the overexpression of the gene (*ptk2*) and a poor disease prognosis and relapse [305, 306].

Finally, we mapped the metabolic reactions identified by the ssCCA as highly correlated to the risk classification task into established metabolic pathways. The number of identified reactions and the pathways they were mapped to are presented in Figure 4.29. The most significantly represented pathway in our data was that of fatty acid oxidation. Together with the pathways of cholesterol metabolism, bile acid synthesis, sphingolipid and glycerophospholipid metabolism, the results indicate an important role of lipid metabolism in different AML risk groups. There is increasing interest on fatty acid oxidation since it has been associated with progression and survival of AML cells [307, 308], as well as development of drug resistance and stem-like phenotype [309].

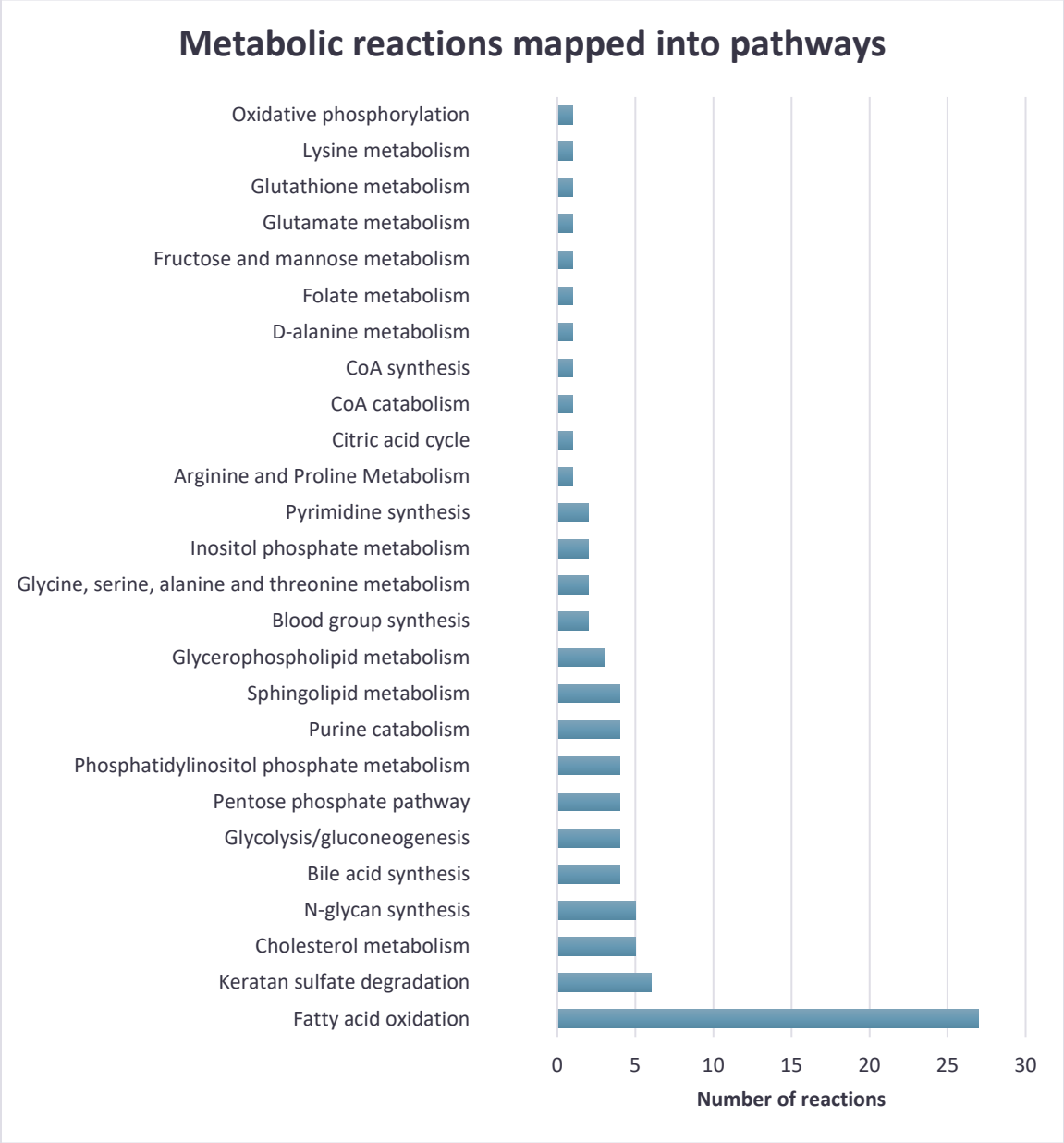


Figure 4.29: The metabolic reactions identified as significantly correlated to the risk classification of AML patients were mapped into metabolic pathways.

4.3.3.2 Random Forest (RF) classifier

Following the feature reduction step using ssCCA, we built a Random Forest (RF) classifier in order to perform risk stratification of AML patients. The dataset was randomly split into training (60%) and test (40%) sets. The RF was trained on classifying the risk category of each patient based on transcript expression (3,685 features), metabolic reaction fluxes (267 features), cytogenetic abnormalities (1 feature), age(1 feature), days to death (1 feature), ethnicity (1 feature) and gender (1 feature) (see section 3.4 of Materials and Methods). Important to note, gene expression represents roughly 93% and the metabolic reaction fluxes represent 6.3% of the total number of features in the dataset.

4.3.3.2.1 Evaluation of the model

The performance of the RF was evaluated on the test set. The goal was for the RF to successfully predict the risk group for AML patients and to identify a set of genes and metabolic reactions responsible for their classification. The RF grouped the patients into three risk groups (predefined in the TCGA dataset): Class 0-

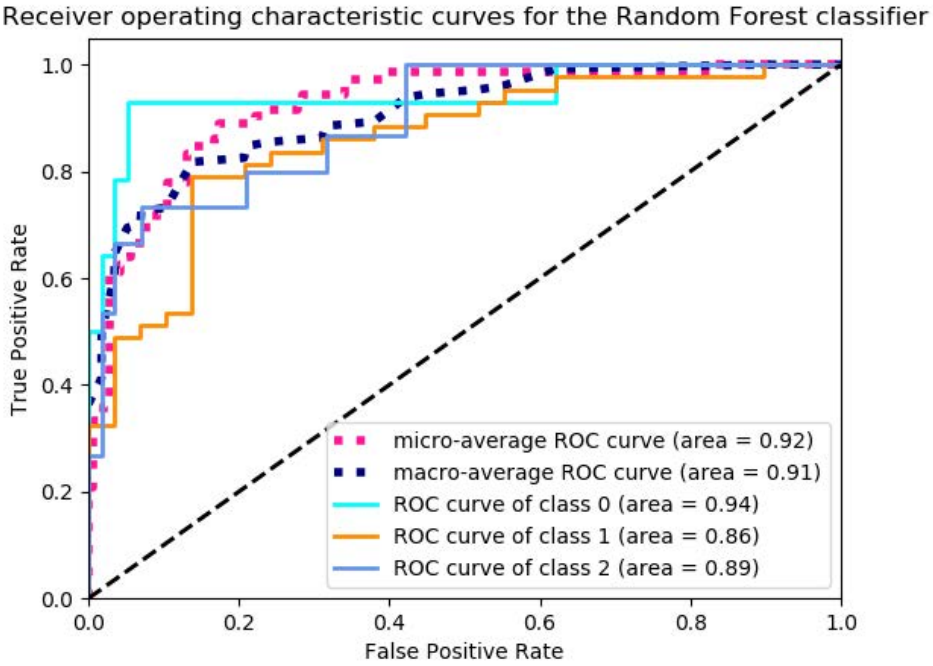


Figure 4.30: Receiver Operating Characteristic (ROC) curves were calculated appropriately for the multi-class Random Forest classifier. The micro -average ROC curve depicts the overall performance of the model across all classes. The macro-average ROC curve illustrates the overall performance of the model for each individual class. The corresponding Area Under the Curve (AUC) is presented in the legend.

Favourable risk category (19.5%), Class 1-Intermediate/Normal risk category (57,2%) and Class 2-Poor risk category (23.3%).

To evaluate the performance of the classifier, the Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) scores were computed, together with the micro- and macro-average performance of the classifier (see Materials and Methods section 3.7.2). The computed AUC scores for the individual classes 0, 1 and 2 were 0.94, 0.86 and 0.89, respectively (Figure 4.30). In addition, the micro-average ROC AUC score was 0.92. These results indicate that the classifier can predict the three classes individually with high sensitivity and specificity. The macro-average ROC AUC score was 0.91, depicting the almost equally high overall prediction accuracy of the model across classes (Figure 4.30).

In addition, the precision-recall curves and their AUC were calculated, in order to evaluate the trade-off between the false positive rate and the false negative rate (see Materials and Methods section 3.7.2). The precision-recall AUC scores for classes 0, 1 and 2 were 0.88, 0.90 and 0.77, respectively (Figure 4.31). The classifier seems to have greater accuracy in the classification of patients of the largest class in the dataset, that of class 1-Intermediate/Normal risk patients. Nonetheless, the model performs considerably well for the two smaller classes 0

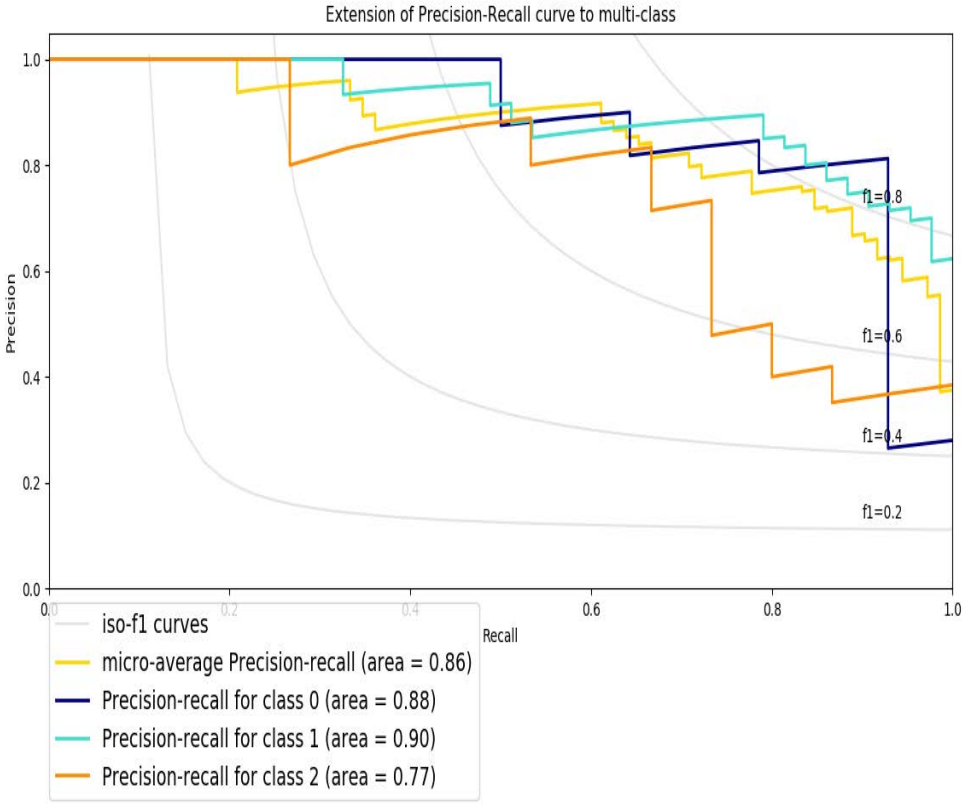


Figure 4.31: Precision and recall curves for all three classes and the micro-average precision-recall curve representing the overall accuracy of the classifier were constructed and the Area Under the Curve (AUC) scores were calculated.

and 2. Additionally, the overall precision score of the classifier micro-averaged for all the classes was calculated and was equal to 0.86 (Figure 4.31), suggesting the overall high predictive power of the classifier.

4.3.3.2.2 FEATURE EXTRACTION

The RF classifier was used to perform feature selection and identify the most important features regarding the classification of AML patients to the Favourable, Intermediate/Normal and Poor risk prognosis groups. In total, 94 features were selected as the most significant for the task and this list of features consisted of both genes (n=86) and metabolic features (i.e. features selected from the metabolic reaction flux dataset) (m=7), as well as the cytogenetics of the patients. In fact, the type of cytogenetic abnormality/ies in patients had the second highest significance score (importance=0.562), according to the RF algorithm, reflecting the already established correlation of karyotyping with AML prognosis. Equally important to highlight, the representation of gene expression and metabolic reaction fluxes over the 94 features identified by the RF corresponds to a proportion of 91% and 7.4% respectively, as expected. Thus, suggesting a well-balanced representation of features in the classification process with the RF. The list of the identified gene expression and the cytogenetics features, together with their importance score calculated by the Random Forest algorithm and a short description of their function are presented in Table 12. Likewise, the identified metabolic reaction fluxes are presented in Table 13.

Table 12: The most important gene expression and cytogenetic features for the risk stratification of AML patients were identified using the Random Forest algorithm.

feature	importance	description
PPP2R2A	0.570	protein phosphatase 2 regulatory subunit Balpha
cytogenetic_abnormality	0.562	type of cytogenetic abnormality/ies
PDE4DIP	0.525	phosphodiesterase 4D interacting protein
IL7	0.379	interleukin 7
CSNK1A1	0.378	casein kinase 1 alpha 1
SLC44A1	0.372	solute carrier family 44 member 1
FGFR1	0.348	fibroblast growth factor receptor 1
RBFOX3	0.299	RNA binding fox-1 homolog 3
TJP2	0.290	tight junction protein 2

AP1S2	0.261	adaptor related protein complex 1 subunit sigma 2
HOXA7	0.259	homeobox A7
PRELID1	0.254	PRELI domain containing 1
MTDH	0.254	metadherin
AK4	0.249	adenylate kinase 4
CPNE8	0.240	copine 8
TMEM38B	0.233	Transmembrane protein 38B
ATP6V1C1	0.229	ATPase H+ transporting V1 subunit C1
ZFHX3	0.227	zinc finger homeobox 3
ASPTA	0.226	aspartoacylase
RIT1	0.218	Ras like without CAAX 1
MPO	0.213	myeloperoxidase
HOXA10	0.213	homeobox A10
ESYT1	0.202	extended synaptotagmin 1
CHMP7	0.201	charged multivesicular body protein7
VPS41	0.200	VPS41 subunit of HOPS complex
CLEC11A	0.199	C-type lectin domain containing11A
HOXA5	0.198	homeobox A5
VEGFA	0.192	vascular endothelial growth factor A
FBXL17	0.192	F-box and leucine rich repeat protein 17
TP53TG1	0.191	TP53 target 1
SIX3	0.184	SIX homeobox 3
KRT17	0.182	keratin 17
SLC40A1	0.181	solute carrier family 40 member 1
PRKDC	0.179	protein kinase, DNA-activated, catalytic subunit
SLC22A16	0.178	solute carrier family 22 member 16

TGIF1	0.178	TGFB induced factor homeobox 1
KCTD9	0.170	potassium channel tetramerization domain containing 9
TAX1BP1	0.168	Tax1 binding protein 1
AK2	0.160	adenylate kinase 2
MTERF3	0.159	mitochondrial transcription termination factor 3
RAD21	0.159	RAD21 cohesin complex component
HOXA9	0.158	homeobox A9
ANP32A	0.148	acidic nuclear phosphoprotein 32 family member A
RAB2A	0.143	RAB2A, member RAS oncogene family
HOXA3	0.140	homeobox A3
RASA1	0.136	RAS p21 protein activator 1
UBR5	0.134	ubiquitin protein ligase E3 component n-recognin 5
FNTA	0.133	farnesyltransferase, CAAX box, alpha
RBPMS	0.133	RNA binding protein, mRNA processing factor
CSF3R	0.132	colony stimulating factor 3 receptor
FUT7	0.130	fucosyltransferase 7
HTATIP2	0.130	HIV-1 Tat interactive protein 2
UBE2B	0.130	ubiquitin conjugating enzyme E2 B
MFN1	0.130	mitofusin 1
RCHY1	0.129	ring finger and CHY zinc finger domain containing 1
ADAM28	0.127	ADAM metallopeptidase domain 28
SLC25A29	0.124	solute carrier family 25 member 29
NUP210	0.123	nucleoporin 210
IKZF1	0.123	IKAROS family zinc finger 1
NAALADL1	0.122	N-acetylated alpha-linked acidic dipeptidase like 1

[156]

RPL8	0.122	ribosomal protein L8
PXK	0.121	PX domain containing serine/threonine kinase like
STIM2	0.121	stromal interaction molecule 2
C11orf21	0.120	chromosome 11 open reading frame 21
FNDC3B	0.119	fibronectin type III domain containing 3B
USP1	0.118	ubiquitin specific peptidase 1
RMND5B	0.116	required for meiotic nuclear division 5 homolog B
HSPA4	0.115	heat shock protein family A (Hsp70) member 4
NPDC1	0.115	neural proliferation, differentiation and control 1
LINC00674	0.114	
ZMIZ1	0.113	zinc finger MIZ-type containing 1
CD84	0.112	CD84 molecule
IGDCC4	0.110	immunoglobulin superfamily DCC subclass member 4
ZBED4	0.110	zinc finger BED-type containing 4
P4HTM	0.109	prolyl 4-hydroxylase, transmembrane
TSPAN32	0.109	tetraspanin 32
KAT6A	0.108	lysine acetyltransferase 6A
IFI35	0.105	interferon induced protein 35
PDLIM5	0.105	PDZ and LIM domain 5
RAD50	0.104	RAD50 double strand break repair protein
IDH3A	0.104	isocitrate dehydrogenase (NAD(+)) 3 catalytic subunit alpha
IRX1	0.104	iroquois homeobox 1
PAWR	0.103	pro-apoptotic WT1 regulator
DERL1	0.100	derlin 1
RHBDF1	0.100	rhomboid 5 homolog 1

EIF4H	0.100	eukaryotic translation initiation factor 4H

Table 13: The most important metabolic features for the risk stratification of AML patients were identified using the Random Forest algorithm.

feature	importance	description
O2t	0.177	o2 transport (diffusion)
SQLSr	0.165	Squalene synthase
ASPGLUm	0.143	aspartate-glutamate mitochondrial shuttle
r0016	0.114	Fe(II):oxygen oxidoreductase Porphyrin metabolism (EC:1.16.3.1)
r0330	0.111	5,6-Dihydrothymine: NAD+ oxidoreductase Pyrimidine metabolism (EC:1.3.1.1)
EX_bilglcur (e)	0.107	exchange reaction for bilirubin mono-glucuronide
FE2t	0.103	iron (II) transport
DHCR7	0.102	7-dehydrocholesterol reductase

In order to obtain some insight on the molecular functions of the subset of the 86 gene expression features identified by the RF, we performed a functional annotation analysis using PANTHER 9.0 (Figure 4.32) [310]. The majority of the genes were associated to the ‘binding’ GO term, including transcription factors, DNA and RNA binding molecules, GTP-binding proteins and histone acetyltransferases, among others. This more general term included the more specific GO terms of ‘transcription regulator activity’ representing eight genes and the ‘molecular function regulator’ term consisting of three genes in our results. The second most abundant GO term among the list of genes was that of ‘catalytic activity’. Indeed, 19 of the genes identified by the RF encoded enzymes involved in several signalling and metabolic pathways. In addition, the GO term ‘transporter

activity' included five genes encoding solute carriers, calcium and ion transporters. The analysis highlighted two genes having 'molecular transducer activity', namely the signal transducer Fibroblast growth factor receptor 1 (FGFR1) and the signal receptor Granulocyte colony-stimulating factor receptor (CSF3R). Finally, the last GO term from this analysis was the 'structural molecule activity' and it included the 60S ribosomal protein L8 (RPL8).

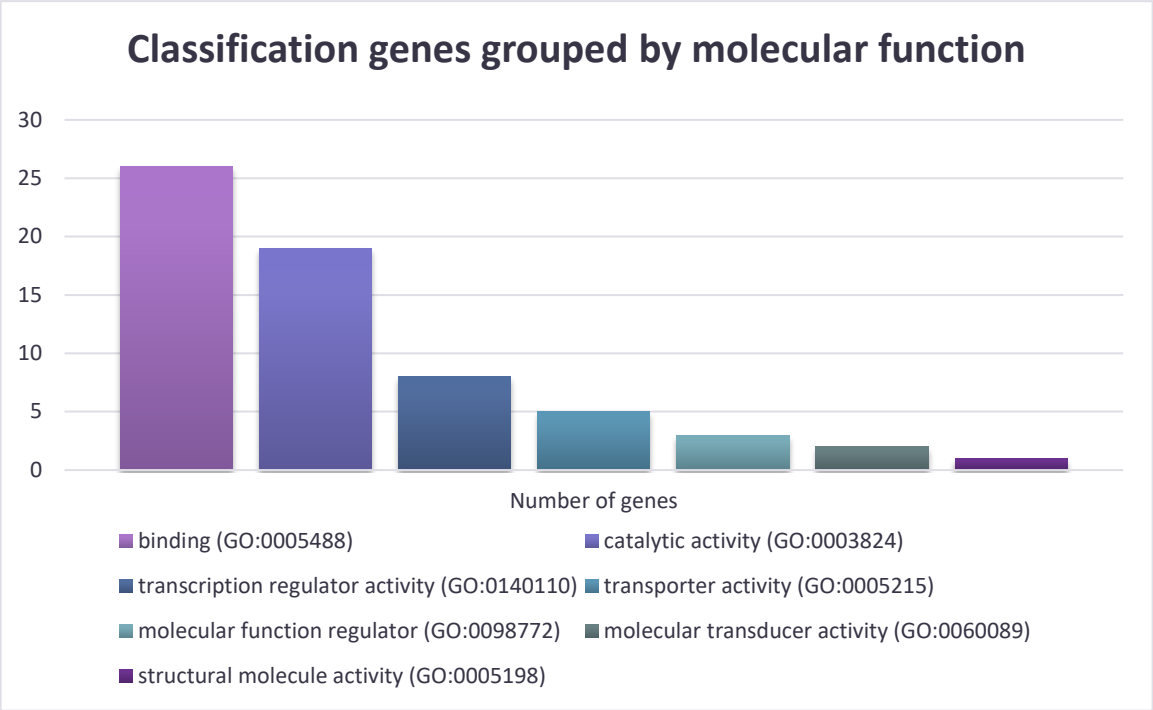


Figure 4.32: A functional annotation of the genes identified by the Random Forest algorithm was performed, to elucidate the molecular functions the genes responsible for the risk stratification of AML patients.

Taking a closer look at the results (Table 12), five members of the HOXA genes, namely the HOXA3, 5, 7, 9 and 10, were found to be critical in the risk stratification of the patients. The HOX genes belong to a family of homeodomain transcription factors highly conserved from Drosophila to humans and are grouped into four classes: A, B, C and D [311]. In general, HOXA genes are expressed in myeloid cells and are crucial in normal hematopoiesis [312]. Several studies have determined the overexpression of HOXA5-10 in AML cell lines and patients, investigated their implication in the disease pathogenesis and suggested their

therapeutic potential as drug targets [313-316]. In particular, HOXA9 has been found to be the most abundantly expressed HOXA member in HSCs, contributing in their proliferation and ultimately in the development of AML [312]. Faber et al. have demonstrated that HOXA9 positively regulates the expression of HOXA7 and HOXA10 genes in AML [317]. The overexpression of both HOXA9 and HOXA10 has been experimentally linked to poor AML patient survival [318, 319]. Recently, Chen et al. performed a bioinformatics analysis of different AML datasets and suggested that the overexpression of HOXA3-10 genes could potentially be used as prognostic biomarkers to distinguish poor survival AML patients [320]. Our results show that only the expression levels of HOXA3, 5, 7, 9 and 10 play a crucial role in the classification of AML patients into the three risk categories. Interestingly, Chen et al. has been the only publication linking HOXA3 to AML prognosis, before our observations. Thus, our work not only falls in line and supports already published findings, but also provides a more specific panel of potential biomarkers within the HOXA family of genes.

Additionally, genes encoding two isoforms of adenylate kinase (AK), the AK2 and AK4, were identified as significant for the risk stratification of AML patients (Table 12). AK has a vital role in metabolic signalling and cellular energy homeostasis. This enzyme catalyses the reversible interconversion of adenosine triphosphate (ATP) and adenosine monophosphate (AMP) into two adenosine diphosphate (ADP) molecules [321, 322]. In this way, AK constantly monitors the levels of ATP, ADP and AMP and therefore acts as a reporter of the cell's energetic state and demands by affecting AMP-sensitive metabolic sensors and enzymes [323, 324]. Nine AK isoforms, named AK1-9, have been identified in humans and they are located in different cellular compartments [325]. Several studies have uncovered a potential role of isoforms AK1, AK2, AK4 and AK6 in the development, progression and metastasis of different types of cancer, e.g. lung, breast or colorectal cancer [326-329]. The AK2 and AK4 isoforms that the RF algorithm identified as potential prognosis biomarkers are located in the mitochondria [325]. AK2 is expressed in the intermembrane mitochondrial space and plays a crucial role in the exchange of nucleotides between the mitochondria and the cytosol, as

well as the export of ATP to the cytosol [321, 330]. AK4 is found in the mitochondrial matrix and it has been implicated in carcinogenesis, drug resistance, regulation of the TCA cycle and protection of cancer cells from oxidative stress [330-332]. To our knowledge, there are no reports linking AML to any of the AK2, AK4 or AK6 isoforms so far. However, Qin et al. demonstrated a correlation of AK1, an isoform located in the cytosol, with a poor clinical outcome of patients receiving chemotherapy or allogeneic hematopoietic stem cell transplantation [333]. In addition, Frejno et al. demonstrated that lower levels of AK1 protein correlated with a longer survival of AML patients receiving AraC treatment, whereas according to Qin et al. higher AK1 protein levels contributed to the development of drug resistance [333, 334]. AK1 is not included in the list of results from our biomarker discovery for AML risk stratification tool. We believe the reason behind this to be that our study has been conducted on data from patients that have never received or were not undergoing any chemotherapy treatment at the time of data collection. Preferably, we hypothesise that the mitochondrial AK2 and AK4 isoforms could serve as more suitable prognostic biomarkers for *de novo* AML patients.

The importance of metabolic features in the risk stratification of AML was prominently highlighted in our results not only by the presence of numerous catalytic enzymes, but also by the inclusion of specific metabolic reactions (Table 13). The exchange reaction of bilirubin mono-glucuronide (EX_bilglcur), the enzyme oxygen oxidoreductase of the Porphyrin metabolic pathway (r0016) and the reactions for iron (II) (FE2t) and oxygen transport (O2t) were identified by the RF algorithm. Haemes are naturally occurring iron (Fe^{2+}) containing porphyrins [335]. They are primarily synthesised in the bone marrow and the liver. Approximately 80% of haem in the human body can be found in the blood [336], since haems are components of the protein haemoglobin, the protein inside red blood cells, and they are required for the reversible binding and transport of oxygen from the lungs to every tissue in the body through the blood stream [335]. When senescent red blood cells or premature red blood cells in the bone marrow are degraded, haem gets converted into bilirubin, through a two-step enzymatic process which requires three molecules of oxygen and NADPH and releases Fe^{2+}

and carbon monoxide [337]. Bilirubin is toxic for the human organism and can cause acute liver failure, jaundice and other pathological issues [338, 339]. For its neutralisation, bilirubin needs to bind to albumin and get transported to the liver for its subsequent excretion to bile [337]. Once in hepatocytes, glucuronidation of bilirubin with glucuronic acid takes place and bilirubin monoglucuronide is formed [340]. We believe that the importance of the final step of haem catabolism to bilirubin in AML patient classification is depicted in our results by the calculation of a high significance score to oxygen oxidoreductase of the Porphyrin metabolic pathway and the reactions for Fe^{2+} and oxygen transport. Moreover, the reaction representing the excretion of bilirubin monoglucuronide to the bile or plasma (EX_bilglcur) was also identified as an important feature in the risk prognosis of AML patients. Only eight cases of patients presenting acute liver failure caused by AML with a very high mortality rate have been reported and a suggestion has been made from Sun et al. for clinicians to consider a link between AML and jaundice or liver failure [341]. In addition, several cases of AML patients developing jaundice or hyperbilirubinemia as a side-effect of receiving different types of chemotherapeutics have been reported [342]. Altogether, we propose the consideration of bilirubin concentration in AML patient samples as a prognostic biomarker and a selection criterion for subsequent treatment.

Also, the correlation of two members of the cholesterol biosynthesis pathway with the disease prognosis of AML was identified by our analysis, namely the enzyme squalene synthase (SQLSr) and the gene encoding the enzyme 7-dehydrocholesterol reductase (DHCR7) (Table 13). Cholesterol is vital for the cell membrane structure and permeability, as well as the synthesis of steroid hormones, bile acids and vitamin D [343, 344]. Squalene synthase catalyses a two-step reaction that results in the formation of squalene, which is first converted to lanosterol and ultimately to cholesterol. The transformation of squalene to cholesterol requires 23 steps and the enzyme 7-dehydrocholesterol reductase catalyses the last of theses reactions, the conversion of 7-dehydrocholesterol to cholesterol [345, 346]. In general, cholesterol biosynthesis is increased in different cancers, including AML, since it is required for the high demands of fast

proliferating cells [347, 348]. Neither squalene synthase nor 7-dehydrocholesterol reductase have been linked to AML prognosis yet. Studies have shown that the use of cholesterol-modulating drugs, in principle statins, sensitise AML drug resistant cells to the chemotherapeutics and provide encouraging results for poor risk patients [349, 350]. However, statins target an enzyme up-stream squalene synthase and inhibit not only cholesterol synthesis, but the non-sterol branch of the pathway as well, leading to severe cytotoxic effects to some patients [351, 352]. Hence, squalene synthase inhibitors are being investigated and proposed as a safer alternative to statins for the inhibition of cholesterol formation [353, 354]. Consequently, we propose the investigation of concentrations of 7-dehydrocholesterol and cholesterol as potential prognostic biomarkers in AML patients and the consideration of squalene synthase inhibitors instead of other cholesterol biosynthesis modulating agents as a co-therapeutic regime for the suitable subgroup of patients.

4.3.4 Discussion

AML is a highly heterogeneous disease, both genetically and phenotypically, which poses great challenges in the clinical aspects of accurate disease prognosis and appropriate treatment selection. The risk stratification of AML patients has been primarily based on the type of chromosomal abnormalities and gene mutations identified [296]. However, approximately 40% of AML patients have a normal karyotype and therefore are misrepresented by the current stratification regime [298, 355]. On the other hand, the need for more personalised therapeutic regimes dependant on the genetic and phenotypic traits of individual patients is becoming increasingly evident, given the high percentage of patients not benefitting from standard chemotherapy drugs due to acquired or innate drug resistance and the high rate of disease relapse among AML patients [356, 357]. An accurate panel of prognosis biomarkers and a more representative stratification of AML patients will provide new insights on the development of better therapeutics for the different disease subgroups.

We tried to address AML patient risk stratification and biomarker discovery by developing a new computational approach that combines personalised GSMMs [163]

and machine learning techniques. The integration of patient derived omic data onto a genome scale metabolic reconstruction led to patient-specific metabolic models and the simulation of the metabolic potential of each patient. We hypothesised that the combination of these individual metabolic profiles of the patients with their gene expression profiles and their demographic data would provide an accurate risk classification method and new prognostic biomarkers accounting for the metabolic changes that occur in AML patients.

To achieve this task we faced two obstacles: first, the absence of publicly available AML datasets that included both transcriptomics and genome scale metabolomics of the patients and secondly, the limitation of most machine learning tools to handle datasets where the number of features to use for a classification problem far exceeds the number of available samples. We managed to overcome both challenges successfully. In the first case we reconstructed a consensus GSMM for AML using metabolomic data from highly phenotypically heterogeneous AML cell-lines. In the second case we used a sparse supervised extension to the Canonical Correlation Analysis to perform feature reduction prior to building a machine learning classifier.

Random Forests was the algorithm chosen to perform the risk stratification of patients and the identification of potential AML prognosis biomarkers. RFs can inherently handle well large and complex datasets that comprise of different types of parameters, such as continuous or categorical. Finally, we built a highly accurate classifier for AML patients that considers chromosomal abnormalities, gene mutations, gene expression profiles, demographic data and the metabolic profiles of the patients.

The results from our work included some already established prognostic biomarkers and identified some new genetic and metabolic features, whose prognostic value in AML is worth exploring. The type of cytogenetic abnormality/ies of each patient was identified as highly significant in the risk stratification of AML patients, as already well determined and applied by the scientific community. Additionally, the aberrant expression of specific genes of the HOXA family and of

two adenylate kinase isoforms had already been proposed as poor risk disease prognostic markers by several studies and was highlighted by our analysis as well [311, 315, 317, 318, 333]. Interestingly, the implication of several, otherwise suspected for the pathogenesis of the disease, metabolites in the risk stratification of the patients was suggested by our method. Haem catabolism and its end product bilirubin, as well as the sterol branch of cholesterol metabolism were distinctly represented in our results; thus, suggesting their implication in AML prognosis and their putative value as biomarkers.

Given the accentuation of metabolic genes and reactions by our biomarker discovery method, we believe that the consideration of metabolic features in biomarker panels for AML patients would be greatly beneficial in the early diagnosis stages. Equally important, we propose that the incorporation of integration methods for patient omic data, such as the reconstruction of personalised GSMMs that we demonstrated, coupled with the use of ML based decision support tools for the risk stratification of patients in clinical practice could provide meaningful indications for the planning of a more patient-specific and efficient therapeutic strategy.

5.

GENERAL DISCUSSION

5 General discussion

Systems biology provides a holistic approach to studying the complex and interwinding interactions among the genetic, proteomic, metabolic and cellular components of organisms [73]. To do so, tools and methods from different disciplines must come together to generate biological data, analyse them and mathematically transform them into descriptive and predictive models of cell, tissue, organism or disease systems. These multiscale models can then be employed in the discovery of condition specific traits, biomarkers, drug targets or even the stratification of patients.

Metabolic modelling methods exhibit significant advantages in the systems biology field. Metabolism represents the closest state of a system to its phenotype and it is the only omic layer that can be fully modelled at genome scale [120]. Metabolic models are mathematical representations of the complex network of all the biochemical reactions that take place in a cell and provide an excellent platform for the integration of other omic datasets, e.g. transcriptomics, fluxomics or proteomics. Thus, providing a strong connection between the genotype and phenotype of a cell or an organism, one of the main objectives in systems biology and biomedical research [119].

Throughout this thesis, new bioinformatics workflows based on the reconstruction and interrogation of Genome Scale Metabolic Models (GSMMs) have been designed to explore various aspects of Acute Myeloid Leukaemia (AML), a highly heterogeneous cancer, both genetically and phenotypically [297]. The functional characterisation and the therapeutic potential of the putative oncogene Transketolase-like 1 (TKTL1) (Chapter 4.1), the metabolic rewiring caused by drug resistance and strategies to overcome it (Chapter 4.2), as well as AML patient risk stratification and the expansion of the prognostic biomarkers for the disease (Chapter 4.3) were all addressed in our work.

Over the last years, increasing interest has been demonstrated towards deciphering the role of TKTL-1 in cancer development, progression and resistance to chemotherapeutics [358-360]. It is a key enzyme in the non-oxidative branch of the Pentose Phosphate Pathway (PPP), one of the most crucial metabolic pathways for cancer cell survival and proliferation [224, 361]. The overexpression of TKTL-1 observed in a large number of AML patients brought forward the question of deciphering the implication of this gene in the pathogenesis and treatment of AML. Towards that end, in Chapter 4.1 we selected THP-1, an AML cell-line that naturally exhibits high TKTL1 gene expression and performed an extensive bioinformatics analysis and the reconstruction of GSMMs integrating genome wide transcriptomic and metabolomic data from the parental THP-1 cells and THP-1 cells with a TKTL1 gene knock down.

Our work revealed a potential mechanism of survival of AML cells, orchestrated by TKTL1 by regulating Ser/Thr kinase Akt phosphorylation, which in turn controls mTOR Complex 2 activity and blocks the Glycogen Synthase Kinase 3 (GSK3)-mediated degradation of several cell survival or proliferation proteins, e.g. that of antiapoptotic Myeloid Cell Leukaemia 1 (MCL-1) [30].

The inhibition of TKTL1 expression in several cancer types and in AML has been shown to reduce cancer cell proliferation significantly, however it has not proven entirely effective as a sole therapeutic agent [358, 362, 363]. For this reason, we performed high-throughput gene knock-out simulations using the reconstructed GSMMs of the THP-1 parental and the THP-1 TKTL1 knock down cells. Hence, we managed to identify putative metabolic vulnerabilities that could be targeted in combination with TKTL1 inhibition and would provide a new therapeutic approach against AML. Our analysis showed a strong dependency of the TKTL1 knock down cells to Complex III and Complex IV of the respiratory chain and to oxidative phosphorylation. We considered that the evaluation of drugs used in cancer therapy directly targeting these complexes, for example isothiocyanates and arsenic trioxide, to be outside the scopes of this work due to their high toxicity [364]. However, targeting the electron transport chain and oxidative

phosphorylation is considered an emerging promising strategy in cancer treatment and therefore we believe that more efforts should be directed towards this area of research [365].

Additionally, our results showed an upregulation of several genes regulated by promoter hypomethylation in AML following the knock down of TKTL1, suggesting that its inhibition had the same effect as the administration of hypomethylating agents to the cells. At the same time, the GSMM based simulations determined an increase of flux through one carbon metabolism and the folate cycle, crucial pathways for DNA methylation [233, 235]. Therefore, we hypothesised that TKTL1 has an important implication in regulating epigenetic methylation events in AML.

Venetoclax (or ABT-199) is a B-cell lymphoma 2 (BCL2)-selective inhibitor [366]. BCL2 belongs to a family of apoptosis regulatory proteins that is often deregulated in cancer [367, 368]. BCL2, as a pro-survival protein, is localised in the outer mitochondrial membrane and prevents proteins BAX and BAK from activating apoptosis by causing outer membrane permeabilization and the release of cytochrome c and other mitochondrial proteins into the cytosol [369]. Cancer cells have been found to overexpress BCL2 to resist apoptosis and enhance their survival [370]. Thus, targeting BCL2 has met a lot of interest and its specific inhibitor Venetoclax is now an approved and widely used chemotherapeutic in Chronic Lymphoid Leukaemia (CLL) and in AML [371-373]. In fact, a synergetic effect of hypomethylating agents and Venetoclax in AML has been established and exploited as combination therapy [374]. What is more, a recently published work by Roca-Portoles et al., 2020, demonstrated that Venetoclax inhibited the mitochondrial TCA cycle and extensively inhibited mitochondrial respiratory complex function, in an indirect manner and completely independently of its BCL2 inhibitory action [375]. Moreover, in their work, the authors obtained these results consistently across different cancer cell lines, suggesting that this effect of Venetoclax on mitochondrial metabolism is not dependent on the cell model. Also, their results indicated a consistent inhibition of respiratory Complex IV using different Venetoclax concentrations, whereas at lower concentrations the function

of Complex I was not inhibited and the function of Complex II only partially inhibited. For these reasons, we decided to validate our hypothesis regarding the role of TKTL1 as an hypomethylating agent and the hypothesis generated by the GSMM gene essentiality analysis highlighting respiratory Complexes III and IV as putative metabolic vulnerabilities of TKTL1 knock down cells by treating them with Venetoclax. Indeed, Venetoclax exhibited a selective inhibitory effect on the THP-1 knock down cells compared to the parental, thus supporting our hypotheses. The inhibition of TKTL1 in combination with Venetoclax could serve as a potential new therapeutic approach for AML subtypes naturally exhibiting high TKTL1 expression.

In Chapter 4.2 we employed GSMMs in order to characterise the metabolic rewiring that occurs in AML cells due to drug resistance, the greatest challenge in AML treatment until today. We focused on two commonly used chemotherapy drugs, Cytarabine (AraC) and Doxorubicin (DOX) and their effect on two phenotypically different AML cell lines, the THP-1 and the HL60. In order to increase the accuracy and robustness of the GSMM reconstructions and better model each cell line before and after the establishment of drug resistance we integrated measurements of basal and maximal mitochondrial respiration of the cells. It has been shown that different types of cancer depend on glycolytic and mitochondrial ATP production at varying degrees [376-378]. Moreover, different subtypes of triple-negative breast cancer exhibited high heterogeneity in respect to their bioenergetic profiles [379]. Given the high phenotypic diversity of AML subtypes, we believe the introduction of such data to have been critical in the accuracy of our metabolic modelling of AML cell lines.

The workflow we applied for our analysis in Chapter 4.2 demonstrated both common and distinct gene expression and metabolic signatures, illustrating the complexity and heterogeneity of AML. We identified a manifestation of traits attributed to Cancer Stem Cells (CSCs) and Leukemic Stem Cells (LSCs) in THP-1 and HL60 drug resistant cells. The upregulation of gene sets representing cell adhesion molecules and cell-cell adhesion events is characteristic of LSCs and

has been known to contribute to AML pathogenesis, progression and disease relapse following chemotherapy treatment [270-273]. In addition, the significant upregulation of gene sets related to inflammatory response, leukocyte activation and migration, as well as the upregulation in the gene expression of C-X-C chemokine receptor type 4 (CXCR4) observed in our results corresponds to the mechanism employed by metastatic CSCs [276-281]. Thus, we have presented evidence of the suspected metastatic potential of drug resistant AML cells.

Furthermore, several putative metabolic vulnerabilities were determined based on high throughput gene knock-out simulations performed on the GSMM reconstructions of the AML untreated and drug resistant cells. The cholesterol biosynthesis pathway was characterised as a THP-1 DOX resistant cell specific essential gene and was successfully evaluated experimentally with the use of a squalene synthase inhibitor. Our analysis led to the identification of common metabolic targets that would compromise the viability of both THP-1 and HL60 drug resistant cells. Indeed, the administration of antifolate drugs, namely Methotrexate and Pemetrexate, proved effective against all drug resistant AML cells. Also, cardiolipin synthesis inhibitor, Alexidine, showed some efficiency against HL60 DOX resistant cells compared to the parental.

Our analysis highlighted several additional metabolic vulnerabilities that have gained scientific interest regarding their role in AML and that are being tested as potential drug targets in clinical trials. We present a summary of our results in this context, more specifically the name, description and pathway affected by the essential gene identified from our analysis, the respecting cell model and the bibliographic reference to the study of interest, in Table 14.

Table 14: Essential genes identified in our work that are being investigated as therapeutic targets in AML in clinical trials.

Essential gene	Description	Pathway affected	Identified as essential gene from GSMMs for	Literature reference
CAD	Dihydroorotase	De novo pyrimidine biosynthesis	All conditions	Phase I/II [382] Phase Ib [384]
DHODH	Dihydroorotate dehydrogenase (quinone)			
UMPS	Orotidine-5-phosphate decarboxylase			
HMGCR	3-hydroxy-3-methylglutaryl-coenzyme A reductase	Mevalonate biosynthesis	THP-1 DOX-resistant HL60 Parental HL60 AraC-resistant	Phase I/II [389, 390]
MVD	Diphosphomevalonate decarboxylase			
MVK	Mevalonate kinase			
PMVK	Phosphomevalonate kinase			
ASS1	Argininosuccinate synthetase-1	Urea cycle	HL60 AraC-resistant	Phase II [391]
ASL	Argininosuccinate lyase			

We identified three members from the *de novo* pyrimidine biosynthesis pathway as putative targets for all cell line models (i.e. THP-1 and HL60 parental, AraC- and DOX-resistant), namely dihydroorotase (CAD), dihydroorotate dehydrogenase (DHODH) and orotidine-5-phosphate decarboxylase (UMPS). The *de novo* pyrimidine biosynthesis pathway (or the orotate pathway) provides pyrimidine nucleotides that are crucial in DNA and RNA synthesis and are also involved in biosynthesis of glycogen and phospholipids [6]. Even though pyrimidines are synthesised in the cytosol, DHODH, the limiting enzyme of the pathway that catalyses its fourth step, is localised in the inner mitochondrial membrane [380]. Therefore, DOHDH connects pyrimidine biosynthesis to mitochondrial bioenergetics and provides a new approach in targeting the electron transfer chain and oxidative phosphorylation [381]. The use of different DOHDH inhibitors in AML has provided very promising results in several clinical and preclinical studies [380, 382-384]. In the latest of them, Cao et al., 2019, presented a novel inhibitory compound that displayed increased efficacy and low cytotoxicity in preclinical AML trials [385].

Moreover, our analysis suggested that the mevalonate biosynthesis pathway constitutes a putative vulnerability for the THP-1 DOX-resistant, the HL60 parental untreated and the HL60 AraC-resistant cells. Our analysis proposed 3-hydroxy-3-methylglutaryl-coenzyme A reductase (HMGCR), diphosphomevalonate decarboxylase (MVD), mevalonate kinase (MVK) and phosphomevalonate kinase (PMVK) as essential genes of the aforementioned cell models. The mevalonate pathway plays a crucial role in metabolism since it results in the production of cholesterol and ubiquinone, among many others [386, 387]. Therefore, the mevalonate pathway fuels major cellular functions, from cell membrane function to the electron transfer and epigenetic modifications [387]. It is targeted via HMGCR inhibition, which is the rate limiting enzyme of the pathway. HMGCR is the target of the widely used class of drugs called statins [388]. In AML, statins and specific targeting of the mevalonate pathway through HMGCR is being widely used as a combination therapy with chemotherapeutics, including AraC [392, 393]. The response rate to this combinatorial approach has been estimated around 75% for

relapse patients, however studies showed that de novo diagnosed AML patients do not benefit equally from it [394]. Our results highlighted HMGCR and the mevalonate pathway as vulnerabilities for the THP-1 DOX-resistant and the HL60 AraC-resistant cells, both cell lines but resistant to different drugs. Additionally, HMGCR was an essential gene for the HL60 untreated cells. Thus, it could be concluded that AML subgroups with HL60-like phenotypes could potentially benefit from the combination regiment of AraC along with statins as a synergic therapy.

Furthermore, argininosuccinate synthase 1 (ASS1) and argininosuccinate lyase (ASL) were highlighted as putative metabolic vulnerabilities specific for the HL60 AraC-resistant cells. Arginine is a vital amino acid for a wide range of cellular processes, such as synthesis of proteins and urea, immunological response and nitric oxide synthesis [395]. Arginine can be consumed from extracellular sources or it can by synthesised endogenously from the cells, thus it is a semi-essential amino acid [396]. The synthesis of arginine involves two reactions. The first reaction is catalysed by ASS1, which converts citrulline and aspartic acid to argininosuccinate and the second reaction is catalysed by ASL, which converts argininosuccinate to arginine and fumaric acid. Many different types of cancers display a deficiency of the enzyme ASS1, they are unable to synthesize arginine and therefore rely predominantly on the uptake of arginine from the environment [397]. Arginine addiction has been well documented in AML and the majority of AML patients present with a deficiency of ASS1. In such cases, arginine deprivation using arginine depleting enzymes has been explored as a potential therapeutic approach [397]. However, Miraki-Moud et al., 2015, demonstrated that a subpopulation of primary AML samples were resistant to arginine depletion treatment, since they expressed relatively higher levels of ASS1 and concluded that ASS1 could be used as a predictive biomarker for patient response to arginine deprivation treatment [398]. Our results suggest that HL60 AraC resistant cells have a greater dependency on the endogenous pathway of arginine synthesis, than the other cell models we examined. Actually, the gene expression analysis revealed an upregulation of the ASS1 gene by a log2FC of 2.28 in the HL60 AraC-resistant cells compared to the parental untreated. It could be therefore

hypothesised that AraC resistance triggers an upregulation of ASS1, inhibition of which could be exploited as a complementary treatment.

We decided to explore the phenotypic heterogeneity of AML further and address the pressing issue of proper AML patient risk stratification by designing a workflow combining patient specific GSMMs and machine learning tools. Currently, the classification of AML patients into different subgroups and subsequently into prognosis related risk categories is primarily based on their molecular and cytogenetic characterisation [296, 297, 299]. This approach might be inadequate in efficiently characterising the big part of AML patients with a normal karyotype, which is considered as the most genetically heterogeneous subgroup and only has an intermediate response to chemotherapeutics [298, 355]. A more accurate and extensive panel of biomarkers for the classification of patients into risk categories could provide valuable insight on designing personalised therapies for AML patients. Such advancements would potentially result in the identification of new drug targets, improve the suitability of chemotherapeutics at the patient level and challenge the issue of intrinsic and acquires drug resistance. We believe that by expanding the existing selection of prognostic biomarkers with the addition of metabolic features would contribute towards that aim. Especially, given the already established crucial role of metabolism in the pathogenesis and progression of cancer and the response to drug treatment [286, 399-401].

For the purposes of Chapter 4.3, a large collection of patient data, namely their chromosomal abnormalities, gene mutations, gene expression profiles, demographic data taken from The Cancer Genome Atlas (TCGA) database and the simulated metabolic profiles of the patients were used to build Random Forest classifiers and perform feature extraction to identify the most important features responsible for the risk stratification of the patients (Figure 5.1). Due to the lack of genome wide metabolomic data on patients, we reconstructed a consensus GSMM of AML based on cell line metabolomic data generated in house and then used patient derived genome wide transcriptomics to build patient specific GSMMs.

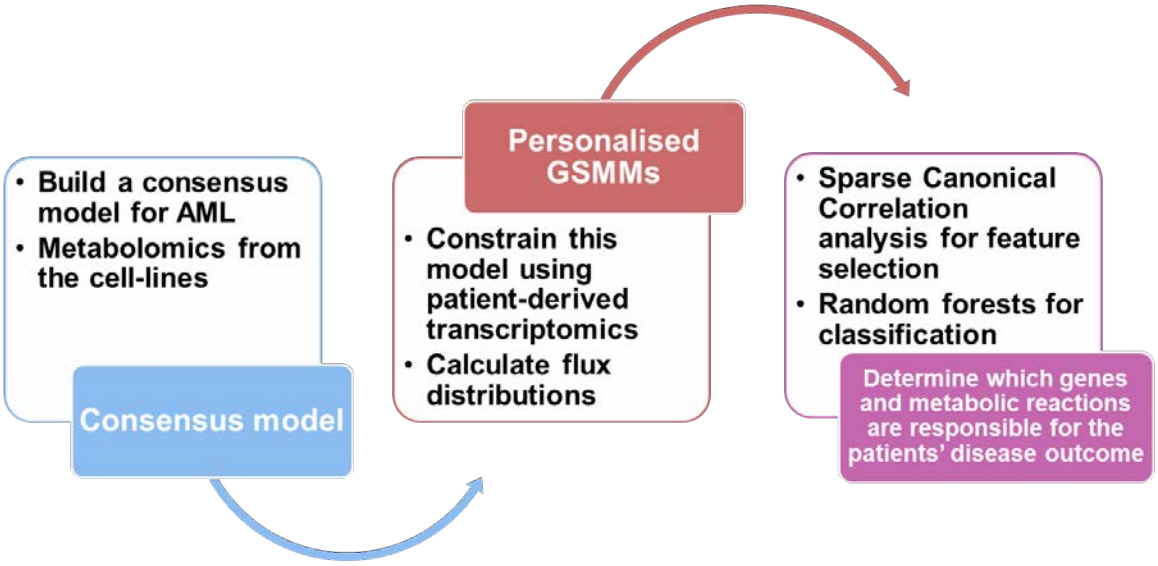


Figure 5.1: Bioinformatics workflow designed to extract genomic and metabolomic biomarkers for AML prognosis. Through this work we introduce a consensus Genome Scale Metabolic Model for AML and the tools to reconstruct patient specific models. Combining constraint-based modelling with machine learning tools, we managed to identify already reported and new biomarkers to serve in the risk stratification of AML patients.

According to our results, the cytogenetics of the patients had the second highest importance score for risk stratification in AML. Additionally, the expression of several genes that have already been correlated with AML prognosis according

to the literature were identified by our classification tool, e.g. five members of the HOXA genes, namely the HOXA3, 5, 7, 9 and 10, thus validating the congruity of our method and findings.

The value of the mitochondrial adenylate kinase (AK) isoforms 2 and 4 was implied by our work. AK plays a crucial role in cellular metabolism and energy homeostasis, by catalysing the reversible interconversion of adenosine triphosphate (ATP) and adenosine monophosphate (AMP) into two adenosine diphosphate (ADP) molecules and serving as a reporter of the energy demands and state of the cell [321-324]. Isoforms AK1, AK2, AK4 and AK6 have been implicated in the pathogenesis and metastatic potential of lung, breast and colorectal cancer and AK1 has been reported as a prognostic biomarker for AML patients undergoing AraC treatment [326-328, 330-334]. Therefore, and to our knowledge, our study is the first to report the putative prognostic value of AK isoforms 2 and 4 for *de novo* diagnosed AML patients. Another important point is that AK2 and AK4 are the two isoforms of AK situated in the mitochondria, in the intermembrane mitochondrial space and in the mitochondrial matrix respectively [325]. Hence, our analysis points once more to the importance of mitochondrial bioenergetics in AML.

In addition, key enzymes and metabolic reactions from haem catabolism and bilirubin excretion were outlined in our results as highly correlated to the prognosis of AML patients. Upon the degradation of red blood cells, which hold the majority of haem in the human body, haem gets converted to bilirubin, a toxic compound for the human body that can cause liver failure and jaundice, among others [335-339]. Hyperbilirubinemia and elevated liver transaminases have been reported as side effects of chemotherapy treatment to AML patients [70]. However, rare cases where liver failure or jaundice were the first manifestations of newly diagnosed AML patients have been reported and were linked with very poor prognosis and high mortality rates [402-404]. Since our analysis was conducted using data from *de novo* diagnosed AML patients that had not undergone chemotherapy treatment, we believe that our results and the identified significance of bilirubin as a prognostic

biomarker strengthens the hypothesis that links liver physiology with AML and has direct implications in selecting appropriate therapeutic strategies for patients with predisposition for acute liver failure [342, 405].

Finally, we identified an implication of two key enzymes of the cholesterol pathway, squalene synthase (SQLS) and 7-dehydrocholesterol reductase (DHCR7), in the risk stratification of AML patients. The deregulation of cholesterol biosynthesis has been reported in different cancer types and AML, while inhibition of this pathway using statins or mevalonate-pathway inhibitors has been widely used as combination therapy with chemotherapeutics re-sensitising drug resistant cells [347, 350]. However, our results point to the potential prognostic value of enzymes in the sterol branch of the cholesterol synthesis pathway. Interestingly, SQLS was identified and validated as a specific metabolic vulnerability of the THP-1 DOX resistant cells in Chapter 4.2. Consequently, we propose the consideration of DHCR7 as an appropriate prognosis biomarker for AML patients and the investigation of SQLS inhibitors as more specific combination therapy for AML.

Overall, the work performed as part of this Ph.D. provides insights on the metabolic reprogramming and mechanisms employed by AML cells to overcome different treatments and how metabolic modelling can be used in the discovery of putative metabolic vulnerabilities and potential new drug targets, as well as in the risk stratification of AML patients. We have introduced for the first time GSMM reconstructions integrating a variety of genome wide omic data and cell respiration data for AML and we have demonstrated the effectiveness of our modelling workflow to capture distinct metabolic signatures in different subgroups of such a heterogeneous cancer as AML. Additionally, we have designed a workflow for the reconstruction of AML patient specific GSMMs, taking one step closer towards personalised therapy in AML. Such reconstructions could find use in clinical practice for the simulation of different therapeutics for the identification of suitable therapeutic strategies and prevention of potential side effects. Our proposal for combining GSMMs and machine learning tools resulted in a robust classification tool for the risk stratification of AML patients and the identification of new metabolic

biomarkers with a potential prognostic value. Future works could focus on the evaluation of the putative metabolic targets and biomarkers identified as part of this work in more AML cell lines and in animal models. The further refinement of our classification tool would also be of interest and for this genome scale omic data from a greater number of AML patients would be needed.

6.

CONCLUSIONS

6 CONCLUSIONS

1. Genome Scale Metabolic Models (GSMMs) and constraint-based modelling provide an appropriate platform for the successful and biologically interpretable integration of multimodal data, including metabolomics, transcriptomics and cell respiration data.
2. Our results suggest that Transketolase-like 1 (TKTL1) is a moon light enzyme, which affects major signalling pathways and has a direct implication in regulating methylation events in AML.
3. The combination of TKTL1 inhibition and Venetoclax is a putative promising therapeutic strategy in Acute Myeloid Leukaemia (AML).
4. The upregulation of genes related to inflammatory response, leukocyte activation and migration, as well as C-X-C chemokine receptor type 4 (CXCR4) observed provides evidence of the suspected metastatic potential of AML.
5. Our workflow resulted in the reconstruction of highly accurate condition-specific GSMMs able to identify the differences in the metabolic reprogramming of subgroups of AML.
6. Inhibition of squalene synthase severely compromised THP-1 DOX resistant cells compared to the untreated AML cells and showed promise as a combination therapy *in vitro*.
7. A common dependency of AML drug-resistant and untreated cells on Folate metabolism was identified *in silico* and successfully validated *in vitro* with the administration of Methotrexate and Pemetrexate.
8. The use of gene expression and flux distribution datasets in machine learning classification problems greatly benefits from dimensionality reduction.
9. The combination of constraint-based modelling and Random Forests provided a highly accurate and robust classifier for AML patient risk-stratification.

10. Adenylate kinases 2 and 4, bilirubin, squalene synthase and 7-dehydrocholesterol reductase were identified as putative new prognosis biomarkers for de novo diagnosed AML patients.

7 ABREVIATIONS

AK: Adenylate Kinase

AM: Adhesion Molecules

AML: Acute Myeloid Leukaemia

AraC: Cytarabine

DE: Differential Expression

DHCR7: 7-Dehydrocholesterol Reductase

DOX: Doxorubicin

EMDR: Environment-Mediated Drug Resistance

FBA: Flux Balance Analysis

GIMME: Gene Inactivity Moderated by Metabolism and Expression

GO: Gene Ontology

GSEA: Gene Set Enrichment Analysis

GSMM: Genome Scale Metabolic Model

KEGG: Kyoto Encyclopaedia of Genes and Genomes

LSC: Leukemic Stem Cell

MICM: Morphologic, Immunologic, Cytogenetic and Molecular

ML: Machine Learning

MOMA: Minimization Of Metabolic Adjustment

mTORC: Mammalian Target Of Rapamycin Complex

NADPH: Nicotinamide Adenine Dinucleotide Phosphate

OXPHOS: Oxidative Phosphorylation

PPP: Pentose Phosphate Pathway

RF: Random Forest

ROS: Reactive Oxygen Species

SQLS: Squalene Synthase

ssCCA: supervised sparse Canonical Correlation Analysis

TKT: Transketolase

TKTL1: Transketolase-like 1

WHO: World Health Organisation

8 REFERENCES

1. Tavassoli, M., *The cell theory: a foundation to the edifice of biology.* The American journal of pathology, 1980. **98**(1): p. 44-44.

2. Clark, D.P., N.J. Pazdernik, and M.R. McGehee, *Molecular Biology*. 2018: Elsevier Science.

3. Alberts, B., Molecular biology of the cell. 2015.

4. Watson, J.D. and F.H.C. Crick, Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. Nature, 1953. **171**(4356): p. 737-738.

5. Crick, F., *Central Dogma of Molecular Biology.* Nature, 1970. **227**(5258): p. 561-563.

6. Berg, J.M., J.L. Tymoczko, and L. Stryer, *Biochemistry, Fifth Edition*. 2002: W.H. Freeman.

7. Clark, D.P., N.J. Pazdernik, and M.R. McGehee, *Chapter 1 - Cells and Organisms*, in *Molecular Biology (Third Edition)*, D.P. Clark, N.J. Pazdernik, and M.R. McGehee, Editors. 2019, Academic Cell. p. 2-37.

8. Fell, D., Understanding the control of metabolism. 1996: Portland.

9. Mitsis, T., et al., Transcription factors and evolution: An integral part of gene expression (Review). World Acad Sci J, 2020. **2**(1): p. 3-8.

10. Cooper, G.M., *The cell : a molecular approach*. 2000, Washington, D.C.; Sunderland, Mass.: ASM Press ; Sinauer Associates.

11. Ryšlavá, H., et al., Effect of posttranslational modifications on enzyme function and assembly. Journal of Proteomics, 2013. **92**: p. 80-109.

12. Cornish-Bowden, A., *Fundamentals of Enzyme Kinetics*. 2013: Wiley.

13. Nelson, D.L., A.L. Lehninger, and M.M. Cox, *Lehninger principles of biochemistry*. 2008: Macmillan.

14. Hanahan, D. and R.A. Weinberg, *The Hallmarks of Cancer.* Cell, 2000. **100**(1): p. 57-70.

15. Hanahan, D. and R.A. Weinberg, *Hallmarks of cancer: the next generation.* Cell, 2011. **144**(5): p. 646-74.

16. Lee, J.K., et al., Mechanisms and Consequences of Cancer Genome Instability: Lessons from Genome Sequencing Studies. Annu Rev Pathol, 2016. **11**: p. 283-312.

17. Lodish H, B.A., Zipursky SL, et al, *Proto-Oncogenes and Tumor-Suppressor Genes*, in *Molecular Cell Biology*. 2000, W. H. Freeman: New York.

18. Fouad, Y.A. and C. Aanei, *Revisiting the hallmarks of cancer.* American Journal of Cancer Research, 2017. **7**(5): p. 1016-1036.

19. Fernández-Medarde, A. and E. Santos, *Ras in cancer and developmental diseases.* Genes & cancer, 2011. **2**(3): p. 344-358.

20. Koretzky, G.A., *The legacy of the Philadelphia chromosome.* The Journal of clinical investigation, 2007. **117**(8): p. 2030-2032.

21. Peiris, M.N., F. Li, and D.J. Donoghue, *BCR: a promiscuous fusion partner in hematopoietic disorders.* Oncotarget, 2019. **10**(28): p. 2738-2754.

22. Vélez-Cruz, R. and D.G. Johnson, *The Retinoblastoma (RB) Tumor Suppressor: Pushing Back against Genome Instability on Multiple Fronts.* International journal of molecular sciences, 2017. **18**(8): p. 1776.

23. Liu, Y., et al., Deletions linked to TP53 loss drive cancer through p53-independent mechanisms. Nature, 2016. **531**(7595): p. 471-475.

24. Perri, F., S. Pisconti, and G. Della Vittoria Scarpati, *P53 mutations and cancer: a tight linkage.* Annals of translational medicine, 2016. **4**(24): p. 522-522.

25. Choi, J.D. and J.-S. Lee, *Interplay between Epigenetics and Genetics in Cancer.* Genomics & informatics, 2013. **11**(4): p. 164-173.

26. Bennett, R.L. and J.D. Licht, *Targeting Epigenetics in Cancer.* Annu Rev Pharmacol Toxicol, 2018. **58**: p. 187-207.

27. Sanchez-Vega, F., et al., *Oncogenic Signaling Pathways in The Cancer Genome Atlas.* Cell, 2018. **173**(2): p. 321-337.e10.

28. Yang, Y., et al., *Emerging agents that target signaling pathways in cancer stem cells.* Journal of Hematology & Oncology, 2020. **13**(1): p. 60.

29. Yang, J., et al., Targeting PI3K in cancer: mechanisms and advances in clinical trials. Molecular Cancer, 2019. **18**(1): p. 26.

30. Manning, B.D. and A. Toker, *AKT/PKB Signaling: Navigating the Network.* Cell, 2017. **169**(3): p. 381-405.

31. Jiramongkol, Y. and E.W.F. Lam, *FOXO transcription factor family in cancer and metastasis.* Cancer and Metastasis Reviews, 2020. **39**(3): p. 681-709.

32. Zimta, A.-A., et al., The Role of Nrf2 Activity in Cancer Development and Progression. Cancers, 2019. **11**(11): p. 1755.

33. Chen, B., et al., HIF1A expression correlates with increased tumor immune and stromal signatures and aggressive phenotypes in human cancers. Cellular Oncology, 2020. **43**(5): p. 877-888.

34. Tian, T., X. Li, and J. Zhang, *mTOR Signaling in Cancer and mTOR Inhibitors in Solid Tumor Targeting Therapy.* International journal of molecular sciences, 2019. **20**(3): p. 755.

35. Forbes, S.A., et al., *COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer.* Nucleic acids research, 2011. **39**(Database issue): p. D945-D950.

36. Hsieh, A.C., et al., The translational landscape of mTOR signalling steers cancer initiation and metastasis. Nature, 2012. **485**(7396): p. 55-61.

37. Laplante, M. and D.M. Sabatini, *mTOR signaling in growth control and disease.* Cell, 2012. **149**(2): p. 274-93.

38. Lamming, Dudley W. and David M. Sabatini, *A Central Role for mTOR in Lipid Homeostasis.* Cell Metabolism, 2013. **18**(4): p. 465-469.

39. Kim, S.-Y., *Cancer Energy Metabolism: Shutting Power off Cancer Factory.* Biomolecules & therapeutics, 2018. **26**(1): p. 39-44.

40. Lu, J., *The Warburg metabolism fuels tumor metastasis.* Cancer and Metastasis Reviews, 2019. **38**(1): p. 157-164.

41. Tarrado-Castellarnau, M., et al., De novo MYC addiction as an adaptive response of cancer cells to CDK4/6 inhibition. Mol Syst Biol, 2017. **13**(10): p. 940.

42. Huber, V., et al., Cancer acidity: An ultimate frontier of tumor immune escape and a novel target of immunomodulation. Seminars in Cancer Biology, 2017. **43**: p. 74-89.

43. Teicher, B.A., W.M. Linehan, and L.J. Helman, *Targeting Cancer Metabolism.* Clinical Cancer Research, 2012. **18**(20): p. 5537-5545.

44. Deberardinis, R.J., et al., *Brick by brick: metabolism and tumor cell growth.* Current opinion in genetics & development, 2008. **18**(1): p. 54-61.

45. Zheng, J., Energy metabolism of cancer: Glycolysis versus oxidative phosphorylation (Review). Oncology letters, 2012. **4**(6): p. 1151-1157.

46. Warburg, O., *The metabolism of carcinoma cells.* The Journal of Cancer Research, 1925. **9**(1): p. 148-163.

47. Warburg, O., F. Wind, and E. Negelein, *Ueber den stoffwechsel von tumoren im körper.* Klinische Wochenschrift, 1926. **5**(19): p. 829-832.

48. Salway, J.G., *Metabolism at a Glance*. 2017: Wiley.

49. Liberti, M.V. and J.W. Locasale, *The Warburg Effect: How Does it Benefit Cancer Cells?* Trends in biochemical sciences, 2016. **41**(3): p. 211-218.

50. Abdel-Haleem, A.M., et al., *The Emerging Facets of Non-Cancerous Warburg Effect.* Frontiers in endocrinology, 2017. **8**: p. 279-279.

51. Ward, P.S. and C.B. Thompson, Metabolic reprogramming: a cancer hallmark even warburg did not anticipate. Cancer cell, 2012. **21**(3): p. 297-308.

52. Estrella, V., et al., *Acidity generated by the tumor microenvironment drives local invasion.* Cancer research, 2013. **73**(5): p. 1524-1535.

53. Ho, P.C., et al., Phosphoenolpyruvate Is a Metabolic Checkpoint of Anti-tumor T Cell Responses. Cell, 2015. **162**(6): p. 1217-28.

54. Ahmad, R., et al., Induction of ROS-mediated cell death and activation of the JNK pathway by a sulfonamide derivative. Int J Mol Med, 2019. **44**(4): p. 1552-1562.

55. Kamarajugadda, S., et al., Manganese superoxide dismutase promotes anoikis resistance and tumor metastasis. Cell Death Dis, 2013. **4**(2): p. e504.

56. Ge, T., et al., The Role of the Pentose Phosphate Pathway in Diabetes and Cancer. Frontiers in Endocrinology, 2020. **11**(365).

57. Li, D., et al., A new G6PD knockdown tumor-cell line with reduced proliferation and increased susceptibility to oxidative stress. Cancer Biother Radiopharm, 2009. **24**(1): p. 81-90.

58. Patra, K.C. and N. Hay, *The pentose phosphate pathway and cancer.* Trends Biochem Sci, 2014. **39**(8): p. 347-54.

59. Kuehne, A., et al., Acute Activation of Oxidative Pentose Phosphate Pathway as First-Line Response to Oxidative Stress in Human Skin Cells. Mol Cell, 2015. **59**(3): p. 359-71.

60. Benito, A., et al., Glucose-6-phosphate dehydrogenase and transketolase modulate breast cancer cell metabolic reprogramming and correlate with poor patient outcome. Oncotarget, 2017. **8**(63): p. 106693-106706.

61. Wang, Y.-Y., et al., *Nrf2-Mediated Metabolic Reprogramming in Cancer.* Oxidative medicine and cellular longevity, 2018. **2018**: p. 9304091-9304091.

62. Jiang, P., et al., p53 regulates biosynthesis through direct inactivation of glucose-6-phosphate dehydrogenase. Nature cell biology, 2011. **13**(3): p. 310-316.

63. Newman, A.C. and O.D.K. Maddocks, *One-carbon metabolism in cancer.* British journal of cancer, 2017. **116**(12): p. 1499-1504.

64. Labuschagne, C.F., et al., Serine, but not glycine, supports one-carbon metabolism and proliferation of cancer cells. Cell Rep, 2014. **7**(4): p. 1248-58.

65. Oppenheim, E.W., et al., Heavy chain ferritin enhances serine hydroxymethyltransferase expression and de novo thymidine biosynthesis. J Biol Chem, 2001. **276**(23): p. 19855-61.

66. Vazquez, A., et al., *Cancer metabolism at a glance.* J Cell Sci, 2016. **129**(18): p. 3367-73.

67. Lewis, D.E. and S.E. Blutt, *2 - Organization of the Immune System*, in *Clinical Immunology (Fifth Edition)*, R.R. Rich, et al., Editors. 2019, Elsevier: London. p. 19-38.e1.

68. Kondo, M., Lymphoid and myeloid lineage commitment in multipotent hematopoietic progenitors. Immunological reviews, 2010. **238**(1): p. 37-46.

69. Agarwal, A., et al., *Differentiation of leukemic blasts is not completely blocked in acute myeloid leukemia.* Proceedings of the National Academy of Sciences, 2019. **116**(49): p. 24593-24599.

70. De Kouchkovsky, I. and M. Abdul-Hay, *'Acute myeloid leukemia: a comprehensive review and 2016 update'.* Blood cancer journal, 2016. **6**(7): p. e441-e441.

71. Drack, M., *Ludwig von Bertalanffy's early system approach.* Systems Research and Behavioral Science, 2009. **26**(5): p. 563-572.

72. Trewavas, A., A brief history of systems biology. "Every object that biology studies is a system of systems." Francois Jacob (1974). The Plant cell, 2006. **18**(10): p. 2420-2430.

73. Kesić, S., *Systems biology, emergence and antireductionism.* Saudi Journal of Biological Sciences, 2016. **23**(5): p. 584-591.

74. Klipp, E., et al., *Systems Biology: A Textbook*. 2016: Wiley.

75. Lightbody, G., et al., Review of applications of high-throughput sequencing in personalized medicine: barriers and facilitators of future progress in research and clinical application. Briefings in Bioinformatics, 2019. **20**(5): p. 1795-1811.

76. Crick, F., *Central dogma of molecular biology.* Nature, 1970. **227**(5258): p. 561-3.

77. McGettigan, P.A., *Transcriptomics in the RNA-seq era.* Current Opinion in Chemical Biology, 2013. **17**(1): p. 4-11.

78. Golub, T.R., et al., Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science, 1999. **286**(5439): p. 531.

79. International Human Genome Sequencing, C., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**: p. 860.

80. Govindarajan, R., et al., *Microarray and its applications.* Journal of Pharmacy & Bioallied Sciences, 2012. **4**(Suppl 2): p. S310-S312.

81. Martis, M.M., et al., RNA-Seq de novo assembly and differential transcriptome analysis of the nematode Ascaridia galli in relation to in vivo exposure to flubendazole. PLOS ONE, 2017. **12**(11): p. e0185182.

82. Zhao, S., et al., Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. PLoS ONE, 2014. **9**(1): p. e78644.

83. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nature reviews. Genetics, 2009. **10**(1): p. 57-63.

84. McLachlan, G.J., *Analyzing microarray gene expression data / Geoffrey J. McLachlan, Kim-Anh Do, Christopher Ambroise*. Wiley series in probability and statistics., ed. K.A. Do and C. Ambroise. 2004, Hoboken, N.J: Wiley-Interscience.

85. Yang, Y.H., et al., Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Research, 2002. **30**(4): p. e15-e15.

86. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.

87. Trapnell, C., L. Pachter, and S.L. Salzberg, *TopHat: discovering splice junctions with RNA-Seq.* Bioinformatics, 2009. **25**(9): p. 1105-1111.

88. Sims, D., et al., Sequencing depth and coverage: key considerations in genomic analyses. Nature Reviews Genetics, 2014. **15**: p. 121.

89. Rapaport, F., et al., Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biology, 2013. **14**(9): p. R95-R95.

90. Oshlack, A. and M.J. Wakefield, *Transcript length bias in RNA-seq data confounds systems biology.* Biology Direct, 2009. **4**: p. 14-14.

91. Dillies, M.-A., et al., A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Briefings in Bioinformatics, 2013. **14**(6): p. 671-683.

92. Anders, S. and W. Huber, *Differential expression analysis for sequence count data.* Genome biology, 2010. **11**(10): p. R106.

93. Love, M.I., W. Huber, and S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 2014. **15**(12): p. 550.

94. Robinson, M.D. and A. Oshlack, A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology, 2010. **11**(3): p. R25-R25.

95. Huang, D.W., B.T. Sherman, and R.A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Protocols, 2008. **4**: p. 44.

96. Huang, D.W., B.T. Sherman, and R.A. Lempicki, Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Research, 2009. **37**(1): p. 1-13.

97. Croft, D., et al., *The Reactome pathway knowledgebase.* Nucleic Acids Research, 2014. **42**(Database issue): p. D472-D477.

98. Fabregat, A., et al., *The Reactome Pathway Knowledgebase.* Nucleic Acids Research, 2018. **46**(Database issue): p. D649-D655.

99. Rafael, A.I., et al., *Gene set enrichment analysis made simple.* Statistical Methods in Medical Research, 2009. **18**(6): p. 565-575.

100. Subramanian, A., et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA, 2005. **102**.

101. Shulaev, V., *Metabolomics technology and bioinformatics.* Briefings in Bioinformatics, 2006. **7**(2): p. 128-139.

102. Roberts, L.D., et al., *Targeted metabolomics.* Current protocols in molecular biology, 2012. **Chapter 30**: p. Unit30.2-30.2.24.

103. Ramirez, T., et al., *Metabolomics in toxicology and preclinical research.* Altex, 2013. **30**(2): p. 209-25.

104. Wang, P., A.I. Shehu, and X. Ma, *The Opportunities of Metabolomics in Drug Safety Evaluation.* Current pharmacology reports, 2017. **3**(1): p. 10-15.

105. Vermeersch, K.A. and M.P. Styczynski, *Applications of metabolomics in cancer research.* Journal of carcinogenesis, 2013. **12**: p. 9-9.

106. Tebani, A. and S. Bekri, *Paving the Way to Precision Nutrition Through Metabolomics.* Frontiers in Nutrition, 2019. **6**(41).

107. Schrimpe-Rutledge, A.C., et al., *Untargeted Metabolomics Strategies-Challenges and Emerging Directions.* Journal of the American Society for Mass Spectrometry, 2016. **27**(12): p. 1897-1905.

108. Madsen, R., T. Lundstedt, and J. Trygg, *Chemometrics in Metabolomics—A Review in Human Disease Diagnosis.* Analytica chimica acta, 2010. **659**: p. 23-33.

109. Tebani, A., C. Afonso, and S. Bekri, Advances in metabolome information retrieval: turning chemistry into biology. Part I: analytical chemistry of the metabolome. Journal of Inherited Metabolic Disease, 2018. **41**(3): p. 379-391.

110. Dudley, E., et al., *Targeted metabolomics and mass spectrometry.* Adv Protein Chem Struct Biol, 2010. **80**: p. 45-83.

111. Jocelyn Paré, J.R. and V. Yaylayan, *Chapter 7 Mass spectrometry: Principles and applications*, in *Techniques and Instrumentation in Analytical Chemistry*, J.R.J. Paré and J.M.R. Bélanger, Editors. 1997, Elsevier. p. 239-266.

112. Karasek, F.W. and R.E. Clement, Basic Gas Chromatography-Mass Spectrometry: Principles and Techniques. 2012: Elsevier Science.

113. Welthagen, W., et al., Comprehensive two-dimensional gas chromatography–time-of-flight mass spectrometry (GC × GC-TOF) for high resolution metabolomics: biomarker discovery on spleen tissue extracts of obese NZO compared to lean C57BL/6 mice. Metabolomics, 2005. **1**(1): p. 65-73.

114. Pitt, J.J., Principles and applications of liquid chromatography-mass spectrometry in clinical biochemistry. The Clinical biochemist. Reviews, 2009. **30**(1): p. 19-34.

115. Günther, H., NMR Spectroscopy: Basic Principles, Concepts and Applications in Chemistry. 2013: Wiley.

116. Emwas, A.-H., et al., *NMR spectroscopy for metabolomics research.* Metabolites, 2019. **9**(7): p. 123.

117. Zhang, A., et al., *Modern analytical techniques in metabolomics analysis.* Analyst, 2012. **137**(2): p. 293-300.

118. Segers, K., et al., *Analytical techniques for metabolomic studies: a review.* Bioanalysis, 2019. **11**(24): p. 2297-2318.

119. Yurkovich, J.T. and B.O. Palsson, *Solving Puzzles With Missing Pieces: The Power of Systems Biology.* Proceedings of the IEEE, 2016. **104**(1): p. 2-7.

120. Angione, C., Human Systems Biology and Metabolic Modelling: A Review—From Disease Metabolism to Precision Medicine. BioMed Research International, 2019. **2019**: p. 8304260.

121. Johnson, C.H., J. Ivanisevic, and G. Siuzdak, *Metabolomics: beyond biomarkers and towards mechanisms.* Nature reviews Molecular cell biology, 2016. **17**(7): p. 451-459.

122. Zampieri, M. and U. Sauer, *Metabolomics-driven understanding of genotype-phenotype relations in model organisms.* Current Opinion in Systems Biology, 2017. **6**: p. 28-36.

123. Handakumbura, P.P., et al., *Metabotyping as a stopover in genome-to-Phenome mapping.* Scientific reports, 2019. **9**(1): p. 1-12.

124. Harrison, B.R., et al., The metabolome as a link in the genotype-phenotype map for peroxide resistance in the fruit fly, Drosophila melanogaster. BMC genomics, 2020. **21**(1): p. 341-341.

125. Zhang, C. and Q. Hua, Applications of Genome-Scale Metabolic Models in Biotechnology and Systems Medicine. Frontiers in Physiology, 2016. **6**(413).

126. Bordbar, A. and B.O. Palsson, Using the reconstructed genome-scale human metabolic network to study physiology and pathology. Journal of Internal Medicine, 2012. **271**(2): p. 131-141.

127. Ma'ayan, A., *Introduction to Network Analysis in Systems Biology.* Science signaling, 2011. **4**(190): p. tr5-tr5.

128. Karakitsou, E., et al., *Metabolomics in systems medicine: an overview of methods and applications.* Current Opinion in Systems Biology, 2019. **15**: p. 91-99.

129. Savageau, M.A., Biochemical systems analysis: I. Some mathematical properties of the rate law for the component enzymatic reactions. Journal of Theoretical Biology, 1969. **25**(3): p. 365-369.

130. Alvarez-Vasquez, F., et al., Simulation and validation of modelled sphingolipid metabolism in Saccharomyces cerevisiae. Nature, 2005. **433**: p. 425.

131. Atkinson, M.R., et al., Development of Genetic Circuitry Exhibiting Toggle Switch or Oscillatory Behavior in Escherichia coli. Cell, 2003. **113**(5): p. 597-607.

132. Visser, D., et al., Optimal re-design of primary metabolism in Escherichia coli using linlog kinetics. Metabolic Engineering, 2004. **6**(4): p. 378-390.

133. Wu, L., et al., A new framework for the estimation of control parameters in metabolic pathways using lin-log kinetics. European Journal of Biochemistry, 2004. **271**(16): p. 3348-3359.

134. Kresnowati, M.T.A.P., W.A. van Winden, and J.J. Heijnen, Determination of elasticities, concentration and flux control coefficients from transient metabolite data using linlog kinetics. Metabolic Engineering, 2005. **7**(2): p. 142-153.

135. Joshi, A. and B.O. Palsson, *Metabolic dynamics in the human red cell: Part I—A comprehensive kinetic model.* Journal of Theoretical Biology, 1989. **141**(4): p. 515-528.

136. Joshi, A. and B.O. Palsson, *Metabolic dynamics in the human red cell. Part III—Metabolic reaction rates.* Journal of Theoretical Biology, 1990. **142**(1): p. 41-68.

137. Visser, D., et al., Tendency Modeling: A New Approach to Obtain Simplified Kinetic Models of Metabolism Applied to Saccharomyces cerevisiae. Metabolic Engineering, 2000. **2**(3): p. 252-275.

138. Teusink, B., et al., Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry. European Journal of Biochemistry, 2003. **267**(17): p. 5313-5329.

139. Chassagnole, C., et al., *Dynamic modeling of the central carbon metabolism of Escherichia coli.* Biotechnology and Bioengineering, 2002. **79**(1): p. 53-73.

140. Fell, D. and A. Cornish-Bowden, *Understanding the control of metabolism*. Vol. 2. 1997: Portland press London.

141. Klipp, E., W. Liebermeister, and C. Wierling, *Inferring dynamic properties of biochemical reaction networks from structural knowledge.* Genome Inform, 2004. **15**(1): p. 125-37.

142. O'Brien, E.J., J.M. Monk, and B.O. Palsson, *Using Genome-scale Models to Predict Biological Capabilities.* Cell, 2015. **161**(5): p. 971-987.

143. Yizhak, K., et al., *Modeling cancer metabolism on a genome scale.* Mol Syst Biol, 2015. **11**(6): p. 817.

144. Monk, J., J. Nogales, and B.O. Palsson, *Optimizing genome-scale network reconstructions.* Nature biotechnology, 2014. **32**(5): p. 447-452.

145. Thiele, I. and B.Ø. Palsson, A protocol for generating a high-quality genome-scale metabolic reconstruction. Nature Protocols, 2010. **5**(1): p. 93-121.

146. Gu, C., et al., Current status and applications of genome-scale metabolic models. Genome Biology, 2019. **20**(1): p. 121.

147. Edwards, J.S. and B.O. Palsson, *Systems properties of the Haemophilus influenzaeRd metabolic genotype.* Journal of Biological Chemistry, 1999. **274**(25): p. 17410-17416.

148. Edwards, J. and B. Palsson, *The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities.* Proceedings of the National Academy of Sciences, 2000. **97**(10): p. 5528-5533.

149. Förster, J., et al., Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. Genome research, 2003. **13**(2): p. 244-253.

150. Durot, M., P.-Y. Bourguignon, and V. Schachter, *Genome-scale models of bacterial metabolism: reconstruction and applications.* FEMS microbiology reviews, 2009. **33**(1): p. 164-190.

151. Yuan, H., et al., A genome-scale metabolic network reconstruction of tomato (Solanum lycopersicum L.) and its application to photorespiratory metabolism. The Plant Journal, 2016. **85**(2): p. 289-304.

152. Price, N.D., J. Schellenberger, and B.O. Palsson, Uniform sampling of steady-state flux spaces: means to design experiments and to interpret enzymopathies. Biophysical journal, 2004. **87**(4): p. 2172-2186.

153. Duarte, N.C., et al., *Global reconstruction of the human metabolic network based on genomic and bibliomic data.* Proceedings of the National Academy of Sciences, 2007. **104**(6): p. 1777-1782.

154. Thiele, I., et al., *A community-driven global reconstruction of human metabolism.* Nature biotechnology, 2013. **31**(5): p. 419-425.

155. Agren, R., et al., Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT. PLoS Comput Biol, 2012. **8**(5): p. e1002518.

156. Swainston, N., et al., *Recon 2.2: from reconstruction to model of human metabolism.* Metabolomics : Official journal of the Metabolomic Society, 2016. **12**: p. 109-109.

157. Brunk, E., et al., Recon3D enables a three-dimensional view of gene variation in human metabolism. Nature biotechnology, 2018. **36**(3): p. 272.

158. Feist, A.M. and B.O. Palsson, *The biomass objective function.* Current Opinion in Microbiology, 2010. **13**(3): p. 344-349.

159. Orth, J.D., I. Thiele, and B.Ø. Palsson, *What is flux balance analysis?* Nature biotechnology, 2010. **28**(3): p. 245-248.

160. Johnson, A.T., *Teaching the principle of biological optimization.* Journal of biological engineering, 2013. **7**(1): p. 6-6.

161. Gudmundsson, S. and I. Thiele, *Computationally efficient flux variability analysis.* BMC Bioinformatics, 2010. **11**(1): p. 489.

162. Mahadevan, R. and C.H. Schilling, The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. Metabolic Engineering, 2003. **5**(4): p. 264-276.

163. Thiele, I., et al., Functional characterization of alternate optimal solutions of Escherichia coli's transcriptional and translational machinery. Biophysical journal, 2010. **98**(10): p. 2072-2081.

164. Bushell, M.E., et al., The use of genome scale metabolic flux variability analysis for process feed formulation based on an investigation of the effects of the zwf mutation on antibiotic production in Streptomyces coelicolor. Enzyme and Microbial Technology, 2006. **39**(6): p. 1347-1353.

165. Pharkya, P. and C.D. Maranas, An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems. Metabolic engineering, 2006. **8**(1): p. 1-13.

166. Feist, A.M., et al., Model-driven evaluation of the production potential for growth-coupled products of Escherichia coli. Metabolic engineering, 2010. **12**(3): p. 173-186.

167. Shlomi, T., et al., *Network-based prediction of human tissue-specific metabolism.* Nature biotechnology, 2008. **26**(9): p. 1003-1010.

168. Becker, S.A. and B.O. Palsson, *Context-Specific Metabolic Networks Are Consistent with Experiments.* PLOS Computational Biology, 2008. **4**(5): p. e1000082.

169. Schmidt, B.J., et al., GIM3E: condition-specific models of cellular metabolism developed from metabolomics and expression data. Bioinformatics, 2013. **29**(22): p. 2900-2908.

170. Fouladiha, H. and S.-A. Marashi, *Biomedical applications of cell- and tissue-specific metabolic network models.* Journal of Biomedical Informatics, 2017. **68**: p. 35-49.

171. McDonald, E.R., III, et al., Project DRIVE: A Compendium of Cancer Dependencies and Synthetic Lethal Relationships Uncovered by Large-Scale, Deep RNAi Screening. Cell, 2017. **170**(3): p. 577-592.e10.

172. Folger, O., et al., Predicting selective drug targets in cancer through metabolic networks. Mol Syst Biol, 2011. **7**: p. 501.

173. Agren, R., et al., Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling. Molecular Systems Biology, 2014. **10**(3): p. 721.

174. Segre, D., D. Vitkup, and G.M. Church, Analysis of optimality in natural and perturbed metabolic networks. PNAS, 2002. **99**.

175. Sandhu, C., A. Qureshi, and A. Emili, *Panomics for precision medicine.* Trends in molecular medicine, 2018. **24**(1): p. 85-101.

176. Larrañaga, P., et al., *Machine learning in bioinformatics.* Brief Bioinform, 2006. **7**(1): p. 86-112.

177. Yang, A., M. Troup, and J.W. Ho, *Scalability and validation of big data bioinformatics software.* Computational and structural biotechnology journal, 2017. **15**: p. 379-386.

178. D'Argenio, V., The High-Throughput Analyses Era: Are We Ready for the Data Struggle? High-throughput, 2018. **7**(1): p. 8.

179. Xu, C. and S.A. Jackson, *Machine learning and complex biological data.* Genome Biology, 2019. **20**(1): p. 76.

180. Libbrecht, M.W. and W.S. Noble, *Machine learning applications in genetics and genomics.* Nature Reviews Genetics, 2015. **16**(6): p. 321-332.

181. Ching, T., et al., *Opportunities and obstacles for deep learning in biology and medicine.* Journal of The Royal Society Interface, 2018. **15**(141): p. 20170387.

182. Zampieri, G., et al., *Machine and deep learning meet genome-scale metabolic modeling.* PLOS Computational Biology, 2019. **15**(7): p. e1007084.

183. He, Y., et al., *A survey on deep learning in DNA/RNA motif mining.* Briefings in Bioinformatics, 2020.

184. Poursheikhali Asghari, M. and P. Abdolmaleki, *Prediction of RNA- and DNA-Binding Proteins Using Various Machine Learning Classifiers.* Avicenna journal of medical biotechnology, 2019. **11**(1): p. 104-111.

185. Bhattacharjee, A. and M.S. Bayzid, Machine learning based imputation techniques for estimating phylogenetic trees from incomplete distance matrices. BMC genomics, 2020. **21**(1): p. 497-497.

186. Lena, P.D., et al., Methylation data imputation performances under different representations and missingness patterns. BMC bioinformatics, 2020. **21**(1): p. 268-268.

187. Zhou, X., et al., Imputing missing RNA-sequencing data from DNA methylation by using a transfer learning–based neural network. GigaScience, 2020. **9**.

188. Shaked, I., et al., *Metabolic network prediction of drug side effects.* Cell systems, 2016. **2**(3): p. 209-213.

189. Li, L., et al., Predicting enzyme targets for cancer drugs by profiling human metabolic reactions in NCI-60 cell lines. BMC bioinformatics, 2010. **11**(1): p. 1-16.

190. Choudhary, R. and H. Gianey, Comprehensive Review On Supervised Machine Learning Algorithms. 2017. 37-43.

191. Wittek, P., *5 - Unsupervised Learning*, in *Quantum Machine Learning*, P. Wittek, Editor. 2014, Academic Press: Boston. p. 57-62.

192. Liu, C., et al., High-dimensional omics data analysis using a variable screening protocol with prior knowledge integration (SKI). BMC systems biology, 2016. **10**(Suppl 4): p. 118-118.

193. Goodfellow, I., Y. Bengio, and A. Courville, *Deep Learning*. 2016: MIT Press.

194. Hotelling, H., Relations between two sets of variates, in Breakthroughs in statistics. 1992, Springer. p. 162-190.

195. Witten, D.M., R. Tibshirani, and T. Hastie, A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics, 2009. **10**(3): p. 515-534.

196. Ho, T.K., Random decision forests, in Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1. 1995, IEEE Computer Society. p. 278.

197. Tin Kam, H., *The random subspace method for constructing decision forests.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998. **20**(8): p. 832-844.

198. Amit, Y. and D. Geman, *Shape quantization and recognition with randomized trees.* Neural Comput., 1997. **9**(7): p. 1545–1588.

199. Breiman, L., *Random Forests.* Machine Learning, 2001. **45**(1): p. 5-32.

200. Konukoglu, E. and B. Glocker, *Chapter 19 - Random forests in medical image computing*, in *Handbook of Medical Image Computing and Computer Assisted Intervention*, S.K. Zhou, D. Rueckert, and G. Fichtinger, Editors. 2020, Academic Press. p. 457-480.

201. Probst, P. and A.-L. Boulesteix, *To tune or not to tune the number of trees in random forest.* The Journal of Machine Learning Research, 2017. **18**(1): p. 6673-6690.

202. Hastie, T., R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. 2009: Springer Science & Business Media.

203. Li, B. and C.N. Dewey, RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics, 2011. **12**(1): p. 323.

204. Ratain, M.J., et al., The cancer and leukemia group B pharmacology and experimental therapeutics committee: a historical perspective. Clin Cancer Res, 2006. **12**(11 Pt 2): p. 3612s-6s.

205. Team, R.D.C., *R: A Language and Environment for Statistical Computing*. 2011: Vienna, Austria : the R Foundation for Statistical Computing.

206. D. Bates, V.C., M. Dettling, S. Dudoit, B. Ellis, L. Gautier, R. Gentleman, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, M. Maechler, A. Rossini, G. Sawitzki, L. Tierney, J.Y.H. Yang, J. Zhang., *Bioconductor*. 2002-.

207. Carvalho, B.S. and R.A. Irizarry, *A framework for oligonucleotide microarray preprocessing.* Bioinformatics, 2010. **26**(19): p. 2363-7.

208. Bolstad, B.M., et al., A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics, 2003. **19**(2): p. 185-93.

209. Irizarry, R.A., et al., *Summaries of Affymetrix GeneChip probe level data.* Nucleic Acids Research, 2003. **31**(4): p. e15-e15.

210. Irizarry, R.A., et al., Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics, 2003. **4**(2): p. 249-264.

211. Ritchie, M.E., et al., limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res, 2015. **43**(7): p. e47.

212. Phipson, B., et al., Robust Hyperparameter Estimation Protects against Hypervariable Genes and Improves Power to Detect Differential Expression. Ann Appl Stat, 2016. **10**(2): p. 946-963.

213. Luo, W., et al., GAGE: generally applicable gene set enrichment for pathway analysis. BMC Bioinformatics, 2009. **10**(1): p. 161.

214. Gene Ontology, C., *Gene Ontology Consortium: going forward.* Nucleic Acids Res, 2015. **43**(Database issue): p. D1049-56.

215. Kanehisa, M. and S. Goto, *KEGG: Kyoto Encyclopedia of Genes and Genomes.* Nucleic Acids Research, 2000. **28**(1): p. 27-30.

216. Kanehisa, M., et al., *KEGG: new perspectives on genomes, pathways, diseases and drugs.* Nucleic Acids Research, 2017. **45**(Database issue): p. D353-D361.

217. Ebrahim, A., et al., *COBRApy: COnstraints-Based Reconstruction and Analysis for Python.* BMC Systems Biology, 2013. **7**(1): p. 74.

218. Swainston, N., et al., *Recon 2.2: from reconstruction to model of human metabolism.* Metabolomics, 2016. **12**(7): p. 109.

219. Schmidt, B.J., et al., GIM3E: condition-specific models of cellular metabolism developed from metabolomics and expression data. Bioinformatics (Oxford, England), 2013. **29**(22): p. 2900-2908.

220. Segrè, D., D. Vitkup, and G.M. Church, *Analysis of optimality in natural and perturbed metabolic networks.* Proceedings of the National Academy of Sciences, 2002. **99**(23): p. 15112-15117.

221. Surhone, L.M., M.T. Timpledon, and S.F. Marseken, Student's T-Test: Student's T-Distribution, Probability Distribution, Normal Distribution, Probability, Statistics, Generalised Hyperbolic Distribution, Guinness Brewery, William Sealy Gosset. 2010: Betascript Publishing.

222. Daniela, M.W. and J.T. Robert, *Extensions of Sparse Canonical Correlation Analysis with Applications to Genomic Data.* Statistical Applications in Genetics and Molecular Biology, 2009. **8**(1).

223. Pedregosa, F., et al., *Scikit-learn: Machine Learning in Python.* J. Mach. Learn. Res., 2011. **12**(null): p. 2825–2830.

224. Patra, K.C. and N. Hay, *The pentose phosphate pathway and cancer.* Trends in Biochemical Sciences, 2014. **39**(8): p. 347-354.

225. Jin, L. and Y. Zhou, *Crucial role of the pentose phosphate pathway in malignant tumors.* Oncology letters, 2019. **17**(5): p. 4213-4221.

226. Zhao, J. and C.-J. Zhong, *A review on research progress of transketolase.* Neuroscience bulletin, 2009. **25**(2): p. 94-99.

227. Agledal, L., M. Niere, and M. Ziegler, *The phosphate makes a difference: cellular functions of NADP.* Redox Report, 2010. **15**(1): p. 2-10.

228. Lu'o'ng, K.v.q. and L.T.H. Nguyễn, *Role of Thiamine in Alzheimer's Disease.* American Journal of Alzheimer's Disease & Other Dementias®, 2011. **26**(8): p. 588-598.

229. Xu, I.M.-J., et al., *Transketolase counteracts oxidative stress to drive cancer development.* Proceedings of the National Academy of Sciences of the United States of America, 2016. **113**(6): p. E725-E734.

230. Ziegler, D., et al., Association of transketolase polymorphisms with measures of polyneuropathy in patients with recently diagnosed diabetes. Diabetes/Metabolism Research and Reviews, 2017. **33**(4): p. e2811.

231. Deshpande, G.P., H.-G. Patterton, and M. Faadiel Essop, *The human transketolase-like proteins TKTL1 and TKTL2 are bona fide transketolases.* BMC Structural Biology, 2019. **19**(1): p. 2.

232. Maslova, A.O., L.E. Meshalkina, and G.A. Kochetov, *Computer modeling of transketolase-like protein, TKTL1, a marker of certain tumor tissues.* Biochemistry (Moscow), 2012. **77**(3): p. 296-299.

233. Yang, X., M.P.M. Wong, and R.K. Ng, *Aberrant DNA Methylation in Acute Myeloid Leukemia and Its Clinical Implications.* International journal of molecular sciences, 2019. **20**(18): p. 4576.

234. Tato, I., et al., Amino acids activate mammalian target of rapamycin complex 2 (mTORC2) via PI3K/Akt signaling. J Biol Chem, 2011. **286**(8): p. 6128-42.

235. Li, Y., et al., Clinical implications of genome-wide DNA methylation studies in acute myeloid leukemia. J Hematol Oncol, 2017. **10**(1): p. 41.

236. Saied, M.H., et al., Genome Wide Analysis of Acute Myeloid Leukemia Reveal Leukemia Specific Methylome and Subtype Specific Hypomethylation of Repeats. PLOS ONE, 2012. **7**(3): p. e33213.

237. Bukowski, K., M. Kciuk, and R. Kontek, *Mechanisms of Multidrug Resistance in Cancer Chemotherapy.* Int J Mol Sci, 2020. **21**(9).

238. Kikuchi, H., et al., Chemopreventive and anticancer activity of flavonoids and its possibility for clinical use by combining with conventional chemotherapeutic agents. American journal of cancer research, 2019. **9**(8): p. 1517-1535.

239. Lichota, A. and K. Gwozdzinski, Anticancer Activity of Natural Compounds from Plant and Marine Environment. Int J Mol Sci, 2018. **19**(11).

240. Spaczyńska, E., et al., Design and synthesis of anticancer 1-hydroxynaphthalene-2-carboxanilides with a p53 independent mechanism of action. Scientific Reports, 2019. **9**(1): p. 6387.

241. Vemuri, S.K., et al., Novel biosynthesized gold nanoparticles as anti-cancer agents against breast cancer: Synthesis, biological evaluation, molecular modelling studies. Materials Science and Engineering: C, 2019. **99**: p. 417-429.

242. *https://www.cancer.org/cancer/acute-myeloid-leukemia/treating/chemotherapy.html*. [cited 2020.

243. Missailidis, S., *Anticancer Therapeutics*. 2008: Wiley.

244. Momparler, R.L., *Optimization of cytarabine (ARA-C) therapy for acute myeloid leukemia.* Experimental hematology & oncology, 2013. **2**: p. 20-20.

245. Arcamone, F., et al., Adriamycin, 14-hydroxydaunomycin, a new antitumor antibiotic from S. peucetius var. caesius. Biotechnol Bioeng, 1969. **11**(6): p. 1101-10.

246. Taymaz-Nikerel, H., et al., Doxorubicin induces an extensive transcriptional and metabolic rewiring in yeast cells. Scientific Reports, 2018. **8**(1): p. 13672.

247. Box, V.G.S., *The intercalation of DNA double helices with doxorubicin and nagalomycin.* Journal of Molecular Graphics and Modelling, 2007. **26**(1): p. 14-19.

248. Silva, E.F., et al., DNA-doxorubicin interaction: New insights and peculiarities. Biopolymers, 2017. **107**(3).

249. Carvalho, C., et al., *Doxorubicin: the good, the bad and the ugly effect.* Curr Med Chem, 2009. **16**(25): p. 3267-85.

250. Thorn, C.F., et al., *Doxorubicin pathways: pharmacodynamics and adverse effects.* Pharmacogenetics and genomics, 2011. **21**(7): p. 440-446.

251. Si, W., et al., The role and mechanisms of action of microRNAs in cancer drug resistance. Clinical Epigenetics, 2019. **11**(1): p. 25.

252. Wang, X., H. Zhang, and X. Chen, *Drug resistance and combating drug resistance in cancer.* Cancer Drug Resistance, 2019. **2**(2): p. 141-160.

253. Zaal, E.A. and C.R. Berkers, *The Influence of Metabolism on Drug Response in Cancer.* Frontiers in oncology, 2018. **8**: p. 500-500.

254. Zhao, Y., E.B. Butler, and M. Tan, *Targeting cellular metabolism to improve cancer therapeutics.* Cell Death & Disease, 2013. **4**(3): p. e532-e532.

255. Rahman, M. and M.R. Hasan, *Cancer Metabolism and Drug Resistance.* Metabolites, 2015. **5**(4): p. 571-600.

256. Ghosh, A. and S. Barman, Application of Euclidean distance measurement and principal component analysis for gene identification. Gene, 2016. **583**(2): p. 112-120.

257. Zhang, J., Y. Gu, and B. Chen, *Mechanisms of drug resistance in acute myeloid leukemia.* OncoTargets and therapy, 2019. **12**: p. 1937-1945.

258. Goldsmith, Z.G. and D.N. Dhanasekaran, *G Protein regulation of MAPK networks.* Oncogene, 2007. **26**(22): p. 3122-3142.

259. Pabst, T., et al., The plasma lipidome in acute myeloid leukemia at diagnosis in relation to clinical disease features. BBA clinical, 2017. **7**: p. 105-114.

260. Farge, T., et al., Chemotherapy-Resistant Human Acute Myeloid Leukemia Cells Are Not Enriched for Leukemic Stem Cells but Require Oxidative Metabolism. Cancer Discovery, 2017. **7**(7): p. 716-735.

261. Yang, F., et al., *Doxorubicin, DNA torsion, and chromatin dynamics.* Biochimica et biophysica acta, 2014. **1845**(1): p. 84-89.

262. Halim, V.A., et al., Doxorubicin-induced DNA Damage Causes Extensive Ubiquitination of Ribosomal Proteins Associated with a Decrease in Protein Translation. Molecular & cellular proteomics : MCP, 2018. **17**(12): p. 2297-2308.

263. Xiang, M., et al., Expression of Th17 and CD4(+) CD25(+) T regulatory cells in peripheral blood of acute leukemia patients and their prognostic significance. Pak J Pharm Sci, 2016. **29**(6 Suppl): p. 2405-2410.

264. Li, P., et al., *Th17 related cytokines in acute myeloid leukemia.* Front Biosci (Landmark Ed), 2012. **17**: p. 2284-94.

265. Tian, T., et al., Aberrant T Helper 17 Cells and Related Cytokines in Bone Marrow Microenvironment of Patients with Acute Myeloid Leukemia. Clinical and Developmental Immunology, 2013. **2013**: p. 915873.

266. Musuraca, G., et al., IL-17/IL-10 double-producing T cells: new link between infections, immunosuppression and acute myeloid leukemia. Journal of Translational Medicine, 2015. **13**(1): p. 229.

267. Kumar, B., et al., Acute myeloid leukemia transforms the bone marrow niche into a leukemia-permissive microenvironment through exosome secretion. Leukemia, 2018. **32**(3): p. 575-587.

268. Kalinkovich, A., et al., Functional CXCR4-expressing microparticles and SDF-1 correlate with circulating acute myelogenous leukemia cells. Cancer Res, 2006. **66**(22): p. 11013-20.

269. Dvořáková, M., et al., DNA released by leukemic cells contributes to the disruption of the bone marrow microenvironment. Oncogene, 2013. **32**(44): p. 5201-9.

270. Kupsa, T., J.M. Horacek, and L. Jebavy, The role of adhesion molecules in acute myeloid leukemia and (hemato)oncology: A systematic review. Biomedical papers, 2015. **159**(1): p. 001-011.

271. Windisch, R., et al., Oncogenic Deregulation of Cell Adhesion Molecules in Leukemia. Cancers, 2019. **11**(3): p. 311.

272. Gruszka, A.M., et al., *Adhesion Deregulation in Acute Myeloid Leukaemia.* Cells, 2019. **8**(1): p. 66.

273. Matsunaga, T., et al., Interaction between leukemic-cell VLA-4 and stromal fibronectin is a decisive factor for minimal residual disease of acute myelogenous leukemia. Nature Medicine, 2003. **9**(9): p. 1158-1165.

274. Li, S., et al., HA117 endows HL60 cells with a stem-like signature by inhibiting the degradation of DNMT1 via its ability to down-regulate expression of the GGL domain of RGS6. PloS one, 2017. **12**(6): p. e0180142-e0180142.

275. Peitzsch, C., et al., *Cancer stem cells: The root of tumor recurrence and metastases.* Seminars in Cancer Biology, 2017. **44**: p. 10-24.

276. Muller, W.A., *Getting leukocytes to the site of inflammation.* Veterinary pathology, 2013. **50**(1): p. 7-22.

277. Strell, C. and F. Entschladen, *Extravasation of leukocytes in comparison to tumor cells.* Cell Communication and Signaling, 2008. **6**(1): p. 10.

278. Zhou, W., et al., *Targeting CXCL12/CXCR4 Axis in Tumor Immunotherapy.* Current medicinal chemistry, 2019. **26**(17): p. 3026-3041.

279. Tulotta, C., et al., CXCR4 signaling regulates metastatic onset by controlling neutrophil motility and response to malignant cells. Scientific Reports, 2019. **9**(1): p. 2399.

280. Mannelli, F., et al., *CXCR4 expression accounts for clinical phenotype and outcome in acute myeloid leukemia.* Cytometry Part B: Clinical Cytometry, 2014. **86**(5): p. 340-349.

281. Huang, X., et al., Evidence that high-migration drug-surviving MOLT4 leukemia cells exhibit cancer stem cell-like properties. International Journal of Oncology, 2016. **49**.

282. Trendowski, M., *The inherent metastasis of leukaemia and its exploitation by sonodynamic therapy.* Critical Reviews in Oncology/Hematology, 2015. **94**(2): p. 149-163.

283. Yamaguchi, H. and J. Condeelis, *Regulation of the actin cytoskeleton in cancer cell migration and invasion.* Biochimica et Biophysica Acta (BBA) - Molecular Cell Research, 2007. **1773**(5): p. 642-652.

284. Fagerberg, L., et al., Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. Mol Cell Proteomics, 2014. **13**(2): p. 397-406.

285. Chen, W.-L., et al., A distinct glucose metabolism signature of acute myeloid leukemia with prognostic value. Blood, 2014. **124**(10): p. 1645-1654.

286. Kreitz, J., et al., *Metabolic Plasticity of Acute Myeloid Leukemia.* Cells, 2019. **8**(8): p. 805.

287. Sanderson, S.M., et al., Methionine metabolism in health and cancer: a nexus of diet and precision medicine. Nature Reviews Cancer, 2019. **19**(11): p. 625-637.

288. Chu, E. and A. Gollerkeri, *Resistance to Inhibitor Compounds of Thymidylate Synthase*, in *Encyclopedia of Cancer (Second Edition)*, J.R. Bertino, Editor. 2002, Academic Press: New York. p. 107-113.

289. Gangjee, A., et al., Dual inhibitors of thymidylate synthase and dihydrofolate reductase as antitumor agents: design, synthesis, and biological evaluation

of classical and nonclassical pyrrolo[2,3-d]pyrimidine antifolates(1). Journal of medicinal chemistry, 2006. **49**(3): p. 1055-1065.

290. Gonen, N. and Y.G. Assaraf, Antifolates in cancer therapy: Structure, activity and mechanisms of drug resistance. Drug Resistance Updates, 2012. **15**(4): p. 183-210.

291. El-Kouhen, K. and Michel L. Tremblay, *PTPMT1: Connecting Cardiolipin Biosynthesis to Mitochondrial Function.* Cell Metabolism, 2011. **13**(6): p. 615-617.

292. Zhang, J., et al., Mitochondrial phosphatase PTPMT1 is essential for cardiolipin biosynthesis. Cell metabolism, 2011. **13**(6): p. 690-700.

293. Niemi, N.M., et al., Downregulation of the mitochondrial phosphatase PTPMT1 is sufficient to promote cancer cell death. PloS one, 2013. **8**(1): p. e53803-e53803.

294. Lee, D. and C.C.-L. Wong, *The folate cycle is a new metabolic weakness of cancer.* Molecular & Cellular Oncology, 2017. **4**(6): p. e1327004.

295. Sica, V., et al., Oxidative phosphorylation as a potential therapeutic target for cancer therapy. Int J Cancer, 2020. **146**(1): p. 10-17.

296. Pourrajab, F., et al., *Genetic Characterization and Risk Stratification of Acute Myeloid Leukemia.* Cancer management and research, 2020. **12**: p. 2231-2253.

297. Charrot, S., et al., *AML through the prism of molecular genetics.* Br J Haematol, 2020. **188**(1): p. 49-62.

298. Zaidi, S.Z., et al., *The challenge of risk stratification in acute myeloid leukemia with normal karyotype.* Hematology/Oncology and Stem Cell Therapy, 2008. **1**(3): p. 141-158.

299. Papaemmanuil, E., et al., *Genomic Classification and Prognosis in Acute Myeloid Leukemia.* New England Journal of Medicine, 2016. **374**(23): p. 2209-2221.

300. Huang da, W., B.T. Sherman, and R.A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc, 2009. **4**(1): p. 44-57.

301. Huang da, W., B.T. Sherman, and R.A. Lempicki, Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res, 2009. **37**(1): p. 1-13.

302. Magrane, M. and U. Consortium, *UniProt Knowledgebase: a hub of integrated protein data.* Database, 2011. **2011**.

303. Hosack, D.A., et al., *Identifying biological themes within lists of genes with EASE.* Genome biology, 2003. **4**(10): p. R70-R70.

304. Kornblau, S.M., et al., Recurrent expression signatures of cytokines and chemokines are present and are independently prognostic in acute myelogenous leukemia and myelodysplasia. Blood, 2010. **116**(20): p. 4251-4261.

305. Carter, B.Z., et al., *Focal Adhesion Kinase as a Potential Target in AML and MDS.* Molecular cancer therapeutics, 2017. **16**(6): p. 1133-1144.

306. Pallarès, V., et al., Focal Adhesion Genes Refine the Intermediate-Risk Cytogenetic Classification of Acute Myeloid Leukemia. Cancers, 2018. **10**(11): p. 436.

307. Tabe, Y., et al., Bone Marrow Adipocytes Facilitate Fatty Acid Oxidation Activating AMPK and a Transcriptional Network Supporting Survival of Acute Monocytic Leukemia Cells. Cancer Res, 2017. **77**(6): p. 1453-1464.

308. Maher, M., et al., Epigenetic-Transcriptional Regulation of Fatty Acid Metabolism and Its Alterations in Leukaemia. Frontiers in Genetics, 2018. **9**(405).

309. Kuo, C.Y. and D.K. Ann, When fats commit crimes: fatty acid metabolism, cancer stemness and therapeutic resistance. Cancer Commun (Lond), 2018. **38**(1): p. 47.

310. Thomas, P.D., et al., PANTHER: a library of protein families and subfamilies indexed by function. Genome Res, 2003. **13**(9): p. 2129-41.

311. Eklund, E., The role of Hox proteins in leukemogenesis: insights into key regulatory events in hematopoiesis. Critical reviews in oncogenesis, 2011. **16**(1-2): p. 65-76.

312. Alharbi, R.A., et al., *The role of HOX genes in normal hematopoiesis and acute leukemia.* Leukemia, 2013. **27**(5): p. 1000-1008.

313. Musialik, E., et al., Promoter DNA methylation and expression levels of HOXA4, HOXA5 and MEIS1 in acute myeloid leukemia. Mol Med Rep, 2015. **11**(5): p. 3948-3954.

314. Zhao, P., et al., *Aberrant Expression of HOXA5 and HOXA9 in AML.* Asian Pac J Cancer Prev, 2015. **16**(9): p. 3941-4.

315. McGonigle, G.J., et al., *HOXA6: A Novel Candidate Gene in AML.* Blood, 2006. **108**(11): p. 2312-2312.

316. Mullighan, C.G., et al., Pediatric acute myeloid leukemia with NPM1 mutations is characterized by a gene expression profile with dysregulated HOX gene expression distinct from MLL-rearranged leukemias. Leukemia, 2007. **21**(9): p. 2000-9.

317. Faber, J., et al., HOXA9 is required for survival in human MLL-rearranged acute leukemias. Blood, 2009. **113**(11): p. 2375-2385.

318. Li, D.-P., et al., *HOXA9 Gene Expression in Acute Myeloid Leukemia.* Cell biochemistry and biophysics, 2013. **67**.

319. Xu, H. and Q. Wen, Downregulation of miR-135a predicts poor prognosis in acute myeloid leukemia and regulates leukemia progression via modulating HOXA10 expression. Mol Med Rep, 2018. **18**(1): p. 1134-1140.

320. Chen, S.-L., et al., The Role of the HOXA Gene Family in Acute Myeloid Leukemia. Genes, 2019. **10**(8): p. 621.

321. Dzeja, P. and A. Terzic, Adenylate kinase and AMP signaling networks: metabolic monitoring, signal communication and body energy sensing. International journal of molecular sciences, 2009. **10**(4): p. 1729-1772.

322. Rogne, P., et al., *Molecular mechanism of ATP versus GTP selectivity of adenylate kinase.* Proceedings of the National Academy of Sciences, 2018. **115**(12): p. 3012-3017.

323. Neumann, D., U. Schlattner, and T. Wallimann, *A molecular approach to the concerted action of kinases involved in energy homoeostasis.* Biochem Soc Trans, 2003. **31**(Pt 1): p. 169-74.

324. Carrasco, A.J., et al., *Adenylate kinase phosphotransfer communicates cellular energetic signals to ATP-sensitive potassium channels.* Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**(13): p. 7623-7628.

325. Panayiotou, C., N. Solaroli, and A. Karlsson, *The many isoforms of human adenylate kinases.* The International Journal of Biochemistry & Cell Biology, 2014. **49**: p. 75-83.

326. Jan, Y.-H., et al., Adenylate Kinase-4 Is a Marker of Poor Clinical Outcomes That Promotes Metastasis of Lung Cancer by Downregulating the Transcription Factor ATF3. Cancer Research, 2012. **72**(19): p. 5119-5129.

327. Speers, C., et al., Identification of Novel Kinase Targets for the Treatment of Estrogen Receptor–Negative Breast Cancer. Clinical Cancer Research, 2009. **15**(20): p. 6327-6340.

328. Ji, Y., et al., Adenylate kinase hCINAP determines self-renewal of colorectal cancer stem cells by facilitating LDHA phosphorylation. Nature Communications, 2017. **8**(1): p. 15308.

329. Bai, D., et al., The ATPase hCINAP regulates 18S rRNA processing and is essential for embryogenesis and tumour growth. Nature Communications, 2016. **7**(1): p. 12310.

330. Klepinin, A., et al., *Adenylate Kinase and Metabolic Signaling in Cancer Cells.* Frontiers in oncology, 2020. **10**: p. 660-660.

331. Liu, R., et al., *Enzymatically inactive adenylate kinase 4 interacts with mitochondrial ADP/ATP translocase.* The international journal of biochemistry & cell biology, 2009. **41**(6): p. 1371-1380.

332. Jan, Y.-H., et al., Adenylate kinase 4 modulates oxidative stress and stabilizes HIF-1α to drive lung adenocarcinoma metastasis. Journal of Hematology & Oncology, 2019. **12**(1): p. 12.

333. Qin, T., et al., High expression of AK1 predicts inferior prognosis in acute myeloid leukemia patients undergoing chemotherapy. Bioscience reports, 2020. **40**(6): p. BSR20200097.

334. Frejno, M., et al., Proteome activity landscapes of tumor cell lines determine drug responses. Nature Communications, 2020. **11**(1): p. 3639.

335. Zhang, L., Heme Biology: The Secret Life Of Heme In Regulating Diverse Biological Processes. 2011: World Scientific Publishing Company.

336. Hooda, J., A. Shah, and L. Zhang, Heme, an essential nutrient from dietary proteins, critically impacts diverse physiological and pathological processes. Nutrients, 2014. **6**(3): p. 1080-1102.

337. Wang, X., J.R. Chowdhury, and N.R. Chowdhury, *Bilirubin metabolism: Applied physiology.* Current Paediatrics, 2006. **16**(1): p. 70-74.

338. Zhou, J., T.S. Tracy, and R.P. Remmel, *Bilirubin glucuronidation revisited: proper assay conditions to estimate enzyme kinetics with recombinant UGT1A1.* Drug metabolism and disposition: the biological fate of chemicals, 2010. **38**(11): p. 1907-1911.

339. Alexandra Brito, M., R.F.M. Silva, and D. Brites, *Bilirubin toxicity to human erythrocytes: A review.* Clinica Chimica Acta, 2006. **374**(1): p. 46-56.

340. Sticova, E. and M. Jirsa, *New insights in bilirubin metabolism and their clinical implications.* World journal of gastroenterology, 2013. **19**(38): p. 6398-6407.

341. Sun, K., et al., Acute myeloid leukaemia presenting as acute liver failure-a case report and literature review. Ecancermedicalscience, 2019. **13**: p. 960-960.

342. DiNardo, C.D. and A.H. Wei, *How I treat acute myeloid leukemia in the era of new drugs.* Blood, 2020. **135**(2): p. 85-96.

343. Maxfield, F.R. and G. van Meer, *Cholesterol, the central lipid of mammalian cells.* Current opinion in cell biology, 2010. **22**(4): p. 422-429.

344. Cortes, V.A., et al., *Physiological and pathological implications of cholesterol.* Front Biosci (Landmark Ed), 2014. **19**: p. 416-28.

345. Menys, V.C. and P.N. Durrington, *Squalene synthase inhibitors.* Br J Pharmacol, 2003. **139**(5): p. 881-2.

346. Boland, M.R. and N.P. Tatonetti, Investigation of 7-dehydrocholesterol reductase pathway to elucidate off-target prenatal effects of pharmaceuticals: a systematic review. Pharmacogenomics J, 2016. **16**(5): p. 411-29.

347. Huang, B., B.-l. Song, and C. Xu, *Cholesterol metabolism in cancer: mechanisms and therapeutic opportunities.* Nature Metabolism, 2020. **2**(2): p. 132-141.

348. Yang, J., L. Wang, and R. Jia, *Role of de novo cholesterol synthesis enzymes in cancer.* Journal of Cancer, 2020. **11**(7): p. 1761-1767.

349. Burke, L.P. and C.A. Kukoly, Statins induce lethal effects in acute myeloblastic leukemia [corrected] cells within 72 hours. Leuk Lymphoma, 2008. **49**(2): p. 322-30.

350. Advani, A.S., et al., Report of the relapsed/refractory cohort of SWOG S0919: A phase 2 study of idarubicin and cytarabine in combination with pravastatin for acute myelogenous leukemia (AML). Leuk Res, 2018. **67**: p. 17-20.

351. Thompson, P.D., P. Clarkson, and R.H. Karas, *Statin-associated myopathy.* Jama, 2003. **289**(13): p. 1681-90.

352. Levy, H.B. and H.K. Kohlhaas, *Considerations for supplementing with coenzyme Q10 during statin therapy.* Ann Pharmacother, 2006. **40**(2): p. 290-4.

353. Do, R., et al., Squalene synthase: a critical enzyme in the cholesterol biosynthesis pathway. Clin Genet, 2009. **75**(1): p. 19-29.

354. Kourounakis, A.P. and E. Bavavea, New applications of squalene synthase inhibitors: Membrane cholesterol as a therapeutic target. Archiv der Pharmazie, 2020. **353**(9): p. 2000085.

355. Nimer, S.D., *Is it important to decipher the heterogeneity of "normal karyotype AML"?* Best practice & research. Clinical haematology, 2008. **21**(1): p. 43-52.

356. Patel, C., et al., Multidrug resistance in relapsed acute myeloid leukemia: evidence of biological heterogeneity. Cancer, 2013. **119**(16): p. 3076-83.

357. Horibata, S., et al., *Heterogeneity in refractory acute myeloid leukemia.* Proceedings of the National Academy of Sciences, 2019. **116**(21): p. 10494-10503.

358. Xu, X., et al., Transketolase-like protein 1 (TKTL1) is required for rapid cell growth and full viability of human tumor cells. International Journal of Cancer, 2009. **124**(6): p. 1330-1337.

359. Schwaab, J., et al., Expression of Transketolase like gene 1 (TKTL1) predicts disease-free survival in patients with locally advanced rectal cancer receiving neoadjuvant chemoradiotherapy. BMC cancer, 2011. **11**: p. 363-363.

360. Dong, Y. and M. Wang, Knockdown of TKTL1 additively complements cisplatin-induced cytotoxicity in nasopharyngeal carcinoma cells by regulating the levels

of NADPH and ribose-5-phosphate. Biomedicine & Pharmacotherapy, 2017. **85**: p. 672-678.

361. Jiang, P., W. Du, and M. Wu, *Regulation of the pentose phosphate pathway in cancer.* Protein Cell, 2014. **5**(8): p. 592-602.

362. Hu, L.-H., et al., The TKTL1 gene influences total transketolase activity and cell proliferation in human colon cancer LoVo cells. Anti-Cancer Drugs, 2007. **18**(4): p. 427-433.

363. Diaz-Moralli, S., et al., *A key role for transketolase-like 1 in tumor metabolic reprogramming.* Oncotarget, 2016. **7**(32): p. 51875-51897.

364. Olszewska, A. and A. Szewczyk, *Mitochondria as a pharmacological target: Magnum overview.* IUBMB Life, 2013. **65**(3): p. 273-281.

365. Ashton, T.M., et al., *Oxidative Phosphorylation as an Emerging Target in Cancer Therapy.* Clinical Cancer Research, 2018. **24**(11): p. 2482-2490.

366. Souers, A.J., et al., ABT-199, a potent and selective BCL-2 inhibitor, achieves antitumor activity while sparing platelets. Nature Medicine, 2013. **19**(2): p. 202-208.

367. Yip, K.W. and J.C. Reed, *Bcl-2 family proteins and cancer.* Oncogene, 2008. **27**(50): p. 6398-6406.

368. Frenzel, A., et al., *Bcl2 family proteins in carcinogenesis and the treatment of cancer.* Apoptosis : an international journal on programmed cell death, 2009. **14**(4): p. 584-596.

369. Chipuk, J.E., L. Bouchier-Hayes, and D.R. Green, *Mitochondrial outer membrane permeabilization during apoptosis: the innocent bystander scenario.* Cell Death & Differentiation, 2006. **13**(8): p. 1396-1402.

370. Certo, M., et al., Mitochondria primed by death signals determine cellular addiction to antiapoptotic BCL-2 family members. Cancer cell, 2006. **9**(5): p. 351-365.

371. Reed, J.C., *Bcl-2 on the brink of breakthroughs in cancer treatment.* Cell Death & Differentiation, 2018. **25**(1): p. 3-6.

372. DiNardo, C.D., et al., Clinical experience with the BCL 2-inhibitor venetoclax in combination therapy for relapsed and refractory acute myeloid leukemia and related myeloid malignancies. American journal of hematology, 2018. **93**(3): p. 401-407.

373. Roberts, A.W., et al., *Targeting BCL2 with venetoclax in relapsed chronic lymphocytic leukemia.* New England Journal of Medicine, 2016. **374**(4): p. 311-322.

374. Mei, M., et al., Hypomethylating agents in combination with venetoclax for acute myeloid leukemia: Update on clinical trial data and practical considerations for use. American Journal of Hematology, 2019. **94**(3): p. 358-362.

375. Roca-Portoles, A., et al., *Venetoclax causes metabolic reprogramming independent of BCL-2 inhibition.* Cell Death & Disease, 2020. **11**(8): p. 616.

376. Pastò, A., et al., Cancer stem cells from epithelial ovarian cancer patients privilege oxidative phosphorylation, and resist glucose deprivation. Oncotarget, 2014. **5**(12): p. 4305-4319.

377. Lanning, N.J., et al., Metabolic profiling of triple-negative breast cancer cells reveals metabolic vulnerabilities. Cancer & metabolism, 2017. **5**: p. 6-6.

378. Romero, N., et al., Abstract 3487: Bioenergetic profiling of cancer cell lines: Quantifying the impact of glycolysis on cell proliferation. Cancer Research, 2018. **78**(13 Supplement): p. 3487-3487.

379. Guha, M., et al., Aggressive triple negative breast cancers have unique molecular signature on the basis of mitochondrial genetic and functional defects. Biochimica et biophysica acta. Molecular basis of disease, 2018. **1864**(4 Pt A): p. 1060-1071.

380. Christian, S., et al., The novel dihydroorotate dehydrogenase (DHODH) inhibitor BAY 2402234 triggers differentiation and is effective in the treatment of myeloid malignancies. Leukemia, 2019. **33**(10): p. 2403-2415.

381. Beuneu, C., et al., Indirect inhibition of mitochondrial dihydroorotate dehydrogenase activity by nitric oxide. Free Radical Biology and Medicine, 2000. **28**(8): p. 1206-1213.

382. Sykes, D.B., et al., Inhibition of Dihydroorotate Dehydrogenase Overcomes Differentiation Blockade in Acute Myeloid Leukemia. Cell, 2016. **167**(1): p. 171-186.e15.

383. Ladds, M.J.G.W., et al., A DHODH inhibitor increases p53 synthesis and enhances tumor cell killing by p53 degradation blockage. Nature Communications, 2018. **9**(1): p. 1107.

384. Dang, W., et al., Pharmacological inhibition of dihydroorotate dehydrogenase induces apoptosis and differentiation in acute myeloid leukemia cells. Haematologica, 2018. **103**(9): p. 1472-1483.

385. Cao, L., et al., Targeting of Hematologic Malignancies with PTC299, A Novel Potent Inhibitor of Dihydroorotate Dehydrogenase with Favorable Pharmaceutical Properties. Molecular Cancer Therapeutics, 2019. **18**(1): p. 3-16.

386. Cerqueira, N.M.F.S.A., et al., *Cholesterol Biosynthesis: A Mechanistic Overview.* Biochemistry, 2016. **55**(39): p. 5483-5506.

387. Karlic, H. and F. Varga, *Mevalonate Pathway*, in *Encyclopedia of Cancer (Third Edition)*, P. Boffetta and P. Hainaut, Editors. 2019, Academic Press: Oxford. p. 445-457.

388. Stancu, C. and A. Sima, *Statins: mechanism of action and effects.* J Cell Mol Med, 2001. **5**(4): p. 378-87.

389. Kornblau, S.M., et al., Blockade of adaptive defensive changes in cholesterol uptake and synthesis in AML by the addition of pravastatin to idarubicin + high-dose Ara-C: a phase 1 study. Blood, 2006. **109**(7): p. 2999-3006.

390. Advani, A.S., et al., Report of the relapsed/refractory cohort of SWOG S0919: A phase 2 study of idarubicin and cytarabine in combination with pravastatin for acute myelogenous leukemia (AML). Leukemia Research, 2018. **67**: p. 17-20.

391. Tsai, H.-J., et al., A Phase II Study of Arginine Deiminase (ADI-PEG20) in Relapsed/Refractory or Poor-Risk Acute Myeloid Leukemia Patients. Scientific Reports, 2017. **7**(1): p. 11253.

392. Stirewalt, D.L., et al., Mevastatin can increase toxicity in primary AMLs exposed to standard therapeutic agents, but statin efficacy is not simply associated with ras hotspot mutations or overexpression. Leukemia Research, 2003. **27**(2): p. 133-145.

393. Lee, J.S., et al., *Statins enhance efficacy of venetoclax in blood cancers.* Science translational medicine, 2018. **10**(445): p. eaaq1240.

394. Shadman, M., et al., Idarubicin, cytarabine, and pravastatin as induction therapy for untreated acute myeloid leukemia and high-risk myelodysplastic syndrome. Am J Hematol, 2015. **90**(6): p. 483-6.

395. Riess, C., et al., Arginine-Depleting Enzymes – An Increasingly Recognized Treatment Strategy for Therapy-Refractory Malignancies. Cellular Physiology and Biochemistry, 2018. **51**(2): p. 854-870.

396. De Santo, C., et al., The arginine metabolome in acute lymphoblastic leukemia can be targeted by the pegylated-recombinant arginase I BCT-100. International Journal of Cancer, 2018. **142**(7): p. 1490-1502.

397. Plunkett, W., *Arginine addiction in AML.* Blood, 2015. **125**(26): p. 3971-3972.

398. Miraki-Moud, F., et al., Arginine deprivation using pegylated arginine deiminase has activity against primary acute myeloid leukemia cells in vivo. Blood, 2015. **125**(26): p. 4060-8.

399. Stuani, L., M. Sabatier, and J.-E. Sarry, Exploiting metabolic vulnerabilities for personalized therapy in acute myeloid leukemia. BMC Biology, 2019. **17**(1): p. 57.

400. Song, K., et al., Resistance to chemotherapy is associated with altered glucose metabolism in acute myeloid leukemia. Oncol Lett, 2016. **12**(1): p. 334-342.

401. Stäubert, C., et al., Rewired Metabolism in Drug-resistant Leukemia Cells a metabolic switch hallmarked by reduced dependence on exogenous glutamine. The Journal of biological chemistry, 2015. **290**.

402. Wandroo, F.A., et al., *Acute myeloid leukaemia presenting as cholestatic hepatitis.* Journal of clinical pathology, 2004. **57**(5): p. 544-545.

403. Murakami, J. and Y. Shimizu, *Hepatic Manifestations in Hematological Disorders.* International Journal of Hepatology, 2013. **2013**: p. 484903.

404. Eisen, A., et al., *Jaundice and Acute Liver Failure as the First Manifestation of Acute Myeloid Leukemia.* The Israel Medical Association journal : IMAJ, 2008. **10**: p. 733-5.

405. Bruguera, M. and R. Miquel, *The Effect of Haematological and Lymphatic Diseases on the Liver.* Textbook of Hepatology, 2007: p. 1662-1670.

APPENDIX I

Essential genes identified by essentiality analysis of the THP-1 and HL60 parental, AraC- and DOX-resistant Genome Scale Metabolic Models reconstructions.

THP-1 Parental

Gene_ symbol	reaction_ id	reaction_name	structural_ KO
ADSL	ADSL1, ADSL2	adenylosuccinate lyase, adenylosuccinate lyase	TRUE
ASNS	ASNS1	asparagine synthase (glutamine-hydrolysing)	TRUE
ATIC	IMPC, AICART	IMP cyclohydrolase, phosphoribosylaminoimidazolecarboxamide formyltransferase	TRUE
CAD	ASPCTr, CBPS, DHORTS	aspartate carbamoyltransferase (reversible), carbamoyl-phosphate synthase (glutamine-hydrolysing), dihydroorotase	TRUE
DHFR	r0512, r0224, FOLR2, r0226, DHFR, r0514	Dihydrofolate:NAD+ oxidoreductase Folate biosynthesis EC:1.5.1.3, 5,6,7,8-Tetrahydrofolate:NAD+ oxidoreductase One carbon pool by folate EC:1.5.1.3, folate reductase, 5,6,7,8-Tetrahydrofolate:NADP+ oxidoreductase One carbon pool by folate / Folate biosynthesis EC:1.5.1.3, dihydrofolate reductase, Dihydrofolate:NADP+ oxidoreductase One carbon pool by folate / Folate biosynthesis EC:1.5.1.3	TRUE
DHODH	DHORD9	dihydoorotic acid dehydrogenase (quinone10)	TRUE
KDSR	3DSPHR	3-Dehydrosphinganine reductase	TRUE
GART	PRAGSr, r0666, GARFT	phosphoribosylglycinamide synthase, 2-(Formamido)-N1-(5-phosphoribosyl)acetamidine cyclo-ligase (ADP-forming) Purine metabolism EC:6.3.3.1, phosphoribosylglycinamide formyltransferase	TRUE

GPI	PGI	glucose-6-phosphate isomerase	FALSE
LCAT	LCAT1e	Lecithin-cholesterol acyltransferase	TRUE
LIPA	r1172, r1179, CHOLESTle, r1177	EC:3.1.1.13, EC:3.1.1.13, hydrolysis of cholesterol ester by cholesterol esterase , EC:3.1.1.13	TRUE
PNP	PNP, PUNP2, NP1, PUNP3, PUNP7, PUNP1, PUNP4, PUNP5, DURIPP, PUNP6	purine-nucleoside phosphorylase, purine-nucleoside phosphorylase (Deoxyadenosine), nucleotide phosphatase, purine-nucleoside phosphorylase (Guanosine), purine-nucleoside phosphorylase (Xanthosine), purine-nucleoside phosphorylase (Adenosine), purine-nucleoside phosphorylase (Deoxyguanosine), purine-nucleoside phosphorylase (Inosine), deoxyuridine phosphorylase, purine-nucleoside phosphorylase (Deoxyinosine)	TRUE
PFAS	PRFGS	phosphoribosylformylglycinamidine synthase	TRUE
PGD	GNDer, GND, GNDc	phosphogluconate dehydrogenase, endoplasmic reticulum, phosphogluconate dehydrogenase, phosphogluconate dehydrogenase	FALSE
PPAT	GLUPRT	glutamine phosphoribosyldiphosphate amidotransferase	TRUE
TXNRD1	TRDR3, r1433, r0027, TRDR, TRDR2	Thioredoxin (ubiquinone 10) reductase (NADH), NADPH:oxidized-thioredoxin oxidoreductase Pyrimidine metabolism EC:1.8.1.9, NADPH:CoA-glutathione oxidoreductase EC:1.8.1.9, thioredoxin reductase (NADPH), Thioredoxin (ubiquinone 10) reductase (NADPH)	TRUE
TYMS	TMDS	thymidylate synthase	TRUE
UMPS	OMPDC, ORPT	"orotidine-5-phosphate decarboxylase", orotate phosphoribosyltransferase	TRUE
PGS1	PGPPT	phosphatidyl-CMP: glycerophosphate phosphatidyltransferase	TRUE
SPTLC2	SERPT	serine C-palmitoyltransferase	TRUE
SPTLC1	SERPT	serine C-palmitoyltransferase	TRUE
PAICS	AIRCr, PRASCS	phosphoribosylaminoimidazole carboxylase, phosphoribosylaminoimidazolesuccinocarboxamide synthase	TRUE

[218]

RPIA	RPI, r0249	ribose-5-phosphate isomerase, D-Ribose-5-phosphate ketol-isomerase Pentose phosphate pathway EC:5.3.1.6	TRUE
CMPK1	CYTK10, CYTK1n, UMPK7, CYTK13n, UMPK6, CYTK14n, UMPK3, CYTK2n, CYTK5n, CYTK13, CYTK12, CYTK9n, UMPK6n, CYTK1, CYTK8n, UMPK2n, UMPK7n, UMPK3n, CYTK2, CYTK7n, CYTK6, CYTK11n, CYTK6n, UMPK5n, UMPK, CYTK3n, CYTK8, CYTK10n, CYTK5, CYTK14, CYTK11, CYTK4n, UMPK5, UMPK4, CYTK7, UMPK4n, UMPK2, CYTK12n, UMPKn, CYTK9	cytidylate kinase (CMP,dGTP), cytidylate kinase (CMP),nuclear, UMP kinase (dGTP), cytidylate kinase (dCMP,dATP),nuclear, UMP kinase (dCTP), cytidylate kinase (dCMP,CTP),nuclear, UMP kinase (UTP), cytidylate kinase (dCMP),nuclear, cytidylate kinase (CMP),nuclear, cytidylate kinase (dCMP,dATP), cytidylate kinase (dCMP,dCTP), cytidylate kinase (CMP,dCTP),nuclear, UMP kinase (dCTP),nuclear, cytidylate kinase (CMP), cytidylate kinase (CMP,dATP),nuclear, UMP kinase (CTP),nuclear, UMP kinase (dGTP),nuclear, UMP kinase (UTP),nuclear, cytidylate kinase (dCMP), cytidylate kinase (CMP,UTP),nuclear, cytidylate kinase (CMP,CTP), cytidylate kinase (dCMP,dGTP),nuclear, cytidylate kinase (CMP,CTP),nuclear, UMP kinase (dATP),nuclear, UMP kinase, cytidylate kinase (dCMP,CTP),nuclear, cytidylate kinase (CMP,dATP), cytidylate kinase (CMP,dGTP),nuclear, cytidylate kinase (dCMP), cytidylate kinase (dCMP,UTP), cytidylate kinase (dCMP,dGTP), cytidylate kinase (dCMP,GTP),nuclear, UMP kinase (dATP), UMP kinase (GTP), cytidylate kinase (CMP,UTP), UMP kinase (GTP),nuclear, UMP kinase (CTP), cytidylate kinase (dCMP,dCTP),nuclear, UMP kinase, nuclear, cytidylate kinase (CMP,dCTP)	TRUE
CRLS1	CLS_hs	cardiolipin synthase	TRUE
SPTLC3	SERPT	serine C-palmitoyltransferase	TRUE

[219]

PTPMT1	PGPP_hs	Phosphatidylglycerol phosphate phosphatase	TRUE
SGMS1	SMS	Sphingomyelin synthase	TRUE

THP-1 AraC-resistant

gene_symbol	reaction_id	reaction_name	structural_KO
ADSL	ADSL1, ADSL2	adenylosuccinate lyase, adenylosuccinate lyase	TRUE
ASNS	ASNS1	asparagine synthase (glutamine-hydrolysing)	TRUE
ATIC	IMPC, AICART	IMP cyclohydrolase, phosphoribosylaminoimidazolecarboxamide formyltransferase	TRUE
CAD	CBPS, ASPCTr, DHORTS	carbamoyl-phosphate synthase (glutamine-hydrolysing), aspartate carbamoyltransferase (reversible), dihydroorotase	TRUE
DHFR	r0224, r0512, DHFR, r0514, FOLR2, r0226	5,6,7,8-Tetrahydrofolate:NAD+ oxidoreductase One carbon pool by folate EC:1.5.1.3, Dihydrofolate:NAD+ oxidoreductase Folate biosynthesis EC:1.5.1.3, dihydrofolate reductase, Dihydrofolate:NADP+ oxidoreductase One carbon pool by folate / Folate biosynthesis EC:1.5.1.3, folate reductase, 5,6,7,8-Tetrahydrofolate:NADP+ oxidoreductase One carbon pool by folate / Folate biosynthesis EC:1.5.1.3	TRUE
DHODH	DHORD9	dihydoorotic acid dehydrogenase (quinone10)	TRUE
KDSR	3DSPHR	3-Dehydrosphinganine reductase	TRUE
GART	GARFT, PRAGSr, r0666	phosphoribosylglycinamide formyltransferase, phosphoribosylglycinamide synthase, 2-(Formamido)-N1-(5-phosphoribosyl)acetamidine cyclo-ligase (ADP-forming) Purine metabolism EC:6.3.3.1	TRUE
GPI	PGI	glucose-6-phosphate isomerase	FALSE

[220]

LCAT	LCAT1e	Lecithin-cholesterol acyltransferase	TRUE
LIPA	r1172, CHOLESTle, r1179, r1177	EC:3.1.1.13, hydrolysis of cholesterol ester by cholesterol esterase , EC:3.1.1.13, EC:3.1.1.13	TRUE
PFAS	PRFGS	phosphoribosylformylglycinamidine synthase	TRUE
PPAT	GLUPRT	glutamine phosphoribosyldiphosphate amidotransferase	TRUE
TXNRD1	TRDR2, TRDR3, r0027, r1433, TRDR	Thioredoxin (ubiquinone 10) reductase (NADPH), Thioredoxin (ubiquinone 10) reductase (NADH), NADPH:CoA-glutathione oxidoreductase EC:1.8.1.9, NADPH:oxidized-thioredoxin oxidoreductase Pyrimidine metabolism EC:1.8.1.9, thioredoxin reductase (NADPH)	TRUE
TYMS	TMDS	thymidylate synthase	TRUE
UMPS	ORPT, OMPDC	orotate phosphoribosyltransferase, "orotidine-5-phosphate decarboxylase"	TRUE
PGS1	PGPPT	phosphatidyl-CMP: glycerophosphate phosphatidyltransferase	TRUE
SPTLC2	SERPT	serine C-palmitoyltransferase	TRUE
SPTLC1	SERPT	serine C-palmitoyltransferase	TRUE
PAICS	AIRCr, PRASCS	phosphoribosylaminoimidazole carboxylase, phosphoribosylaminoimidazolesuccinocarboxamide synthase	TRUE
RPIA	RPI, r0249	ribose-5-phosphate isomerase, D-Ribose-5-phosphate ketol-isomerase Pentose phosphate pathway EC:5.3.1.6	TRUE
CMPK1	UMPK4n, CYTK8, CYTK8n, CYTK9, CYTK5n, CYTK9n, UMPK7, UMPK7n, CYTK1, UMPK5, CYTK10, CYTK10n, CYTK6,	UMP kinase (GTP),nuclear, cytidylate kinase (CMP,dATP), cytidylate kinase (CMP,dATP),nuclear, cytidylate kinase (CMP,dCTP), cytidylate kinase (CMP),nuclear, cytidylate kinase (CMP,dCTP),nuclear, UMP kinase (dGTP), UMP kinase (dGTP),nuclear, cytidylate kinase (CMP), UMP kinase (dATP), cytidylate kinase (CMP,dGTP), cytidylate kinase (CMP,dGTP),nuclear, cytidylate kinase (CMP,CTP), cytidylate kinase (dCMP,dGTP), cytidylate kinase (dCMP,dGTP),nuclear, cytidylate kinase	TRUE

	CYTK11, CYTK11n, CYTK12, CYTK12n, UMPK5n, CYTK13, CYTK13n, CYTK6n, CYTK14, CYTK14n, UMPKn, CYTK1n, UMPK6, CYTK2, CYTK2n, CYTK7, CYTK3n, CYTK4n, UMPK6n, CYTK5, UMPK2, CYTK7n, UMPK2n, UMPK3, UMPK, UMPK3n, UMPK4	(dCMP,dCTP), cytidylate kinase (dCMP,dCTP),nuclear, UMP kinase (dATP),nuclear, cytidylate kinase (dCMP,dATP), cytidylate kinase (dCMP,dATP),nuclear, cytidylate kinase (CMP,CTP),nuclear, cytidylate kinase (dCMP,UTP), cytidylate kinase (dCMP,CTP),nuclear, UMP kinase, nuclear, cytidylate kinase (CMP),nuclear, UMP kinase (dCTP), cytidylate kinase (dCMP), cytidylate kinase (dCMP),nuclear, cytidylate kinase (CMP,UTP), cytidylate kinase (dCMP,CTP),nuclear, cytidylate kinase (dCMP,GTP),nuclear, UMP kinase (dCTP),nuclear, cytidylate kinase (dCMP), UMP kinase (CTP), cytidylate kinase (CMP,UTP),nuclear, UMP kinase (CTP),nuclear, UMP kinase (UTP), UMP kinase, UMP kinase (UTP),nuclear, UMP kinase (GTP)	
CRLS1	CLS_hs	cardiolipin synthase	TRUE
SPTLC3	SERPT	serine C-palmitoyltransferase	TRUE
PTPMT1	PGPP_hs	Phosphatidylglycerol phosphate phosphatase	TRUE
SGMS1	SMS	Sphingomyelin synthase	TRUE

THP-1 DOX-resistant

gene_symbol	reaction_id	reaction_name	structural_KO
ADSL	ADSL2, ADSL1	adenylosuccinate lyase, adenylosuccinate lyase	TRUE
ASNS	ASNS1	asparagine synthase (glutamine-hydrolysing)	TRUE

[222]

ATIC	AICART, IMPC	phosphoribosylaminoimidazolecarboxamide formyltransferase, IMP cyclohydrolase	TRUE
ATP5F1A	ATPS4m	ATP synthase (four protons for one ATP)	TRUE
ATP5F1B	ATPS4m	ATP synthase (four protons for one ATP)	TRUE
ATP5F1C	ATPS4m	ATP synthase (four protons for one ATP)	TRUE
ATP5F1D	ATPS4m	ATP synthase (four protons for one ATP)	TRUE
ATP5F1E	ATPS4m	ATP synthase (four protons for one ATP)	TRUE
ATP5PB	ATPS4m	ATP synthase (four protons for one ATP)	TRUE
ATP5ME	ATPS4m	ATP synthase (four protons for one ATP)	TRUE
ATP5PF	ATPS4m	ATP synthase (four protons for one ATP)	TRUE
CAD	DHORTS, ASPCTr, CBPS	dihydroorotase, aspartate carbamoyltransferase (reversible), carbamoyl-phosphate synthase (glutamine-hydrolysing)	TRUE
COX4I1	CYOOm3	cytochrome c oxidase, mitochondrial Complex IV	TRUE
COX5B	CYOOm3	cytochrome c oxidase, mitochondrial Complex IV	TRUE
COX6A1	CYOOm3	cytochrome c oxidase, mitochondrial Complex IV	TRUE
COX6A2	CYOOm3	cytochrome c oxidase, mitochondrial Complex IV	TRUE
COX6B1	CYOOm3	cytochrome c oxidase, mitochondrial Complex IV	TRUE
COX6C	CYOOm3	cytochrome c oxidase, mitochondrial Complex IV	TRUE
COX7A1	CYOOm3	cytochrome c oxidase, mitochondrial Complex IV	TRUE
COX7A2	CYOOm3	cytochrome c oxidase, mitochondrial Complex IV	TRUE
COX7B	CYOOm3	cytochrome c oxidase, mitochondrial Complex IV	TRUE
COX7C	CYOOm3	cytochrome c oxidase, mitochondrial Complex IV	TRUE
COX8A	CYOOm3	cytochrome c oxidase, mitochondrial Complex IV	TRUE

SLC25A10	r0834, r0835, r0836, FUMSO3tm, MALSO4tm, FUMSO4tm, r0819, FUMtm, FUMTSULtm, r0885, MALTSULtm, MALtm, MALSO3tm, r0821, r2419, AKGMALtm, r2420, r0822, r0829, r0830, SUCCt2m	Mitochondrial Carrier (MC) TCDB:2.A.29.2.2, Mitochondrial Carrier (MC) TCDB:2.A.29.2.2, Mitochondrial Carrier (MC) TCDB:2.A.29.2.2, Fumarate:sulfite antiport, mitochondrial, Malate:sulfate antiport, mitochondrial, Fumarate:sulfate antiport, mitochondrial, Mitochondrial Carrier (MC) TCDB:2.A.29.2.2, fumarate transport, mitochondrial, Fumarate:thiosulfate antiport, mitochondrial, Mitochondrial Carrier (MC) TCDB:2.A.29.2.7, Malate:thiosulfate antiport, mitochondrial, malate transport, mitochondrial, Malate:sulfite antiport, mitochondrial, Mitochondrial Carrier (MC) TCDB:2.A.29.2.2, Mitochondrial Carrier (MC) TCDB:2.A.29.2.7, alpha-ketoglutarate/malate transporter, Mitochondrial Carrier (MC) TCDB:2.A.29.2.7, Mitochondrial Carrier (MC) TCDB:2.A.29.2.2, Mitochondrial Carrier (MC) TCDB:2.A.29.2.2, Mitochondrial Carrier (MC) TCDB:2.A.29.2.2, succinate transport, mitochondrial	TRUE
CYC1	CYOR_u10m	ubiquinol-6 cytochrome c reductase, Complex III	TRUE
CYP51A1	r0781	Lanosterol,NADPH:oxygen oxidoreductase (14-methyl cleaving) Biosynthesis of steroids EC:1.14.13.70	TRUE
DHCR7	DHCR71r, DHCR72r	7-dehydrocholesterol reductase, 7-dehydrocholesterol reductase	TRUE
DHFR	DHFR, r0224, r0226, r0512, r0514, FOLR2	dihydrofolate reductase, 5,6,7,8-Tetrahydrofolate:NAD+ oxidoreductase One carbon pool by folate EC:1.5.1.3, 5,6,7,8-Tetrahydrofolate:NADP+ oxidoreductase One carbon pool by folate / Folate biosynthesis EC:1.5.1.3, Dihydrofolate:NAD+ oxidoreductase Folate biosynthesis EC:1.5.1.3, Dihydrofolate:NADP+	TRUE

		oxidoreductase One carbon pool by folate / Folate biosynthesis EC:1.5.1.3, folate reductase	
DHODH	DHORD9	dihydoorotic acid dehydrogenase (quinone10)	TRUE
FDFT1	r0170, SQLSr, r0575	Farnesyl-diphosphate:farnesyl-diphosphate farnesyltransferase Biosynthesis of steroids EC:2.5.1.21, Squalene synthase, Presqualene diphosphate:farnesyl-diphosphate farnesyltransferase Biosynthesis of steroids EC:2.5.1.21	TRUE
KDSR	3DSPHR	3-Dehydrosphinganine reductase	TRUE
GART	GARFT, PRAGSr, r0666	phosphoribosylglycinamide formyltransferase, phosphoribosylglycinamide synthase, 2-(Formamido)-N1-(5-phosphoribosyl)acetamidine cyclo-ligase (ADP-forming) Purine metabolism EC:6.3.3.1	TRUE
GPI	PGI	glucose-6-phosphate isomerase	TRUE
HMGCR	r0488, HMGCOARc	(R)-Mevalonate:NADP+ oxidoreductase (CoA acylating) Biosynthesis of steroids EC:1.1.1.34, Hydroxymethylglutaryl CoA reductase (ir) in cytosol	TRUE
HSD17B4	FAOXC6C4x, FAOXC6DCC4DCx, FAOXC200180x, FAOXC14C12x, FAOXC8C6x, RE3005M, FAOXC226205x, FAOXC16DCC14DCx, FAOXC10C8x, FAOXC2242046x, FAOXC16C14x, FAOXC10DCC8DCx, FAOXC16080x, HSD17B4x,	fatty acid beta oxidation(C6-->C4)x, fatty acid beta oxidation(C6DC-->C4DC)x, Beta oxidation of long chain fatty acid, fatty acid beta oxidation(C14-->C12)x, fatty acid beta oxidation(C8-->C6)x, RE3005, Beta oxidation of long chain fatty acid, fatty acid beta oxidation(C16DC-->C14DC)x, fatty acid beta oxidation(C10-->C8)x, Beta oxidation of long chain fatty acid, fatty acid beta oxidation(C16-->C14)x, fatty acid beta oxidation(C10DC-->C8DC), Beta oxidation of long chain fatty acid, hydroxysteroid (17-beta) dehydrogenase 4, fatty acid beta oxidation(C8DC-->C6DC)x, fatty acid beta xoidation(C12:1-->C10)x, hydroxysteroid (17-beta) dehydrogenase 4, fatty acid beta	TRUE

	FAOXC8DCC6DCx, FAOXC121C10x, HSD17B42x, FAOXC14DCC12DCx, FAOXC12C10x, FAOXC180x, FAOXC12DCC10DCx, C3STKR2r, r0744, RE1526M, r0743	oxidation(C14DC-->C12DC)x, fatty acid beta oxidation(C12-->C10)x, Beta oxidation of long chain fatty acid, fatty acid beta oxidation(C12DC-->C10DC)x, C-3 sterol keto reductase (zymosterol), (24R,25R)-3alpha,7alpha,12alpha,24-tetrahydroxy-5beta-cholestanoyl- CoA hydro-lyase Bile acid biosynthesis EC:4.2.1.107, RE1526, hydroxysteroid (17-beta) dehydrogenase 4 Bile acid biosynthesis EC:1.1.1.35	
LCAT	LCAT1e	Lecithin-cholesterol acyltransferase	TRUE
LIPA	r1177, r1179, r1172, CHOLESTle	EC:3.1.1.13, EC:3.1.1.13, EC:3.1.1.13, hydrolysis of cholesterol ester by cholesterol esterase	TRUE
LSS	LNSTLSr	lanosterol synthase	TRUE
MT-CO1	CYOOm3	cytochrome c oxidase, mitochondrial Complex IV	TRUE
MT-CO2	CYOOm3	cytochrome c oxidase, mitochondrial Complex IV	TRUE
MT-CO3	CYOOm3	cytochrome c oxidase, mitochondrial Complex IV	TRUE
MT-CYB	CYOR_u10m	ubiquinol-6 cytochrome c reductase, Complex III	TRUE
MTHFR	r0792, MTHFD, MTHFR3	5-methyltetrahydrofolate:NAD+ oxidoreductase One carbon pool by folate / Methane metabolism EC:1.5.1.20, methylenetetrahydrofolate dehydrogenase (NADP), 5,10-methylenetetrahydrofolatereductase (NADPH)	TRUE
MTR	METS	methionine synthase	TRUE
MVD	DPMVDx, DPMVDc	diphosphomevalonate decarboxylase, diphosphomevalonate decarboxylase, cytosol	TRUE
MVK	MEVK1c, MEVK1x	mevalonate kinase (atp) cytosol, mevalonate kinase (atp)	TRUE
NSF	ATPS4m	ATP synthase (four protons for one ATP)	TRUE
PFAS	PRFGS	phosphoribosylformylglycinamidine synthase	TRUE

PLD2	PCHOLP_hs, RE3273C, RE3301C	choline phosphatase, RE3273, RE3301	TRUE
PPAT	GLUPRT	glutamine phosphoribosyldiphosphate amidotransferase	TRUE
MSMO1	C4STMO1r, C4STMO2r	C-4 sterol methyl oxidase (4,4-dimethylzymosterol), C-4 methyl sterol oxidase	TRUE
SQLE	SQLEr	Squalene epoxidase, endoplasmic reticular (NADP)	TRUE
TM7SF2	C14STRr, r0780	C-14 sterol reductase, 4,4-dimethyl-5a-cholesta-8,24-dien-3b-ol:NADP+ D14-oxidoreductase Biosynthesis of steroids EC:1.3.1.70	TRUE
TXNRD1	TRDR2, r0027, r1433, TRDR3, TRDR	Thioredoxin (ubiquinone 10) reductase (NADPH), NADPH:CoA-glutathione oxidoreductase EC:1.8.1.9, NADPH:oxidized-thioredoxin oxidoreductase Pyrimidine metabolism EC:1.8.1.9, Thioredoxin (ubiquinone 10) reductase (NADH), thioredoxin reductase (NADPH)	TRUE
TYMS	TMDS	thymidylate synthase	TRUE
UMPS	OMPDC, ORPT	"orotidine-5-phosphate decarboxylase", orotate phosphoribosyltransferase	TRUE
UQCRB	CYOR_u10m	ubiquinol-6 cytochrome c reductase, Complex III	TRUE
UQCRC1	CYOR_u10m	ubiquinol-6 cytochrome c reductase, Complex III	TRUE
UQCRC2	CYOR_u10m	ubiquinol-6 cytochrome c reductase, Complex III	TRUE
UQCRFS1	CYOR_u10m	ubiquinol-6 cytochrome c reductase, Complex III	TRUE
UQCRH	CYOR_u10m	ubiquinol-6 cytochrome c reductase, Complex III	TRUE
COX7A2L	CYOOm3	cytochrome c oxidase, mitochondrial Complex IV	TRUE
COX5A	CYOOm3	cytochrome c oxidase, mitochondrial Complex IV	TRUE
PGS1	PGPPT	phosphatidyl-CMP: glycerophosphate phosphatidyltransferase	FALSE
SPTLC2	SERPT	serine C-palmitoyltransferase	TRUE

[227]

ATP5MF	ATPS4m	ATP synthase (four protons for one ATP)	TRUE
PTDSS1	PSSA1_hs	Phosphatidylserine synthase homo sapiens	TRUE
ATP5PD	ATPS4m	ATP synthase (four protons for one ATP)	TRUE
SPTLC1	SERPT	serine C-palmitoyltransferase	TRUE
PAICS	AIRCr, PRASCS	phosphoribosylaminoimidazole carboxylase, phosphoribosylaminoimidazolesuccinocarboxamide synthase	TRUE
ATP5MG	ATPS4m	ATP synthase (four protons for one ATP)	TRUE
PMVK	PMEVKc, PMEVKx	phosphomevalonate kinase, cytosol, phosphomevalonate kinase	TRUE
EBP	EBP1r, EBP2r, r1381	3-beta-hydroxysteroid-delta(8),delta(7)-isomerase, 3-beta-hydroxysteroid-delta(8),delta(7)-isomerase, 5alpha-Cholest-7-en-3beta-ol delta7-delta8-isomerase Biosynthesis of steroids EC:5.3.3.5	TRUE
UQCR11	CYOR_u10m	ubiquinol-6 cytochrome c reductase, Complex III	TRUE
RPIA	RPI, r0249	ribose-5-phosphate isomerase, D-Ribose-5-phosphate ketol-isomerase Pentose phosphate pathway EC:5.3.1.6	TRUE
PISD	PSDm_hs	Phosphatidylserine decarboxylase	TRUE
PPA2	PPAm, r0009	inorganic diphosphatase, Pyrophosphate phosphohydrolase EC:3.6.1.1	TRUE
UQCRQ	CYOR_u10m	ubiquinol-6 cytochrome c reductase, Complex III	TRUE
UQCR10	CYOR_u10m	ubiquinol-6 cytochrome c reductase, Complex III	TRUE
NSDHL	C4STMO2r, C4STMO2Pr, C3STDH1r, C3STDH1Pr	C-4 methyl sterol oxidase, C-4 methyl sterol oxidase, C-3 sterol dehydrogenase (4-methylzymosterol), C-3 sterol dehydrogenase (4-methylzymosterol)	TRUE
CMPK1	CYTK1n, CYTK2, CYTK9n, CYTK2n, CYTK7,	cytidylate kinase (CMP),nuclear, cytidylate kinase (dCMP), cytidylate kinase (CMP,dCTP),nuclear, cytidylate kinase (dCMP),nuclear, cytidylate kinase (CMP,UTP),	TRUE

	CYTK3n, CYTK4n, CYTK5, CYTK5n, CYTK6, CYTK7n, CYTK6n, UMPK3n, UMPK, UMPK4, CYTK8, CYTK8n, CYTK9, UMPK4n, UMPK6, UMPK6n, UMPK2, UMPK7, UMPK7n, CYTK1, CYTK10, UMPK5, CYTK10n, CYTK11, UMPK2n, CYTK11n, CYTK12, CYTK12n, CYTK13, UMPK5n, CYTK13n, CYTK14, UMPK3, CYTK14n, UMPKn	cytidylate kinase (dCMP,CTP),nuclear, cytidylate kinase (dCMP,GTP),nuclear, cytidylate kinase (dCMP), cytidylate kinase (CMP),nuclear, cytidylate kinase (CMP,CTP), cytidylate kinase (CMP,UTP),nuclear, cytidylate kinase (CMP,CTP),nuclear, UMP kinase (UTP),nuclear, UMP kinase, UMP kinase (GTP), cytidylate kinase (CMP,dATP), cytidylate kinase (CMP,dATP),nuclear, cytidylate kinase (CMP,dCTP), UMP kinase (GTP),nuclear, UMP kinase (dCTP), UMP kinase (dCTP),nuclear, UMP kinase (CTP), UMP kinase (dGTP), UMP kinase (dGTP),nuclear, cytidylate kinase (CMP), cytidylate kinase (CMP,dGTP), UMP kinase (dATP), cytidylate kinase (CMP,dGTP),nuclear, cytidylate kinase (dCMP,dGTP), UMP kinase (CTP),nuclear, cytidylate kinase (dCMP,dGTP),nuclear, cytidylate kinase (dCMP,dCTP), cytidylate kinase (dCMP,dCTP),nuclear, cytidylate kinase (dCMP,dATP), UMP kinase (dATP),nuclear, cytidylate kinase (dCMP,dATP),nuclear, cytidylate kinase (dCMP,UTP), UMP kinase (UTP), cytidylate kinase (dCMP,CTP),nuclear, UMP kinase, nuclear	
CRLS1	CLS_hs	cardiolipin synthase	TRUE
SPTLC3	SERPT	serine C-palmitoyltransferase	TRUE
COX4I2	CYOOm3	cytochrome c oxidase, mitochondrial Complex IV	TRUE
PTPMT1	PGPP_hs	Phosphatidylglycerol phosphate phosphatase	FALSE
COX6B2	CYOOm3	cytochrome c oxidase, mitochondrial Complex IV	TRUE
COX7B2	CYOOm3	cytochrome c oxidase, mitochondrial Complex IV	TRUE

SGMS1	SMS	Sphingomyelin synthase	FALSE
COX8C	CYOOm3	cytochrome c oxidase, mitochondrial Complex IV	TRUE

HL60 Parental

gene_ symbol	reaction_ id	reaction_name	structu ral KO
ADSL	ADSL2, ADSL1	adenylosuccinate lyase, adenylosuccinate lyase	TRUE
ASNS	ASNS1	asparagine synthase (glutamine-hydrolysing)	TRUE
ATIC	AICART, IMPC	phosphoribosylaminoimidazolecarboxamide formyltransferase, IMP cyclohydrolase	TRUE
CAD	DHORTS, ASPCTr, CBPS	dihydroorotase, aspartate carbamoyltransferase (reversible), carbamoyl-phosphate synthase (glutamine-hydrolysing)	TRUE
CYP51 A1	r0781	Lanosterol,NADPH:oxygen oxidoreductase (14-methyl cleaving) Biosynthesis of steroids EC:1.14.13.70	TRUE
DHCR7	DHCR71r, DHCR72r	7-dehydrocholesterol reductase, 7-dehydrocholesterol reductase	TRUE
DHFR	r0224, DHFR, r0512, r0226, FOLR2, r0514	5,6,7,8-Tetrahydrofolate:NAD+ oxidoreductase One carbon pool by folate EC:1.5.1.3, dihydrofolate reductase, Dihydrofolate:NAD+ oxidoreductase Folate biosynthesis EC:1.5.1.3, 5,6,7,8-Tetrahydrofolate:NADP+ oxidoreductase One carbon pool by folate / Folate biosynthesis EC:1.5.1.3, folate reductase, Dihydrofolate:NADP+ oxidoreductase One carbon pool by folate / Folate biosynthesis EC:1.5.1.3	TRUE
DHOD H	DHORD9	dihydoorotic acid dehydrogenase (quinone10)	TRUE
FDFT1	r0170, SQLSr, r0575	Farnesyl-diphosphate:farnesyl-diphosphate farnesyltransferase Biosynthesis of steroids EC:2.5.1.21, Squalene synthase, Presqualene	TRUE

[230]

		diphosphate:farnesyl-diphosphate farnesyltransferase Biosynthesis of steroids EC:2.5.1.21	
KDSR	3DSPHR	3-Dehydrosphinganine reductase	TRUE
GART	r0666, PRAGSr, GARFT	2-(Formamido)-N1-(5-phosphoribosyl)acetamidine cyclo-ligase (ADP-forming) Purine metabolism EC:6.3.3.1, phosphoribosylglycinamide synthase, phosphoribosylglycinamide formyltransferase	TRUE
GOT2	TYRTAm, ASPTAm, 3SALATAim, VITD3t2	tyrosine transaminase, mitochondrial, aspartate transaminase, 3-sulfino-alanine transaminase (irreversible), mitochondrial, Vitamin D3 uptake	TRUE
GPI	PGI	glucose-6-phosphate isomerase	TRUE
HMGCR	r0488, HMGCOARc	(R)-Mevalonate:NADP+ oxidoreductase (CoA acylating) Biosynthesis of steroids EC:1.1.1.34, Hydroxymethylglutaryl CoA reductase (ir) in cytosol	TRUE
HMGCS1	r0463, HMGCOASi	(S)-3-Hydroxy-3-methylglutaryl-CoA acetoacetyl-CoA-lyase (CoA-acetylating) Synthesis and degradation of ketone bodies / Valine, leucine and isoleucine degradation / Butanoate metabolism EC:2.3.3.10, Hydroxymethylglutaryl CoA synthase (ir)	TRUE
HSD17B4	HSD17B4x, FAOXC226205x, FAOXC180x, FAOXC12C10x, r0744, RE1526M, FAOXC8DCC6DCx, FAOXC6C4x, FAOXC16DCC14DCx, FAOXC6DCC4DCx, FAOXC200180x, r0743, FAOXC8C6x, RE3005M, FAOXC2242046x, C3STKR2r, FAOXC10C8x,	hydroxysteroid (17-beta) dehydrogenase 4, Beta oxidation of long chain fatty acid, Beta oxidation of long chain fatty acid, fatty acid beta oxidation(C12-->C10)x, (24R,25R)-3alpha,7alpha,12alpha,24-tetrahydroxy-5beta-cholestanoyl- CoA hydro-lyase Bile acid biosynthesis EC:4.2.1.107, RE1526, fatty acid beta oxidation(C8DC-->C6DC)x, fatty acid beta oxidation(C6-->C4)x, fatty acid beta oxidation(C16DC-->C14DC)x, fatty acid beta oxidation(C6DC-->C4DC)x, Beta oxidation of long chain fatty acid, hydroxysteroid (17-beta) dehydrogenase 4 Bile acid biosynthesis EC:1.1.1.35, fatty acid beta oxidation(C8-->C6)x, RE3005,	TRUE

	FAOXC14DCC12DCx, FAOXC14C12x, FAOXC10DCC8DCx, FAOXC16080x, FAOXC121C10x, HSD17B42x, FAOXC12DCC10DCx, FAOXC16C14x	Beta oxidation of long chain fatty acid, C-3 sterol keto reductase (zymosterol), fatty acid beta oxidation(C10-->C8)x, fatty acid beta oxidation(C14DC-->C12DC)x, fatty acid beta oxidation(C14-->C12)x, fatty acid beta oxidation(C10DC-->C8DC), Beta oxidation of long chain fatty acid, fatty acid beta xoidation(C12:1-->C10)x, hydroxysteroid (17-beta) dehydrogenase 4, fatty acid beta oxidation(C12DC-->C10DC)x, fatty acid beta oxidation(C16-->C14)x	
IDH2	r0422, r0084, r0423, r0424, ICDHyrm, r0425, r0083	Isocitrate:NADP+ oxidoreductase (decarboxylating) Citrate cycle (TCA cycle) EC:1.1.1.42, Oxalosuccinate:NADP+ oxidoreductase (decarboxylating) Citrate cycle (TCA cycle) EC:1.1.1.42, Isocitrate:NADP+ oxidoreductase (decarboxylating) Citrate cycle (TCA cycle) EC:1.1.1.42, Isocitrate:NADP+ oxidoreductase (decarboxylating) Citrate cycle (TCA cycle) EC:1.1.1.42, Isocitrate dehydrogenase (NADP+), EC:1.1.1.42, Oxalosuccinate:NADP+ oxidoreductase (decarboxylating) Citrate cycle (TCA cycle) EC:1.1.1.42	TRUE
LCAT	LCAT1e	Lecithin-cholesterol acyltransferase	TRUE
LIPA	r1177, r1172, r1179, CHOLESTle	EC:3.1.1.13, EC:3.1.1.13, EC:3.1.1.13, hydrolysis of cholesterol ester by cholesterol esterase	TRUE
LSS	LNSTLSr	lanosterol synthase	TRUE
MVD	DPMVDx, DPMVDc	diphosphomevalonate decarboxylase, diphosphomevalonate decarboxylase, cytosol	TRUE
MVK	MEVK1c, MEVK1x	mevalonate kinase (atp) cytosol, mevalonate kinase (atp)	TRUE
PFAS	PRFGS	phosphoribosylformylglycinamidine synthase	TRUE
PPAT	GLUPRT	glutamine phosphoribosyldiphosphate amidotransferase	TRUE
MSMO1	C4STMO1r, C4STMO2r	C-4 sterol methyl oxidase (4,4-dimethylzymosterol), C-4 methyl sterol oxidase	TRUE

SQLE	SQLEr	Squalene epoxidase, endoplasmic reticular (NADP)	TRUE
TM7SF2	C14STRr, r0780	C-14 sterol reductase, 4,4-dimethyl-5a-cholesta-8,24-dien-3b-ol:NADP+ D14-oxidoreductase Biosynthesis of steroids EC:1.3.1.70	TRUE
TXNRD1	TRDR, TRDR3, TRDR2, r0027, r1433	thioredoxin reductase (NADPH), Thioredoxin (ubiquinone 10) reductase (NADH), Thioredoxin (ubiquinone 10) reductase (NADPH), NADPH:CoA-glutathione oxidoreductase EC:1.8.1.9, NADPH:oxidized-thioredoxin oxidoreductase Pyrimidine metabolism EC:1.8.1.9	TRUE
TYMS	TMDS	thymidylate synthase	TRUE
UMPS	OMPDC, ORPT	"orotidine-5-phosphate decarboxylase", orotate phosphoribosyltransferase	TRUE
PGS1	PGPPT	phosphatidyl-CMP: glycerophosphate phosphatidyltransferase	FALSE
SPTLC2	SERPT	serine C-palmitoyltransferase	TRUE
SPTLC1	SERPT	serine C-palmitoyltransferase	TRUE
PAICS	AIRCr, PRASCS	phosphoribosylaminoimidazole carboxylase, phosphoribosylaminoimidazolesuccinocarboxamide synthase	TRUE
PMVK	PMEVKc, PMEVKx	phosphomevalonate kinase, cytosol, phosphomevalonate kinase	TRUE
EBP	EBP1r, EBP2r, r1381	3-beta-hydroxysteroid-delta(8),delta(7)-isomerase, 3-beta-hydroxysteroid-delta(8),delta(7)-isomerase, 5alpha-Cholest-7-en-3beta-ol delta7-delta8-isomerase Biosynthesis of steroids EC:5.3.3.5	TRUE
RPIA	RPI, r0249	ribose-5-phosphate isomerase, D-Ribose-5-phosphate ketol-isomerase Pentose phosphate pathway EC:5.3.1.6	TRUE
PISD	PSDm_hs	Phosphatidylserine decarboxylase	TRUE
NSDHL	C3STDH1Pr, C4STMO2Pr, C4STMO2r, C3STDH1r	C-3 sterol dehydrogenase (4-methylzymosterol), C-4 methyl sterol oxidase, C-4 methyl sterol oxidase, C-3 sterol dehydrogenase (4-methylzymosterol)	TRUE
CMPK1	CYTK13n, CYTK14,	cytidylate kinase (dCMP,dATP),nuclear, cytidylate kinase (dCMP,UTP),	TRUE

	CYTK7n, CYTK14n, CYTK5, CYTK1n, CYTK2, CYTK2n, CYTK8, CYTK3n, CYTK5n, CYTK4n, UMPK, UMPK2, CYTK6, CYTK8n, CYTK6n, CYTK7, UMPK2n, UMPK4, UMPK4n, UMPK5, CYTK9, UMPK5n, CYTK9n, UMPK6n, UMPK3, UMPK7, UMPK7n, CYTK10, CYTK1, UMPKn, UMPK6, CYTK10n, CYTK11, UMPK3n, CYTK11n, CYTK12, CYTK12n, CYTK13	cytidylate kinase (CMP,UTP),nuclear, cytidylate kinase (dCMP,CTP),nuclear, cytidylate kinase (dCMP), cytidylate kinase (CMP),nuclear, cytidylate kinase (dCMP), cytidylate kinase (dCMP),nuclear, cytidylate kinase (CMP,dATP), cytidylate kinase (dCMP,CTP),nuclear, cytidylate kinase (CMP),nuclear, cytidylate kinase (dCMP,GTP),nuclear, UMP kinase, UMP kinase (CTP), cytidylate kinase (CMP,CTP), cytidylate kinase (CMP,dATP),nuclear, cytidylate kinase (CMP,CTP),nuclear, cytidylate kinase (CMP,UTP), UMP kinase (CTP),nuclear, UMP kinase (GTP), UMP kinase (GTP),nuclear, UMP kinase (dATP), cytidylate kinase (CMP,dCTP), UMP kinase (dATP),nuclear, cytidylate kinase (CMP,dCTP),nuclear, UMP kinase (dCTP),nuclear, UMP kinase (UTP), UMP kinase (dGTP), UMP kinase (dGTP),nuclear, cytidylate kinase (CMP,dGTP), cytidylate kinase (CMP), UMP kinase, nuclear, UMP kinase (dCTP), cytidylate kinase (CMP,dGTP),nuclear, cytidylate kinase (dCMP,dGTP), UMP kinase (UTP),nuclear, cytidylate kinase (dCMP,dGTP),nuclear, cytidylate kinase (dCMP,dCTP), cytidylate kinase (dCMP,dCTP),nuclear, cytidylate kinase (dCMP,dATP)	
CRLS1	CLS_hs	cardiolipin synthase	TRUE
SPTLC3	SERPT	serine C-palmitoyltransferase	TRUE
PTPMT1	PGPP_hs	Phosphatidylglycerol phosphate phosphatase	FALSE
SGMS1	SMS	Sphingomyelin synthase	FALSE

HL60 AraC-resistant

gene_ id	reaction_ id	reaction_name	structu ral KO
ACACA	ACCOAC	acetyl-CoA carboxylase	TRUE
ADSL	ADSL1, ADSL2	adenylosuccinate lyase, adenylosuccinate lyase	TRUE
ASL	ARGSL	argininosuccinate lyase	TRUE
ASNS	ASNS1	asparagine synthase (glutamine-hydrolysing)	TRUE
ASS1	ARGSS	argininosuccinate synthase	TRUE
ATIC	IMPC, AICART	IMP cyclohydrolase, phosphoribosylaminoimidazolecarboxamide formyltransferase	TRUE
CAD	ASPCTr, DHORTS, CBPS	aspartate carbamoyltransferase (reversible), dihydroorotase, carbamoyl-phosphate synthase (glutamine-hydrolysing)	TRUE
CPS1	r0034, CBPSam	Carbon-dioxide:ammonia ligase (ADP-forming,carbamate-phosphorylating) Urea cycle and metabolism of amino groups / Nitrogen metabolism EC:6.3.4.16, carbamoyl-phosphate synthase (ammonia) (mitochondria	TRUE
CYP51 A1	r0781	Lanosterol,NADPH:oxygen oxidoreductase (14-methyl cleaving) Biosynthesis of steroids EC:1.14.13.70	TRUE
DHCR7	DHCR72r, DHCR71r	7-dehydrocholesterol reductase, 7-dehydrocholesterol reductase	TRUE
DHFR	r0512, r0514, FOLR2, r0224, DHFR, r0226	Dihydrofolate:NAD+ oxidoreductase Folate biosynthesis EC:1.5.1.3, Dihydrofolate:NADP+ oxidoreductase One carbon pool by folate / Folate biosynthesis EC:1.5.1.3, folate reductase, 5,6,7,8-Tetrahydrofolate:NAD+ oxidoreductase One carbon pool by folate EC:1.5.1.3, dihydrofolate reductase, 5,6,7,8-Tetrahydrofolate:NADP+ oxidoreductase One carbon pool by folate / Folate biosynthesis EC:1.5.1.3	TRUE
DHOD H	DHORD9	dihydoorotic acid dehydrogenase (quinone10)	TRUE
FASN	FAS160COA, FAS140COA,	fatty-acyl-CoA synthase (n-C16:0CoA), fatty-acyl-CoA synthase (n-C14:0CoA),	TRUE

[235]

	KAS8, FAS100COA, FAS180COA, FAS80COA_L, FAS120COA	b-ketoacyl synthetase (palmitate, n-C16:0), fatty acyl-CoA synthase (n-C10:0CoA), fatty-acyl-CoA synthase (n-C18:0CoA), fatty acyl-CoA synthase (n-C8:0CoA), lumped reaction, fatty-acyl-CoA synthase (n-C12:0CoA)	
	SQLSr, r0170, r0575	Squalene synthase, Farnesyl-diphosphate:farnesyl-diphosphate farnesyltransferase Biosynthesis of steroids EC:2.5.1.21, Presqualene diphosphate:farnesyl-diphosphate farnesyltransferase Biosynthesis of steroids EC:2.5.1.21	TRUE
FDFT1	SQLSr, r0170, r0575	Squalene synthase, Farnesyl-diphosphate:farnesyl-diphosphate farnesyltransferase Biosynthesis of steroids EC:2.5.1.21, Presqualene diphosphate:farnesyl-diphosphate farnesyltransferase Biosynthesis of steroids EC:2.5.1.21	TRUE
KDSR	3DSPHR	3-Dehydrosphinganine reductase	TRUE
GART	r0666, GARFT, PRAGSr	2-(Formamido)-N1-(5-phosphoribosyl)acetamidine cyclo-ligase (ADP-forming) Purine metabolism EC:6.3.3.1, phosphoribosylglycinamide formyltransferase, phosphoribosylglycinamide synthase	TRUE
GOT2	3SALATAim, VITD3t2, ASPTAm, TYRTAm	3-sulfino-alanine transaminase (irreversible), mitochondrial, Vitamin D3 uptake, aspartate transaminase, tyrosine transaminase, mitochondrial	TRUE
GPI	PGI	glucose-6-phosphate isomerase	TRUE
HMGCR	r0488, HMGCOARc	(R)-Mevalonate:NADP+ oxidoreductase (CoA acylating) Biosynthesis of steroids EC:1.1.1.34, Hydroxymethylglutaryl CoA reductase (ir) in cytosol	TRUE
HMGCS1	HMGCOASi, r0463	Hydroxymethylglutaryl CoA synthase (ir), (S)-3-Hydroxy-3-methylglutaryl-CoA acetoacetyl-CoA-lyase (CoA-acetylating) Synthesis and degradation of ketone bodies / Valine, leucine and isoleucine degradation / Butanoate metabolism EC:2.3.3.10	TRUE
HSD17B4	HSD17B42x, FAOXC180x,	hydroxysteroid (17-beta) dehydrogenase 4, Beta oxidation of	TRUE

	FAOXC12C10x, FAOXC14DCC12DCx, FAOXC12DCC10DCx, FAOXC16080x, r0744, RE1526M, FAOXC16DCC14DCx, FAOXC6C4x, HSD17B4x, FAOXC200180x, FAOXC226205x, FAOXC14C12x, FAOXC8C6x, FAOXC2242046x, RE3005M, C3STKR2r, FAOXC6DCC4DCx, FAOXC10C8x, FAOXC10DCC8DCx, r0743, FAOXC16C14x, FAOXC8DCC6DCx, FAOXC121C10x	long chain fatty acid, fatty acid beta oxidation(C12-->C10)x, fatty acid beta oxidation(C14DC-->C12DC)x, fatty acid beta oxidation(C12DC-->C10DC)x, Beta oxidation of long chain fatty acid, (24R,25R)-3alpha,7alpha,12alpha,24-tetrahydroxy-5beta-cholestanoyl- CoA hydro-lyase Bile acid biosynthesis EC:4.2.1.107, RE1526, fatty acid beta oxidation(C16DC-->C14DC)x, fatty acid beta oxidation(C6-->C4)x, hydroxysteroid (17-beta) dehydrogenase 4, Beta oxidation of long chain fatty acid, Beta oxidation of long chain fatty acid, fatty acid beta oxidation(C14-->C12)x, fatty acid beta oxidation(C8-->C6)x, Beta oxidation of long chain fatty acid, RE3005, C-3 sterol keto reductase (zymosterol), fatty acid beta oxidation(C6DC-->C4DC)x, fatty acid beta oxidation(C10-->C8)x, fatty acid beta oxidation(C10DC-->C8DC), hydroxysteroid (17-beta) dehydrogenase 4 Bile acid biosynthesis EC:1.1.1.35, fatty acid beta oxidation(C16-->C14)x, fatty acid beta oxidation(C8DC-->C6DC)x, fatty acid beta xoidation(C12:1-->C10)x	
IDH2	r0425, r0084, r0422, r0083, r0423, r0424, ICDHyrm	EC:1.1.1.42, Oxalosuccinate:NADP+ oxidoreductase (decarboxylating) Citrate cycle (TCA cycle) EC:1.1.1.42, Isocitrate:NADP+ oxidoreductase (decarboxylating) Citrate cycle (TCA cycle) EC:1.1.1.42, Oxalosuccinate:NADP+ oxidoreductase (decarboxylating) Citrate cycle (TCA cycle) EC:1.1.1.42, Isocitrate:NADP+ oxidoreductase (decarboxylating) Citrate cycle (TCA cycle) EC:1.1.1.42, Isocitrate:NADP+ oxidoreductase (decarboxylating) Citrate cycle (TCA cycle) EC:1.1.1.42, Isocitrate dehydrogenase (NADP+)	TRUE
LCAT	LCAT1e	Lecithin-cholesterol acyltransferase	TRUE

LIPA	r1179, r1172, CHOLESTle, r1177	EC:3.1.1.13, EC:3.1.1.13, hydrolysis of cholesterol ester by cholesterol esterase , EC:3.1.1.13	TRUE
LSS	LNSTLSr	lanosterol synthase	TRUE
MTHFR	MTHFD, r0792, MTHFR3	methylenetetrahydrofolate dehydrogenase (NADP), 5-methyltetrahydrofolate:NAD+ oxidoreductase One carbon pool by folate / Methane metabolism EC:1.5.1.20, 5,10-methylenetetrahydrofolatereductase (NADPH)	TRUE
MTR	METS	methionine synthase	TRUE
MVD	DPMVDc, DPMVDx	diphosphomevalonate decarboxylase, cytosol, diphosphomevalonate decarboxylase	TRUE
MVK	MEVK1x, MEVK1c	mevalonate kinase (atp), mevalonate kinase (atp) cytosol	TRUE
OAT	ORNTArm	ornithine transaminase reversible (m)	TRUE
OTC	OCBTm	ornithine carbamoyltransferase, irreversible	TRUE
PFAS	PRFGS	phosphoribosylformylglycinamidine synthase	TRUE
PPAT	GLUPRT	glutamine phosphoribosyldiphosphate amidotransferase	TRUE
MSMO1	C4STMO1r, C4STMO2r	C-4 sterol methyl oxidase (4,4-dimethylzymosterol), C-4 methyl sterol oxidase	TRUE
SQLE	SQLEr	Squalene epoxidase, endoplasmic reticular (NADP)	TRUE
TM7SF2	C14STRr, r0780	C-14 sterol reductase, 4,4-dimethyl-5a-cholesta-8,24-dien-3b-ol:NADP+ D14-oxidoreductase Biosynthesis of steroids EC:1.3.1.70	TRUE
TXNRD1	TRDR2, TRDR, TRDR3, r0027, r1433	Thioredoxin (ubiquinone 10) reductase (NADPH), thioredoxin reductase (NADPH), Thioredoxin (ubiquinone 10) reductase (NADH), NADPH:CoA-glutathione oxidoreductase EC:1.8.1.9, NADPH:oxidized-thioredoxin oxidoreductase Pyrimidine metabolism EC:1.8.1.9	TRUE
TYMS	TMDS	thymidylate synthase	TRUE
UMPS	ORPT, OMPDC	orotate phosphoribosyltransferase, "orotidine-5-phosphate decarboxylase"	TRUE

[238]

PGS1	PGPPT	phosphatidyl-CMP: glycerophosphate phosphatidyltransferase	FALSE
SPTLC2	SERPT	serine C-palmitoyltransferase	TRUE
SPTLC1	SERPT	serine C-palmitoyltransferase	TRUE
PAICS	AIRCr, PRASCS	phosphoribosylaminoimidazole carboxylase, phosphoribosylaminoimidazolesuccinocarboxamide synthase	TRUE
PMVK	PMEVKx, PMEVKc	phosphomevalonate kinase, phosphomevalonate kinase, cytosol	TRUE
EBP	EBP2r, EBP1r, r1381	3-beta-hydroxysteroid-delta(8),delta(7)-isomerase, 3-beta-hydroxysteroid-delta(8),delta(7)-isomerase, 5alpha-Cholest-7-en-3beta-ol delta7-delta8-isomerase Biosynthesis of steroids EC:5.3.3.5	TRUE
RPIA	RPI, r0249	ribose-5-phosphate isomerase, D-Ribose-5-phosphate ketol-isomerase Pentose phosphate pathway EC:5.3.1.6	TRUE
PISD	PSDm_hs	Phosphatidylserine decarboxylase	TRUE
NSDHL	C3STDH1r, C4STMO2Pr, C3STDH1Pr, C4STMO2r	C-3 sterol dehydrogenase (4-methylzymosterol), C-4 methyl sterol oxidase, C-3 sterol dehydrogenase (4-methylzymosterol), C-4 methyl sterol oxidase	TRUE
CMPK1	CYTK13, CYTK13n, CYTK7, CYTK14, CYTK14n, CYTK1n, CYTK2, CYTK2n, CYTK7n, CYTK3n, CYTK5, CYTK4n, UMPK, CYTK5n, CYTK8, CYTK6, CYTK6n, UMPK2, UMPK3n, UMPK4, UMPK4n, CYTK8n, UMPK5, CYTK9, CYTK9n, UMPK2n,	cytidylate kinase (dCMP,dATP), cytidylate kinase (dCMP,dATP),nuclear, cytidylate kinase (CMP,UTP), cytidylate kinase (dCMP,UTP), cytidylate kinase (dCMP,CTP),nuclear, cytidylate kinase (CMP),nuclear, cytidylate kinase (dCMP), cytidylate kinase (dCMP),nuclear, cytidylate kinase (CMP,UTP),nuclear, cytidylate kinase (dCMP,CTP),nuclear, cytidylate kinase (dCMP), cytidylate kinase (dCMP,GTP),nuclear, UMP kinase, cytidylate kinase (CMP),nuclear, cytidylate kinase (CMP,dATP), cytidylate kinase (CMP,CTP), cytidylate kinase (CMP,CTP),nuclear, UMP kinase (CTP), UMP kinase (UTP),nuclear, UMP kinase (GTP), UMP kinase (GTP),nuclear, cytidylate	TRUE

	UMPK6n, UMPK7, UMPK7n, CYTK1, UMPK5n, CYTK10, CYTK10n, UMPK3, CYTK11, CYTK11n, UMPKn, CYTK12, CYTK12n, UMPK6	kinase (CMP,dATP),nuclear, UMP kinase (dATP), cytidylate kinase (CMP,dCTP), cytidylate kinase (CMP,dCTP),nuclear, UMP kinase (CTP),nuclear, UMP kinase (dCTP),nuclear, UMP kinase (dGTP), UMP kinase (dGTP),nuclear, cytidylate kinase (CMP), UMP kinase (dATP),nuclear, cytidylate kinase (CMP,dGTP), cytidylate kinase (CMP,dGTP),nuclear, UMP kinase (UTP), cytidylate kinase (dCMP,dGTP), cytidylate kinase (dCMP,dGTP),nuclear, UMP kinase, nuclear, cytidylate kinase (dCMP,dCTP), cytidylate kinase (dCMP,dCTP),nuclear, UMP kinase (dCTP)	
CRLS1	CLS_hs	cardiolipin synthase	TRUE
SPTLC3	SERPT	serine C-palmitoyltransferase	TRUE
NIT2	r0085, r0086	2-Oxoglutaramate amidohydrolase Glutamate metabolism EC:3.5.1.3, 2-Oxoglutaramate amidohydrolase Glutamate metabolism EC:3.5.1.3	TRUE
PTPMT1	PGPP_hs	Phosphatidylglycerol phosphate phosphatase	FALSE
SGMS1	SMS	Sphingomyelin synthase	TRUE

HL60 DOX-resistant

gene_symbol	reaction_id	reaction_name	structural KO
ADSL	ADSL1, ADSL2	adenylosuccinate lyase, adenylosuccinate lyase	TRUE
ASNS	ASNS1	asparagine synthase (glutamine-hydrolysing)	TRUE
ATIC	IMPC, AICART	IMP cyclohydrolase, phosphoribosylaminoimidazolecarboxamide formyltransferase	TRUE
CAD	CBPS, ASPCTr, DHORTS	carbamoyl-phosphate synthase (glutamine-hydrolysing), aspartate carbamoyltransferase (reversible), dihydroorotase	TRUE

[240]

DHFR	r0224, r0226, r0512, DHFR, r0514, FOLR2	5,6,7,8-Tetrahydrofolate:NAD+ oxidoreductase One carbon pool by folate EC:1.5.1.3, 5,6,7,8-Tetrahydrofolate:NADP+ oxidoreductase One carbon pool by folate / Folate biosynthesis EC:1.5.1.3, Dihydrofolate:NAD+ oxidoreductase Folate biosynthesis EC:1.5.1.3, dihydrofolate reductase, Dihydrofolate:NADP+ oxidoreductase One carbon pool by folate / Folate biosynthesis EC:1.5.1.3, folate reductase	TRUE
DHODH	DHORD9	dihydoorotic acid dehydrogenase (quinone10)	TRUE
KDSR	3DSPHR	3-Dehydrosphinganine reductase	TRUE
GART	GARFT, PRAGSr, r0666	phosphoribosylglycinamide formyltransferase, phosphoribosylglycinamide synthase, 2-(Formamido)-N1-(5-phosphoribosyl)acetamidine cyclo-ligase (ADP-forming) Purine metabolism EC:6.3.3.1	TRUE
LCAT	LCAT1e	Lecithin-cholesterol acyltransferase	TRUE
LIPA	r1172, r1179, r1177, CHOLESTle	EC:3.1.1.13, EC:3.1.1.13, EC:3.1.1.13, hydrolysis of cholesterol ester by cholesterol esterase	TRUE
PFAS	PRFGS	phosphoribosylformylglycinamidine synthase	TRUE
PPAT	GLUPRT	glutamine phosphoribosyldiphosphate amidotransferase	TRUE
TXNRD1	TRDR, TRDR2, TRDR3, r1433, r0027	thioredoxin reductase (NADPH), Thioredoxin (ubiquinone 10) reductase (NADPH), Thioredoxin (ubiquinone 10) reductase (NADH), NADPH:oxidized-thioredoxin oxidoreductase Pyrimidine metabolism EC:1.8.1.9, NADPH:CoA-glutathione oxidoreductase EC:1.8.1.9	TRUE
TYMS	TMDS	thymidylate synthase	TRUE
UMPS	ORPT, OMPDC	orotate phosphoribosyltransferase, "orotidine-5-phosphate decarboxylase"	TRUE
PGS1	PGPPT	phosphatidyl-CMP: glycerophosphate phosphatidyltransferase	FALSE
SPTLC2	SERPT	serine C-palmitoyltransferase	TRUE

[241]

SPTLC 1	SERPT	serine C-palmitoyltransferase	TRUE
PAICS	AIRCr, PRASCS	phosphoribosylaminoimidazole carboxylase, phosphoribosylaminoimidazolesuccinocarboxamide synthase	TRUE
RPIA	RPI, r0249	ribose-5-phosphate isomerase, D-Ribose-5-phosphate ketol-isomerase Pentose phosphate pathway EC:5.3.1.6	TRUE
CMPK 1	CYTK2, CYTK9n, CYTK2n, CYTK7, CYTK3n, CYTK4n, UMPK, CYTK5, CYTK5n, CYTK6, CYTK7n, CYTK6n, UMPK4, UMPK2, UMPK4n, CYTK8, CYTK8n, CYTK9, UMPK5, UMPK6n, UMPK7, UMPK2n, UMPK7n, CYTK1, CYTK10, CYTK10n, UMPK5n, CYTK11, CYTK11n, UMPK3, CYTK12, CYTK12n, UMPKn, CYTK13, CYTK13n, UMPK6, CYTK14,	cytidylate kinase (dCMP), cytidylate kinase (CMP,dCTP),nuclear, cytidylate kinase (dCMP),nuclear, cytidylate kinase (CMP,UTP), cytidylate kinase (dCMP,CTP),nuclear, cytidylate kinase (dCMP,GTP),nuclear, UMP kinase, cytidylate kinase (dCMP), cytidylate kinase (CMP),nuclear, cytidylate kinase (CMP,CTP), cytidylate kinase (CMP,UTP),nuclear, cytidylate kinase (CMP,CTP),nuclear, UMP kinase (GTP), UMP kinase (CTP), UMP kinase (GTP),nuclear, cytidylate kinase (CMP,dATP), cytidylate kinase (CMP,dATP),nuclear, cytidylate kinase (CMP,dCTP), UMP kinase (dATP), UMP kinase (dCTP),nuclear, UMP kinase (dGTP), UMP kinase (CTP),nuclear, UMP kinase (dGTP),nuclear, cytidylate kinase (CMP), cytidylate kinase (CMP,dGTP), cytidylate kinase (CMP,dGTP),nuclear, UMP kinase (dATP),nuclear, cytidylate kinase (dCMP,dGTP), cytidylate kinase (dCMP,dGTP),nuclear, UMP kinase (UTP), cytidylate kinase (dCMP,dCTP), cytidylate kinase (dCMP,dCTP),nuclear, UMP kinase, nuclear, cytidylate kinase (dCMP,dATP), cytidylate kinase (dCMP,dATP),nuclear, UMP kinase (dCTP), cytidylate kinase (dCMP,UTP), cytidylate kinase (dCMP,CTP),nuclear, UMP kinase (UTP),nuclear, cytidylate kinase (CMP),nuclear	TRUE

[242]

	CYTK14n, UMPK3n, CYTK1n		
CRLS1	CLS_hs	cardiolipin synthase	TRUE
SPTLC3	SERPT	serine C-palmitoyltransferase	TRUE
PTPMT1	PGPP_hs	Phosphatidylglycerol phosphate phosphatase	FALSE
SGMS1	SMS	Sphingomyelin synthase	FALSE

APPENDIX II

We performed statistical enrichment analysis on the list of genes identified by the supervised sparse Canonical Correlation Analysis (ssCCA) (Chapter 4.3) and identified the most enriched UniProt Knowledgebase (UniProtKB) keywords and Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathway terms.

Functional annotation analysis - UniProtKB keywords

This table complements Figure 4.27, by linking the enriched UniProtKB keywords, the penalised Fisher's probability score and the genes in the ssCCA results that belong in each UniProtKB keyword.

UniProtKB_keyword	penalised_Fisher's	genes
Phosphoprotein	1.83E-54	ITSN2, USP6NL, POP5, JRK, PWWP2B, AVIL, TESK1, TMEM94, ATP5C1, MYLK, ZC3H12C, KIAA1109, TBK1, MPRIP, AKT2, FAM110B, FGFR1OP2, ANKFY1, PSMD1, BBX, PRKACB, FNBP4, GABARAPL2, MEF2C, PRKAB2, RBFOX2, CSNK2A1, C5ORF15, DNTTIP1, HDGFRP3, PHKA1, PHKA2, PRKAB1, SND1, CD2AP, ZFYVE16, RUFY3, ZFYVE19, PSME3, PSME4, ORAI3, DDIT4, RFX5, AGTR1, ORAI1, CFAP97, OPTN, SKAP2, ZNF275, ZBTB8A, TTYH3, CDCA2, CEP85L, NPRL3, CDCA7, AGAP1, GATA3, THY1, ZC3HAV1, HACD3, SLC5A3, GATA1, OAZ2, RNF214, RNF213, PRPSAP2, PRKAR2B, LTA4H, ZBTB7A, ATG7, PPARGC1A, HEXIM1, LYN, ZNF264, ZNF263, PRPF38A, CCZ1B, PMM1, LRBA, LARP7, EIF2AK2, NETO2, EIF2AK4, GRIN2B, FERMT1, UCK2, QARS, AGO2, FXYD1, HYOU1, PSMG1, DCHS1, STRN, FERMT3, ARHGAP9, ZNF496, MTMR12, ATF7IP2, RCSD1, PEAR1, SHB, PRKAG3, ARHGAP4, NHSL2, TPSAB1, JAK2, IKBKE, CGNL1, DLGAP4, JAK1,

		CHST7, RIPK2, TEX10, LRRC40, TMC8, SIRT7, EPS15L1, GMIP, MED27, MED26, DIEXF, PGM2L1, LRRC42, PSMA5, MMP14, PSMA4, PSMA1, MMP15, RILPL1, PSMA2, CDC42EP4, CDC42EP3, DPYD, RAPGEF1, MMP19, VAMP1, RAPGEF2, HCST, PAFAH1B3, DLD, KANK2, FH, TMF1, SZRD1, ECHS1, STAU1, GSTP1, CIPC, ATL2, KCNA3, HTT, ZNF518A, CCNDBP1, SDAD1, PRAM1, FAM117B, PSMA7, FAM117A, GGA2, RECQL4, MYCT1, TUBA3C, C18ORF25, MICALL2, FLNA, FLNB, MEF2D, SRSF10, SRSF11, ATAD1, ZNF462, EGLN2, UTP3, TRAPPC10, ATAD2, SNX24, SAMD4B, PSMB8, NUDT21, CENPF, FABP5, SLMAP, SNX15, CCDC6, ADGRL1, MPLKIP, AGRN, EBAG9, C11ORF58, CRLF1, ADGRL3, RIF1, UBE2D3, GMNN, FAM13A, CASC4, SLA, MSI2, APEH, FBXO22, YY1, SLK, FNTA, SESN1, FNTB, PIM1, FBRSL1, MAP3K8, KPNA4, NOMO1, MAP3K5, BNIP3L, SH3GLB2, AGTPBP1, TICAM2, KHNYN, ZBTB38, STX7, FBXW9, ATP11C, ATP11A, ZBTB33, FBXO10, EMSY, SLC9B2, SAP30, MSH6, ACLY, SLC9A7, C18ORF54, SRSF3, PKP4, SRSF5, SNRNP200, IL6ST, SRSF8, KPNB1, LRRC58, KHDRBS1, RNASEH2B, CREM, FOXO6, SPATA5, CORO1B, HSP90B1, CLTCL1, SMYD2, EIF4EBP2, THEM6, EVI5, ZBTB18, NEK9, SDF2L1, BRPF1, GOT1, RBPMS, API5, NEK6, FOXN3, DEFA1, ZC3HC1, GGCT, IGSF10, RGCC, STK24, DMXL1, SP3, DMXL2, AAK1, ZNF655, LRCH4, ZNF654,

		LRCH3, NUMA1, MTCL1, ITGB4, FLT4, ITGB3, ITGB2, RASGRF1, FHL2, PTEN, APIP, ETFA, LRMP, PIK3CB, FUNDC2, NADK2, FSD1L, SPR, DDI2, YAF2, BCAS2, NCK1, SH2D3A, VPS13C, FAM207A, VPS13B, RAD23A, RAD23B, NSMCE4A, TMBIM1, PCCB, LCOR, ITGA6, DSG2, ITGA5, GAPDH, AHDC1, TRIM56, SRC, C1R, ZBTB46, SRSF1, MAK16, MLLT1, ATP10A, MLLT3, CXXC5, RCHY1, IL2RG, MFF, LRP8, FYB, LRP6, PARD6B, GANAB, ZNF629, CECR6, PHKG2, EIF4H, SH3BP4, IGF2BP3, IGF2BP2, LRRC8B, EIF4E, TRIM44, PTPN1, NOP58, SLC12A4, HNRNPA3, SLC12A5, SSB, GPKOW, NDUFA4, MID1IP1, DONSON, IMMT, BAIAP2, TPD52L1, SORL1, HNRNPAB, MYO1D, EIF3K, TRIM38, P4HB, ATXN10, PTPN7, EIF3C, SPG21, EIF3A, EIF3B, CPNE8, AHCYL1, ZAK, ICAM3, MT1X, CCNC, IKZF1, IKZF2, NUDT5, CCDC102A, CRKL, MT2A, EFR3A, HERC1, PIEZO2, FBXO4, CPNE3, KIF21B, CCNL2, SLC12A6, SCAPER, ACVR1, SFMBT1, ACTN1, NSMCE2, NSMCE1, TALDO1, ANK2, CDC40, GAREM1, PPA1, LACTB2, TAGLN2, TRIB3, KCTD12, MXRA7, LAP3, BCORL1, SPATA18, EIF1B, TP53, KCTD15, TMEM87A, KIAA1211L, TMEM87B, GGNBP2, TBC1D9B, TMEM63A, MYCBP2, PIK3R3, PTTG1IP, TTBK2, TGOLN2, C7ORF50, SEC14L1, SNX2, LARS, ERI1, CD19, HIVEP3, APOE, KCNN4, N4BP1, ZBED4, SEC31A, DECR1, PPP1R26, APOBEC3G, XRCC5, RANBP17, RAB27A, PTK6, PILRA, HSPA12B,

		PAICS, PTK2, ICK, AIMP2, ALDH6A1, TF, POLA2, ARMC10, NDUFAF4, GRB2, DHX29, CALM2, SPAG16, INTS12, FKBP15, INTS10, AKR1B1, PTPRJ, DOC2B, SLC7A11, AATF, CD3D, PPP1R18, LARP1B, HERPUD1, PTPRG, SYNCRIP, SCML2, PPP3CC, SCML4, NUDCD1, ARIH2, MLX, KIF1B, SPTAN1, CD34, CD33, CAP1, NCOA2, SREBF1, GRIPAP1, PBRM1, STARD3NL, IL16, GAB2, SYN1, CTTNBP2NL, LAT2, MN1, IL1A, ARHGEF9, PITPNM1, KIF2A, CSDE1, PITPNM2, CNKSR3, CANX, KIT, SDE2, TAX1BP1, GLCCI1, CD46, DEPDC5, STX12, FOXC1, DERL1, NEXN, RELA, CAMKK2, PGRMC1, PAK1, PGRMC2, CCT8, HEATR6, RAE1, NPDC1, CD74, GIT2, NPM1, MAP3K1, ERCC6L2, PLEKHA2, MIS18A, NOL4L, NEDD4, STK17B, TTC4, KIF26B, HNRNPA2B1, TCEA2, TCEA1, AGGF1, ESYT1, CD69, KIAA1958, MAD2L1, CARS, CD84, DCLRE1C, MPG, GDA, RPP30, WWC2, ZFAND5, BCCIP, UBE3B, SRP14, NR3C1, ETS2, CLGN, CCND3, PPIP5K1, CAPN7, PALM, TEKT2, ACAA1, RGS8, PTRF, SBF2, BORCS5, ADPRHL2, CYBB, TBC1D2B, CYBA, SLC30A1, TTC7A, SDHA, LAX1, CCNB1IP1, TANC1, TMX4, TYROBP, DOK4, CD8B, TMX1, HNRNPUL2, GPD2, EHBP1L1, DNAJB11, TBC1D24, ANKS6, VCL, PPID, RTN2, UBA7, PSMD14, ARHGEF28, DDX1, UTP23, SLC41A1, SLC1A4, SLC1A5, PDS5B, CACNA1C, PLD1, HIGD1A, ACACB, RTN4, NCKIPSD, CNN2, DEAF1, GRK6, NPHP4, CYC1, QSER1, YARS,

		IVNS1ABP, EEFSEC, ZFHX3, SYT1, PHF10, CAV2, CRIM1, NFXL1, HMGA2, NUP153, SSH2, GCLC, PQLC1, CCPG1, TMEM237, FAM175A, UBA2, TCF4, FGF13, UBE2K, LSM14A, FBN1, PRR5, RERE, RBM27, GSK3B, PRR7, PRKCSH, CHD7, PLEK, ADAR, ATP2A1, PCTP, CHD3, DNPH1, AFF2, LYL1, BCL7A, YWHAQ, BCKDHA, RBM17, SUN2, ARHGEF10, NAA30, ARHGEF12, NCBP1, EED, PDPK1, MMP2, DDX56, SHROOM3, PLA2G4A, SSTR2, TTC38, TBC1D1, LMBRD1, RCN2, ELF2, TBC1D4, GORASP2, ELF4, CAT, BLVRB, NCBP3, NAA38, MVD, IRF5, FBP1, SHANK3, NUP58, OTUB1, RBM22, DCTN6, CCDC189, CCDC186, DCTN4, GOSR1, RNF8, ACVR1B, ADD3, TROVE2, ADD1, PARL, PSMB10, ADD2, TAF1D, DNPEP, OXCT1, SV2A, CLMN, MKNK1, TGS1, PTK2B, TMEM209, PDLIM5, MTA3, TTC14, ARFGEF1, MBTPS1, RBM39, TIMM8A, CDK11A, GALNT2, EYA3, PDAP1, ACVR2B, SCIMP, EPOR, SOD1, IRGQ, TNIP1, ABHD17B, ACKR3, CLNK, VIM, PI4K2B, RPL5, RPL30, IPO11, RNH1, RPL31, MYT1L, C2ORF88, ARPC5L, SMC5, MSANTD3, ENO1, SMC3, TPGS1, SLC4A4, SMC4, GLS, ELK3, FAM122B, ELK4, MSANTD2, EVA1B, MYC, SOX15, CSPP1, WDR91, CREB3L2, RPL35, HOMER3, ANKS1A, VPS18, PRKCH, NUP210, DAPK1, PRKCE, FNDC3B, PRKCA, MED4, SYTL5, ATG13, SYTL4, CLPX, GTPBP4, NUP93, ARMC1, ADCY9, ZC2HC1A, ARMC5, SDPR, BIN1, PRKD3, HOXB4,

		HOXB2, PRKCQ, NUP98, ASF1A, LRPAP1, NUP205, PRKDC, HSPA4L, NEDD4L, SLAIN1, SKA2, MTDH, FAM124B, CRACR2B, PRDX1, RAD21, RRS1, PLCG2, PNPO, CHST13, PROM1, RREB1, CEP78, SLC19A2, PROM2, NAA50, PLK4, BARD1, CDK18, NDFIP2, JUND, ARPP21, H3F3B, HOMER1, ST13, PLK2, FZD6, STARD10, NFATC3, NFATC2, CDC7, SLC16A10, BRAF, UPF3A, NR1D2, RMDN2, MEX3D, MEX3C, PHC3, C9ORF85, WDR55, GLUD1, CYLD, CPEB1, HNRNPK, CEP85, CDK14, PHF3, CEP57, GPI, SETD5, SERPINA1, PHF20, DENND1A, CLTA, ASH2L, LYST, IFIH1, MRC2, TRIM9, AIFM1, IMPA1, TIMP1, AP3S1, LDLRAP1, JARID2, HAVCR2, UBL7, HELLS, WDR36, ANXA1, MIOS, ANXA4, NFAM1, PAWR, RNF168, INF2, PLCB4, HEATR5A, TAL1, MTF2, TCP1, PLIN3, EXOC5, GART, PLCB2, ENG, GRAMD1B, NBAS, PHLPP1, PLPPR1, SGMS1, PLPPR3, GPS1, ADCY3, ADCY2, TMEM163, ADCY6, FGD2, PPP1R2, SRP72, RNF139, ALOX5, PPP1R7, RPL14, USP1, SC5D, GPAT3, TAF9B, DTD1, RNF130, SF3B1, BRD4, SF3A3, SF3A1, KIDINS220, RIOK3, CSNK1A1, CARD9, KIAA0196, WWP2, C1ORF21, RRP1B, HIPK1, GPR6, FMO5, SCAF1, WDR13, SCAF4, HIPK2, RNF146, NFIA, HYLS1, NUCB2, PF4, CCDC71L, ANKRD13C, ZNF330, DGKD, ANKRD13D, RAB3A, ATP8A1, C1ORF131, TNF, CDC14A, PROSC, TUBB6, DEPTOR, HMGN5, APMAP, PHYHD1, CFL1, TUBB1, ZNF329,

		PAPOLA, SLC39A6, NBN, LZTS2, TMSB10, WHSC1L1, MB21D1, GTF2I, PLS1, TPD52, TLE3, C8ORF59, MREG, TLE1, CEP135, ANKRD49, TPM4, USP4, RALGAPA2, F2R, TPM1, DYRK1A, VRK1, VRK2, VRK3, LEMD3, C1ORF123, RAB32, GPATCH11, CLIP1, CACNB3, ILF3, CACNB4, STIM1, STIM2, KAT6B, KAT6A, MADD, ZNF318, CLIP4, ERF, ERH, AAGAB, ADNP, FKBP4, SNRPC, NOTCH2, ANKRD17, DNMT1, SLC26A2, CALCRL, FAM114A1, CBFB, FAM114A2, GLO1, CCDC25, AK1, AK2, ITPR1, ILDR2, TARS, AK3, ITPR2, ITPR3, RASAL2, KLC1, GTF2E2, BAG4, BAG2, OARD1, TBXA2R, UBR5, IGFBP7, CD300LF, SRGAP1, ZNF784, MAP2K5, OSBPL9, SEC16B, AKIRIN2, CD163, OSBPL5, ATP8B2, LSM1, ASNS, SYNJ2, ABHD15, MOB1B, CDK9, SLC7A6OS, NBEA, CPD, PLXNB2, SERINC1, ZNF777, IARS, DHCR7, EIF4G3, C8ORF33, MCM9, ZC3H3, AMD1, LSP1, SYNE1, PSTPIP2, EFHD2, DNMT3B, PHACTR1, GRASP, TNPO3, PHACTR4, IDH3A, PLEKHG3, HMGCS1, RFC2, TRPC1, SECISBP2L, C15ORF52, HAUS6, KCNAB2, PLRG1, REXO2, DUSP27, PTBP3, DUSP22, DNAJC7, DNAJC6, POLR1B, LANCL2, PECAM1, POLR1D, MCM4, MCM5, EVL, SNRPA1, DNAJC9, MCM6, DAGLB, EIF4E2, SSX2IP, SFT2D1, TRABD, HDAC4, SLC24A3, SMARCD3, HDAC1, CEBPE, EI24, KIAA1211, HDAC9, HDAC6, HDAC7, TTN, FBXL20, ABLIM1, VPS50, ZNF507, RSBN1L, VPS53, VPS52, RBBP8, DNAJB9,

		XIRP2, SNRPB2, RBBP6, RBBP7, WASF1, FNIP2, WASF3, SMARCE1, IRX1, PTPN18, IRX5, SEMA4B, PLCL2, ADRM1, DDHD2, PTPN11, CDC42BPB, CDC42BPA, GFER, PTPN14, OXR1, ASS1, ASXL1, RAD50, VAPA, POLR3D, DRG1, ESF1, EVI2B, PFKM, APP, SLC35B2, CXORF23, PANK1, GFI1, PITPNB, TIAL1, PNN, FRAT1, GOLGA1, CHEK1, DAG1, TRMT44, TNS3, SLC35A5, SKP1, SRRM2, IL15RA, CPT1A, C17ORF49, SP110, SLC35C2, ACSL5, SRRM1, INPP4A, RRAGA, KIAA0922, NEIL2, KARS, PDE12, VDAC3, TFAM, VDAC1, SIK2, EZR, MACF1, TBCEL, SLC22A5, NECAB3, TSHZ3, DCUN1D5, ANP32A, STXBP4, NLK, HNRNPLL, RPAP2, CRHR1, STRIP2, PTDSS1, PRRC1, MAP7, GSE1, HIST1H3H, TNRC18, STXBP5, MAP4, TRPM6, MAP9, VAT1, ZRSR2, ZRANB2, TMEM184B, TREX1, SIAH1, RAB39B, AMMECR1, PYCR2, PARVB, PARP14, PUM1, TRERF1, MANF, CCDC88B, CCT6A, CCDC88A, HAL, ZZZ3, HIST1H4H, RPS20, MAGED2, RNPS1, ANP32B, ROBO3, COL18A1, TRAM1, AHCTF1, CLIC4, SLC44A1, SH3KBP1, C2CD2, SRA1, COIL, XPO1, XPO5, KIF13B, UBXN8, RALGPS2, CBR1, TSC22D4, TSC22D1, PARP2, PRMT3, PPHLN1, UCHL3, ITFG2, MLF1, CEACAM1, CDCP1, ALDH5A1, RBM12B, CLDN12, WAC, RAB7B, MAPRE2, MAGEH1, PTGER1, NOL7, TRAK1, ZNF7, TYMP, CST3, PURB, LIMA1, PPP2CB, CUX1, TCEAL1, APBB2, TBL1X, KANSL1L, MAP4K3, MPZL1, LAIR1,

		MAP4K4, PURG, MPP1, RANBP2, RANBP1, MYEF2, SLC31A2, AP3D1, XPO7, LILRB2, KRT10, MPP5, PDZK1IP1, SVIP, HIST3H2A, KRT17, CAPZA2, KBTBD2, NBR1, F2RL1, RAB3GAP2, CTNNB1, PRPF31, COPG1, NUPL2, RAB9B, SIPA1, RAB3GAP1, TJP2, HDLBP, CIB1, COMT, ARHGAP35, SACS, HIST1H2AD, AZI2, TMEM38B, PDE8A, PDIA3, CD300A, ARID5B, KCTD1, LIFR, LIG3, PDIA6, TUFM, DPM1, TMEM134, OXA1L, KCTD9, TBL1XR1, SPECC1L, HIST1H2BE, ELMO2, CHMP4B, RELL1, LRIF1, XPR1, ATP6V1A, DTX2, ADRB2, ZFP36L2, ARHGAP22, FAM188A, PDZD2, ATP6V1H, RPRD1A, LONRF1, ASIC1, SPECC1, HSPA9, CBX5, SMAD3, MLKL, HSPA5, HSPA4, SMURF1, CBX2, PTCH1, CPSF2, TBCC, DNAJC12, ARHGAP27, ESR1, DNAJC17, ARHGAP23, ARHGAP32, DIAPH1, KIF18A, MAVS, DNMBP, DLG4, CHMP2B, CASS4, BRWD1, FGFR2, FGFR1, GPSM1, DOCK5, GPSM2, ACADVL, ABCD3, USP32, DOCK9, PDE3B, USP39, ELAVL1, MFSD6, SPRED1, SIX4, H2AFJ, CASP6, TRPS1, CASP1, FLOT1, SPEG, HIST1H1E, RBM5, HIST1H1C, KDM6A, PROSER2, RHBDF1, MAPK1IP1L, HGSNAT, H2AFY, PDE4D, SPATS2, ARAP3, PPP2R5C, OXSR1, SENP3, DCK, TGFBR2, DNM3, LATS2, RASA3, IGDCC4, RASA1, ALDH1A1, SAG, CHMP3, CHMP7, DOCK1, MCPH1, LTF, HEMGN, HCN3, VAC14, KRT80, LTK, RGS14, PALM2, ETFDH, MTIF2, FAM169A, HSPD1, MAPK9, UGP2,

		ERICH1, RPS6KC1, NRIP1, FAM120A, KLRG2, MAPK1, STOM, ABCF2, CAMK2G, SNCA, ZNF286A, USP24, USP25, ZNF184, CMTM6, PDE4DIP, NAP1L1, DEK, KANSL1, SCD, SSBP1, LPIN1, PDE7A, SSBP4
Alternative splicing	6.03E-44	ITSN2, POP5, PWWP2B, ATP5C1, ACCS, NDST2, NDST1, ZC3H12C, KIAA1109, AKT2, NDST3, FGFR1OP2, RAVER2, EPSTI1, VSTM1, ANKFY1, BBX, FAM19A5, PRKACB, WLS, FNBP4, PRKAB2, PHKA1, ZNF12, VOPP1, MPPE1, OPTN, CDCA2, SHMT1, CDCA7, GATA3, ZC3HAV1, GATA1, RNF214, RNF213, PRPSAP2, RHOT1, PDGFD, ADAMTS18, ATG7, JAG1, PRPF38A, MTERF3, TEX30, EIF2AK2, LARP7, EIF2AK4, PBX1, QARS, AGO2, HYOU1, STRN, DZIP3, ARHGAP9, ZNF496, ARHGAP8, MTMR12, SHB, PRKAG3, ARHGAP4, PCMTD2, PANX1, M1AP, ZNF41, TPSAB1, ENOPH1, PCMTD1, DLGAP4, DUSP3, TEX10, SLC39A11, EPS15L1, MED27, GMIP, MED26, SLC39A14, MED29, MED20, DPYD, RAPGEF1, MMP19, PPARA, HCST, OAT, SZRD1, SAMD8, ATL2, SDAD1, FAM117B, FAM117A, ADAM28, CST11, MFN1, STX2, P4HTM, ATAD1, ZNF462, LIMS3, TRAPPC10, ATAD2, SNX24, PASD1, KIAA2013, C20ORF196, SNX18, SLCO3A1, SLMAP, SNX15, ADGRL1, AGRN, EBAG9, ADGRL3, RIF1, PDCD6, CASC4, SLA, MSI2, FBXO21, FBXO22, FADS2, LAPTM4B, HHAT, SLK, FNTA, FNTB, POTEE, MTPAP, FADS1, MAP3K5, BNIP3L, FST, STX7, C21ORF58, FBXW9,

		ATP11C, FBXO10, FBXO11, MSH6, ACLY, MINPP1, MFNG, ZNF438, SRSF3, PKP4, SRSF5, SLC27A3, IL6ST, SRSF8, CDKL1, SPATA9, KHDRBS1, UQCRB, CREM, SPATA5, NPAS3, MED30, SMYD2, KDSR, API5, NEK6, HPS3, IGSF10, SP3, TBXAS1, DMXL2, LRCH3, PIWIL4, GPR89A, NUMA1, MTCL1, ITGB4, ITGB3, PTEN, APIP, ETFA, PTER, DDI2, TRIM68, ITGB6, YAF2, SLC36A1, BCAS4, CERS5, RNASET2, VPS13C, VPS13B, FAM207A, NRG1, PCCB, LCOR, ITGA6, FXN, GAPDH, TRIM56, PRPS1, SRC, SRSF1, ATP10A, SRI, MLLT3, RCHY1, IL2RG, TSPAN32, FYB, KCTD20, PHKG2, EIF4H, IGF2BP3, IGF2BP2, EIF4E, TBX1, HNRNPA3, EXOG, TM6SF1, TTC39C, HNRNPAB, TTC39B, EIF3K, TSPAN17, TRIM36, OCIAD2, ATXN10, EIF3C, EIF3A, EIF3B, TRIM34, MYO1G, CPNE8, IKZF1, IKZF2, NUDT6, EFR3A, FBXO4, PIEZO2, FBXO3, CPNE2, ATXN7L1, LGALS9, LGALS8, SCAPER, TRIM22, ACTN1, GAREM1, CRISPLD1, RIN3, BCORL1, TP53, KCTD15, GGNBP2, TBC1D9B, MYCBP2, TGFA, ACTR3B, GSPT1, TTBK2, NKAIN2, ACTR3C, SEC14L1, SARAF, DECR1, NTNG2, APOBEC3G, RANBP17, RAB27A, NBEAL1, PTK6, WDYHV1, IFFO1, PTK2, ALDH6A1, POLA2, NDUFAF6, NDUFAF7, MYO5C, ECHDC1, ECHDC3, GRB2, CD200, FBXL7, SPAG16, IGHM, ECM1, PYROXD1, HERPUD1, SYNCRIP, PPP3CC, NUDCD1, MUC15, SPTAN1, BBS4, HTATIP2, CAP1, SREBF1, PBRM1, IL16, GAB2,

		SYN1, RNF144B, ARHGEF9, PITPNM1, PITPNM2, CANX, KIT, SDE2, TAX1BP1, DEPDC5, NEXN, PRICKLE4, RELA, OS9, PGGT1B, TCEB1, GIT2, CFAP157, VEGFA, RIT1, NOL4L, KLHL5, KLHL7, HNRNPA2B1, CACHD1, TCEA2, AGGF1, TCEA1, RTN4IP1, KIAA1958, CARS, SYS1, GDA, RPP30, LOXL3, UBE3B, NR3C1, PPIP5K1, TBC1D31, ENPP2, IGFLR1, PTRF, BORCS5, TMOD1, TBC1D2B, SDHD, TTC7A, SDHA, SLFN13, TYROBP, TBC1D24, SIDT1, PPIF, ANKS6, FBN2, COX15, DDX1, HTRA2, CACNA1C, PLD1, HIGD1A, CNN2, NCKIPSD, ODF2L, ADGRG1, SYCE1, GRK6, QSER1, UBE2F, EEFSEC, TCAF1, CAV2, CNOT10, NUP153, UEVLD, CCPG1, TMEM237, MKRN1, UBA2, NABP1, TCF4, FGF13, UBE2K, GALK2, RERE, MS4A3, ADAR, PCTP, ADGRE2, MAN1A1, MYBL1, TGIF1, SUN2, PDPK1, SHROOM3, C4ORF48, QRSL1, EEF1G, ELF2, SETBP1, OTUB1, RNF8, PKD1L1, RETN, ADD3, TROVE2, ADD1, AGPAT3, ADD2, CAND2, CYTH2, CTAG1A, OXCT1, SV2A, SGK494, MKNK1, MXI1, PTK2B, TMEM209, MTA3, TTC14, CDK11A, EYA3, EPOR, UFM1, TNIP1, IPO11, TES, ANKLE1, WIPF2, MSANTD3, HNRNPR, SLC4A4, SMC4, SMC2, GLS, NYNRIN, ELK4, FAM122B, MSANTD2, ZMIZ1, ZMIZ2, CSPP1, WDR91, FAM122C, AP1S2, HOMER3, AP1S3, PRKCH, NUP210, COL23A1, ATG10, FNDC3B, HUS1, GTPBP8, CEL, GTPBP2, ATG13, GTPBP4, NUP93, ARMC1, ARMC5, ZSCAN20, PRKD3,

		PRKCQ, ZSCAN25, NUP98, PRKDC, FAM124B, BLOC1S2, LDHD, PRDM16, GAR1, ST3GAL5, ST3GAL6, PROM1, CEP78, ST3GAL3, PROM2, BARD1, NAA50, PLK4, PRELID1, NDFIP1, HOMER1, ARPP21, GRTP1, NFATC3, NFATC2, RMDN2, MEX3D, PHC3, WDR55, HNRNPK, TCN2, CEP85, XKR6, CEP57, PHF3, SETD5, GPI, PHF20, DENND1A, ARPC1A, WDR41, ASH2L, LYST, IFIH1, AIFM1, MTG2, JARID2, ZIK1, DENND2D, HELLS, COL2A1, HEATR5A, MTF2, PLIN3, GART, GRAMD1C, GRAMD1B, NBAS, SGMS1, AP5Z1, ADCY3, ADCY2, METTL21A, ADCY6, FGD2, ERMP1, ALOX5, PPP1R7, TP53TG1, APOL3, WDR17, KIDINS220, CSNK1A1, PCGF5, ELP4, RRP1B, HIPK1, FAM213A, SLC8B1, WDR13, SCAF4, HIPK2, MKLN1, NFIA, DHRS9, JKAMP, TMEM39A, NUCB2, CCDC71L, ANKRD13C, MTFMT, ANKRD13D, LCLAT1, CDC14A, DEPTOR, LAMP1, PHYHD1, SLC39A6, PAPOLA, SLC39A4, MB21D1, GTF2I, RAB2A, TLE3, C8ORF59, CEP135, SNRPN, COL27A1, TPM4, USP4, RALGAPA2, TPM1, DYRK1A, VRK2, VRK3, LBHD1, MORF4L1, ILF3, KAT6B, RAB37, ZNF318, AAGAB, SNRPB, RAB5B, CBFB, FAM114A1, CCDC25, INSIG1, GLT8D1, ORC4, UBR5, ZNF786, ZNF302, CD300LF, SRGAP1, MAP2K5, OSBPL9, SEC16B, OSBPL5, FANCL, SYNJ2, P4HA1, NBEA, REEP5, CPD, ZNF778, TPP1, ZNF777, SLC26A9, CNIH4, C8ORF33, SLC48A1, ADPGK, MCM9, ZC3H3, AMD1, SYNE1,

		TRO, ADAMTS2, SCRN1, PSTPIP2, PHACTR1, TNPO2, TNPO3, PHACTR4, IDH3A, PLEKHG3, RFC5, SLC13A3, GGPS1, RFC2, SECISBP2L, TRPC1, HAUS6, PLRG1, NPEPL1, DNAJC7, DNAJC6, POLR1B, PECAM1, POLR1D, HDAC4, EI24, NOD1, DENND3, HDAC9, COCH, HDAC6, HDAC7, TTN, LMF1, FBXL20, STOX2, ABLIM1, VPS50, VPS53, VPS52, XIRP2, SLC13A5, ANGPT2, SLC35A3, ANGPT1, ASCC1, FBXL17, GFER, ASXL1, P2RX5, P2RX4, VAPA, VPS41, EVI2B, VEZT, SLC35B2, ISCA1, TRAF3IP2, PANK1, PITPNB, IFIT5, PPP2R2A, CALML4, TIAL1, PNN, PPP2R1B, C10ORF128, SRRM2, IL15RA, CPT1A, C17ORF49, SLC35C2, ARRDC2, P3H1, YOD1, UHMK1, SRRM1, KIAA0922, NEIL2, CYP2U1, KARS, MACF1, SLC22A5, STXBP4, STXBP2, SLC35D2, HNRNPLL, RPAP2, CRHR1, STRIP2, ZNF827, GSE1, TNRC18, STXBP5, TRPM5, TRPM6, VAT1, ZRANB2, PARVB, PUM1, GBAS, TRERF1, SMU1, HAL, ZZZ3, RPS20, MAGED2, RNPS1, CCSER1, TECPR2, DET1, ROBO3, COL18A1, AHCTF1, ZCCHC10, SLC44A1, FAM45A, C2CD2, CITED2, MAGT1, SPPL3, SLC22A16, GSG1, UBXN8, RALGPS1, TMED5, RALGPS2, LRRCC1, TOP3B, CBR1, ZYG11B, PARP3, EDN3, PARP2, ZDHHC13, PPHLN1, GFRA3, ITFG2, MLF1, VWA8, CDCP1, ALDH5A1, TLR6, C9ORF3, GDPD1, NOL7, TYMP, ZNF7, LIMA1, TCEAL1, PDPN, SUSD3, APBB2, GABRE, KANSL1L, APBB3, EMC8,

		LAIR1, PURG, MPP1, LMLN, MYEF2, FAM46A, AIG1, MPP5, NR6A1, CAPZA2, NBR1, NUPL2, KBTBD8, LZIC, HOXA10, SACS, AZI2, HADH, HGF, LMO2, ALG13, VWDE, LIFR, LIG3, SPECC1L, CERCAM, HRASLS5, XPR1, RASL12, PXK, CHRNA6, CLEC16A, FBLN1, HRH2, TCTEX1D1, RPRD1A, CDH26, LONRF1, SPECC1, MLKL, RHBDD2, RHBDD1, CBX2, DIAPH1, SELL, CHMP2B, CASS4, BRWD1, LRP12, FGFR2, FGFR1, GPSM1, CASP6, H2AFJ, TOP1MT, TRPS1, CASP1, MFSD1, RBM5, PROSER2, H2AFY, PLAUR, PPP2R5C, DNM3, IQCK, IGDCC4, RASA3, RASA1, DNASE2, LOC440434, VAC14, LTK, ATF6B, NRXN3, ETFDH, NRXN2, FAM169A, MAPK9, UGP2, RPS6KC1, LEPROTL1, FAM120A, MAPK1, CYYR1, SNCA, ZNF286A, DEK, KANSL1, LPIN1, USP6NL, C5ORF22, CHIC1, AVIL, TMEM94, RBPJ, ZFYVE1, LIPA, MYLK, PSMD6, MPRIP, PSMD1, CHID1, MEF2C, RBFOX2, CSNK2A1, RBFOX3, HDHD2, EID1, ZFYVE16, PSME3, RUFY3, ZFYVE19, RFX5, PSME4, PSME1, CFAP97, ZNF275, ZBTB8A, ABCB1, PARPBP, TTYH3, CEP85L, AGAP1, SLCO5A1, HACD3, VTI1A, PPARGC1A, LYN, COL24A1, LRBA, BBIP1, NETO2, MRPL22, FERMT1, UCK2, TMEM99, PSMG1, ZNF256, FERMT3, ARF3, ZNF253, PIGT, ATF7IP2, HEXA, TMEM53, RCSD1, ZNF248, IKBKE, CGNL1, LYPLA1, C5ORF56, RIPK2, NIPA1, VWA5A, TMC8, MANEAL, SIRT7, TMC4, CCSAP, SUMF2, SRP9, SIRT3, PSMA5, MRPL50, PSMA4,

		PSMA1, RILPL1, CDC42EP4, TMEM106C, VAMP1, DLD, TMEM106A, KANK2, RNASEK, TMF1, MYRF, STAU1, ZNF518A, CCNDBP1, PSMA7, PRAM1, FSTL4, PTGS1, TUBA3C, C18ORF25, CARNS1, TMEM68, MICALL2, FLNA, FLNB, ZNF227, MEF2D, SRSF10, SRSF11, ZNF222, C11ORF63, TMEM64, TMCO4, PSMB8, FAM13B, C11ORF71, FAM13A, UBE2D3, ZC3HAV1L, HDDC3, SESN1, ZNF688, TCP11, SH3GLB2, ZNF681, LAG3, TICAM2, AGTPBP1, ZBTB37, EMSY, SLC9B2, C18ORF54, SNRNP200, KPNB1, RNASEH2B, FBLIM1, GLIS3, TMEM26, ATXN7, CLTCL1, HYI, SLC38A9, EVI5, ZBTB18, BRPF1, GOT1, RBPMS, TNFRSF10B, FOXN3, FOXN2, GGCT, ZC3HC1, RGCC, STK24, AAK1, ZNF655, TUBGCP4, TMEM19, SCHIP1, FLT4, RASGRF1, FHL1, FHL2, LRMP, NADK2, FSD1L, SUMO3, NCK1, SH2D3A, SIGMAR1, RAD23A, RAD23B, DDOST, NSMCE4A, SNURF, SLC25A12, PPM1L, NDUFB5, PRCP, CXXC5, MFF, LRP8, SLC25A27, SLC25A29, PARD6B, GANAB, NCLN, SLC25A28, CECR5, CECR6, TP53INP1, BEND4, SLITRK5, SH3BP4, BEND6, CTHRC1, SLC12A4, SLC12A5, DONSON, IMMT, BAIAP2, TPD52L1, TMLHE, PTPN7, ZNF610, SPG21, EPHB6, CCNJ, AHCYL1, ZAK, OGFOD3, THUMPD2, CCNC, IFI35, SPATA21, KIF21B, CCNL2, SLC12A6, CALN1, ENTPD1, SFMBT1, PHYKPL, PAQR3, ANK2, TAGLN2, KDELR2, MXRA7, SPATA18, TMEM87A, TMEM87B, PIK3R3,

		APH1A, TGOLN2, APH1B, SNX2, PCBP3, LARS, CD19, DHX35, ST8SIA4, HIVEP3, ARSD, ARSB, RAB11FIP4, SEC31A, SURF4, PILRA, PAICS, ICK, ARMC10, ACER3, ID1, ANKRD18A, INTS9, TMEM151B, FKBP15, N4BP2L1, PTPRJ, IDE, CD3D, CHRDL1, N4BP2L2, PPP1R18, LARP1B, PTPRG, IMMP2L, SCML2, SCML4, HYAL3, MLX, CD37, KIF1B, CD34, CD33, CTSC, GRIPAP1, TMEM150A, COMMD2, ANO6, COMMD7, LAT2, KIF2A, CSDE1, CRELD1, CD46, CRELD2, CD63, DERL1, CAMKK2, PAK1, PGRMC1, ALS2CL, PGRMC2, CD58, CCT8, CD74, NPM1, MOG, ERCC6L2, AKR1C2, GFI1B, DPY19L1, DPY19L2, NEDD4, KIF26B, ESYT1, MAD2L1, CD84, DCLRE1C, CD82, MPG, WWC2, MPL, BCCIP, OGDHL, MPO, CLGN, CCND3, KIF5C, PALM, BAALC, ACAA1, RGS8, SBF2, KIRREL3, CD96, SLC30A5, SSU72, LAX1, CAMKMT, TANC1, CD8B, GPD2, UBE2V1, VCL, RTN2, CCDC149, ARHGEF28, NPR3, UTP23, KLHL12, SLC1A4, SLC1A5, PDS5B, ACACB, RTN4, DEAF1, NPHP4, PRODH, C16ORF52, ZFHX3, PHF10, VTA1, NFXL1, HMGA2, SSH2, RCL1, PQLC1, IL7, FAM175A, GCLM, LSM14A, PRR5, GSK3B, PRR4, MCUR1, PRR7, NAA40, PRKCSH, CHD7, SIRPB2, ATP2A1, CHD3, DNPH1, AFF2, BCL7A, UBE2Q1, UBE2Q2, ZNF385C, BCKDHA, ARHGEF10, HSDL2, NAA30, HSDL1, ARHGEF12, EED, MMP2, DDX56, TBC1D8, SSTR2, CCNA1, TBC1D1, LMBRD1, RCN2, TBC1D4, MMRN1, GORASP2,

		NDUFS2, NUP54, ATG4A, NCBP3, NAA38, IRF5, NUP58, RBM22, CCDC189, GOSR2, DCTN4, GOSR1, ACVR1B, CLCN3, PARL, RAB40C, PDLIM5, RBM38, RBM39, GALNT3, EGF, GALNT2, SOD2, ACVR2B, SCIMP, ACVR2A, ABHD17B, GLB1, FAM131A, CLNK, FAM98B, CCDC171, SERPINE2, RPL31, MYT1L, SERPINE1, IPP, TPGS1, DCAF7, MYC, TEPP, SOX15, CREB3L2, MS4A14, HOXA1, ANKS1A, DUS4L, ACOT9, DAPK1, IFT80, MED4, SYTL5, SYTL4, BIN1, LGALS12, HOXB3, NHLRC2, NEDD4L, A1BG, SLAIN1, SKA2, CRACR2B, PNPO, CHST13, RREB1, SLC19A2, CDK18, NSUN5P1, FZD6, CDC7, UPF3A, C9ORF85, CYLD, GLUD1, CPEB1, TNFSF4, CDK14, HOXD8, ZNF133, SERPINA1, CLTA, ING1, MS4A4A, C16ORF70, TRIM9, ING3, IMPA1, ANXA8, HAVCR2, TIGD7, MIOS, ANXA4, C1ORF43, MS4A6A, PLCB4, MRAS, INF2, TAL1, PLCB2, ENG, CCDC90B, ASAH1, PHLPP1, PLPPR3, GPS1, AOAH, TMEM163, RHD, RNF135, SRP72, RNF130, SF3B1, BRD4, AOC3, SF3A1, RIOK3, CCL20, CARD8, CARD9, NMNAT3, WWP2, GPR6, FMO5, RNF146, RNF145, RFWD2, LPCAT3, LPCAT2, LAMTOR2, ZNF577, LAMTOR3, CCL28, SLC28A3, DGKG, DGKD, CSF3R, ATP8A1, ITGA2B, C1ORF131, STKLD1, APMAP, CLEC5A, ZNF567, STK32B, WHSC1L1, ZNF562, TPD52, BIVM, TOR3A, MREG, GPT2, TAP2, ARMCX4, GPATCH11, CACNB3, CLIP1, CACNB4, STIM1, STIM2, C6ORF136,

		MADD, CLIP4, ERF, ZNF557, ZNF555, PHLDB3, ANKRD17, ZNF550, DNMT1, ENY2, FECH, GLO1, AK2, ITPR1, AK3, TARS, ITPR2, RASAL2, KLC1, BAG4, BAG2, PITRM1, TBXA2R, BTBD3, ZNF548, ANKRD10, UGT3A1, IGFBP7, MPDU1, STRA6, CD164, CD163, TOR1A, ANKRD24, ATP8B2, ANKRD23, ASNS, HOPX, CDK9, MOB1B, CNOT7, FAM26F, CNOT8, EIF4G3, LSP1, DNMT3B, LINC01547, INSL3, GRASP, VWF, C15ORF52, KCNAB2, PTBP3, DUSP22, PTP4A3, TRAF5, EFHC1, EVL, DAGLB, EIF4E2, ZNF511, SSX2IP, TRABD, SMARCD3, RPE, MRM1, JAKMIP2, ENOX2, AP3M2, ZNF507, RSBN1L, RBBP8, EEF2KMT, RBBP6, RBBP7, ZNF500, ACSS1, FNIP2, DCAF13, WASF3, SMARCE1, PTPN18, SEMA4A, IRX5, SPAG6, SEMA4B, PLCL2, DDHD2, PTPN11, CDC42BPA, OXR1, PNRC2, KLF6, RAD50, MGAT4B, PFKM, CDRT1, APP, CMAS, CXORF23, CHEK1, TRMT44, TNS3, SKP1, SP110, ACSL5, EBF3, EBF4, MIF4GD, INPP4A, INPP4B, PDE12, CD109, BOLA3, VDAC3, TFAM, TMEM161B, BEX2, HKR1, NECAB3, LMAN2L, MGST1, MGST2, CD99L2, PTDSS1, PRRC1, MAP7, ASL, MAP4, MAP9, SIGLEC15, TREX1, SIAH1, AMMECR1, PARP14, UBAC2, CCDC88B, CCT6A, CCDC88A, ANP32B, TRAM1, SH3KBP1, FAM69B, OLA1, STAB1, KIF13B, TSC22D4, ATP6V0B, TSC22D1, STRBP, TSC22D2, PRMT3, IL17RE, RHOF, IL17RC, NAALADL1, CEACAM1, NME6, WAC, B3GNT2, MRRF, MAPRE2, SPAM1,

		AKAP7, TRAK1, CUX1, TBL1X, DISC1, MAP4K3, MPZL1, ANKMY1, MAP4K4, RANBP1, SCD5, AP3D1, LILRB2, RAB3GAP2, CTNNB1, PRPF31, B9D1, GHRL, RAB3GAP1, TJP2, GALNT12, HDLBP, CIB1, IL18RAP, TMEM147, PDE8A, TMEM14B, CD300A, EPDR1, ARID5B, PDIA6, OXA1L, TMEM134, TMEM135, TMEM136, COL4A1, FAM185A, ELMO2, LRIF1, ABCG1, ABCG2, ATP6V1A, DTX2, DTX3, LTBP3, NID2, ARHGAP22, TMEM123, FAM188A, PDZD2, ATP6V1H, ASIC1, BTNL8, SMAD3, HSPA4, SMURF1, PTCH1, TDRD9, DNAJC12, ARHGAP27, ARHGAP26, ESR1, ARHGAP23, ARHGAP32, MAVS, DNMBP, DLG4, DNAJC10, MAN2B1, NFKBID, B4GALT2, DOCK5, ACADVL, ABCD3, USP32, DOCK9, PDE3B, SLC2A5, USP39, ELAVL1, KYNU, CYP4V2, TMEM107, FLOT1, SPEG, GUSB, PDE4D, ARAP3, TRNAU1AP, TGFBR2, DDX19A, AGAP9, AGAP6, CHMP3, CXORF57, CHMP7, MCPH1, KRT80, TM2D2, RGS14, HPGD, PALM2, TFPI, HSPD1, RMND5B, RMND5A, KLRG2, STOM, ABCF2, CAMK2G, KLRG1, ZNF189, CMTM7, CMTM8, USP25, RWDD3, PDE4DIP, NAP1L1, GDAP1, TSHR, MEIS1, PDE7A, SSBP4
Acetylation	4.85E-37	USP6NL, RPL5, RPL30, IPO11, RNH1, RPL31, ATP5C1, HNRNPR, ENO1, RBPJ, SMC3, SMC4, SMC2, MYLK, GLS, FAM122B, PSMD7, AKT2, MYC, RAVER2, ANKFY1, PSMD1, RPL35, ANKS1A, MT1HL1, FNBP4,

		ACOT9, GABARAPL2, VPS18, MEF2C, PRKCA, MED4, ATG13, GTPBP4, SND1, CLPX, ARMC1, SDPR, BIN1, PSME3, RFX5, NUP98, ZNF277, NUP205, LXN, ANAPC16, PRKDC, SHMT1, RPN1, HSPA4L, ENDOD1, AGAP1, NEDD4L, WDR61, ZC3HAV1, SDSL, HACD3, GATA1, HACD2, MTDH, RNF214, CYB5R4, PRPSAP2, BLOC1S2, PRDX1, LDHD, RRS1, LTA4H, PROM1, ATG7, PPARGC1A, SLC19A2, NAA50, NDFIP1, PMM1, ARPP21, CCZ1B, H3F3B, ST13, LRBA, STARD10, ERLIN1, NFATC3, LARP7, EIF2AK2, BRAF, NR1D2, EIF2AK4, C9ORF85, GLUD1, UCK2, HNRNPK, QARS, PKIA, PSMG3, HYOU1, PSMG1, SFXN1, GPI, PHF20, CLTA, ASH2L, ING3, TMEM43, AIFM1, SH3BGRL3, LDLRAP1, JARID2, JAK1, LYPLA1, ANXA1, ANXA4, EPS15L1, CCSAP, MED29, SRP9, PSMA5, PLCB4, PSMA4, INF2, PSMA1, PSMA2, DPYD, TCP1, PLIN3, EXOC5, C5ORF51, GART, DLD, PAFAH1B3, FH, NBAS, SZRD1, PHLPP1, OAT, ECHS1, STAU1, MYRF, GSTP1, HTT, CCNDBP1, PSMA7, PPP1R2, ERMP1, TUBA3C, SRP72, RNF139, PPP1R7, RPL14, FLNA, FLNB, TAF9B, MEF2D, SF3B1, SRSF11, BRD4, SF3A3, SF3A1, CSNK1A1, UTP3, SNX24, RRP1B, WDR13, SCAF4, MKLN1, NUDT21, CENPF, FABP5, LPCAT3, CCDC6, C11ORF58, LAMTOR5, PDCD6, GMNN, CASC4, C1ORF131, ZC3HAV1L, SAT2, AHR, MSI2, APEH, LCLAT1, FBXO22, HDDC3, PTAR1, DEPTOR, FNTA, APMAP, UROD, MPC1, CFL1,

		MACROD1, PAPOLA, KPNA4, MTPAP, TMSB10, WHSC1L1, FADS1, PLS1, MB21D1, GTF2I, RAB2A, SH3GLB2, TLE3, C8ORF59, TPM4, GPT2, TPM1, STX7, FBXW9, LEMD3, RAB32, MSH6, ACLY, MORF4L1, GPATCH11, ILF3, KAT6B, RAB37, KAT6A, SRSF3, ERH, SRSF5, ADNP, FKBP4, SNRPC, SNRNP200, SRSF8, KPNB1, ANKRD17, KHDRBS1, ENY2, RAB5B, DNMT1, RNASEH2B, FECH, UQCRB, GLO1, CCDC25, AK1, AK2, AK3, TARS, AK4, UQCRH, GTF2E2, HSP90B1, MED30, PITRM1, BAG2, OARD1, CLTCL1, ESD, UBR5, MAP2K5, MPDU1, OSBPL9, NEK9, BRPF1, RBPMS, API5, FANCL, NEK6, ASNS, LSM4, CDK9, ZC3HC1, MOB1B, MFHAS1, LSM8, COPS5, STK24, UBLCP1, SP3, AAK1, IARS, C8ORF33, NUMA1, FHL1, PTEN, ETFA, LSP1, NADK2, RPL7A, FSD1L, SPR, SCRN1, EFHD2, CCDC53, TNPO2, BCAS2, TNPO3, IDH3A, NCK1, RFC5, ACTR3, HMGCS1, GGPS1, RFC2, VPS13C, VEZF1, PLRG1, KCNAB2, REXO2, PTBP3, DNAJC7, DNAJC6, PCCB, POLR1D, MCM4, MCM5, SNRPA1, MCM6, EIF4E2, SLC25A12, GAPDH, AHDC1, TRABD, SMARCD3, HDAC1, EI24, RPE, SRSF1, MAK16, MLLT1, FYB, VPS50, VPS53, RSBN1L, VPS52, EIF4H, POLR2E, RBBP8, EEF2KMT, SNRPB2, RBBP6, PCBD1, PCBD2, IGF2BP2, RBBP7, ACSS1, SEC23B, EIF4E, DCAF13, NDUFA9, PTPN1, HNRNPA3, SSB, GPKOW, NDUFA4, MID1IP1, PLCL2, ADRM1, IMMT, PTPN11, HNRNPAB, MYO1D, RAD50,

		TMLHE, VAPA, EIF3K, POLR3D, DRG1, ESF1, OCIAD2, P4HB, EIF3C, PFKM, EIF3A, EIF3B, MYO1G, AHCYL1, CMAS, PITPNB, CISD1, PPP2R2A, MT1X, IKZF2, NUDT5, TIAL1, PNN, MT2A, RPS19, PPP2R1B, PPAT, CCNL2, SRRM2, CPT1A, ACTN1, NSMCE2, ACSL5, TALDO1, MIF, SRRM1, NEIL2, KARS, PPA1, LACTB2, VDAC3, TAGLN2, VDAC1, SUCLG1, KCTD12, SIK2, LAP3, EZR, EIF1B, TP53, MMADHC, MACF1, MGST1, ACTR3B, RPAP2, NLN, PTDSS1, GNG10, SNX2, LARS, ATP5D, MAP7, GSE1, HIST1H3H, ASL, MAP4, MAP9, ATP6V1C1, DECR1, VAT1, ZRANB2, XRCC5, TMEM50B, RANBP17, RAB27A, PYCR2, RAB27B, PUM1, PAICS, TRERF1, UBAC1, PTK2, CCT6A, SMU1, RPS25, ALDH6A1, ZZZ3, MYO5C, ECHDC1, HIST1H4H, RPS20, MAGED2, RNPS1, GRB2, ANP32B, CALR, CALM2, CLIC4, FKBP15, HIBADH, PYROXD1, OLA1, AKR1B1, AATF, PPP1R18, HERPUD1, SYNCRIP, XPO1, XPO5, RAC2, KIF1B, SPTAN1, HTATIP2, CAP1, NCOA2, GRIPAP1, CBR1, PBRM1, ZER1, STARD3NL, PARP2, PRMT3, ZDHHC13, RHOF, PPHLN1, MN1, ALDH5A1, KIF2A, CSDE1, RBM12B, CANX, WAC, MAPRE2, STX12, CSTA, MRPS33, DERL1, RELA, CAMKK2, PURB, LIMA1, PAK1, EXOSC4, EMC2, CCT8, KANSL1L, MAP4K3, MAP4K4, MPP1, RANBP2, RANBP1, NPM1, MAP3K1, YIPF2, AP3D1, XPO7, SPCS2, HIST3H2A, DPY19L4, TTC4, NEDD4, RPA3, HNRNPA2B1, CAPZA2, CTNNB1, TCEA1, AGGF1,

		PRPF31, ESYT1, TOMM6, MAD2L1, CARS, RPP30, ZFAND5, ROGDI, HDLBP, UBE3B, NR3C1, CLGN, CAPN7, PALM, SACS, HIST1H2AD, HADH, PTRF, PSPH, PDIA3, SLC30A5, SDHA, SDHB, TUFM, PDIA4, DPM1, TANC1, TBL1XR1, UBE2V2, PPIF, HIST1H2BE, CHMP4B, UBE2V1, PPIB, VCL, PPID, ATP6V1A, DDX1, DTX2, SLC1A4, SLC1A5, PDS5B, HIGD1A, RTN4, CNN2, RPRD1A, PRODH, QSER1, HSPA9, YARS, UBE2F, SMAD3, HSPA5, PHF10, HSPA4, VTA1, DNAJC12, TBCC, HMGA2, CNOT10, NUP153, DIAPH1, GCLC, DNMBP, CHMP2B, UBA2, GCLM, UBE2K, LSM14A, RERE, DOCK5, ACADVL, ABCD3, PLEK, PCTP, SLC2A5, DNPH1, ELAVL1, MFSD6, SPRED1, UBE2Q1, SIX4, YWHAQ, H2AFJ, KYNU, HIST1H1E, MYBL1, HIST1H1C, MAPK1IP1L, BCKDHA, RBM17, HSDL2, NAA30, HSDL1, ARHGEF12, H2AFY, NCBP1, PDPK1, EED, PEX2, PPP2R5C, OXSR1, TTC38, ACTA2, EEF1G, DDX19A, SETBP1, TBC1D4, RASA1, ALDH1A1, CAT, NDUFS2, NAA38, MVD, FBP1, OTUB1, RBM22, VAC14, CCDC186, ATF6B, GOSR2, DCTN4, GOSR1, ETFDH, ADD3, TROVE2, ADD1, PSMB10, ADD2, HSPD1, RMND5B, CAND2, UGP2, RMND5A, DNPEP, ERICH1, GNG7, NRIP1, FAM120A, MAPK1, PDLIM5, ABCF2, SNCA, RBM39, EYA3, CMTM6, PDAP1, NAP1L1, DEK, S100B, SOD2, GDAP1, SOD1, KANSL1, SPRY1, VIM, SSBP1, SSBP4

RNA-binding	1.30E-06	RPL5, AHCYL1, IFIT5, THUMPD2, HDLBP, THUMPD3, HNRNPR, MSI2, SRP14, NR3C1, NUDT5, RPL8, TIAL1, RAVER2, PAPOLA, MTPAP, PTRF, SRRM2, SNRPN, RBFOX2, RBFOX3, EIF1AX, IMP3, SRRM1, UHMK1, ILF3, LACTB2, SRSF3, SNRPE, SRSF5, SNRPC, SRSF8, SNRPB, ANKRD17, KHDRBS1, ARHGEF28, DDX1, ZC3HAV1, DDX60, HNRNPLL, ZFP36L2, RPF1, PCBP3, ERI1, DND1, GAR1, PPARGC1A, YARS, ZRSR2, NSUN5P1, ZRANB2, RBPMS, MTERF4, CPSF2, LSM1, LARP7, EIF2AK2, UPF3A, SYNJ2, EIF2AK4, MEX3D, PUM1, MEX3C, LSM4, DNAJC17, LSM8, CPEB1, HNRNPK, CNOT7, AGO4, AGO2, RNPS1, CNOT8, EIF4G3, DZIP3, RBM27, MRPS17, PIWIL4, ADAR, TFB1M, ELAVL1, LARP1B, AFF2, IFIH1, SYNCRIP, XPO1, XPO5, RBM5, RBM17, STRBP, DDX56, TRNAU1AP, SRP9, PTBP3, DDX19A, CSDE1, RBM12B, SNURF, NCBP3, SNRPA1, EIF4E2, RBM22, STAU1, SRSF1, TROVE2, EXOSC4, NOCT, FAM120A, EIF4H, IGF2BP3, SNRPB2, IGF2BP2, SRSF10, EIF4E, SRSF11, RANBP2, RBM38, RBM39, NPM1, HNRNPA3, SSB, MYEF2, HNRNPAB, SCAF1, SCAF4, NUDT21, HNRNPA2B1, PRPF31, EIF3A, EIF3B
ATP-binding	1.05E-05	EPHB6, PI4K2B, PANK3, ZAK, PANK1, TESK1, SMC5, SMC3, SMC4, MYLK, SMC2, TBK1, AKT2, CHEK1, KIF21B, PRKACB, ACVR1, ENTPD1, PRKCH, CSNK2A1, DAPK1, PRKCE, ACSL5, PRKCA, CLPX, UHMK1,

		ADCY9, KARS, PRKD3, PRKCQ, SIK2, SLC22A4, ABCB1, SLC22A5, PRKDC, HSPA4L, NLK, ACTR3B, TTBK2, ACTR3C, LARS, DHX35, TRPM6, LYN, PLK4, CDK18, XRCC5, PLK2, EIF2AK2, PTK6, CDC7, BRAF, EIF2AK4, HSPA12B, PAICS, PTK2, ICK, CCT6A, GLUD1, UCK2, QARS, MYO5C, HYOU1, DHX29, CDK14, OLA1, IDE, PRKAG3, IFIH1, KIF13B, KIF1B, JAK2, IKBKE, JAK1, HELLS, RIPK2, VWA8, KIF2A, NME6, KIT, TCP1, GART, ADCY3, ADCY2, HSPA13, ADCY6, CAMKK2, RECQL4, PAK1, CARNS1, CCT8, MAP4K3, MAP4K4, ATAD1, MAP3K1, RIOK3, CSNK1A1, ERCC6L2, ATAD2, GATC, NMNAT3, HIPK1, HIPK2, STK17B, KIF26B, DGKG, CARS, DGKD, ATP8A1, UBE2D4, UBE2D3, STKLD1, SLK, PPIP5K1, KIF5C, PAPOLA, PIM1, MAP3K8, MTPAP, STK32B, MAP3K5, MB21D1, TOR3A, TAP2, DYRK1A, ATP11C, VRK1, LIG3, ATP11A, VRK2, MSH6, ACLY, SLFN13, SNRNP200, ABCG1, CDKL1, ABCG2, ATP6V1A, UBA7, DDX1, AK1, AK2, RHOBTB3, TARS, AK4, DDX60, SPATA5, ACACB, AK7, HSP90B1, ORC4, GRK6, MAP2K5, HSPA9, YARS, UBE2F, NEK9, TOR1A, MLKL, UBE2B, HSPA5, ATP8B2, HSPA4, NEK6, TDRD9, ASNS, CDK9, KIF18A, GCLC, STK24, AAK1, UBA2, IARS, UBE2K, GALK2, FGFR2, FGFR1, GSK3B, ABCD3, FLT4, MCM9, CHD7, ATP2A1, CHD3, PIK3CB, NADK2, UBE2Q1, UBE2Q2, SPEG, RFC5, ACTR3, PDPK1, RFC2, DDX56, OXSR1, QRSL1, DCK, TGFBR2, ACTA2, DDX19A, LATS2,

		PCCB, MCM4, MCM5, MVD, MCM6, PRPS1, LTK, SRC, NOD1, ATP10A, ACVR1B, CLCN3, TTN, HSPD1, MAPK9, RPS6KC1, SGK494, MKNK1, PHKG2, PTK2B, MAPK1, ACSS1, ABCF2, CAMK2G, CDK11A, CDC42BPB, CDC42BPA, ACVR2B, ACVR2A, ASS1, MYO1D, RAD50, PFKM, MYO1G
Mitochondrion	1.41E-05	ISCA1, CXORF23, CISD1, ATP5C1, GHITM, ALKBH7, NUDT6, GLS, GCSH, ACOT9, CPT1A, ENDOG, PHYKPL, STARD7, CMC2, ACSL5, PRKCA, CLPX, KARS, PDE12, LACTB2, DDIT4, VDAC3, TFAM, SUCLG1, VDAC1, SPATA18, TP53, MMADHC, MGST1, MRPL16, MRPL14, MRPL15, NLN, RHOT1, BLOC1S1, ATP5D, LDHD, COX3, DECR1, VAT1, PRELID1, MTERF3, MTERF4, PYCR2, MRPL22, GLUD1, ALDH6A1, NDUFAF6, NDUFAF7, NDUFAF4, SFXN1, ECHDC3, MRPS17, CLIC4, MRPS15, HIBADH, SLC44A1, MRPS12, HCCS, TFB1M, MRPL41, IMMP2L, AIFM1, MTG2, KIF1B, ARG2, MRPS24, TIMM23, SIRT3, RNF144B, MRPL50, ALDH5A1, VAMP1, MICU3, SDHAF4, DLD, MRRF, CCDC90B, KANK2, FH, MYCBP, OAT, ECHS1, GSTP1, MRPS33, TRAK1, MFN1, ERCC6L2, GATC, NMNAT3, SYNJ2BP, SLC8B1, BLID, RTN4IP1, TOMM6, MTFMT, MPG, OGDHL, NR3C1, COX6A2, CHCHD2, MPC1, PMPCB, MTPAP, HADH, PTRF, FADS1, BNIP3L, AGTPBP1, ADPRHL2, LIG3, SDHD, VRK2, SDHA, SDHB, TUFM, SLC9B2, RAB32, OXA1L, GPD2, PPIF, FKBP4,

		SLC27A3, ABCG2, FECH, UQCRB, COX15, ATP6, AK2, AK3, HTRA2, ABHD6, AK4, HIGD1A, SPATA5, ACACB, UQCRH, PITRM1, COX14, CYC1, PRODH, HSPA9, PRNP, UQCC2, RHBDD1, MAVS, ACADVL, MCUR1, ETFA, C7ORF73, METTL12, NADK2, TOP1MT, IDH3A, BCKDHA, HSDL1, VPS13C, MMP2, NDUFC1, REXO2, QRSL1, NDUFS8, PCCB, NDUFS2, SLC25A12, FXN, AGPAT5, SRC, NDUFB5, ETFDH, MTIF2, MFF, MRM1, PARL, HSPD1, SLC25A27, SLC25A29, SLC25A28, OXCT1, ACSS1, NDUFA9, TIMM8A, NDUFA4, IMMT, EXOG, SOD2, GDAP1, GFER, OXR1, SOD1, TMLHE, SLC25A32, SSBP1
Nucleotide-binding	1.79E-05	EPHB6, PI4K2B, PANK3, ZAK, PANK1, TESK1, SMC5, SMC3, SMC4, MYLK, SMC2, TBK1, AKT2, CHEK1, KIF21B, PRKACB, ACVR1, ENTPD1, PRKCH, CSNK2A1, DAPK1, PRKCE, ACSL5, GTPBP8, PRKCA, GTPBP2, GTPBP4, CLPX, UHMK1, RRAGA, ADCY9, KARS, PRKD3, VDAC3, SUCLG1, PRKCQ, SIK2, SLC22A4, ABCB1, SLC22A5, PRKDC, HSPA4L, AGAP1, NLK, GSPT2, ACTR3B, GSPT1, TTBK2, ACTR3C, RHOT1, LARS, PRKAR2B, DHX35, TRPM6, LYN, PLK4, CDK18, XRCC5, PLK2, RAB39B, RAB27A, EIF2AK2, PTK6, CDC7, BRAF, RAB27B, EIF2AK4, HSPA12B, PAICS, PTK2, ICK, CCT6A, GLUD1, UCK2, QARS, MYO5C, HYOU1, DHX29, CDK14, ARF3, OLA1, IDE, PRKAG3, IFIH1, MTG2, RAC2, KIF13B, KIF1B, JAK2, IKBKE, JAK1, HELLS, RIPK2, RHOF, RHOC, RAB33A,

[196]

[272]

		VWA8, MRAS, KIF2A, NME6, KIT, DPYD, TCP1, RAB7B, GART, ARF5, RHOQ, ATL2, ADCY3, AKAP7, ADCY2, HSPA13, ADCY6, CAMKK2, RECQL4, PAK1, TUBA3C, CARNS1, MFN1, CCT8, MAP4K3, MAP4K4, ATAD1, MAP3K1, RIOK3, CSNK1A1, ERCC6L2, ATAD2, GATC, NMNAT3, HIPK1, HIPK2, RIT1, STK17B, KIF26B, RAB9B, DGKG, CARS, RAB3A, DGKD, ATP8A1, UBE2D4, UBE2D3, STKLD1, THG1L, TUBB6, HHAT, SLK, PPIP5K1, KIF5C, TUBB1, PAPOLA, PIM1, MAP3K8, MTPAP, STK32B, MAP3K5, MB21D1, RAB2A, TOR3A, TAP2, DYRK1A, ATP11C, VRK1, LIG3, ATP11A, VRK2, VRK3, TUFM, RAB32, MSH6, ACLY, SLFN13, RAB37, SLC27A3, SNRNP200, ABCG1, CDKL1, ABCG2, ATP6V1A, RASL12, RAB5B, UBA7, DDX1, AK1, AK2, RHOBTB3, AK3, TARS, AK4, DDX60, SPATA5, ACACB, AK7, HSP90B1, ORC4, RAP1A, GRK6, MAP2K5, HSPA9, EEFSEC, YARS, UBE2F, NEK9, TOR1A, MLKL, UBE2B, HSPA5, ATP8B2, HSPA4, NEK6, TDRD9, ASNS, CDK9, KIF18A, GCLC, STK24, AAK1, UBA2, IARS, UBE2K, GALK2, FGFR2, FGFR1, GPSM2, GSK3B, ABCD3, FLT4, MCM9, CHD7, ATP2A1, CHD3, PIK3CB, NADK2, ARL5B, UBE2Q1, UBE2Q2, SPEG, RFC5, ACTR3, PDPK1, RFC2, DDX56, OXSR1, QRSL1, DCK, TGFBR2, ACTA2, DNM3, DDX19A, LATS2, PCCB, MCM4, MCM5, MVD, MCM6, HCN3, PRPS1, LTK, SRC, NOD1, ATP10A, MTIF2, ACVR1B, CLCN3, TTN, HSPD1, MAPK9, RAB40C, RPS6KC1,

		SGK494, MKNK1, PHKG2, PTK2B, MAPK1, ACSS1, ABCF2, CAMK2G, GBP4, CDK11A, CDC42BPB, CDC42BPA, ACVR2B, ACVR2A, ASS1, MYO1D, RAD50, DRG1, DRG2, PFKM, MYO1G
Chromosomal rearrangement	9.14E-05	CARS, RERE, IKZF1, MSI2, CCDC28A, LYL1, RPL7A, HOXA9, BCL7A, MYC, FGFR1OP2, ZNF41, CREB3L2, MACROD1, JAK2, CGNL1, WHSC1L1, KIRREL3, BCAS4, CD96, ARHGEF12, LMO2, LIFR, NRG1, OLIG2, MLF1, MN1, ELF4, KAT6B, TAL1, KAT6A, ZFYVE19, NUP98, SHANK3, C11ORF95, CBFB, CEP85L, FBLN1, MLLT1, MLLT3, HDAC9, RELA, NPAS3, NCKIPSD, NKAIN2, RNF213, RNF139, PRDM16, KDSR, DISC1, EVI5, EIF4E, BRD4, SEC31A, SPECC1, NPM1, HMGA2, PDE4DIP, BRAF, DEK, ARHGAP26, GRIN2B, PBX1, MFHAS1, CCDC6, TCEA1, HIST1H4H, CTNNB1, FGFR1
Serine/threonine-protein kinase	1.22E-04	GSK3B, ZAK, TESK1, MYLK, SLK, TBK1, AKT2, CHEK1, PIM1, SPEG, MAP3K8, IKBKE, STK32B, PRKACB, MAP3K5, ACVR1, PRKCH, CSNK2A1, PDPK1, RIPK2, DAPK1, PRKCE, DYRK1A, VRK1, PRKCA, VRK2, OXSR1, TGFBR2, UHMK1, LATS2, PRKD3, PRKCQ, SIK2, CDKL1, PRKDC, ACVR1B, NLK, TTBK2, CAMKK2, TTN, MAPK9, PAK1, SGK494, RPS6KC1, MKNK1, PHKG2, GRK6, MAPK1, TRPM6, CAMK2G, MAP4K3, MAP2K5, MAP4K4, PLK4, CDK18, NEK9, CDK11A, MAP3K1, CSNK1A1, RIOK3, PLK2, NEK6, EIF2AK2, BRAF, CDC7, EIF2AK4, CDC42BPB,

		CDC42BPA, HIPK1, ACVR2B, ACVR2A, HIPK2, CDK9, ICK, STK24, STK17B, AAK1, CDK14
Kinase	2.61E-04	DGKG, PI4K2B, PANK3, DGKD, ZAK, PANK1, TESK1, MYLK, SLK, TBK1, PPIP5K1, AKT2, CHEK1, PIM1, MAP3K8, STK32B, PRKACB, MAP3K5, ACVR1, PRKAB2, PRKCH, CSNK2A1, DAPK1, PRKCE, DYRK1A, VRK1, PRKCA, VRK2, VRK3, PHKA2, PRKAB1, UHMK1, PRKD3, PRKCQ, SIK2, CDKL1, PXK, PRKDC, AK1, AK2, PIK3R3, AK3, AK4, NLK, AK7, TTBK2, PRKAR2B, GRK6, TRPM6, MAP2K5, LYN, PLK4, CDK18, NEK9, PLK2, NEK6, EIF2AK2, PTK6, CDC7, BRAF, EIF2AK4, PTK2, CDK9, ICK, MOB1B, UCK2, STK24, PKIA, AAK1, CALM2, CDK14, GALK2, FGFR2, FGFR1, GSK3B, SH3KBP1, PRKCSH, ADPGK, FLT4, PIK3CB, PRKAG3, NADK2, SPEG, JAK2, IKBKE, JAK1, PDPK1, RIPK2, PLAUR, OXSR1, DCK, TGFBR2, LATS2, NME6, KIT, PRPS1, LTK, SGMS1, SRC, AKAP7, ACVR1B, TTN, CAMKK2, MAPK9, PAK1, RPS6KC1, SGK494, MKNK1, PHKG2, PTK2B, MAPK1, CAMK2G, MAP4K3, MAP4K4, GIT2, CDK11A, MAP3K1, RIOK3, KIDINS220, CSNK1A1, CDC42BPB, CDC42BPA, HIPK1, ACVR2B, ACVR2A, HIPK2, STK17B, CDK2AP2, PFKM
Transferase	3.51E-04	PI4K2B, CMAS, PANK3, ZAK, PANK1, TESK1, THUMPD2, THUMPD3, NUDT5, MYLK, NDST2, NDST1, TBK1, HERC1, AKT2, PPAT, CHEK1,

		NDST3, TRMT44, PRKACB, ACVR1, TKTL2, CPT1A, PRKAB2, PRKCH, CSNK2A1, PHYKPL, DAPK1, PRKCE, TALDO1, PRKCA, PHKA2, PRKAB1, UHMK1, PRKD3, PRKCQ, SIK2, PRKDC, SHMT1, RPN1, MGST1, MGST2, NEDD4L, PIK3R3, NLK, TTBK2, UGT8, PTDSS1, PRKAR2B, RNF217, CHST10, ST8SIA4, ST3GAL5, CHST13, ST3GAL6, TRPM6, ST3GAL1, ST3GAL3, BARD1, NAA50, LYN, PLK4, CDK18, NSUN5P1, PLK2, EIF2AK2, PTK6, CDC7, BRAF, EIF2AK4, PARP14, PTK2, ICK, UCK2, NDUFAF7, PKIA, CDK14, CALM2, SH3KBP1, ASH2L, TFB1M, PRKAG3, ARIH2, JAK2, IKBKE, PCMTD1, JAK1, CHST6, PARP3, CHST7, RIPK2, PARP2, PRMT3, ZDHHC13, ELOVL6, ELOVL7, PGM2L1, RNF144B, NME6, KIT, B3GNT2, GART, SAMD8, OAT, SGMS1, GSTP1, ZDHHC23, AKAP7, ZDHHC24, METTL21A, TRAK1, TYMP, CAMKK2, PAK1, PGGT1B, GPAT3, MAP4K3, MAP4K4, GIT2, MAP3K1, RIOK3, KIDINS220, CSNK1A1, WWP2, NMNAT3, HIPK1, HIPK2, DPY19L1, DPY19L2, ZDHHC9, DPY19L4, STK17B, NEDD4, LPCAT3, LPCAT2, DGKG, GALNT12, MTFMT, DGKD, UBE2D4, UBE2D3, SAT2, UBE3B, GXYLT2, COMT, LCLAT1, THG1L, HPGDS, HHAT, PTAR1, SLK, FNTA, PPIP5K1, FNTB, PAPOLA, PIM1, MAP3K8, ACAA1, MTPAP, STK32B, WHSC1L1, MAP3K5, MB21D1, GPT2, DYRK1A, ALG13, VRK1, VRK2, VRK3, CASD1, CAMKMT, DPM1, TPST2, ACLY, KAT6B, MFNG, KAT6A, HRASLS5,

		CDKL1, PXK, DNMT1, AK1, GLT8D1, AK2, AK3, AK4, FUT2, AK7, FUT4, FUT7, UBR5, GRK6, SMYD2, UGT3A1, MAP2K5, NEK9, UBE2B, GOT1, SMURF1, NEK6, CDK9, GGCT, MOB1B, STK24, AAK1, UBE2K, GALK2, FGFR2, FGFR1, B4GALT2, GSK3B, COLGALT2, NAA40, PRKCSH, ADPGK, FLT4, PIK3CB, METTL12, NADK2, UBE2Q1, UBE2Q2, DNMT3B, SPEG, POLI, HGSNAT, CERS5, NAA30, HMGCS1, GGPS1, PDPK1, PLAUR, OXSR1, DCK, DDOST, TGFBR2, LATS2, POLR1B, GAPDH, ST6GALNAC4, AGPAT5, PRPS1, LTK, SRC, ACVR1B, MRM1, AGPAT3, TTN, MAPK9, UGP2, RPS6KC1, OXCT1, SGK494, MKNK1, TGS1, PHKG2, EEF2KMT, PTK2B, MAPK1, CAMK2G, GALNT3, GALNT2, CDC42BPB, CDC42BPA, ACVR2B, ACVR2A, HS3ST1, MGAT4B, CDK2AP2, PFKM
GTPase activation	3.70E-04	ARHGAP9, USP6NL, TBC1D9B, ARHGAP8, RGS14, FAM13B, NPRL3, AGAP11, FAM13A, AGAP1, AGAP4, RASAL2, ARHGAP35, ARHGAP4, ARHGAP22, ALS2CL, RGS8, SRGAP1, RANBP1, GIT2, ARHGEF12, GRTP1, RALGAPA2, TBC1D8, ARAP3, TBC1D2B, ARHGAP27, GMIP, ARHGAP26, ARHGAP23, ARHGAP32, AGAP9, TBC1D1, TBC1D4, RASA3, AGAP6, RASA1, TBC1D24, RAPGEF2, RAB3GAP2, RIN3, DEPDC5, RAB3GAP1, SIPA1
Methylation	4.14E-04	CMAS, RAB3A, WIPF2, TESK1, MSI2, NR3C1, FAM122B, PNN, HOXA9, RPS19, ZMIZ2, PALM, HIST1H2AD, PDIA3, SRRM2, SNRPN,

		RBFOX2, RBFOX3, ARID5B, ILF3, CACNB4, HNRNPUL2, RAB37, HIST1H2BE, PKP4, TAGLN2, ADNP, FKBP4, TP53, SNRPB, ANKRD17, KHDRBS1, BEX2, DNMT1, HSPA4L, ILDR2, DTX2, SLAIN1, NID2, ORC4, BAG4, GNG10, RAP1A, GSE1, HIST1H3H, EEFSEC, HSPA9, ARPP21, H3F3B, HSPA5, API5, HSPA4, CBX2, EPHX1, RAB39B, RAB27A, BRAF, RAB27B, PUM1, ESR1, GNG11, DNAJC17, ARHGAP32, TF, MAVS, HNRNPK, AAK1, HIST1H4H, CALM2, LSM14A, C8ORF33, RBM27, PHF3, GPSM1, PIWIL4, USP32, CHD7, ADAR, ASH2L, ELAVL1, SYNCRIP, SPRED1, UBL3, H2AFJ, RAC2, SPEG, HIST1H1E, GRASP, DLGAP4, HIST1H1C, KDM6A, PLEKHG3, PROSER2, CAP1, NCOA2, H2AFY, EED, STRBP, VPS13C, KCNAB2, SIRT7, RHOF, RHOC, MED27, RAB33A, SYN1, ACTA2, PTP4A3, MRAS, ELF2, PITPNM1, TBC1D4, GORASP2, CDC42EP4, PITPNM2, SNURF, NDUFS2, ZNF511, GAPDH, SHANK3, AHDC1, RHOQ, KANK2, STAU1, HDAC1, PALM2, SRSF1, ATL2, HDAC6, RELA, PURB, PPP2CB, TUBA3C, C18ORF25, CCDC85A, GNG7, FAM120A, EIF4H, WASF1, RANBP2, SMARCE1, ZNF462, HNRNPA3, MYEF2, UTP3, SAMD4B, PDAP1, NAP1L1, CDC42BPB, TPD52L1, FMO5, HNRNPAB, WDR13, NUDT21, CENPF, GFI1B, HIST3H2A, TNIP1, NFIA, POLR3D, HNRNPA2B1, SNX15, CCDC6, ADGRL1, MPLKIP, ATXN10, SSBP4

Fatty acid biosynthesis	6.29E-04	PRKAB2, SCD5, ELOVL6, ELOVL7, HACD3, ACACB, HACD2, PRKAB1, PRKAG3, PTGS1, HACD4, FADS2, HPGDS, SCD, TBXAS1, PTGDS, FADS1
Proteasome	7.94E-04	PSMD14, ADRM1, RAD23A, RAD23B, PSMB8, PSMB10, PSMA7, PSMA5, PSMD6, PSMD7, PSMA4, PSMA1, PSMA2, PSME3, PSME4, PSME1, PSMD1
Repressor	1.00E-03	CCNC, IKZF1, AHR, ENO1, RBPJ, ARHGAP35, ELK3, YY1, ELK4, UXT, TRIM22, TLE3, TLE1, SFMBT1, ZBTB38, KCTD1, ZBTB33, ZNF12, EMSY, SAP30, EID1, TBL1XR1, KAT6B, KAT6A, ZNF318, ZNF438, ERF, BCORL1, TP53, ZNF275, DNMT1, TSHZ3, ANP32A, GLIS3, CREM, TWIST1, GATA1, PRDM16, RREB1, ZBTB7A, HEXIM1, ZBTB18, AKIRIN2, ZNF263, ZFHX3, CBX2, NR1D2, FOXN3, HOPX, PBX1, CPEB1, HNRNPK, CNOT7, ZNF418, SP3, AGO2, ID1, CNOT8, LSM14A, ZNF256, MXD4, ZNF133, ZNF496, RERE, ZNF253, CITED2, ATF7IP2, BMI1, SCML2, SCML4, TRPS1, SIX3, MLX, DNMT3B, JARID2, ZIK1, TGIF1, TSC22D4, TSC22D1, EED, PPHLN1, SIRT7, PTBP3, LCOR, HDAC4, TMF1, HDAC1, CIPC, CXXC5, HDAC9, HDAC6, HDAC7, PURB, CUX1, MXI1, MAPK1, RBBP7, BEND6, MYEF2, PCGF5, SAMD4B, ASXL1, GFI1B, MNDA
Translation regulation	0.002	PIWIL4, SYNCRIP, AKT2, MTG2, MKNK1, EIF4EBP2, IGF2BP3, IGF2BP2, PAIP2, EIF4E, NCK1, NCBP1, DAPK1, CNOT10, EIF2AK4,

		PUM1, MIF4GD, CPEB1, CNOT7, AGO4, AGO2, CNOT8, EIF4E2, LSM14A, EIF4G3, GAPDH
Spliceosome	0.002	SRSF1, HNRNPR, USP39, PNN, SYNCRIP, DHX35, SNRPB2, SF3B1, RBM5, BCAS2, SF3A3, SRRM2, ZRSR2, RBM17, SF3A1, HNRNPA3, PRPF38A, PLRG1, LSM4, CDC40, SRRM1, LSM8, HNRNPK, HNRNPA2B1, PRPF31, SNRPA1, SNRPE, SNRNP200, RBM22, SNRPB
Protein biosynthesis	2.00E-03	CARS, MTFMT, TARS, MTIF2, GSPT2, GSPT1, LARS, EIF4H, EIF4EBP2, TCEB1, EIF4E, EEFSEC, YARS, EIF1AX, GATC, TRNAU1AP, QRSL1, TUFM, EEF1G, AIMP2, EIF3K, KARS, QARS, TCEA2, TCEA1, DHX29, IARS, EIF3C, EIF4E2, EIF4G3, MRRF, EIF1B, EIF3A, EIF3B
Proto-oncogene	3.00E-03	CARS, CBFB, SRC, MLLT1, BMI1, MLLT3, ETS2, NCKIPSD, LYL1, HOXA9, FRAT1, RNF213, MYC, AKT2, PDGFD, MXI1, CREB3L2, PIM1, MAP3K8, KDSR, JAK2, WHSC1L1, SEC31A, LYN, BCAS4, SPECC1, NPM1, ARHGEF12, USP4, LMO2, HMGA2, PRKCA, BRAF, OLIG2, DEK, ARHGAP26, PBX1, HOPX, GFI1B, ELF4, TAL1, ZFYVE19, KAT6A, KIT, CCDC6, EVI2B, FGFR2
Fatty acid metabolism	0.004	ACADVL, ECHS1, HPGD, HACD3, HACD2, ACACB, PRKAG3, HACD4, PTGS1, FADS2, HPGDS, HADH, ACAA1, PTGDS, FADS1, DECR1, LYPLA1, CPT1A, PRKAB2, SCD5, ACSL5, ELOVL6, ELOVL7, PRKAB1, SCD, TBXAS1, LPIN1, SLC27A3

Angiogenesis	0.005	GPI, ECM1, FLT4, PDCD6, PDE3B, NRXN3, CIB1, SHB, TYMP, ARHGAP22, MYDGF, RNF213, PTK2B, HTATIP2, SEMA4A, EGFL7, ANGPT2, ANGPT1, MMP2, PRKCA, PTK2, EREG, VEGFA, COL4A2, COL4A1, AC G
Apoptosis	6.00E-03	APP, SH3KBP1, PDCD6, UBE2D3, APIP, PTEN, SRA1, CIB1, GHITM, SHB, NR3C1, TIAL1, MRPL41, MYDGF, SLK, AIFM1, CASP6, AKT2, PIM1, CASP1, KIF1B, IFT57, RBM5, HTATIP2, MAP3K5, SRGN, BNIP3L, MEF2C, CSNK2A1, RIPK2, DAPK1, PAWR, PRKCA, FBXO10, TGFBR2, MFSD10, RNF144B, RRAGA, AIM2, NME6, PSME3, DDIT4, TRAF5, LGALS12, MADD, TAX1BP1, PPIF, ELMO2, CHMP3, TRIB3, DNASE2, VDAC1, DOCK1, GAPDH, TP53, PPID, KANK2, BEX2, PHLPP1, SGMS1, EI24, CDCA7, ITPR1, HTRA2, HTT, NOD1, FAM188A, PAK1, RAD21, TP53INP1, MAPK1, DRAM2, RNF130, MEF2D, PRELID1, CDK11A, GADD45B, API5, RHBDD1, NEK6, CARD8, SIAH1, TNFRSF10C, TNFRSF10B, HIPK2, AIMP2, STK24, STK17B, BLID, EBAG9, FGFR2
Cytoskeleton	0.006	RIF1, AVIL, WIPF2, IPP, ARPC5L, CIB1, TPGS1, NR3C1, ARHGAP35, CDC14A, MYLK, FAM110A, UXT, TUBB6, MPRIP, KIF5C, CSPP1, CFL1, CHEK1, TUBB1, FAM110B, DAG1, TEKT2, KIF21B, LZTS2, TMSB10, TMOD1, CEP135, TPM4, DAPK1, ACTN1, PRKCE, TPM1, VRK1, ANK2, IFT80, CD2AP, CLIP1, STIM1, ZFYVE19, SPECC1L, PKP4, EZR, FKBP4,

		TRIB2, VCL, MACF1, TBCEL, FBLIM1, CEP85L, KLC1, ACTR3B, CORO1B, SKA2, TTBK2, ATXN7, BLOC1S2, MAP7, NPHP4, MAP4, EVI5, MAP9, RAB11FIP4, CEP78, IVNS1ABP, PLK4, TOR1A, PLK2, NEK6, RMDN2, PARVB, SYNJ2, ARHGAP26, SSH2, PTK2, ICK, CYLD, FERMT1, DIAPH1, KIF18A, RGCC, DNMBP, CASS4, CEP85, TUBGCP4, CALM2, FBXL7, SPAG16, CEP57, GPSM2, CLIC4, NUMA1, SH3KBP1, MTCL1, ARPC1A, CLTA, LRMP, CHD3, PPP1R18, SYNE1, TRIM9, KIF13B, KIF1B, IFT57, SPTAN1, BBS4, LRRCC1, ACTR3, PARP3, PDE4D, SHROOM3, ARAP3, HAUS6, KCNAB2, RHOF, CCSAP, CTTNBP2NL, ACTA2, DNM3, LATS2, DNAJC7, KIF2A, RILPL1, CDC42EP4, TCP1, CDC42EP3, EFHC1, FAM96B, EVL, GAPDH, SSX2IP, MAPRE2, MCPH1, DCTN6, RGS14, SRC, DCTN4, NEXN, ADD3, ADD1, FGD2, ADD2, LIMA1, PPP2CB, ABLIM1, TUBA3C, MICALL2, FLNA, MAPK1, STOM, FLNB, RBBP6, CCT8, DISC1, WASF1, WASF3, NPM1, SPAG6, CSNK1A1, ERCC6L2, MID1IP1, PDE4DIP, BAIAP2, PTPN14, CENPF, KLHL5, KIF26B, SLMAP, CCDC6, MPLKIP, CTNNB1, TRIM36, B9D1, PTPN7, ACTR10, KBTBD8, MAD2L1
Chromatin regulator	0.007	PHF20, CHD7, ASH2L, IKZF1, CHD3, BMI1, NR3C1, ING1, HMGN5, ING3, JAK2, JARID2, WHSC1L1, KDM6A, PBRM1, H2AFY, SFMBT1, C17ORF49, EED, SIRT7, EMSY, RNF168, MORF4L1, KAT6B, TBL1XR1, KAT6A, MTF2, WAC, BCORL1, ASF1A, HDAC4, ENY2, DNMT1,

		SMARCD3, HDAC1, RNF8, HDAC9, HDAC6, HDAC7, SMYD2, RBBP7, BRD4, SMARCE1, BRPF1, UTP3, EYA3, CBX2, DEK, ASXL1, GFI1B, KANSL1, FAM175A, PHF13
Lipid metabolism	0.007	ACADVL, PTEN, HDLBP, LIPA, LCLAT1, PRKAG3, FADS2, HPGDS, GM2A, CYP4V2, ENPP2, LDLRAP1, ACAA1, HADH, PTGDS, FADS1, SREBF1, CERS5, PRKAB2, LYPLA1, CPT1A, HMGCS1, PLA2G4A, ACSL5, ELOVL6, CEL, ELOVL7, CYP7B1, PRKAB1, ACLY, PLCB4, MVD, DAGLB, SLC27A3, PAFAH1B3, PLCB2, AGPAT5, SAMD8, ECHS1, SGMS1, HPGD, ABHD2, INSIG1, PLD4, HACD3, PLD1, ACACB, HACD2, AGPAT3, HACD4, PTGS1, UGT8, PTDSS1, PLCG2, SC5D, GPAT3, APOE, KDSR, DECR1, MBTPS1, MID1IP1, PLCL2, SCD5, DDHD2, ERLIN1, AKR1C2, SORL1, SCD, LPCAT3, APOC2, TBXAS1, LPCAT2, SERINC1, DHCR7, LPIN1
Protein transport	7.00E-03	IPO11, DGKD, SYS1, RAB3A, AP1S2, KPNA4, AP1S3, RAB2A, VPS18, GABARAPL2, NUP210, VPS33A, ATG10, TAP2, NUP93, RAB37, CHMP4B, KDELR2, AAGAB, NUP98, PPID, KPNB1, ENY2, NUP205, RAB5B, STXBP2, AGAP1, LIN7A, SNX2, VTI1A, STXBP5, ATG7, SEC31A, SEC16B, VTA1, RANBP17, RAB39B, NUP153, BBIP1, KIF18A, SLC7A6OS, CHMP2B, CNIH4, ARF3, AHCTF1, TRAM1, GPR89A, DENND1A, I 5, CCDC53, AP3S1, TNPO2, TMED5, BBS4,

		TNPO3, RHBDF1, TIMM23, VPS13B, DDX19A, PITPNM1, RILPL1, CHMP3, ATG4A, NUP54, EXOC5, RAB7B, CHMP7, SFT2D1, NUP58, ARF5, STX12, CD63, NBAS, AP5Z1, GOSR2, GOSR1, TGFBRAP1, DERL1, SDAD1, GGA2, AP3M2, VPS50, VPS53, VPS52, SEC23B, RANBP2, ARFGEF1, TIMM8A, KIAA0196, AP3D1, SNX24, XPO7, SNX18, VPS41, SNX15, COPG1, NUPL2, RAB9B, TOMM6
mRNA transport	0.008	RANBP2, ENY2, AHCTF1, NUP205, NCBP1, NUP210, ZC3H3, RANBP17, SRSF1, XPO7, NUP153, UPF3A, NUP93, DDX19A, XPO1, HNRNPA2B1, SRSF3, IGF2BP3, NUP54, NCBP3, NUP98, IGF2BP2, NUPL2, NUP58
Nuclear pore complex	0.01	RANBP2, ENY2, AHCTF1, NUP205, NUP210, RANBP17, XPO7, NUP153, NUP93, DDX19A, NUP54, NUP98, NUPL2, NUP58
Lipid biosynthesis	0.01	AGPAT5, HACD3, HACD2, ACACB, LCLAT1, PRKAG3, AGPAT3, HACD4, PTGS1, FADS2, HPGDS, PTDSS1, SC5D, GPAT3, PTGDS, FADS1, CERS5, PRKAB2, HMGCS1, MID1IP1, SCD5, ELOVL6, ELOVL7, PRKAB1, ACLY, SCD, LPCAT3, TBXAS1, LPCAT2, SERINC1, MVD, DHCR7
mRNA processing	1.60E-02	ADAR, HNRNPR, USP39, AFF2, PNN, SYNCRIP, PAPOLA, MTPAP, BCAS2, RBM5, SRRM2, RBM17, RBFOX2, RBFOX3, NCBP1, SSU72, PLRG1, CDC40, PTBP3, SRRM1, PDE12, SRSF3, SNRPA1, SNRPE,

		NCBP3, SRSF5, SNRNP200, SRSF8, RBM22, SNRPB, KHDRBS1, DDX1, SRSF1, DHX35, NOCT, SNRPB2, SRSF10, SF3B1, SRSF11, SF3A3, RBM38, ZRSR2, RBM39, SF3A1, HNRNPA3, ZRANB2, PRPF38A, CPSF2, LSM1, SCAF1, LSM4, NUDT21, LSM8, CPEB1, HNRNPK, HNRNPA2B1, PRPF31, RNPS1
Electron transport	2.00E-02	NDUFA9, UQCRB, NDUFB5, NDUFA4, ETFDH, CYBB, NDUFC1, ETFA, CYBA, SDHD, HIGD1A, SDHA, SDHB, UQCRH, CYB561D2, ENOX2, FADS2, TMX4, NDUFS8, TMX1, NDUFS2, CYC1, FADS1
Alternative promoter usage	0.021	COL18A1, CREM, NRXN3, NRXN2, ADAR, IL16, ESR1, PTK2, DUSP7, VEGFA, LIMA1, TMLHE, PRDM16, SPEG, SLC12A6, PPARGC1A, SHANK3, TP53, TJP2, LTF
Lipid transport	0.021	OSBPL9, PRELID1, ATP8B2, ATP8A1, OSBPL5, SIGMAR1, STARD10, HDLBP, ATP11C, ATP10A, ANO6, PCTP, ATP11A, SORL1, APOC2, APOC1, ESYT1, LBP, APOE, APOL3, ABCG1
mRNA splicing	0.026	SRSF1, HNRNPR, USP39, AFF2, PNN, SYNCRIP, DHX35, SNRPB2, SRSF10, SF3B1, RBM5, SRSF11, BCAS2, SF3A3, SRRM2, RBM38, ZRSR2, RBM39, RBM17, SF3A1, HNRNPA3, PRPF38A, RBFOX2, ZRANB2, NCBP1, RBFOX3, LSM1, PLRG1, SCAF1, LSM4, CDC40, PTBP3, SRRM1, LSM8, HNRNPK, HNRNPA2B1, SRSF3, PRPF31, SNRPA1, RNPS1, SNRPE, SRSF5, SNRNP200, SRSF8, RBM22, SNRPB

Threonine protease	3.60E-02	PSMA5, PSMA4, PSMA1, PSMA2, PSMB8, PSMB10, PSMA7
Membrane	0.038	CHIC1, TMEM94, ATP5C1, GHITM, TMEM97, NDST2, NDST1, KIAA1109, AKT2, NDST3, ANKFY1, PSMD1, VSTM1, FAM19A5, PRKACB, WLS, C5ORF15, PHKA1, PHKA2, ZFYVE16, RUFY3, ORAI3, AGTR1, ORAI2, ORAI1, VOPP1, CYP2E1, MPPE1, TTYH3, ABCB1, SLCO5A1, THY1, HACD3, SLC5A3, HACD2, KIAA0040, HACD4, UGT8, RHOT1, PRKAR2B, ADAMTS18, RNF217, COX3, VTI1A, LYN, JAG1, CCZ1B, LRBA, ERLIN1, AZU1, NETO2, GRIN2B, FERMT1, QARS, TMEM99, FXYD1, DCHS1, STRN, SFXN1, PIGT, GPR21, TMEM53, PEAR1, SHB, AREG, PANX1, MMP24, TMEM43, JAK2, DLGAP4, PCMTD1, JAK1, CHST6, LYPLA1, CHST7, TEX10, SLC39A11, NIPA1, TMC8, ELOVL6, ELOVL7, EPS15L1, MANEAL, TMC4, SLC39A14, MMP14, PSMA1, MMP15, CDC42EP4, CDC42EP3, TMEM106C, VAMP1, RAPGEF2, HCST, ARF5, TMEM106A, RNASEK, TMF1, SAMD8, MYRF, ATL2, KCNA3, SLC7A1, PTGS1, GGA2, MYCT1, ADAM28, TMEM68, MFN1, MICALL2, STX2, P4HTM, ATAD1, P2RY14, SNX24, KIAA2013, TMEM64, TMCO4, SNX18, SLCO3A1, SLMAP, SNX15, ADGRL1, AGRN, EBAG9, ADGRL3, PDCD6, UBE2D3, CASC4, PLOD3, LAPTM4B, FADS2, HHAT, PIM1, KPNA4, SLC16A9, TCP11, NOMO1, FADS1, BNIP3L, LAG3, TICAM2, VPS33A, STX7, ATP11C, ATP11A, CASD1, SLC9B2, TPST2, SLC9A7,

		MFNG, PKP4, SLC27A3, IL6ST, SPATA9, KHDRBS1, UQCRB, PEF1, KCNE5, ATP6, GPR85, LIN7A, UQCRH, TMEM26, RAP1A, CLTCL1, SLC38A9, KDSR, HTR1F, TNFRSF10C, TNFRSF10B, STK24, TBXAS1, DMXL2, AAK1, LRCH4, LRCH3, TMEM19, GPR89A, MTCL1, ITGB4, FLT4, ITGB3, ITGB2, LRMP, FUNDC2, FSD1L, ITGB6, SMIM14, SLC36A1, CERS5, VPS13C, ITGA2, SIGMAR1, SMIM19, NDUFC1, NRG1, DDOST, ADORA2A, TMBIM1, ITGA6, DSG2, ITGA5, SLC25A12, ST6GALNAC4, GAPDH, SMIM24, PPM1L, SRC, NDUFB5, ATP10A, LRP3, SRI, IL2RG, MFF, LRP8, TSPAN32, LRP6, SLC25A27, SLC25A29, PARD6B, GANAB, NCLN, SLC25A28, KCTD20, SH3BP4, SLITRK5, LRRC8B, DRAM2, SEC23B, KCNJ2, PTPN1, SLC12A4, SLC12A5, NDUFA4, IMMT, EXOG, TM6SF1, BAIAP2, SORL1, MFSD14A, MFSD14C, TSPAN17, SLC25A32, P4HB, RECK, SPG21, MYO1G, EPHB6, AHCYL1, OGFOD3, ICAM3, EFR3A, HERC1, PPAT, PIEZO2, CPNE3, CPNE2, SLC12A6, CALN1, ACVR1, ENTPD1, ACTN1, PAQR3, ANK2, EREG, MFSD10, NINJ2, KDELR2, KCTD12, MXRA7, SPATA18, TP53, TMEM87A, TMEM87B, TBC1D9B, TMEM63A, TGFA, PTTG1IP, LPAR4, TGOLN2, NKAIN2, APH1A, SNX2, APH1B, PGBD5, CD19, ATP5D, ST8SIA4, SARAF, TSPAN5, ARSD, KCNN4, RAB11FIP4, SEC31A, NTNG2, LRRN2, SURF4, RAB27A, PTK6, RAB27B, PILRA, PTK2, NDUFAF6, ARMC10, NDUFAF4,

		ACER3, ECHDC1, CD200, IGHM, TMEM151B, SECTM1, PTPRJ, DOC2B, PCSK7, SLC7A11, IDE, CD3D, HERPUD1, PTPRG, IMMP2L, CD37, MUC15, CD34, BBS4, CD33, CAP1, SREBF1, TMEM150A, STARD3NL, MMD, ANO6, GAB2, COMMD7, LAT2, RNF144B, PITPNM1, PITPNM2, CANX, KIT, MICU3, CRELD1, CD46, CD63, STX12, DERL1, C7ORF13, PGRMC1, PAK1, PGRMC2, CD59, CD58, NPDC1, CD74, MOG, YIPF2, PLEKHA2, ORMDL1, DPY19L1, RIT1, DPY19L2, ZDHHC9, DPY19L4, NEDD4, STK17B, CACHD1, ESYT1, RTN4IP1, CD69, CCDC167, CD84, SYS1, CD82, MPG, MPL, GXYLT2, COX6A2, TMEM263, CLGN, CCND3, PPIP5K1, PALM, ENPP2, IGFLR1, BAALC, ENPP4, RGS8, BORCS7, PTRF, SBF2, BORCS5, KIRREL3, CD96, SLC30A9, SLC30A5, CYBB, CYBA, SLC30A1, SDHD, TTC7A, SDHA, SDHB, LAX1, TANC1, TMX4, TYROBP, CD8B, TMX1, HNRNPUL2, GPD2, SIDT1, VCL, RTN2, FBN2, PLGRKT, COX15, ARHGEF28, NPR3, SLC41A1, HTRA2, PLD4, SLC1A4, SLC1A5, CACNA1C, FUT2, PLD1, HIGD1A, ACACB, FUT4, RTN4, ADGRG1, FUT7, ADGRG5, DHRS7B, GRK6, COX14, CYC1, C16ORF52, PRNP, TCAF1, SYT1, UBE2B, CAV2, VTA1, SLC52A2, CRIM1, NFXL1, KLHL23, NUP153, GCLC, PQLC1, CCPG1, TMEM237, CNTN3, RTP4, GSK3B, MCUR1, PRR7, MS4A3, SIRPB2, ATP2A1, C7ORF73, ADGRE2, GRM2, MAN1A1, SUN2, ARHGEF12, EED, PDPK1, MMP2, PEX2, SSTR2,

		LMBRD1, PEX3, GORASP2, TMEM218, NDUFS2, NUP54, SHANK3, NUP58, AGPAT5, GOSR2, GOSR1, RNF8, PKD1L1, ACVR1B, ADD3, CLCN3, ADD1, PARL, AGPAT3, ADD2, CYTH2, SV2C, CYTH4, RAB40C, SV2A, CLMN, GNG7, PTK2B, TMEM209, PDLIM5, ARFGEF1, MBTPS1, TIMM8A, GALNT3, GALNT2, EGF, ACVR2B, SCIMP, EPOR, ACVR2A, INAFM2, CD7, ACKR3, S100P, PI4K2B, ANKLE1, C2ORF88, MRGPRF, ENO1, SLC4A4, NYNRIN, GJA3, EVA1B, CREB3L2, AP1S2, MS4A14, HOMER3, AP1S3, C2ORF82, VPS18, NUP210, GLCE, PRKCE, COL23A1, FNDC3B, PRKCA, SYTL5, SYTL4, NUP93, ADCY9, SDPR, WDR82, PRKD3, PRKCQ, NUP98, NUP205, RPN1, MTDH, BLOC1S1, BLOC1S2, CHST10, SLC19A3, ST3GAL5, CHST13, ST3GAL6, PROM1, ST3GAL1, SLC19A2, ST3GAL3, PROM2, NDFIP2, NDFIP1, HOMER1, FZD6, STARD10, SLC16A10, BRAF, RMDN2, GP5, SQLE, CYLD, CPEB1, TNFSF4, CDK14, XKR6, DENND1A, CLTA, MS4A4A, CYB561D2, MRC2, UBL3, AIFM1, MTG2, AP3S1, HAVCR2, ANXA1, SSR4, NFAM1, TIMM23, C1ORF43, MS4A6A, MRAS, PLIN3, GART, CCDC90B, ENG, GRAMD1C, GRAMD1B, NBAS, ASAH1, PHLPP1, PLPPR1, SGMS1, PLPPR3, ADCY3, ADCY2, TMEM163, ADCY6, FGD2, RHD, ERMP1, GPR153, RNF139, ALOX5, SC5D, GPAT3, DTD1, RNF130, AOC3, WDR17, KIDINS220, GABRA5, SYNJ2BP, GPR6, SLC8B1, FMO5, TMEM156, RNF145, DHRS9,

		RFWD2, JKAMP, LPCAT3, LPCAT2, LAMTOR2, TMEM39A, LAMTOR3, NUCB2, SLC28A3, DGKG, ANKRD13C, DGKD, ANKRD13D, CSF3R, RAB3A, ATP8A1, PROS1, ITGA2B, TNF, LCLAT1, APMAP, LAMP1, MPC1, CFL1, CLEC5A, PAPOLA, SLC39A6, SLC39A4, RAB2A, MREG, F2R, TPM1, TAP2, VRK1, VRK2, LEMD3, RAB32, ARMCX4, CLIP1, CLEC4D, STIM1, STIM2, C6ORF136, MADD, SLCO2A1, CLIP4, CLEC4E, NOTCH2, RTN4R, RAB5B, SLC26A2, CALCRL, FECH, ABHD2, INSIG1, GLT8D1, ITPR1, ILDR2, ITPR2, ABHD6, ITPR3, TBXA2R, UBR5, PLXNA1, UGT3A1, CD300LF, STRA6, MPDU1, OSBPL9, CD164, SEC16B, CD163, SLC37A1, TOR1A, OSBPL5, ATP8B2, SYNJ2, NBEA, REEP5, CPD, PLXNB2, SERINC1, FAM26F, SLC26A9, DHCR7, CNIH4, SLC48A1, ADPGK, LSP1, SYNE1, PSTPIP2, EFHD2, LBP, GRASP, SLC13A3, TRPC1, KCNAB2, PLRG1, PTP4A3, DNAJC4, LANCL2, PECAM1, DNAJC9, DAGLB, SFT2D1, SLC24A3, EI24, NOD1, LMF1, ENOX2, AP3M2, P2RY8, P2RY2, VPS53, VPS52, P2RY1, SLC13A5, SEMA4A, SLC35A3, SEMA4B, CDC42BPB, TPCN2, P2RX5, P2RX4, CLEC2B, VAPA, VPS41, MGAT4B, EVI2B, VEZT, APP, SLC35B2, AQP6, CISD1, IFIT5, C10ORF128, GOLGA1, DAG1, PTGDS, SLC35A5, ARRDC4, IL15RA, CPT1A, ATP6V0E1, SLC35C2, ACSL5, INPP4B, KIAA0922, CYP2U1, KARS, CD109, VDAC3, VDAC1, EZR, CD101, SLC22A4, SMIM3, MACF1, TMEM161B, SLC22A5,

		DCUN1D5, LMAN2L, STXBP2, SLC35D2, MGST1, MGST2, TMTC2, TEX261, CD99L2, SMIM7, CRHR1, GNG10, PTDSS1, MAP7, MLEC, STXBP5, TRPM5, TRPM6, VAT1, SIGLEC15, TMEM184B, MAGEE1, TMEM50B, TREX1, RAB39B, PARVB, GNG11, UBAC2, CCDC88B, CCDC88A, VNN1, ROBO3, TRAM1, CLIC4, SLC44A1, SH3KBP1, FAM69B, HCCS, MAGT1, SPPL3, SLC22A16, SLC22A18, STAB1, GSG1, UBXN8, CLIC2, RALGPS1, TMED5, RALGPS2, ATP6V0B, PRMT3, ZDHHC13, IL17RE, RHOF, RHOC, GFRA3, CYP7B1, IL17RC, RAB33A, ITFG1, NAALADL1, CEACAM1, CDCP1, CLDN12, CEACAM4, B3GNT2, RDH5, TLR6, RAB7B, ATP6V0C, SPAM1, RHOQ, PTGER1, ZDHHC23, GDPD1, ZDHHC24, AKAP7, NKD1, ZNF7, CUX1, PDPN, SUSD3, GABRE, DISC1, MPZL1, LAIR1, MPP1, LMLN, RANBP2, SLC31A2, SCD5, AP3D1, LILRB2, AIG1, MPP5, PDZK1IP1, SVIP, SPCS2, CAPZA2, F2RL1, CTNNB1, COPG1, NUPL2, RAB9B, SIPA1, TJP2, TOMM6, GALNT12, CIB1, COMT, MANEA, ARHGAP35, TMEM147, IL18RAP, TMEM38B, TMEM14A, TMEM14B, CD300A, LIFR, PDIA6, TMEM134, OXA1L, TMEM135, TMEM136, TBL1XR1, ELMO2, CHMP4B, RELL1, XPR1, ABCG1, ABCG2, PXK, CHRNA6, CLEC16A, ADRB2, TMEM121, TMEM123, HRH2, CDH26, ASIC1, SPECC1, BTNL8, MLKL, RHBDD2, UQCC2, SMURF1, RHBDD1, PTCH1, EPHX1, ARHGAP27, ESR1, SELP, ARHGAP32, DIAPH1, MAVS,

		SELL, DLG4, DNAJC10, CHMP2B, FGFR2, LRP12, FGFR1, B4GALT2, GPSM1, DOCK5, ACADVL, ABCD3, USP32, DOCK9, PDE3B, SLC40A1, SLC2A5, MFSD6, SPRED1, CYP4V2, TMEM107, FLOT1, TMEM109, MFSD1, RHBDF1, HGSNAT, PDE4D, PLAUR, ARAP3, TGFBR2, DDX19A, RASA3, IGDCC4, CHMP3, DOCK1, HCN3, VAC14, LTK, RGS14, ATF6B, TM2D2, PALM2, POMP, ETFDH, NRXN3, NRXN2, TFPI, FAM169A, RPS6KC1, LEPROTL1, FAM120A, KLRG2, STOM, CYYR1, CAMK2G, GBP4, SNCA, KLRG1, ZNF286A, CMTM7, CMTM8, CMTM6, GDAP1, TSHR, SCD, PMP22, SPRY1, LPIN1, SSBP4
Innate immunity	0.04	ANKRD17, CFD, CD84, C1R, IFIT5, NOD1, ADAR, GATA3, ZC3HAV1, CFP, DDX60, IFIH1, RNF135, TBK1, SEC14L1, C1RL, CLEC5A, LBP, JAK2, HAVCR2, KLRG1, MAP3K5, MB21D1, CHID1, LYN, AKIRIN2, TICAM2, ANXA1, APOBEC3G, RIPK2, CARD9, EIF2AK2, MIF, CYLD, MAVS, CLEC4D, AIM2, SLPI, POLR3D, F2RL1, IRF5, TLR6, TRIM38, CD46, TRIM56
Amino-acid biosynthesis	0.043	GOT1, APIP, ASNS, PYCR2, ASL, ENOPH1, PSPH, ASS1
Mitochondrion outer membrane	0.051	VAT1, PRNP, BNIP3L, CPT1A, SLC44A1, VPS13C, MGST1, CISD1, ACSL5, SYNJ2BP, MFF, GDAP1, RAB32, MAVS, RHOT1, MFN1, VDAC3, VAMP1, VDAC1, RTN4IP1, SPATA18, TOMM6

Activator	0.054	CCNC, IKZF1, AHR, IKZF2, RBPJ, ELK3, YY1, ELK4, UXT, PNN, CHCHD2, MYC, CREB3L2, MEF2C, ZBTB38, ARID5B, EBF3, ZBTB33, EBF4, MED4, TBL1XR1, KAT6B, KAT6A, RFX5, TFAM, TP53, NOTCH2, ENY2, DNMT1, DDX1, GLIS3, CREM, FOXO6, TWIST1, GATA3, GATA1, MED30, PRDM16, RREB1, PPARGC1A, ZFHX3, JUND, BRPF1, RBPMS, XRCC5, ZNF260, NFATC3, NFATC2, NR1D2, EIF2AK4, TRERF1, ESR1, PBX1, CPEB1, HNRNPK, SP3, TCF4, BRWD1, ZNF496, CITED2, ATF7IP2, SRA1, MLX, DNMT3B, MYBL1, NCOA2, SREBF1, PDPK1, VEZF1, MED27, MED26, MED29, MN1, ELF2, ELF4, MED21, MED20, LCOR, PPARA, ATF6B, MYRF, CEBPE, MLLT1, MLLT3, RELA, NRIP1, TP53INP1, PLAGL2, PCBD1, TBL1X, MEF2D, RBM39, EYA3, ATAD2, PNRC2, KLF6, MEIS1, GFI1B, NFIA, CTNNB1, MNDA
Protein phosphatase	0.057	PHLPP1, PPM1L, PTEN, PTPRJ, RPAP2, CDC14A, PTPRG, PPP2CB, PPP3CC, PTPN1, PTPN18, DUSP2, DUSP3, EYA3, SSU72, DUSP28, PTPN11, SSH2, PTPN14, DUSP7, DUSP22, PTP4A3, UBLCP1, DNAJC6, PTPN7
Lipoprotein	0.06	CHIC1, RAB3A, C2ORF88, HDLBP, SLA, CIB1, TNF, EFR3A, HHAT, PALM, BAALC, PRKACB, BORCS5, RAB2A, GABARAPL2, ENTPD1, TICAM2, PHKA1, PHKA2, PRKAB1, RAB32, RAB37, CD109, AGTR1, ABCG1, VC B, TGFA, ADRB2, THY1, PLD1, GNG10,

		RAP1A, HRH2, GRK6, APOE, SPECC1, LYN, NTNG2, PRNP, APOBEC3G, SYT1, TNFRSF10C, RAB39B, RAB27A, RAB27B, ESR1, GNG11, SELP, VNN1, DLG4, CPD, NDUFAF4, FXYD1, SERINC1, CNTN3, ARF3, NAA40, SLC44A1, ITGB4, USP32, HCCS, PCSK7, PCMTD2, ARL5B, UBL3, RAC2, LDLRAP1, PCMTD1, PLAUR, ZDHHC13, RHOF, SSTR2, GFRA3, RHOC, RAB33A, LAT2, DUSP22, IL1A, PTP4A3, MRAS, GORASP2, CANX, PECAM1, LANCL2, TMEM106C, CHMP3, ITGA6, RAB7B, SPAM1, ARF5, RHOQ, CD63, SRC, PALM2, KCNA3, ZDHHC23, ZDHHC24, AKAP7, TFPI, LRP8, NKD1, LRP6, WNT11, RAB40C, GNG7, STOM, CD59, CD58, KCNJ2, APOL3, MPP1, NAP1L1, GPR6, SVIP, SCIMP, SOD1, CENPF, ZDHHC9, APOC2, APOC1, CD7, F2RL1, GHRL, RECK, RAB9B
Prenylation	0.065	RAB5B, RAB3A, USP32, PALM2, GNG10, RAP1A, UBL3, RAB40C, GNG7, PALM, RAC2, RAB2A, RAB39B, RAB27A, NAP1L1, PHKA1, RAB27B, RHOF, RHOC, PHKA2, GNG11, RAB33A, RAB32, PTP4A3, CENPF, MRAS, RAB37, RAB7B, RAB9B, RHOQ
Initiation factor	0.066	EIF1AX, MTIF2, EIF3K, EIF4H, EIF4EBP2, DHX29, EIF3C, EIF4E2, EIF4G3, EIF1B, EIF4E, EIF3A, EIF3B
Oxidoreductase	0.068	ACADVL, HIBADH, PYROXD1, OGFOD3, LOXL3, PLOD3, LOXL4, AKR1B1, OGDHL, ALKBH7, EPX, MPO, FADS2, SPR, AIFM1, CYP4V2, SESN1, PHYHD1, DUS4L, HADH, FADS1, HTATIP2, IDH3A, KDM6A,

		BCKDHA, CBR1, HSDL2, CYBB, P3H1, CYBA, KCNAB2, CYP7B1, SDHA, SDHB, HPDL, ALDH5A1, NDUFS8, LOX, CYP2U1, GPD2, DPYD, CAT, ALDH1A1, BLVRB, RDH5, NDUFS2, CYP2E1, GAPDH, DLD, FXN, HPGD, ETFDH, PTGS1, ENOX2, CYB5R4, PRDX4, PRDX1, ALOX5, DHRS7B, LDHD, PNPO, SC5D, KDSR, PRODH, P4HTM, DECR1, AOC3, VAT1, EGLN2, SCD5, AKR1C2, PYCR2, SOD2, FMO5, GFER, SOD1, UEVLD, SQLE, GLUD1, ALDH6A1, TMLHE, P4HA1, DHRS9, SCD, TBXAS1, DNAJC10, PIR, P4HB, DHCR7
Translocation	7.20E-02	RANBP2, TIMM8A, TRAM1, ENY2, AHCTF1, NUP205, NUP210, RANBP17, TIMM23, XPO7, NUP153, NUP93, DDX19A, NUP54, NUP98, NUPL2, NUP58
Chemotaxis	7.30E-02	CMTM7, CCL13, ROBO3, CMTM8, CXCL8, PLGRKT, CCL20, CMTM6, IL16, PPBP, AZU1, CXCL3, RNASE2, CXCL2, TYMP, ENPP2, LGALS9, CCL28, PF4
Transcription	0.091	MYT1L, GFI1, CCNC, IKZF1, ENO1, IKZF2, RBPJ, ELK3, UXT, ELK4, PNN, HOXA9, ZMIZ1, MYC, ZMIZ2, CREB3L2, HOXA3, BBX, HOXA1, CCNL2, HOXA7, TRIM22, HOXA5, MEF2C, CSNK2A1, SFMBT1, SP110, DNTTIP1, EBF3, EBF4, ZNF12, MED4, SND1, EID1, ZSCAN20, RFX5, HOXB4, HOXB3, VOPP1, TFAM, HOXB2, TRIB3, ZSCAN25, BCORL1, TP53, ASF1A, ZNF277, ZNF275, ZBTB8A, HKR1, TSHZ3, ANP32A,

		MYCBP2, CDCA7, TWIST1, WDR61, GATA3, NLK, GATA1, RPAP2, ZNF709, COMMD8, ZNF827, PRDM16, HIVEP3, RREB1, PPARGC1A, ZBTB7A, HEXIM1, ZNF264, ZNF263, JUND, MTERF3, XRCC5, ZNF260, NFATC3, NFATC2, EIF2AK2, NR1D2, PARP14, TRERF1, PBX1, POU6F1, HNRNPK, ZZZ3, AGO2, ID1, HOXD8, ZNF256, ZNF133, ZNF496, ZNF253, ZNF251, PHF20, CITED2, ATF7IP2, SRA1, ASH2L, BMI1, TFB1M, SCML2, ING3, SCML4, ZNF41, MLX, ZNF248, JARID2, ZIK1, SREBF1, TSC22D4, NCOA2, ZNF121, HELLS, PBRM1, TSC22D1, COMMD2, COMMD3, PAWR, IL16, PPHLN1, SIRT7, MED27, MED26, COMMD7, MED29, MN1, TAL1, MED21, MED20, WAC, PPARA, KANK2, FOXC1, MYCBP, TMF1, MYRF, CIPC, ZNF518A, ZNF8, RELA, ZNF7, PURB, CUX1, TCEAL1, PLAGL2, ZNF107, ZNF227, TCEB1, TBL1X, TAF9B, MEF2D, BRD4, ZNF222, ZNF462, MYEF2, PCGF5, ATAD2, SAMD4B, ELP4, HIPK1, HIPK2, ZNF33B, GFI1B, NR6A1, NFIA, TCEA2, TCEA1, CTNNB1, ZNF577, ZBTB26, AHR, NR3C1, ETS2, ARHGAP35, HOXA10, YY1, CHCHD2, HMGN5, ZNF329, ZNF688, ZNF567, MTPAP, PTRF, WHSC1L1, GTF2I, ZNF562, TLE3, SLC30A9, ZNF681, TLE1, ZBTB38, ZBTB37, KCTD1, ARID5B, ZBTB33, EMSY, SAP30, MORF4L1, ILF3, TBL1XR1, KAT6B, KAT6A, ZNF318, ERF, ZNF438, ZNF557, ADNP, LRIF1, ZNF555, NOTCH2, ZNF671, ZNF550, KHDRBS1, ENY2, DNMT1, DDX1, GLIS3, CREM,

[296]

		FOXO6, GTF2E2, NPAS3, ARHGAP22, MED30, ATXN7, DEAF1, SMYD2, ZNF548, ZNF667, ZNF302, ZNF786, ZNF784, ZBTB18, AKIRIN2, ZFHX3, BRPF1, SMAD3, RBPMS, PHF10, CBX2, HMGA2, CNOT10, FOXN3, FOXN2, ESR1, HOPX, DNAJC17, CDK9, CNOT7, ZNF418, SP3, PIR, ZNF778, ZNF777, TCF4, ZNF655, CNOT8, BRWD1, ZNF654, FGFR1, MXD4, RERE, CHD7, FHL2, CHD3, LYL1, SIX4, ZNF529, TRPS1, SIX3, IFT57, YAF2, MYBL1, TGIF1, EED, PDPK1, VEZF1, OLIG2, ELF2, ZNF519, ELF4, POLR1B, ZNF516, POLR1D, LCOR, IRF5, ZNF511, LTF, HDAC4, SMARCD3, ATF6B, CEBPE, ZBTB46, HDAC1, MLLT1, MLLT3, CXXC5, HDAC9, HDAC6, HDAC7, CAND2, TAF1D, ZNF629, ZNF507, TGS1, NRIP1, TP53INP1, MXI1, POLR2E, MAPK1, PCBD1, RBBP7, ZNF500, ZNF189, BEND6, ZNF286A, TBX1, RBM39, IRX5, ZNF184, ASCC1, EYA3, PTPN14, PNRC2, ASXL1, MEIS1, KLF6, POLR3D, ZNF616, ESF1, MNDA, LPIN1, ZNF610, ZNF850
Cell cycle	0.092	RIF1, ZAK, DSCC1, GMNN, BCCIP, CIB1, SMC5, IKZF1, AHR, NR3C1, SMC3, SMC4, CDC14A, SMC2, CCND3, CHEK1, PIM1, MAP3K8, NBN, LZTS2, CSNK2A1, PRKCE, NSMCE2, VRK1, LIG3, CDC40, CD2AP, EID1, ZFYVE19, PSME3, SPECC1L, ERH, TP53, KHDRBS1, BEX2, CDCA2, ANAPC16, CDCA7, PDS5B, GSPT2, SKA2, SYCE1, RAD21, EVI5, MAP9, NEK9, PRNP, NEK6, SIAH1, HMGA2, CDC7, FOXN3, EIF2AK4, ZC3HC1,

		CYLD, RGCC, CCPG1, PHF13, CDK14, FBXL7, PRR5, AHCTF1, NUMA1, CLTA, USP39, ING1, MRPL41, LRRCC1, HELLS, HAUS6, CCSAP, MLF1, CCNA1, LATS2, KIF2A, MCM4, CHMP3, MCM5, MCM6, MAPRE2, RGS14, SRC, RNF8, CCNDBP1, PARD6B, RBBP8, MAPK1, LMLN, RBM38, CDK11A, CSNK1A1, CENPF, RAD50, SNX18, MIS18A, MPLKIP, MAD2L1
DNA repair	9.60E-02	SWSAP1, ANKLE1, PARPBP, DCLRE1C, PSMD14, AP5Z1, PRKDC, MPG, MCM9, UBE2D3, RNF8, BCCIP, SMC5, SMC3, YY1, RAD21, CHEK1, UBR5, USP1, RBBP8, POLI, NBN, BARD1, UBE2B, ERCC6L2, FANCL, XRCC5, EYA3, NSMCE2, NSMCE1, LIG3, RAD23A, RAD23B, EMSY, CDK9, MSH6, RNF168, MORF4L1, RAD50, NEIL2, NSMCE4A, RPA3, PSME4, RPA4, FAM175A, NABP1, OTUB1
Immunity	0.097	CD84, DCLRE1C, IFIT5, LRMP, ADAR, IFIH1, TBK1, NUDCD1, CLEC5A, C1RL, VSTM1, MAP3K8, LBP, LGALS9, JAK2, HAVCR2, MB21D1, MAP3K5, CHID1, ANXA1, TICAM2, CD300A, RIPK2, PRKCE, TAP2, MIF, LAX1, LAT2, CLEC4D, AIM2, CD8B, PSMA1, SLPI, ORAI1, IRF5, PRKCQ, TLR6, CLEC4E, CD46, TRIM56, OTUB1, LTF, CFD, ANKRD17, SRC, C1R, NOD1, GATA3, ZC3HAV1, DDX60, CFP, RNF135, SEC14L1, PTK2B, CD300LF, KCNN4, LAIR1, KLRG1, LYN, BTNL8, CD74, AKIRIN2, SEMA4A, APOBEC3G, CARD9, EIF2AK2, LILRB2, EIF2AK4, PSMB8, CYLD, MAVS, POLR3D, CD7, F2RL1, TRIM38, MYO1G

Gene Ontology enrichment analysis-KEGG terms

This table complements Figure 4.28, by linking the enriched KEGG terms (signalling and metabolic pathways), the penalised Fisher's probability score and the genes in the ssCCA results that belong in each KEGG term.

KEGG term	Penalised Fisher's	genes
KEGG 04611: Platelet activation	2.33E-05	SRC, ITGB3, ITGA2B, ITPR1, ADCY3, PIK3R3, ITPR2, ADCY2, ITPR3, PIK3CB, ARHGAP35, ADCY6, MYLK, PTGS1, RAP1A, TBXA2R, AKT2, P2RY1, PLCG2, MAPK1, PRKACB, LYN, ARHGEF12, COL27A1, COL24A1, VWF, ITGA2, F2R, PLA2G4A, GP5, COL2A1, PLCB4, ADCY9, STIM1, COL5A1, TBXAS1, ORAI1, PLCB2, FERMT3
KEGG 04510: Focal adhesion	1.45E-04	GSK3B, LAMC3, ITGB4, FLT4, ITGB3, RASGRF1, ITGA2B, PTEN, PIK3CB, ARHGAP35, MYLK, CRKL, CCND3, TNN, AKT2, RAC2, ITGB6, COL27A1, VWF, PDPK1, HGF, ACTN1, ITGA2, PRKCA, COL2A1, COL4A2, COL4A1, RAPGEF1, ITGA6, ITGA5, DOCK1, VCL, SRC, PIK3R3, MAPK9, PAK1, RAP1A, PDGFD, FLNA, MAPK1, FLNB, COL24A1, EGF, CAV2, BRAF, PARVB, PTK2, VEGFA, DIAPH1, COL5A1, CTNNB1, GRB2
KEGG 01200: Carbon metabolism	0.002	GPI, PRPS1, FH, ECHS1, ADPGK, SHMT1, RPE, OGDHL, ENO1, SDSL, ESD, ACSS1, PSPH, IDH3A, TKTL2, GOT1, GPT2, TALDO1, SDHD, SDHA, SDHB, GLUD1, ALDH6A1, CAT, PCCB, SUCLG1, DLD, FBP1, PFKM, GAPDH

KEGG 04015: Rap1 signaling pathway	0.003	RGS14, SRC, FLT4, ITGB3, ITGA2B, ITGB2, ADCY3, PIK3R3, ADCY2, LPAR4, PIK3CB, FYB, ADCY6, CRKL, PARD6B, RAP1A, AKT2, PDGFD, P2RY1, RAC2, MAPK1, ANGPT2, ANGPT1, EGF, HGF, F2R, ARAP3, PRKCA, BRAF, GRIN2B, VEGFA, PLCB4, MRAS, ADCY9, ADORA2A, PRKD3, ID1, KIT, RAPGEF1, CTNNB1, RAPGEF2, FGF13, PLCB2, CALM2, FGF11, FGFR2, SIPA1, FGFR1
KEGG 03050: Proteasome	0.004	PSMD14, POMP, PSMB8, PSMB10, PSMA7, PSMA5, PSMD6, PSMD7, PSMA4, PSMA1, PSMA2, PSME3, PSME4, PSME1, PSMD1
KEGG 05205: Proteoglycans in cancer	0.005	CD63, SRC, ITGB3, ITPR1, PIK3R3, TWIST1, ITPR2, ITPR3, PIK3CB, TNF, PAK1, WNT11, MYC, AKT2, PLCG2, FLNA, MAPK1, TIMP3, FLNB, CAMK2G, PRKACB, TGFB1, ARHGEF12, CAV2, PDPK1, HGF, MMP2, ITGA2, PTCH1, FZD6, PLAUR, PRKCA, PTPN11, ANK2, BRAF, ESR1, PTK2, VEGFA, MRAS, CTNNB1, GRB2, ITGA5, EZR, TP53, FGFR1
KEGG 04142: Lysosome	0.006	CD63, ASAH1, HEXA, CLTA, LIPA, AP3M2, GGA2, LAPTM4B, CTSO, GM2A, LAMP1, CLTCL1, AP1S2, ATP6V1H, AP3S1, AP1S3, AGA, GUSB, ARSB, CTSC, HGSNAT, CD164, ATP6V0B, AP3D1, GALNS, GLB1, MAN2B1, TPP1, DNASE2, ATP6V0C
KEGG 04912: GnRH signaling pathway	0.007	MAP3K1, SRC, MMP2, ITPR1, PLA2G4A, ADCY3, ITPR2, PRKCA, ADCY2, ITPR3, CACNA1C, PLD1, ADCY6, MAPK9, MMP14, ADCY9, PLCB4, PTK2B, MAPK1, GRB2, CALM2, CAMK2G, PLCB2, PRKACB

KEGG 04750: Inflammatory mediator regulation of TRP channels	0.009	SRC, ITPR1, ADCY3, PIK3R3, ITPR2, ADCY2, ITPR3, PIK3CB, ADCY6, MAPK9, P2RY2, PLCG2, CAMK2G, PRKACB, ASIC1, PRKCH, PRKCE, PLA2G4A, PRKCA, PLCB4, ADCY9, F2RL1, PRKCQ, PLCB2, CALM2
KEGG 04010: MAPK signaling pathway	0.010	ZAK, RASGRF1, TNF, CRKL, ELK4, PPP3CC, MYC, AKT2, RAC2, MAP3K8, PRKACB, MAP3K5, DUSP2, MEF2C, DUSP3, PLA2G4A, PRKCA, TGFBR2, DUSP7, IL1A, CACNB3, CACNB4, MRAS, RASA1, RAPGEF2, TP53, CACNA1C, NLK, RELA, MAPK9, PAK1, RAP1A, MKNK1, FLNA, MAPK1, FLNB, MAP4K3, MAP2K5, MAP4K4, TGFB1, JUND, MAP3K1, GADD45B, EGF, NFATC3, BRAF, GRB2, PTPN7, FGF13, LAMTOR3, FGF11, FGFR2, FGFR1
KEGG 04141: Protein processing in endoplasmic reticulum	0.010	TRAM1, UBE2D4, ATF6B, PRKCSH, UBE2D3, RPN1, HSPA4L, DERL1, HSP90B1, HERPUD1, MAPK9, GANAB, OS9, BAG2, MAN1A1, SEC23B, SEC31A, SKP1, MAP3K5, PDIA3, MBTPS1, SSR4, HSPA5, EIF2AK2, YOD1, RAD23A, EIF2AK4, PDIA6, SVIP, RAD23B, DDOST, PDIA4, DNAJB11, DNAJC10, CANX, HYOU1, P4HB, CALR
KEGG 04925: Aldosterone synthesis and secretion	0.015	ATF6B, PRKCE, ITPR1, ADCY3, ITPR2, PRKCA, ADCY2, ITPR3, CACNA1C, ADCY6, ADCY9, PLCB4, PRKD3, CREB3L2, AGTR1, ORAI1, DAGLB, CALM2, CAMK2G, PLCB2, PRKACB

KEGG 01040: Biosynthesis of unsaturated fatty acids	0.015	FADS2, SCD, SCD5, ELOVL6, HACD3, ACAA1, HACD2, FADS1, HACD4
KEGG 00531: Glycosaminoglycan degradation	0.016	GALNS, HGSNAT, GLB1, HEXA, HYAL3, GUSB, SPAM1, ARSB
KEGG 04012: ErbB signaling pathway	0.017	GSK3B, SRC, EGF, PIK3R3, TGFA, BRAF, PRKCA, NRG1, PIK3CB, AREG, PTK2, EREG, CRKL, MAPK9, PAK1, MYC, AKT2, PLCG2, MAPK1, GRB2, CAMK2G, NCK1
KEGG 04922: Glucagon signaling pathway	0.020	PRKAB2, CPT1A, PDE3B, ITPR1, ITPR2, PHKA1, ADCY2, ITPR3, ACACB, PHKA2, PRKAB1, PRKAG3, PLCB4, PPP3CC, AKT2, PHKG2, CREB3L2, SIK2, PPARA, CALM2, PPARGC1A, CAMK2G, PLCB2, PRKACB
KEGG 04910: Insulin signaling pathway	0.021	GSK3B, PDE3B, PIK3R3, PIK3CB, ACACB, PRKAG3, CRKL, MAPK9, PRKAR2B, AKT2, MKNK1, PHKG2, FLOT1, MAPK1, PPARGC1A, EIF4E, PRKACB, SREBF1, PTPN1, PRKAB2, PDPK1, PHKA1, BRAF, PHKA2, PRKAB1, RAPGEF1, GRB2, EIF4E2, FBP1, CALM2, RHOQ
KEGG 01212: Fatty acid metabolism	0.022	ACADVL, CPT1A, ECHS1, SCD5, ACSL5, ELOVL6, HACD3, HACD2, HACD4, FADS2, SCD, HADH, ACAA1, FADS1

KEGG 04151: PI3K-Akt signaling pathway	0.028	GSK3B, CSF3R, LAMC3, ITGB4, FLT4, ITGB3, ITGA2B, PTEN, PPP2R2A, PIK3CB, CCND3, PPP2R1B, YWHAQ, TNN, MYC, AKT2, CREB3L2, ITGB6, JAK2, JAK1, COL27A1, VWF, PDPK1, HGF, ITGA2, F2R, PRKCA, PPP2R5C, COL2A1, COL4A2, COL4A1, DDIT4, KIT, ITGA6, ITGA5, EIF4E2, TP53, PHLPP1, ATF6B, PIK3R3, LPAR4, IL2RG, RELA, HSP90B1, PPP2CB, GNG10, CD19, PDGFD, GNG7, MAPK1, EIF4E, ANGPT2, COL24A1, ANGPT1, EGF, GNG11, EPOR, PTK2, VEGFA, COL5A1, IL7, GRB2, FGF13, FGF11, FGFR2, FGFR1
KEGG 04512: ECM-receptor interaction	0.031	COL27A1, COL24A1, VWF, ITGB4, LAMC3, ITGB3, ITGA2, ITGA2B, GP5, COL2A1, SV2C, COL4A2, COL5A1, COL4A1, SV2A, TNN, DAG1, ITGA6, ITGB6, ITGA5, AGRN
KEGG 03030: DNA replication	0.035	RFC5, POLA2, RNASEH2B, RNASEH1, RFC2, RPA3, RPA4, MCM4, MCM5, SSBP1, MCM6
KEGG 04350: TGF-beta signaling pathway	0.041	TGIF1, ACVR1, TGFB1, SMAD3, SMURF1, FST, INHBA, ACVR1B, TNF, ACVR2B, BMP6, ACVR2A, TGFBR2, PPP2CB, ZFYVE16, PPP2R1B, MYC, ID1, MAPK1, SKP1
KEGG 04152: AMPK signaling pathway	0.042	CAB39, PIK3R3, PPP2R2A, PIK3CB, ELAVL1, ACACB, PRKAG3, CAMKK2, PPP2CB, PPP2R1B, AKT2, CREB3L2, PPARGC1A, RAB2A, SREBF1, CPT1A, PRKAB2, PDPK1, SCD5, PPP2R5C, PRKAB1, CCNA1, TBC1D1, SCD, LEP, FBP1, PFKM

KEGG 04550: Signaling pathways regulating pluripotency of stem cells	0.042	GSK3B, RIF1, PIK3R3, PIK3CB, BMI1, ACVR1B, WNT11, MYC, AKT2, MAPK1, JAK2, JARID2, JAK1, ACVR1, ZFHX3, SMAD3, PCGF5, FZD6, LIFR, INHBA, ACVR2B, ACVR2A, MEIS1, KAT6A, ID1, CTNNB1, GRB2, IL6ST, FGFR2, FGFR1
KEGG 04066: HIF-1 signaling pathway	0.046	EGLN2, ANGPT2, ANGPT1, EGF, SERPINE1, PIK3R3, PRKCA, ENO1, PIK3CB, RELA, VEGFA, TF, AKT2, MKNK1, PLCG2, MAPK1, TCEB1, TIMP1, EIF4E2, GAPDH, CAMK2G, EIF4E
KEGG 04920: Adipocytokine signaling pathway	0.053	PRKAB2, CPT1A, ACSL5, PTPN11, ACACB, TNF, PRKAB1, PRKAG3, RELA, CAMKK2, MAPK9, AKT2, LEP, PRKCQ, JAK2, PPARA, PPARGC1A
KEGG 04640: Hematopoietic cell lineage	0.056	CSF3R, ITGB3, ITGA2, ITGA2B, CD3D, TNF, GP5, EPOR, IL1A, CD8B, IL7, CD19, KIT, CD7, CD59, ITGA6, CD37, ITGA5, CD34, CD33
KEGG 01230: Biosynthesis of amino acids	0.066	TKTL2, PRPS1, ARG2, GOT1, GPT2, SHMT1, RPE, TALDO1, PYCR2, ENO1, SDSL, ASS1, ASL, GAPDH, PFKM, PSPH, IDH3A

KEGG 00220: Arginine biosynthesis	0.068	GLUD1, ARG2, GOT1, GPT2, ASL, ASS1, GLS
KEGG 04144: Endocytosis	0.068	ARF3, RAB5B, SRC, SH3KBP1, WIPF2, ARPC1A, ARPC5L, NEDD4L, AGAP1, CLTA, PLD1, IL2RG, CYTH2, PARD6B, SNX2, CYTH4, KIF5C, CLTCL1, GRK6, CCDC53, LDLRAP1, RAB11FIP4, ARFGEF1, SH3GLB2, GIT2, SMAD3, SMURF1, CAV2, VTA1, KIAA0196, ARAP3, EPS15L1, TGFBR2, DNM3, ZFYVE16, BIN1, DNAJC6, NEDD4, CHMP2B, CAPZA2, CHMP3, CHMP4B, CHMP7, FGFR2, SPG21, ARF5
KEGG 00020: Citrate cycle (TCA cycle)	0.070	ACLY, FH, SUCLG1, OGDHL, SDHD, SDHA, SDHB, DLD, IDH3A
KEGG 04666: Fc gamma R-mediated phagocytosis	0.072	LYN, PRKCE, ARPC1A, ARPC5L, PIK3R3, PRKCA, PIK3CB, GAB2, PLD1, CRKL, PAK1, BIN1, AKT2, CFL1, PLCG2, RAC2, MAPK1, WASF1, WASF3
KEGG 04071: Sphingolipid signaling pathway	0.084	ASAH1, SGMS1, PTEN, PIK3R3, PPP2R2A, PIK3CB, PLD1, TNF, RELA, PPP2CB, MAPK9, PPP2R1B, AKT2, RAC2, MAPK1, MAP3K5, CERS5, PDPK1, PRKCE, PRKCA, PPP2R5C, GAB2, PLCB4, PLCB2, TP53

KEGG 04150: mTOR signaling pathway	0.087	CAB39, PDPK1, PTEN, PIK3R3, BRAF, PRKCA, PIK3CB, TNF, RRAGA, AKT2, DDIT4, MAPK1, EIF4E2, EIF4E
KEGG 04921: Oxytocin signaling pathway	0.088	SRC, ITPR1, ADCY3, ITPR2, ADCY2, ITPR3, CACNA1C, PRKAG3, ADCY6, CAMKK2, MYLK, PPP3CC, MAPK1, CAMK2G, PRKACB, MAP2K5, KCNJ2, MEF2C, PRKAB2, NFATC3, PLA2G4A, NFATC2, PRKCA, PRKAB1, CACNB3, CACNB4, PLCB4, ADCY9, PLCB2, CALM2
KEGG 04062: Chemokine signaling pathway	0.090	CCL13, GSK3B, CXCL8, SRC, ADCY3, PIK3R3, ADCY2, PIK3CB, CXCL3, CXCL2, RELA, ADCY6, CRKL, PAK1, GNG10, RAP1A, AKT2, GNG7, GRK6, RAC2, PTK2B, MAPK1, JAK2, PRKACB, LYN, CCL20, BRAF, PPBP, GNG11, PTK2, PLCB4, ADCY9, GRB2, PLCB2, CCL28, PF4
KEGG 03040: Spliceosome	0.092	SRSF1, USP39, SNRPB2, SRSF10, SF3B1, BCAS2, SF3A3, RBM17, SF3A1, HNRNPA3, PRPF38A, NCBP1, PLRG1, LSM4, CDC40, LSM8, HNRNPK, SRSF3, PRPF31, SNRPA1, SNRPE, SRSF5, SNRPC, SNRNP200, SRSF8, RBM22, SNRPB

APPENDIX III

Metabolomics in systems medicine: an overview of methods and applications

Effrosyni Karakitsou, Carles Foguet, Pedro de Atauri, Kim Kultima, Payam Emami Khoonsari, Vitor A.P. Martins dos Santos, Edoardo Saccenti, Antonio Rosato, Marta Cascante.

Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Systems Biology

Metabolomics in systems medicine: an overview of methods and applications

Effrosyni Karakitsou[1,2,†], Carles Foguet[1,3,†], Pedro de Atauri[1,3], Kim Kultima[4], Payam Emami Khoonsari[4], Vitor A. P. Martins dos Santos[5,6], Edoardo Saccenti[5], Antonio Rosato[7] and Marta Cascante[1,3]

Abstract

Patient-derived metabolomics offers valuable insights into the metabolic phenotype underlying diseases with a strong metabolic component. Thus, these data sets will be pivotal to the implementation of personalized medicine strategies in health and disease. However, to take full advantage of such data sets, they must be integrated with other omics within a coherent pathophysiological framework to enable improved diagnostics, to identify therapeutic interventions, and to accurately stratify patients. Herein, we provide an overview of the state-of-the-art data analysis and modeling approaches applicable to metabolomics data and of their potential for systems medicine.

Addresses

[1] Department of Biochemistry and Molecular Biomedicine and Institute of Biomedicine (IBUB), Faculty of Biology, Universitat de Barcelona (UB), Barcelona, Spain

[2] Institute of Cancer and Genomic Sciences and Centre for Computational Biology, University of Birmingham, B15 2TT, Birmingham, UK

[3] Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBEREHD), Metabolomics Node at INB-Bioinformatics Platform, Instituto de Salud Carlos III (ISCIII), Madrid, Spain

[4] Department of Medical Sciences, Clinical Chemistry, Uppsala University, Uppsala, Sweden

[5] Laboratory of Systems and Synthetic Biology, Wageningen University & Research, Stippeneng 4, 6708WE, Wageningen, the Netherlands

[6] LifeGlimmer GmbH, Markelstraße 38, 12163 Berlin, Germany

[7] Department of Chemistry and Magnetic Resonance Center, University of Florence, Italy

Corresponding authors: Saccenti, Edoardo (edoardo.saccenti@wur.nl); Rosato, Antonio (rosato@cerm.unifi.it); Cascante, Marta (martacascante@ub.edu)

[†] These authors contributed equally to this work.

Current Opinion in Systems Biology 2019, **15**:91–99

This reviews comes from a themed issue on **Gene regulation**

Edited by **Mariko Okada** and **Shinya Kuroda**

For a complete overview see the Issue and the Editorial

Available online 29 March 2019

https://doi.org/10.1016/j.coisb.2019.03.009

Keywords

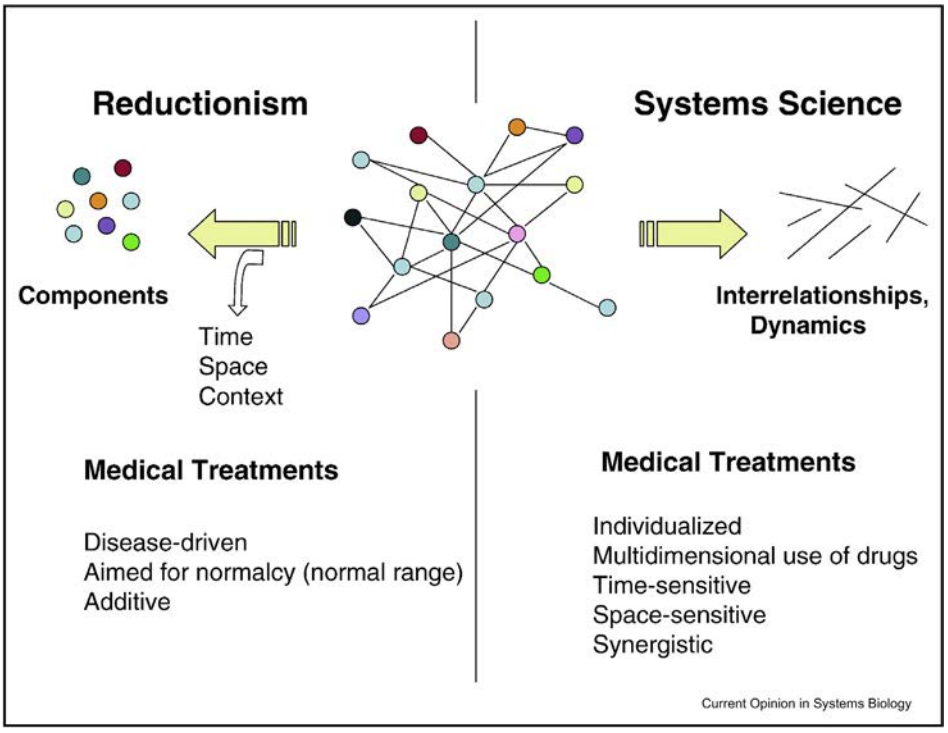
Systems medicine, Metabolomics, Multiomics, Kinetic modellig, Constraint-based modelling, Personalized medicine.

The brave new world of systems medicine

Systems biology treats biological systems as ensembles of networks at multiple levels, starting from the molecular level and from there gradually addressing more complex systems such as cells, tissues, organs, whole organisms, and finally analyzing population dynamics. Systems biology aims to describe and predict the behavior of groups of interacting components. To do so, it uses mathematical and computational tools to analyze measurements collected by systematic high-throughput technologies such as (post)genomics, metabolomics, or proteomics among others. The goal of systems approaches is to boost our understanding of biology by overcoming the limitations of reductive science, which addresses individual genes, proteins, metabolites, pathways, or cells, and thus does not account for the properties emerging from their interactions [1,2].

Current medical science is mostly conducted using the reductionist approach [3,4]. This limits our ability to grasp how multiple variables interact with one another to create emergent effects [3] and hampers our understanding of diseases, as well as our capability of delivering better treatments. Systems medicine can be regarded as the application of systems biology to human physiology in a clinical context [5,6]. It addresses the aforementioned issues by applying iterative and reciprocal feedback between clinical research and practice through computational, statistical, and mathematical multiscale analysis. This includes modeling of disease progression and remission, treatment responses, and adverse events both at the epidemiological and patient level. This new paradigm of systems science and medicine strongly complements the traditional reductionist approach (Figure 1).

Figure 1



Overview of the core differences between reductionism and systems science, when analyzing the properties of a system; figure initially published in Tillmann et al., 2015 [7] under the terms of Creative Commons Attribution 2.0 license.

The functioning of the human body is regulated by the interaction and interdependencies of biological molecules at multiple levels (protein–protein, protein–RNA, and protein–DNA networks and metabolic networks) [8]. Therefore, it can only be efficiently analyzed by examining various omics concurrently. Systems medicine provides the appropriate framework to achieve this goal. The complementary perspectives offered by different data sets allow the genotype of an individual to be linked to its observed phenotype as a function of lifestyle and environmental conditions. Eventually, this could lead to defining how any healthy state can transition into a pathological one and vice versa and pave the way for personalized medicine.

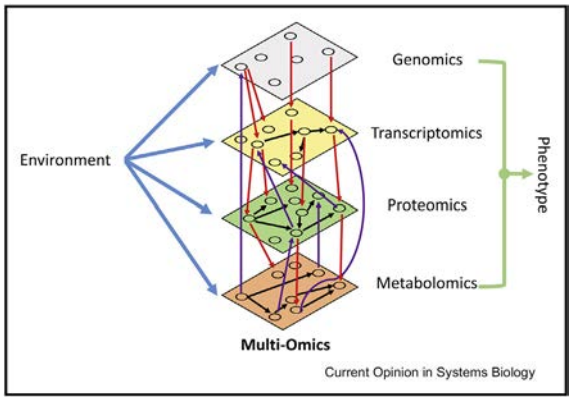
Multi-omics data integration

The integration of multiple omics data (sometimes also called trans-omics) will further enhance the contribution of omics science to our understanding of biomedicine [9]. The example in Figure 2 shows the connections among genomics, transcriptomics, proteomics, and metabolomics, thus providing an overview of the system from its potential (encoded in DNA) to the actual outcome (monitored by metabolomics).

It is commonly accepted that genes, gene products, and metabolites participate in complex, interconnected networks (Figure 2). Various biological molecules can be represented as nodes in a network and the interactions connecting them as edges. For example, in metabolomics, metabolites would be the nodes, and the edges would represent the enzymatic reactions interconnecting them. Graph theory can be applied to analyze the complexity of the interactions within a biological network and link *a priori* knowledge from the literature and databases [11]. The application of network analysis allows the identification of nodes with a high degree of connectivity ('hubs') and groups of highly interconnected nodes ('modules'), identifying molecules functionally related to a disease state [12–14].

It is possible to outline a general strategy to integrate various omics data sets based on network representations. First, the network scaffold is defined by defining how the individual components are interconnected. The structure of the network can be identified based on the data or prior knowledge (i.e., database information). Subsequently, the network itself can be separated into modules. Finally, all the information can be combined with computational models of the whole system to simulate and predict how the network determines the observed phenotype. In practice, if two omics elements share a common driver, or if one perturbs the other, they will exhibit correlation or association. Various specialized

Figure 2



Multi-omics integration across different omics layers. Red arrows highlight the top-down flow of interactions across layers: genes are transcribed, transcripts determine enzyme concentrations, and finally enzymes act on metabolites. Purple arrows highlight the bottom-up interactions, whereby metabolite levels modulate enzyme activities, the DNA/RNA-binding affinities of regulators or DNA methylation. Note that metabolites can also interact directly with transcripts. Black arrows are intra-omics networks, which can be derived based on individual omics data sets (for a review of methods in metabolomics, see Rosato et al., 2018 [10]). Intra-omics networks may describe direct physical interactions (e.g., protein–protein interactions) and correlations between their abundances (e.g., transcript levels or metabolite concentrations). Environmental stimuli (blue arrows) can affect all omics layers. For example, they can trigger DNA mutations and transcriptional events and modify protein activity. In addition, the environment is also a source of metabolites and xenobiotic molecules. Overall, the different omics levels, which are a function of the environment and the omics interactions, determine the phenotype.

statistical approaches can be applied to measure these correlations. For example, a linear model taking into account age, gender, body mass index, and white blood cell count was used to find correlations between DNA methylation and metabolite concentrations in human blood serum [15]. An even broader study analyzed the genome, transcriptome, proteome, metabolome, and metabolic fluxes in *Escherichia coli* to understand how its metabolic state reacted to perturbations [16]. More recently, weighted gene correlation network analysis was used to identify connectivity-based gene modules highly correlated to pathways identified by metabolomics [17].

Connecting the metabolome layer and other omics layers

Metabolomics measures the metabolites present within a cell, tissue, or organism. It is a core experimental omics within systems biology as it delivers an integrated view of biochemistry [18,19]. Current experimental approaches in metabolomics are mostly based on nuclear magnetic resonance and mass spectroscopy [20,21]. Metabolomic studies can be divided into two major groups: targeted and untargeted.

Targeted metabolomics quantitatively measures the abundance of a predefined group of known, well-characterized metabolites in a sample. Usually, the aim is to identify novel associations between metabolites in the context of specific physiological states [22,23]. On the other hand, untargeted metabolomics typically focuses on capturing all the chemical compounds present in a sample, including metabolites of unknown chemical structure, thus generating notably large data sets. By comparing the metabolome of the control and test groups and focusing on the differences between their metabolic profiles, the number of significant detected signals becomes more manageable. Finally, the compounds or metabolites identified are annotated using *in silico* libraries when possible or by applying analytical chemistry methods to explore the newly observed structure [24].

One of the technical challenges in connecting the metabolome with other omics layers is matching the identities of the same objects in different layers (ID conversion). Various databases support this task: the Kyoto Encyclopedia of Genes and Genomes (KEGG) integrates one computationally generated and fifteen manually curated databases, allowing the users to link metabolites to reactions, enzymes, pathways, and genes [25]; BRENDA provides information on enzymes, such as kinetic parameters for enzymatic reactions, allosteric effectors, and association with diseases [26]; Reactome is a database that organizes metabolites into biological pathways and processes, using reactions to define relationships [27]; and MetaCyc is a database of metabolic pathways and enzymes, whereas BioCyc (BioCyc.org) collects organism-specific genomes and computationally predicted metabolic networks [28]. For example, in a multi-omics study on the flow of the insulin signal based on time course data from the metabolome, phosphoproteome, and transcriptome, a global metabolism map was generated by mapping quantitatively changed metabolites and their corresponding metabolic enzymes to the KEGG database [29].

Finally, it is worth mentioning the Investigation/Study/Assay (ISA-Tab) format, which is a convenient standard to store the metadata and the results of experiments across the various omics, is already implemented in metabolomic platforms such as MetaboLights or PhenoMeNal [30–32].

Metabolic models

The metabolic phenotype is defined by two complementary omics, the metabolome and the fluxome. The first offers a static view of metabolism (snapshot-like), whereas the latter represents the rate at which metabolites are interconverted through metabolic pathways and therefore provides a dynamic view of the metabolic phenotype [33]. The fluxome emerges from complex interactions among metabolites, enzymes, and transmembrane carriers. Thus, the fluxome cannot be directly measured and instead needs to be inferred

through the analysis of other omics measurements. One of the most informative techniques to determine the fluxome is stable isotope-resolved metabolomics (SIRM). In SIRM, a biological system is incubated with a substrate labeled with a stable heavy isotope (e.g., ^{13}C) that propagates to metabolites in the network generating characteristic label patterns which are indicative of the underlying flux distribution [34].

Metabolic models, mathematical representations of metabolism, are the tools used by systems biology and systems medicine to integrate multiple layers of data and predict metabolic fluxes. Nowadays, the vast availability of genomic data and the functional annotations allows the reconstruction of genome-scale metabolic models (GSMMs). GSMMs are built starting from genome annotations, which are used to identify enzyme-coding genes. These can then be mapped to reactions using biochemical databases, such as KEGG, BRENDA, or MetaCyc. The resulting network is then curated to account for misannotations and missing reactions. Finally, the built reconstruction is validated by simulating the known metabolic functions of the target organism [35]. In 2007, the first human GSMMs were reconstructed [36,37]. They formed the basis for much more in-depth human genome-scale reconstruction models including Human Metabolic Reaction, Recon 2, and Recon3D [38–40].

Metabolic simulations based on these genome-scale networks, or a subset of them, are usually performed with either kinetic- or constraint-based modeling (CBM) techniques. Kinetic models integrate kinetic properties of enzymes (e.g., their affinity for substrates, the number of catalytic cycles that they can undergo per unit of time, and their regulation by activators or inhibitors) and allow to simulate the dynamic behavior of fluxes and metabolites. However, they are limited by the complexity to build and parametrize kinetic models for large networks. In contrast, CBM uses network stoichiometry and the assumption of the metabolic pseudo-steady state (i.e., intracellular metabolite concentrations are constant in time) to simulate steady-state flux distributions. Although CBM is easily applied to large networks such as GSMMs, it has a more limited capacity when it comes to studying the dynamic behavior of metabolic networks than kinetic models.

Building large-scale kinetic models

Kinetic models are systems of ordinary differential equations (ODEs) where metabolic fluxes are computed as a function of metabolite concentrations through a set of defined kinetic equations. Each metabolite has an ODE equation representing its variation in time, and each reaction has a kinetic equation describing the dependency of reaction fluxes to metabolite and enzyme concentrations. Metabolomic data, taken at multiple time points, are the primary input to validate kinetic models and iteratively fit unknown parameters of the kinetic equations (Figure 3) [41].

There are two approaches to building large-scale kinetic models: the bottom-up or forward reconstruction and the top-down or inverse reconstruction. In the former method, the various subparts of the model are built individually and then put together to form the final model, whereas in the latter, the entire model is reconstructed, and all the parameters are fitted at the same time. The major issues in large-scale kinetic model reconstruction are the many unknown parameters in the model and the lack of knowledge of regulatory information. Indeed, the greatest challenge to build large kinetic models is the parameter inference or fitting step. Over the last few years, approaches such as structural kinetic modeling and mass action stoichiometric simulation (MASS) modeling have been developed to tackle this step.

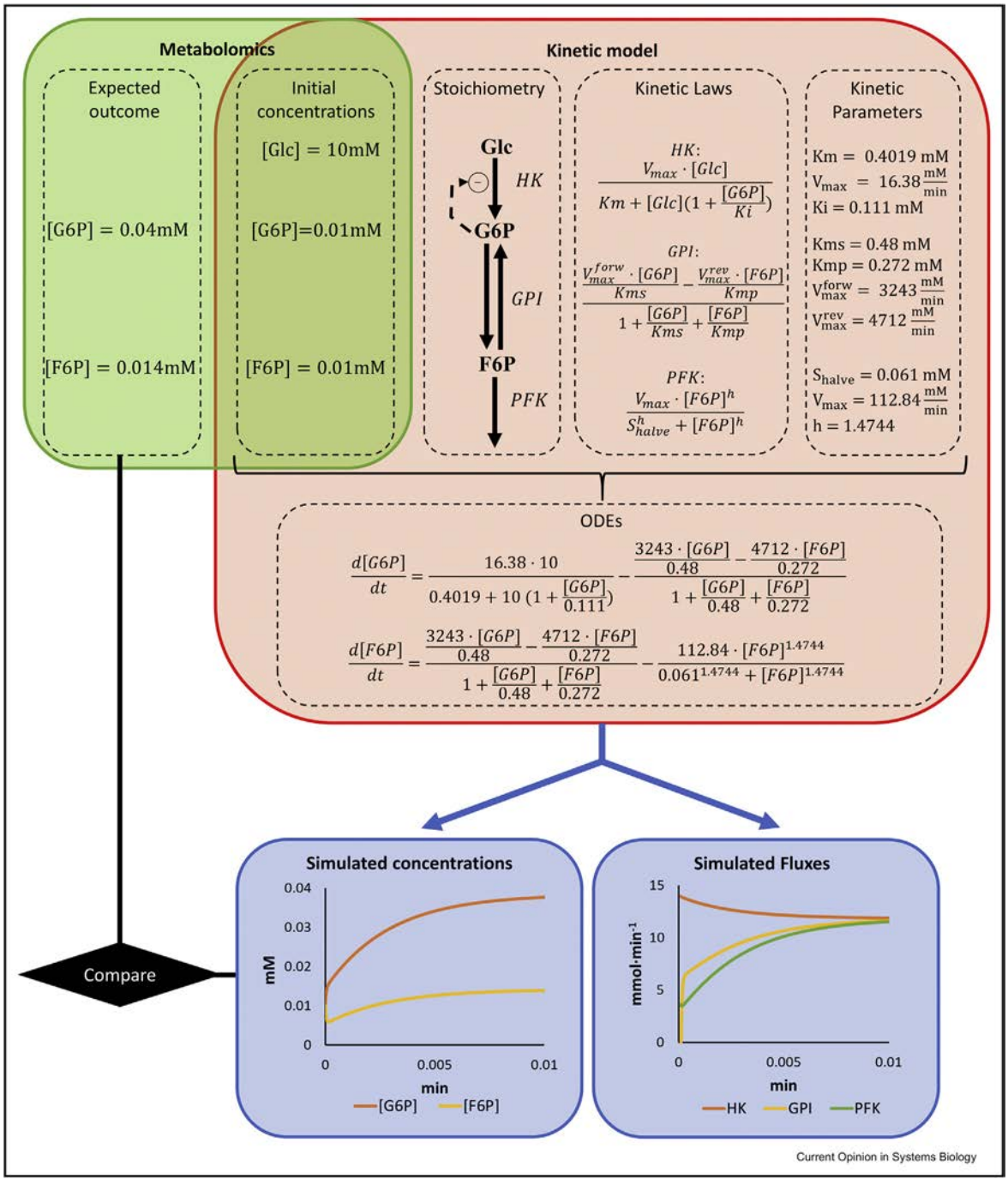
Structural kinetic modeling aims to quantitatively describe the dynamic performance of a system, rather than specifically define kinetic parameters, and constructs local linear approximations for each parameter according to experimental data and feasible biochemical states. Then, the reconstructed local linear models are used for the interrogation of a solution parameter space [43,44]. On the other hand, MASS models try to combine constraint-based stoichiometric reconstructions with matrix-based kinetic modeling. More specifically, MASS uses large-scale stoichiometric network reconstructions as scaffolds, onto which fluxomic and metabolomic data measured *in vivo* are integrated, and then, kinetic parameters, explicit for the modeled steady state of the system, are estimated. If simulations of growth conditions are performed, kinetic constants for the evolution of the system can be calculated, thus describing its dynamic behavior [45].

Constraint-based modeling

CBM assumes a metabolic pseudo-steady state to build mass balance constraints around metabolites and identify valid steady-state flux distributions. In this manner, the stoichiometry of the network can be represented as a system of linear equations, and steady-state flux distributions can be simulated without the need for defining the kinetic equations for each enzyme [36,37,46]. As the resulting system is usually underdetermined, additional constraints and optimizations need to be applied to reduce the solution space toward a unique solution (Figure 4) [38,39].

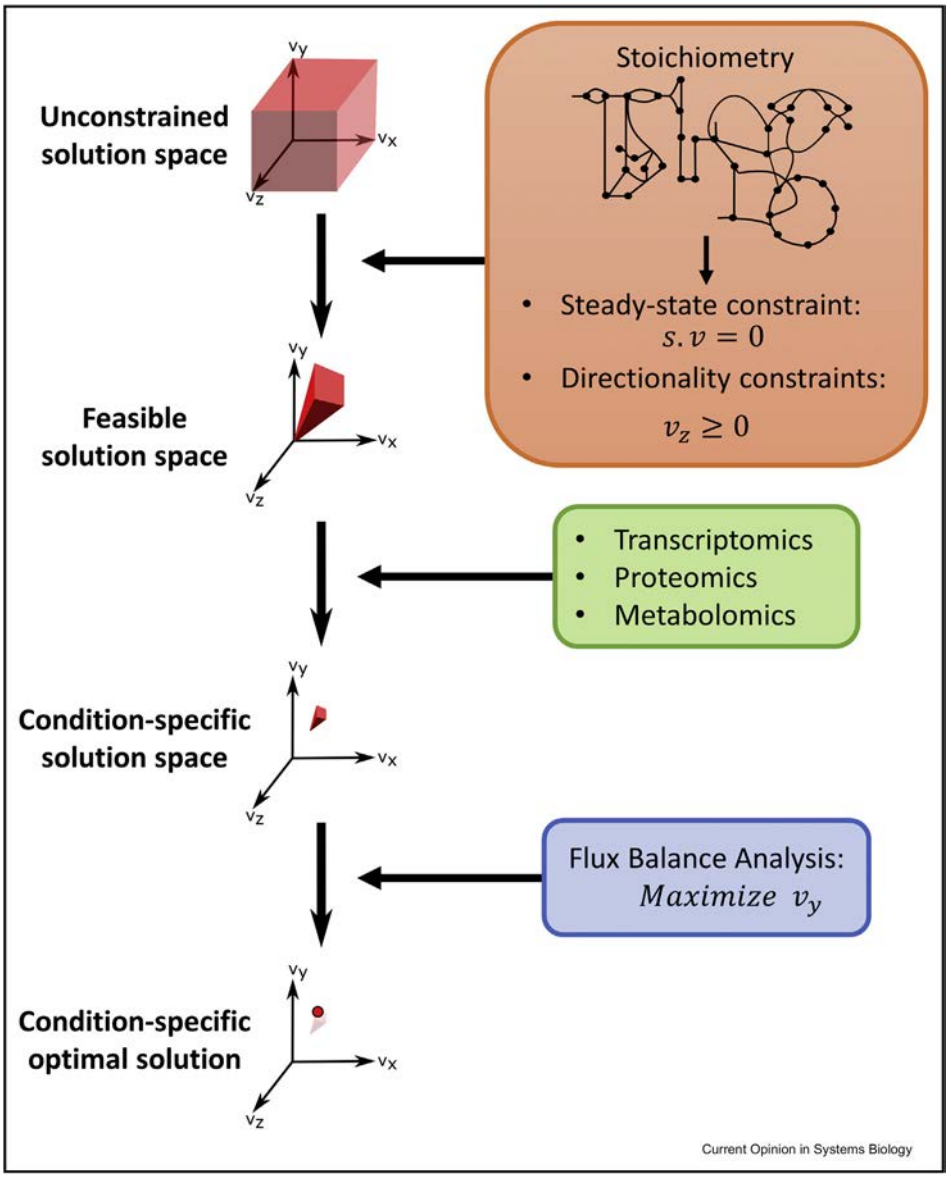
For instance, GSMMs generally need to be constrained by integrating transcriptomics or proteomics data. This need arises because GSMMs define the entire

Figure 3



Kinetic model of upper glycolysis (Puigjaner et al., 1997 [42]). The network has three metabolites (Glc: glucose, G6P: glucose 6-phosphate, F6P: fructose 6-phosphate) connected by three reactions (HK: hexokinase, GPI: glucose 6-phosphate isomerase, PFK: phosphofructokinase). HK has a Michaelis–Menten kinetic law with an uncompetitive inhibition by G6P; GPI, a reversible Michaelis–Menten kinetic law; and PFK, a Hill cooperative kinetic law. Each kinetic law is parametrized from measurements of mice muscle extracts (V_{max}: maximal reaction rate, Km/Kms/Kmp/S_{halve}: concentration at which half of the V_{max} is achieved, Ki: Concentration at which half of the inhibition is achieved, h: Hill cooperativity coefficient). From network stoichiometry, the parametrized kinetic laws are combined to build a system of ODEs, with each equation describing the dependent dynamic of a metabolite concentration. Starting with initial metabolomic values, solving the system of ODEs simulates time courses for metabolite concentrations and reaction fluxes, which can be compared with additional metabolomics data for validation. ODE, ordinary differential equation.

Figure 4



Constraint-based modeling. First, the stoichiometry of the metabolic network is written as a stoichiometric matrix (*s*), where the $s_{m,r}$ element of the matrix is the stoichiometric coefficient of the metabolite *m* in the reaction *r*. From an infinite space of possible flux (*v*) solutions, a feasible solution space which contains possible steady-state solutions is obtained by applying the steady-state constraint ($s.v = 0$) and defining the directionality of reactions. A condition-specific solution space can be obtained by integrating condition-specific omics such as transcriptomics, proteomics, or metabolomics. Finally, an optimization can be performed in the solution space to select the best solution(s). For instance, biomass production can be maximized so that the solution(s) that optimizes growth efficiency can be selected.

metabolic potential for a given organism, whereas at any given cell and time point, only a subset of enzymes are expressed and only a subset of reactions will be active. There are several approaches to integrate such data, but they are generally based on maximizing the consistency between the transcript and protein abundances of enzymes and the flux through reactions catalyzed by them. Integrating transcriptomics and proteomics allows to obtain maps of active/inactive reactions, as well as to characterize the changes in flux distributions between two or more different conditions or time points [47–52].

The range of feasible flux values can be further constrained by metabolomics data. Metabolomics from the extracellular media can be used to constrain

extracellular fluxes (i.e., rates of uptake of secretion for extracellular metabolites). Concerning intracellular metabolomics, if a metabolite is detected, the model can be constrained to have at least one reaction active, where this metabolite is produced [50]. Furthermore, quantitative metabolomics of intracellular metabolites allows setting the rate at which intracellular metabolites must be synthesized to maintain a steady state in proliferating or growing systems [53]. Finally, SIRM-based metabolic flux analysis can be applied to identify the range of flux values underlying a given set of SIRM measurements. The resulting flux ranges can be added to the GSMM as flux bounds [34].

Even after integrating transcriptomics or proteomics and metabolomics, GSMMs are generally still undetermined. Flux balance analysis aims to identify a unique optimal solution by maximizing or minimizing one or more fluxes in the metabolic network [54]. The choice of objective depends on the system under study, for instance, to study rapidly proliferating systems, such as cancer cells, the synthesis of biomass is used as the objective, but other objectives can be set depending on the system of study [54–59].

Applications in systems medicine

The integration of multiple omics data in a systems medicine manner is an emerging field. Nevertheless, it has already provided new insights into the interplay among different regulatory layers.

For example, by studying the associations between single-nucleotide polymorphism and metabolomics measurements, it has been demonstrated that the variability of metabolite concentrations in the blood between individuals is explained to a large extent by common genetic variants [60]. In another study, associations using epigenome-wide association data in combination with cytosine–guanine dinucleotide methylation data and other multi-omics data suggested a causal effect of metabolite levels on methylation of obesity-associated cytosine–guanine dinucleotide sites [61].

Furthermore, even if the reconstruction of large-scale kinetic models still poses a big challenge, several examples of kinetic models in systems medicine demonstrate their great potential. For instance, a kinetic model of human erythrocytes was used to identify metabolic targets that would selectively kill the parasite *Trypanosoma brucei* with minimal collateral damage to human cells [62]. Berndt et al. [63] reconstructed a kinetic model of the liver, and they used it to characterize the metabolic phenotype of hepatocytes and the metabolic reprogramming that they have undergone during carcinogenesis. Bordbar et al. [64] have simulated individual responses to drug exposure including side effect incidence and demonstrated that enzyme activities and cellular dynamics, rather than metabolomics, are the most accurate representation of the genotype.

CBM has also been widely used in systems medicine to perform multi-omics data integration in the framework of GSMMs. For example, Mardinoglu et al. [65] integrated proteomics and transcriptomics to build an adipocyte-specific GSMM and identified several putative therapeutics against obesity. GSMMs have also been widely applied to identify genes or sets of genes that are essential for a disease-related process [59,66–68]. For instance, Folger et al. [69] created a GSMM of cancer metabolism that predicted 52 cytostatic drug targets, 40% of which were targeted by known anticancer drugs. Similarly, Agren et al. [70] built 27 patient-specific GSMMs of hepatocellular carcinoma and identified 101 potential drug targets, many of which had a strong correlation with disease progression. GSMMs have also shown great potential in biomarker discovery, for example, in liver diseases and type 2 diabetes [71,72].

Conclusions and future perspectives

The primary goal of systems medicine is to explain, predict, and prevent the progression of disease based on clinical, environmental, and multi-omics data. Given the inherent network structure of metabolic processes, network modeling and the analysis of multi-omics data provide powerful and flexible inference tools to decipher the complex interactions in biological systems. However, consensus models built from samples from many individuals, albeit informative, might fail to capture the heterogeneity that is present in a population [73]. This limits the elucidation of the molecular drivers for an individual-specific phenotype (either healthy or pathological), which result from the differential regulation or dysfunction of individual-specific networks.

Toward that end, methods are being proposed to build patient-specific networks that capture the subject's specificity of clinical manifestation with the goal of understanding diseases at the individual level and providing targeted and personalized treatments [74–77]. In principle, a personalized database could be generated for each individual, containing his/her omics information (e.g., genomics, urine and blood metabolomics, gut microbiome), together with lifestyle data across time. This information, if properly analyzed, can provide the means to build patient-specific networks to identify the best diagnostic, therapeutic, and prevention strategies for each individual and enable predictive, preventive, personalized and participatory medicine [78,79].

Conflict of interest statement

Nothing declared.

Acknowledgements

AR acknowledges financial support from CIRMMP. VAPMDS and ES acknowledge financial support from the EU FP7 funded project INFECT (contract no. 305340; www.fp7infect.eu/). MC acknowledges support from MINECO-European Commission FEDER funds – "Una manera de hacer Europa" (SAF2017-89673-R and SAF2015-70270-REDT), Instituto de Salud Carlos III and Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBEREHD CB17/04/00023), AGAUR – Generalitat de Catalunya (2017SGR-1033), and through the prize 'ICREA Academia' for excellence in research, funded by ICREA foundation—Generalitat de Catalunya.

References

Papers of particular interest, published within the period of review, have been highlighted as:

* of special interest
** of outstanding interest

1. Ideker T, Galitski T, Hood L: **A new approach to decoding life: systems biology**. *Annu Rev Genom Hum Genet* 2001, **2**: 343–372.

2. Kirschner MW: **The meaning of systems biology**. *Cell* 2005, **121**:503–504.

3. Ahn AC, *et al.*: **The limits of reductionism in medicine: could systems biology offer an alternative?** *PLoS Med* 2006, **3**:e208.

4. Miles A: **On a medicine of the whole person: away from scientistic reductionism and towards the embrace of the complex in clinical practice***. *J Eval Clin Pract* 2009, **15**:941–949.

5. Auffray C, Chen Z, Hood L: **Systems medicine: the future of medical genomics and healthcare**. *Genome Med* 2009, **1**:2.

6. Hood L, Balling R, Auffray C: **Revolutionizing medicine in the 21st century through systems approaches**. *Biotechnol J* 2012, **7**:992–1001.

7. Tillmann T, *et al.*: **Systems Medicine 2.0: potential benefits of combining electronic health care records with systems science models**. *J Med Internet Res* 2015, **17**:e64.

8. Barabási A-L, Gulbahce N, Loscalzo J: **Network medicine: a network-based approach to human disease**. *Nat Rev Genet* 2011, **12**:56–68.

9. Joyce AR, Palsson BØ: **The model organism as a system: integrating 'omics' data sets**. *Nat Rev Mol Cell Biol* 2006, **7**:198.

10. Rosato A, *et al.*: **From correlation to causation: analysis of metabolomics data using systems biology approaches**. *Metabolomics* 2018, **14**:37.

11. Ma'ayan A: **Introduction to network analysis in systems biology**. *Sci Signal* 2011, **4**:tr5.

12. Barabási A-L: **Scale-free networks: a decade and beyond**. *Science* 2009, **325**:412.

13. Vidal M, Cusick ME, Barabási A-L: **Interactome networks and human disease**. *Cell* 2011, **144**:986–998.

14. Emilsson V, *et al.*: **Genetics of gene expression and its effect on disease**. *Nature* 2008, **452**:423.

15. * Petersen A-K, *et al.*: **Epigenetics meets metabolomics: an epigenome-wide association study with blood serum metabolic traits**. *Hum Mol Genet* 2014, **23**:534–545.
This is the first epigenome-wide association study (EWAS) between DNA methylation and metabolic types in human blood, showing that that DNA methylation plays an important role in regulating human metabolism.

16. ** Ishii N, *et al.*: **Multiple high-throughput analyses monitor the response of 'E. coli' to perturbations**. *Science* 2007, **316**: 593–597.
This study integrated multiple omics measurements on *Escherichia coli*. It showed that in most cases the disruption of genes encoding metabolic enzymes led to surprisingly small changes in messenger RNA and proteins and did not sizably alter metabolite levels. This is due to the rerouting of metabolic fluxes. In contrast, *E. coli* actively controlled enzyme levels to maintain a stable metabolic state in response to changes in growth rate.

17. Liu C, *et al.*: **Arachidonic acid metabolism pathway is not only dominant in metabolic modulation but associated with phenotypic variation after acute hypoxia exposure**. *Front Physiol* 2018, **9**.

18. Clish CB: **Metabolomics: an emerging but powerful tool for precision medicine**. *Cold Spring Harb Mol Case Stud* 2015, **1**: a000588.

19. Trivedi DK, Hollywood KA, Goodacre R: **Metabolomics for the masses: the future of metabolomics in a personalized world**. *New Horiz Transl Med* 2017, **3**:294–305.

20. Dettmer K, Aronov PA, Hammock BD: **Mass spectrometry-based metabolomics**. *Mass Spectrom Rev* 2007, **26**:51–78.

21. Wishart DS: **Quantitative metabolomics using NMR**. *Trac Trends Anal Chem* 2008, **27**:228–237.

22. Bingol K: **Recent advances in targeted and untargeted metabolomics by NMR and MS/NMR methods**. *High-Throughput* 2018, **7**:9.

23. Roberts LD, *et al.*: **Targeted metabolomics**. *Curr Protoc Mol Biol* 2012, **98**:30.2.1–30.2.24.

24. Alonso A, Marsal S, Julià A: **Analytical methods in untargeted metabolomics: state of the art in 2015**. *Front Bioeng Biotechnol* 2015, **3**.

25. Kanehisa M, *et al.*: **KEGG: new perspectives on genomes, pathways, diseases and drugs**. *Nucleic Acids Res* 2017, **45**: D353–D361.

26. Placzek S, *et al.*: **BRENDA in 2017: new perspectives and new tools in BRENDA**. *Nucleic Acids Res* 2017, **45**:D380–D388.

27. Fabregat A, *et al.*: **The reactome pathway knowledgebase**. *Nucleic Acids Res* 2018, **46**:D649–D655.

28. Caspi R, *et al.*: **The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases**. *Nucleic Acids Res* 2016, **44**:D471–D480.

29. Yugi K, *et al.*: **Reconstruction of insulin signal flow from phosphoproteome and metabolome data**. *Cell Rep* 2014, **8**: 1171–1183.

30. Sansone S-A, *et al.*: **The first RSBI (ISA-TAB) workshop: "can a simple format work for complex studies?** *OMICS* 2008, **12**: 143–149.

31. Kale NS, *et al.*: **MetaboLights: an open-access database repository for metabolomics data**. *Curr Protoc Bioinformatics* 2016, **53**:14.13.1–14.13.18.

32. Peters K, *et al.*: *PhenoMeNal: processing and analysis of metabolomics data in the cloud*. GigaScience; 2018giy149.

33. Niedenführ S, Wiechert W, Nöh K: **How to measure metabolic fluxes: a taxonomic guide for 13C fluxomics**. *Curr Opin Biotechnol* 2015, **34**:82–90.

34. Balcells C, *et al.*: **Tracing metabolic fluxes using mass spectrometry: stable isotope-resolved metabolomics in health and disease**. *Trac Trends Anal Chem* 2019, https://doi.org/10.1016/j.trac.2018.12.025.

35. Faria JP, *et al.*: *Methods for automated genome-scale metabolic model reconstruction*. Biochemical Society Transactions; 2018. BST20170246.

36. Duarte NC, *et al.*: **Global reconstruction of the human metabolic network based on genomic and bibliomic data**. *Proc Natl Acad Sci Unit States Am* 2007, **104**:1777–1782.

37. Ma H, *et al.*: **The Edinburgh human metabolic network reconstruction and its functional analysis**. *Mol Syst Biol* 2007, **3**:135.

38. Thiele I, *et al.*: **A community-driven global reconstruction of human metabolism**. *Nat Biotechnol* 2013, **31**:419–425.

39. Pornputtapong N, Nookaew I, Nielsen J: **Human metabolic atlas: an online resource for human metabolism**. *Database :*

the journal of biological databases and curation 2015, **2015**: bav068.

40. Brunk E, *et al.*: **Recon3D enables a three-dimensional view of gene variation in human metabolism**. *Nat Biotechnol* 2018, **36**: 272.

41. Foguet C, *et al.*: **HepatoDyn: a dynamic model of hepatocyte metabolism that integrates 13C isotopomer data**. *PLoS Comput Biol* 2016, **12**, e1004899.

42. Puigjaner J, *et al.*: **Comparison of control analysis data using different approaches: modelling and experiments with muscle extract**. *FEBS (Fed Eur Biochem Soc) Lett* 1997, **418**: 47–52.

43. Steuer R, *et al.*: **Structural kinetic modeling of metabolic networks**. *Proc Natl Acad Sci Unit States Am* 2006, **103**:11868.

44. Grimbs S, *et al.*: **The stability and robustness of metabolic states: identifying stabilizing sites in metabolic networks**. *Mol Syst Biol* 2007, **3**.

45. Jamshidi N, Palsson BØ: **Mass action stoichiometric simulation models: incorporating kinetics and regulation into stoichiometric models**. *Biophys J* 2010, **98**:175–185.

46. ** Nilsson A, Nielsen J: **Genome scale metabolic modeling of cancer**. *Metab Eng* 2017, **43**:103–112.
A comprehensive review of GSMMs focused on their application to the study of Cancer metabolism.

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

48. Jensen PA, Papin JA: **Functional integration of a metabolic network model and expression data without arbitrary thresholding**. *Bioinformatics* 2011, **27**:541–547.

49. Agren R, *et al.*: **Reconstruction of genome-scale Active metabolic networks for 69 human cell types and 16 cancer types using INIT**. *PLoS Comput Biol* 2012, **8**, e1002518.

50. Schmidt BJ, *et al.*: **GIM3E: condition-specific models of cellular metabolism developed from metabolomics and expression data**. *Bioinformatics* 2013, **29**:2900–2908.

51. Yizhak K, *et al.*: **Model-based identification of drug targets that revert disrupted metabolism and its application to ageing**. *Nat Commun* 2013, **4**:2632.

52. Galhardo M, *et al.*: **Integrated analysis of transcript-level regulation of metabolism reveals disease-relevant nodes of the human metabolic network**. *Nucleic Acids Res* 2014, **42**: 1474–1496.

53. Reimers A-M, Reimers AC: **The steady-state assumption in oscillating and growing systems**. *J Theor Biol* 2016, **406**: 176–186.

54. Orth JD, Thiele I, Palsson BØ: **What is flux balance analysis?** *Nat Biotechnol* 2010, **28**:245–248.

55. Ramakrishna R, *et al.*: **Flux-balance analysis of mitochondrial energy metabolism: consequences of systemic stoichiometric constraints**. *Am J Physiol Regul Integr Comp Physiol* 2001, **280**:R695–R704.

56. Gille C, *et al.*: **HepatoNet1: a comprehensive metabolic reconstruction of the human hepatocyte for the analysis of liver physiology**. *Mol Syst Biol* 2010, **6**:411.

57. Bordbar A, *et al.*: **Constraint-based models predict metabolic and associated cellular functions**. *Nat Rev Genet* 2014, **15**: 107.

58. Rienksma RA, *et al.*: **Modeling the metabolic state of Mycobacterium tuberculosis upon infection**. *Front Cell Infect Microbiol* 2018, **8**.

59. Wang T, *et al.*: **Gene essentiality profiling reveals gene networks and synthetic lethal interactions with oncogenic ras**. *Cell* 2017, **168**:890–903. e15.

60. Suhre K, *et al.*: **Human metabolic individuality in biomedical and pharmaceutical research**. *Nature* 2011, **477**:54–60.

61. * Zaghlool SB, *et al.*: **Deep molecular phenotypes link complex disorders and physiological insult to CpG methylation**. *Hum Mol Genet* 2018, **27**:1106–1121.
combined Epigenome-Wide ASsociation data (EWAS) with cytosine-guanine dinucleotide (CpG) methylation and other multi-omics datasets and revealed a causal effect of metabolite levels on methylation of obesity-associated CpG sites.

62. * Haanstra JR, *et al.*: **Targeting pathogen metabolism without collateral damage to the host**. *Sci Rep* 2017, **7**:40406.
This study highlights how kinetic modelling can assist drug design by identifying targets against pathogen metabolism with have minimal side effects on the host.

63. * Berndt N, *et al.*: **HEPATOKIN1 is a biochemistry-based model of liver metabolism for applications in medicine and pharmacology**. *Nat Commun* 2018, **9**:2386.
In this study, the authors build a large-scale kinetic model of hepatocyte metabolism and integrated proteomics to highlight the metabolic differences in the metabolic phenotype hepatocytes compared to adenoma and hepatocellular carcinoma cells.

64. ** Bordbar A, *et al.*: **Personalized whole-cell kinetic models of metabolism for discovery in genomics and pharmacodynamics**. *Cell Systems* 2015, **1**:283–292.
constructed personalized whole cell kinetic models of erythrocyte and showed that personalized kinetic rate constants are the best representation of the genotype. They were also able to identify individuals at risk for a drug side effect.

65. Mardinoglu A, *et al.*: **Integration of clinical data with a genome-scale metabolic model of the human adipocyte**. *Mol Syst Biol* 2013, **9**:649.

66. von Kamp A, Klamt S: **Enumeration of smallest intervention strategies in genome-scale metabolic networks**. *PLoS Comput Biol* 2014, **10**, e1003378.

67. Pratapa A, Balachandran S, Raman K: **Fast-SL: an efficient algorithm to identify synthetic lethal sets in metabolic networks**. *Bioinformatics* 2015, **31**:3299–3305.

68. Zhan T, Boutros M: **Towards a compendium of essential genes – from model organisms to synthetic lethality in cancer cells**. *Crit Rev Biochem Mol Biol* 2016, **51**:74–85.

69. Folger O, *et al.*: **Predicting selective drug targets in cancer through metabolic networks**. *Mol Syst Biol* 2011, **7**:501.

70. Agren R, *et al.*: **Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling**. *Mol Syst Biol* 2014, **10**:721.

71. Mardinoglu A, *et al.*: **Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease**. *Nat Commun* 2014, **5**:3083.

72. Väremo L, *et al.*: **Proteome- and transcriptome-driven reconstruction of the human myocyte metabolic network and its use for identification of markers for diabetes**. *Cell Rep* 2015, **11**:921–933.

73. Kuiper JS, *et al.*: **Social relationships and risk of dementia: a systematic review and meta-analysis of longitudinal cohort studies**. *Ageing Res Rev* 2015, **22**:39–57.

74. Hamburg MA, Collins FS: **The path to personalized medicine**. *N Engl J Med* 2010, **363**:301–304.

75. Liu X, *et al.*: **Identifying disease genes and module biomarkers by differential interactions**. *J Am Med Inform Assoc : JAMIA* 2012, **19**:241–248.

76. Liu R, *et al.*: **Early diagnosis of complex diseases by molecular biomarkers, network biomarkers, and dynamical network biomarkers**. *Med Res Rev* 2014, **34**:455–478.

77. Zhang W, *et al.*: **Diagnosing phenotypes of single-sample individuals by edge biomarkers**. *J Mol Cell Biol* 2015, **7**:231–241.

78. Chen R, Snyder M: **Systems biology: personalized medicine for the future?** *Curr Opin Pharmacol* 2012, **12**:623–628.

79. Flores M, *et al.*: **P4 medicine: how systems medicine will transform the healthcare sector and society**. *Pers Med* 2013, **10**:565–576.