Understanding the metabolic signatures of haematological cancers through an integrative multi-Omics approach

Grigorios Papatzikas

BSc DUTH (Greece), MSc UC (Greece)

A thesis submitted to the
University of Birmingham for the degree of
DOCTOR OF PHILOSOPHY

Centre for Computational Biology
Institute of Cancer and Genomic Sciences
College of Medical and Dental Sciences
University of Birmingham

September 2020

ABSTRACT

Haematological cancers are heterogenous diseases caused by a series of events which drive cells to uncontrolled proliferation and tumour progression. Nowadays, our understanding is that one hallmark of cancer cells is to reprogram their normal cellular metabolism to sustain their anabolic requirements for continuous cell growth and proliferation. Despite the remarkable progress in cancer metabolism, the exact mechanisms behind cancer metabolic reprogramming are not yet fully understood. The work presented in this thesis aims to provide novel biological insights into the metabolic reprogramming of haematological cancers and highlight potential metabolic vulnerabilities for therapeutic targeting approached to be investigated in future studies. A multi-Omics data integration approach was selected to achieve such ambitious aims. Herein, recent computational methodologies were applied to integrate and analyse transcriptomic with metabolomic profiles derived from cancer patients, as well as cell lines, mostly from mature B-cell neoplasms.

Mature B-cell neoplasms, such as Chronic Lymphocytic Leukaemia (CLL) and Non-Hodgkin Lymphomas (NHL), rise from the clonal expansion of mature B-cells and they are responsible for most newly diagnosed cases of haematological cancers worldwide. The second chapter of this thesis presents an investigation into the transcriptome profile of CLL patients characterised by a distinct clinical

response. Deregulated metabolic genes and pathways were identified between rare CLL cases that have undergone spontaneous regression versus CLL cases with poor clinical outcome. CLL cells from cases with poor outcome presented a differential reliance on oxidative phosphorylation and mitochondrial respiration compared to spontaneous regressed CLL cells. Going beyond traditional gene expression analysis, we performed an integration of transcriptomics profiles with Genome Scale Metabolic Models to identify metabolic genes as potential vulnerabilities in CLL. Our findings emphasise the important role of metabolic reprogramming in CLL and suggest the possibility of targeting metabolism for future studies and therapeutic approaches.

The third chapter of this thesis describes a study exploring cancer metabolism in aggressive NHL associated with germinal centre development, focusing on endemic Burkitt Lymphoma (BL) and the germinal-centre–like subtype Diffuse Large B-cell Lymphomas (DLBCL). Analysis of the transcriptome of primary tumours revealed that BL cases possessed a distinct gene expression profile compared to DLBCL cases. This BL profile is suggestive of altered function of metabolism with elevated expression in serine metabolic genes, the c-Myc and mTORC1 pathways. On the opposite, DLBCL cases appeared to be dependent on extracellular signals from cytokines (INFγ response) or inflammation, possibly to trigger activation of intracellular signalling pathways that impact metabolism. Furthermore, integrative analysis at the pathway level between transcriptomic and metabolomic datasets from cell lines, indicated a dependency of BL cells on non-

essential amino acid metabolism and particularly on the alanine, aspartate and glutamine metabolic pathways. These results not only highlighted key metabolic regulators in NHL, but most importantly, demonstrated the necessity of understanding and monitoring metabolic properties in these lymphomas.

Finally, chapter four describes work undertaken to explore the transcriptomic and metabolic diversity of cancer cell lines. Machine learning approaches were applied to integrate and analyse Omics datasets retrieved from the Cancer Cell Line Encyclopaedia (CCLE) database. Unsupervised analysis highlighted the distinct transcriptomic and metabolomic profile of haematopoietic cell lines compared to other tumours. Taking a supervised approach enabled us to associate gene expression changes in cytoskeleton and cell adhesion molecules with aberrant metabolites levels, such as xanthine and creatinine. Together, these observations provide proof of concept for the highly dynamic variations between transcriptome and metabolome in different cancers.

In summary, this work portrays the power of multi-Omics data integration to unveil key elements in metabolic reprogramming of haematological cancers and raises numerous questions and new hypotheses for future metabolic studies.

ACKNOWLEDGEMENT

The work in this thesis has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 675790. This work was undertaken from 2016 to 2020 both at University of Birmingham in the Centre for Computational Biology and at the University of Barcelona in the Biology department. At this point, I would like to thank all the staff in both these respective Universities for all their help and support during this period.

Above all, I would like to express my gratitude to my primary supervisor Prof. Jean-Baptiste Cazier for his mentorship and his excellent supervision that enabled me to develop all the essential skills and the acquired scientific knowledge that helps me to pursue a future career in science. Furthermore, I would like to give special thanks to my co-supervisors Prof. Marta Cascante and Prof. Ulrich Günther for sharing their incredible scientific expertise and for the opportunity to be part of such a prestigious European training network.

I would also like to thank Prof. Tatjana Stankovic and Dr. Marwan Kwok for providing, the valuable for my work, CLL transcriptomic data and for their constructive discussion on my work. Furthermore, I gratefully acknowledge Dr. Farhat Khanim for the use of cell line datasets and for her productive feedback.

Special thanks go to Dr. Wayne Croft for reviewing and correcting my thesis. His constructive comments helped me to complete this thesis. I would also like to thank Prof. Francisco Planes, Dr. Carles Foguet, Mr. Luis Vi Valcárcel and Dr. Apaolaza Emparanza Iñigo for their technical troubleshooting and constructive advice on Genome Scale Metabolic Modelling. More specifically, Dr. Carles Foguet's corrections and suggestions highlighted several limitations related to the robust Metabolic Transformation Algorithm. Similarly, Mr. Luis Vi Valcárcel and Dr. Apaolaza Emparanza Iñigo's work on the genetic Minimal Cut Sets significantly assisted me to obtain and improve the results presented in chapter 2. Additionally, I would like to thank Ms Nuria Vilaplana, Ms Zuhal Eraslan, Ms Agata Stodolna and Mr Sam Benkwitz-Bedford for providing data and working together in various projects; to Mrs Jessica Mylchreest and Mrs Jordan McCormick for their help with administrating work; and finally, to everyone in the Advance Research Computing Team of University of Birmingham for the technical support with CaStLeS computational resources.

More importantly, I am deeply thankful to my family, especially to my spouse Natalia and to my son Panagiotis, for their patience and their support to complete a such time intensive task.

All the bioinformatics work in this thesis was performed by the author in a virtual machine in CaStLeS Cloud computing resources of University of Birmingham (Thompson et al., 2019).

PUBLICATIONS

Kwok M, Rawstron A, Goel A, **Papatzikas G**, Oldreive C, Rhiannon EJ, Drennan S, Agathanggelou A, Sharma-Oates A, Evans P, Smith E, Mao J, Beaumont J, Rai J, Hamada M, Dalal S, Gordon N, Davies N, Parry H, Beggs A, Munir T, Moreton P, Paneesha A, Pratt G, Taylor M, Forconni F, Baird MD, Cazier JB, Moss P, Hillmen P, Stankovic T. **Integrative analysis of spontaneous CLL regression highlights genetic and microenvironmental interdependency in CLL.** *Blood* 135, 411–428 (2020). DOI: 10.1182/blood.2019001262

Eraslan Z*, **Papatzikas G***, Cazier JB, Khanim F, Gunther U. Targeting asparagine and serine metabolism in germinal centre-derived B cells Non-Hodgkin lymphomas (B-NHL). *Cells 2021,* 10(10), 2589. DOI: 10.3390/cells10102589

**Share the first authorship*

Manuscript in preparation

Papatzikas G, Foguet C, Kwok M, Valcárcel LV, Planes FJ, Stankovic T, Cascante M, Cazier JB. **Recruiting Genome Scale Metabolic Modelling integration approaches to predict metabolic vulnerabilities in CLL patients.**

TABLE OF CONTENTS

CHAPTER 1
INTRODUCTION

CHAPTER 2

METABOLIC MODELLING INTEGRATION TO REVEAL METABOLIC VULNERABILITIES IN CLL

CHAPTER 3
PATHWAY INTEGRATION TO CHARACTERISE METABOLIC VARIATIONS IN GC-DERIVED LYMPHOMAS

CHAPTER 4
MULTI-OMICS DATA INTEGRATION FOR CANCER CELL LINES WITH MACHINE LEARNING

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Terms:

ABC	Activated B-cell group
ALDHs	Aldehydes Dehydrogenases enzymes
AML	Acute Myeloid Leukaemia
AUC	Area Under the ROC Curve
BER	Balanced Error Rate
BL	Burkitt Lymphoma
CCLE	Cancer Cell Line Encyclopedia
CEs	Cholesterol Esters
CHOP	Cyclophosphamide, Hydroxydaunorubicin, Oncovin, Prednisone
CLL	Chronic Lymphocytic Leukaemia
CNS	Central Nervous System
CNV	Copy Number Variation
DEA	Differential Expression Analysis
DEGs	Differentially Expressed Genes
DIABLO	Data Integration analysis for Biomarker discovery using Latent components
DL	Deep Learning
DLBCL	Diffuse Large B-cell Lymphoma
EBV	Epstein-Barr virus
ECM	Extracellular Matrix
ETC	Electron Transport Chain
FAO	Fatty Acids Oxidation
FBA	Flux Balance Analysis
FBS	Foetal Bovine Serum
FDA	Food and Drug Administration

FDR	False Discovery Rate
FL	Follicular Lymphomas
GCB	Germinal Centre like group
GC-MS	Gas Chromatography Mass Spectrometry
GCs	Germinal Centres
GEP	Gene Expression Profiles studies
gMCSs	genetic Minimal Cut Sets
GPR	Gene-Protein-Reaction
GSEA	Gene Set Enrichment Analysis
GSMMs	Genome Scale Metabolic Models
GUI	Graphical User Interface
ICD-11	International Classification of Diseases 11th revision
iMAT	integration Metabolic Analysis Tool
KEGG	Kyoto Encyclopedia of Genes and Genomes
LASSO	Least Absolute Shrinkage and Selection Operator
LC-MS	Liquid Chromatography Mass Spectrometry
MAS	Malate-Aspartate Shuttle
MDSCs	Myeloid-derived suppressor cells
ML	Machine Learning
MS	Mass Spectroscopy
MTA	Metabolic Transformation Algorithm
NCDs	Noncommunicable Diseases
NES	Normalised Enrichment Score
NGS	Next Generation Sequencing
NHL	Non-Hodgkin Lymphoma
NK	Natural Killer cells
NMR	Nuclear Magnetic Resonance
ORA	Over Representations analysis
OXPHOS	Oxidative Phosphorylation

PC	Principal Component
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
PET	Positron Emission Tomography
PIs	Pathway Inhibitors
PLS	Partial Least Squares
PPP	Pentose Phosphate Pathway
R-CHOP	Chemotherapy and immunotherapy
rMTA	robust Metabolic Transformation Algorithm
RNAseq	RNA sequencing
ROC	Receiver Operating Characteristic
sPLS-DA	sparse Partial Least Squares Discriminant analysis
SRA	Sequence Read Archive
TCA	Tricarboxylic Acid Cycle
TPM	Transcripts Per Million
tSNE	t-Distributed Stochastic Neighbor Embedding
UMAP	Uniform Manifold Approximation and Projection
WHO	World Health Organisation

Genes and proteins:

ACSS2	Acetyl-CoA synthetase 2
AhR	Aryl hydrocarbon receptor
ALDH5A1	acetaldehyde dehydrogenase
AMOTL2	Angiomotin-like protein 2
BCL2	BCL2 apoptosis regulator
BCL6	BCL6 transcription repressor
BCR	B-cell antigen receptor
BTK	Bruton tyrosine kinase

CAIX	Carbonic anhydrase 9
Cas	Carbonic anhydrases
CD30	TNF receptor superfamily member 8
CDA	Cytidine deaminase
c-Myc	cellular Myelocytomatosis oncogene
CREBBP	CREB binding protein
EP300	E1A binding protein p300
EZH2	Enhancer of zeste 2 polycomb repressive complex 2 subunit
FCRL5	Fc receptor like 5
FDFT1	Farnesyl-diphosphate farnesyltransferase 1
FFPE	Formalin-fixed and paraffin-embedded
G6PDH	Glucose-6-phosphate dehydrogenase
GDH	Glutamate dehydrogenase enzyme
GLS	Glutaminases
GLUTs	Glutamine transporters
GOT1/2	Aspartate transaminases
GPXs	Glutathione peroxidases
GR	Glutathione reductase
HIF-1	Hypoxia-inducible factor 1
HK2	Hexokinase 2
IDH	Isocitrate dehydrogenase
IDO1	Indoleamine 2,3-dioxygenase 1
IgHV	Immunoglobulin heavy chain genes
INFγ	Interferon gamma
KRAS	KRAS proto-oncogene, GTPase
LDHA	Lactate dehydrogenase A
MCT	Monocarboxylate transporters
mTOR	Mechanistic target of rapamycin kinase
NEK2	Serine/threonine kinase protein kinase Nek2

NF-kB	Nuclear factor kappa B
PDK	Pyruvate dehydrogenase kinase
PGAM1	Phosphoglycerate mutase 1
PHGDH	Phosphoglycerate dehydrogenase
PI3K	Phosphatidylinositol 3 kinase
PKM2	Pyruvate kinase M2
PSAT1	Phosphoserine aminotransferase 1
PYCR1	Proline-5-carboxylate reductase 1
Ras	Ras GTPase protein
RIMKLB	Ribosomal modification protein rimK like family member B
SHMT1/2	Serine hydromethylotransferases
SLC1A5/ASCT2	Neutral amino acid transporter B(0)
TCF3	Transcription factor 3
TNF	Tumour necrosis factor
VEGF	Vascular endothelial growth factor
ZAP70	Zeta chain of T-cell receptor associated protein kinase 70

Metabolites and other molecules:

2DG	2-deoxy-D-glucose
2HG	2-hydroxyglutarate
3PG	3-phosphoglycerate
AG-221	Enosidenib, IDH2 inhibitor
AG-881	Vorasidenibe, IDH1 and IDH2 inhibitor
ASNase	Asparaginase
AT-101	R-(-)-gossypol acetic acid, gossypol
ATP	Adenosine triphosphate
AZD3965	Monocarboxylate Transporter 1 (MCT1) Inhibitor

BCG	β-citryl-L-glutamate
BIPTES	Bis-2-(5-phenylacetamido-1,2,4-thiadiazol-2-yl)ethyl sulphide
CB-839	Teleglenastat
cDNA	complementary DNA
CHoP	Phosphorylcholine
Cu	Copper
DNA	Deoxyribonucleic acid
$FADH_2$	Flavine adenine dinucleotide
Fe	Iron
G6P	Glucose 6-phosphate
GLU	Glutamate
GSH	Glutathione
GSSG	Glutathione disulphide
H_2O_2	Hydrogen peroxide
lncRNA	long non-coding RNA
mRNA	messenger RNA
NADH	Nicotinamide adenine dinucleotide
NADPH	Nicotinamide adenine dinucleotide phosphate
NH_4^+	Ammonium
RNA	Ribonucleic acid
ROS	Reactive oxygen species
rRNA	ribosomal RNA
siRNA	small interfering RNA
snRNA	small nuclear RNA
TAK-475	Squalene synthase inhibitor
tRNA	transfer RNA
UK5099	acyano—(1-phenylindol-3-yl)-acrylate
YM-5360	Squalene synthase inhibitor
αKG	α-ketoglutarate

CHAPTER 1

INTRODUCTION

1.1. Hallmarks of cancer

Cancer, just like cardiovascular disease, diabetes and chronic respiratory disease, is considered a noncommunicable disease (NCD), i.e. a condition that is not transmit from person to person. In 2016, NCDs were responsible for 41.0 million deaths (71%) of the overall 57 million deaths globally, with cancer being accountable for 9 million of those (15.7%) (World Health Organization, 2018). Cancer is a disease with great diversity, characterised by cells uncontrolled proliferation and tumour progression. The transformation of normal cells to malignant derivatives involves a series of events that disrupt the normal tissue architecture and create tumours (or else malignant neoplasms), in a process that is called tumourigenesis. Cancer has captured the scientific and public interest in the latest century, which saw the beginning of the “war on Cancer”. Historically, the first medical record of this disease lies in a 4500 years old Egyptian papyrus, written by the great Egyptian physician Imhotep. Around 160 AC, another great physician the Greek Claudius Galen accused the black bile to be the cause of cancer, based on Hippocrates’ humoral theory. This theory dominated in medicine for several centuries (**Figure 1.1**) (Mukherjee, 2011). By the second half of the 19th century, Hilário de Gouvêa, a Brazilian doctor, was the first to propose that an inherited intrinsic factor might be accountable for retinoblastoma, a rare eye cancer (Monteiro and Waizbort, 2007). For almost a century, it has been known that cancer has genetic and environmental causes. However, it was in 1971 with Alfred

G. Knudson's "two-hit hypothesis" when scientists focused on the genetic basis of cancer and started to realise that disruption of normal growth is caused by mutations in oncogenes or tumour suppressor genes (**Figure 1.1**) (Knudson, 1971). In 1990, Bert Vogelstein with his discoveries in colorectal cancer, demonstrated cancer's genetic diversity and suggested that cancer arises from accumulation of sequential mutations in a cell (Fearon and Vogelstein, 1990).

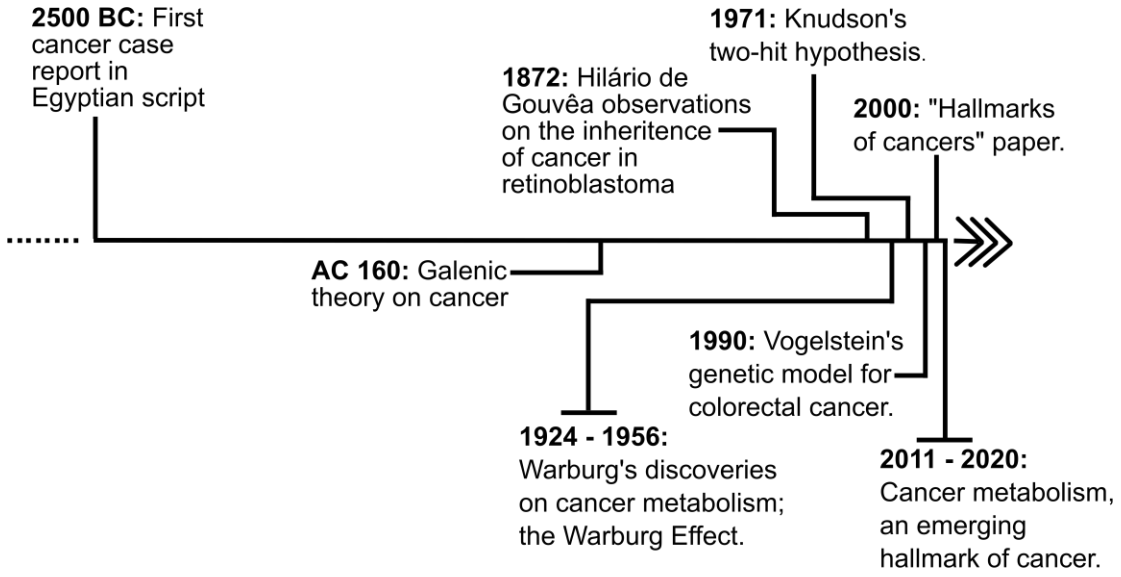


Figure 1.1. Cancer metabolism, an emerging hallmark of cancer. The arrow represents the most significant events and discoveries, overtime, that led to considered metabolism as a hallmark of cancer.

At the turn of the new millennium, Hanahan and Weinberg published one of the most cited papers in *Cell* scientific journal, where they highlighted six hallmark capabilities (**Figure 1.2**) that normal cells can acquire in order to become cancerous (Hanahan and Weinberg, 2000). The most essential of these is the ability of cancer cells to maintain chronic proliferation by disrupting growth-promoting signals and cell cycle. A second hallmark is the inactivation of tumour suppressor genes (e.g. RB, TP53) that act as gatekeepers for cell proliferation. Another feature of some cancer cells is an ability to resist cell death, such as apoptosis. Moreover, cancer cell capability to avoid senescence and gain immortalization is also a hallmark. Cellular senescence is the loss of proliferative ability of normal cells due to the shortening of their telomeric DNA. The capacity of a tumour to induce angiogenesis in order to maintain neoplastic progression, is also consider another hallmark of cancer. The final trait is the invasion of other tissues and the metastatic mechanisms that cancer cells develop in the later stages of tumourigenesis. The same authors in 2011, proposed that the reprogramming of cellular metabolism in cancer, together with the capability of cancer cells to evade immune destruction, are also two new emerging hallmarks (Hanahan and Weinberg, 2011). Since then, cancer research achieved remarkable progress to understand and elaborate the mechanisms that establish cancer metabolism as a hallmark of cancer.

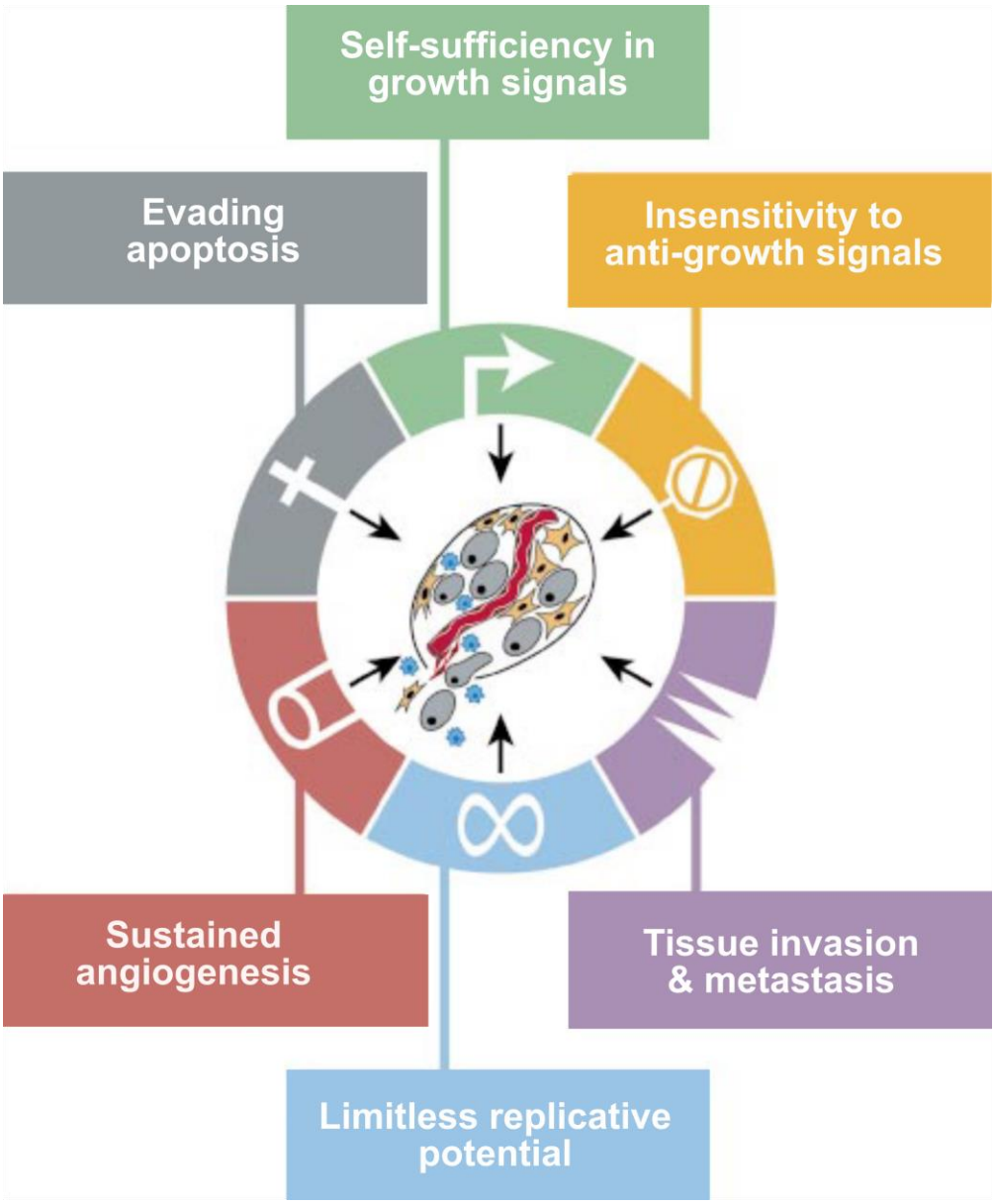


Figure 1.2. The first six hallmark capabilities acquired of cancer (Hanahan and Weinberg, 2000).

1.2. Cancer Metabolism, a hallmark of cancer

1.2.1. Deregulated bioenergetic profile

1.2.1.1. Glycolysis and the Warburg Effect

Cancer cells reprogram their normal cellular metabolism to fulfil their high energy demands and anabolic requirements for continuous cell growth and proliferation. Otto Warburg in the 1920s was the first to observe this metabolic deregulation in cancer. He reported that tumours consumed more glucose compared to other normal tissues (Warburg, 1924; Warburg et al., 1927). Glucose is the most profuse monosaccharide in human body, which is catabolised to generate energy in the form of adenosine triphosphate (ATP), in a process that is called cellular respiration. Generally, cells import glucose through GLUT protein transporters and extract energy via glycolysis by convert it into two three-carbon molecules called pyruvates. Glycolysis consists of ten enzyme-catalysed reactions which generate ten primary metabolites: glucose-6-phosphate, fructose-6-phosphate, fructose-1,6-bisphosphate, dihydroxyacetone-phosphate, glyceraldehyde-3-phopshate, 1,3-bisphosphoglycerate, 3-phosphoglycerate, 2-phosphoglycerate, phosphoenolpyruvate and pyruvate. Glycolysis is distinguished into an energy-requiring phase followed by an energy-releasing phase. During the energy-requiring phase, two ATP molecules are required to breakdown glucose

through a series of reactions into glyceraldehyde-3-phopshate. At the energy releasing phase, two ATP molecules and one NADH molecule is produced when glyceraldehyde-3-phospate is converted ultimately into pyruvate. Since this phase is required twice to convert one molecule of glucose into two molecules of pyruvate, the net products of glycolysis are two molecules of ATP and two molecules of NADH. Pyruvate then degrades in mitochondria with the consumption of oxygen via the tricarboxylic acid cycle (TCA) and the electron transport chain (ETC), where oxidative phosphorylation (OXPHOS) produce ample amount of ATP (**Figure 1.3**). Alternatively, pyruvate can also be converted to lactate under anaerobic conditions, in almost an 18-fold less productive way for energy. Warburg later discovered that even under aerobic conditions, cancers cells convert pyruvate to lactate, which has been termed as aerobic glycolysis or “Warburg Effect” (Warburg, 1956). This less energetically efficient metabolic reprogramming of cancer cells has puzzled scientists for several decades and even today the main reasons remain not fully understood. Still, this knowledge laid the foundation for the successful clinical use of radiolabelled glucose analogues (e.g. ^{18}F-FDG), as radiotracers in positron emission tomography (PET) to diagnose and monitor several different types of cancer (Duhaylongsod et al., 1995).

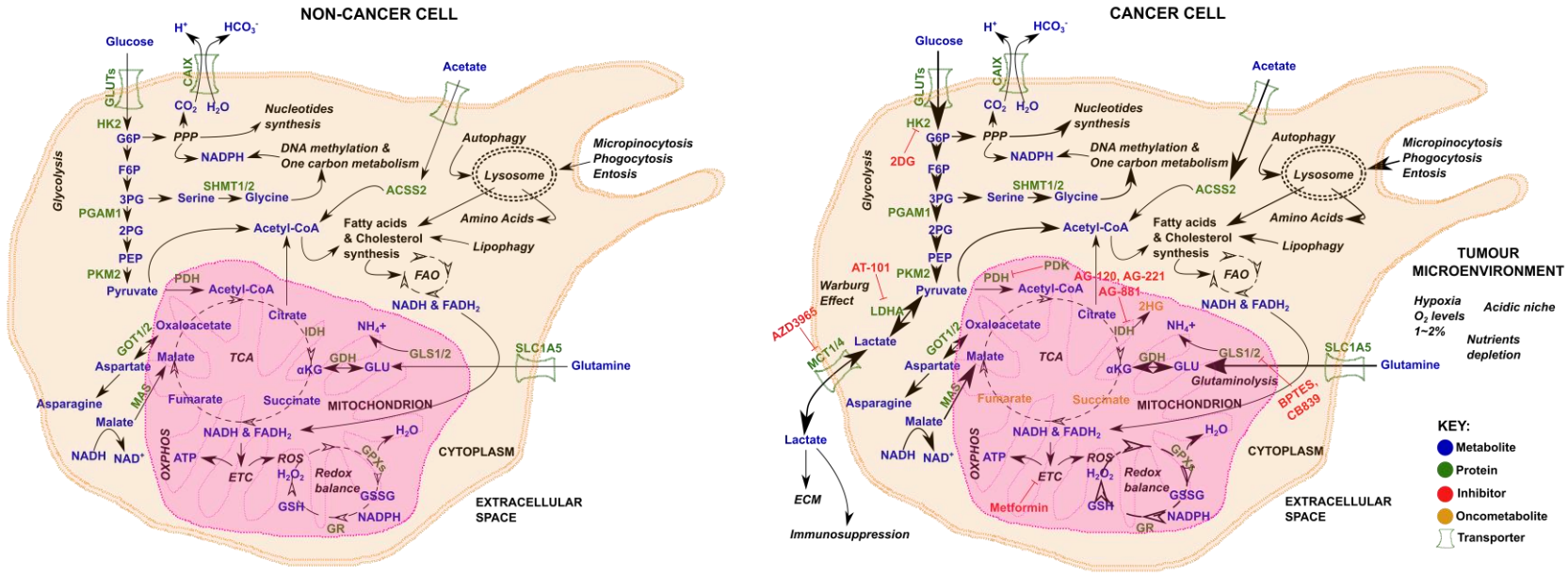


Figure 1.3Key metabolic pathways in a cancer cell. Intrinsic and extrinsic factors that alter the bioenergetic and biosynthetic metabolic profile of a cancer cell to sustain tumour growth and proliferation. Bold arrows represent changes in metabolic fluxes in a cancer cell compared to a non-cancer cell.

Today, our understanding is that cancer cells rely on both intrinsic and extrinsic oncogenic signalling factors to reprogram metabolism and sustain tumour growth and proliferation (Vander Heiden and DeBerardinis, 2017). Intrinsic factors involve any intracellular activities and effectors, such as genetic alterations in metabolic enzymes or elevated expression in transcriptional targets (MYC, KRAS and mTORC1), that carry out conventional metabolic tasks like supporting energetics, generating macromolecules and maintaining redox state for tumour progression. An example are mutations in succinate dehydrogenase (SDH) enzymes that result in the accumulation of succinate, which acts at high levels as an oncometabolite and interferes with dioxygenase function (Selak et al., 2005). Extrinsic factors involve processes outside of the cell membrane, such as access to nutrients and oxygen, attachment to extracellular matrix, interactions with stromal cells and exposure to radiation or chemotherapy. Examples of how cell metabolism is affected by extrinsic factors are presented in section 1.2.4.

The Warburg Effect not only fuels part of cancer cell's energy metabolism, but also activates and supplies the essential substrates for several conjoining biosynthetic pathways through upregulation of glycolysis (Lunt and Vander Heiden, 2011); contributes to the NAD^+ pool with the use of lactate dehydrogenase A (LDHA); exports lactate, which influences the tumour microenvironment promoting metastasis (Gottfried et al., 2006; Walenta et al., 2000); and assists cancers cells to regulate redox homeostasis in tumours' hypoxic environment by decreasing OXPHOS dependency (Gwangwa et al., 2018). Nevertheless, there are some normal cells and tissues, such as activated T-cells or the embryonic tissues, that

are also using the Warburg Effect to support rapid proliferation (Cham and Gajewski, 2005). Notably, the expression of numerous glycolytic enzymes, that are involved in this mechanism is controlled by oncoproteins such as c-Myc, HIF-1 and NEK2 (Kim et al., 2004). Therefore, overexpression of these transcription factors in several cancers alters the cellular metabolism and elevates glycolysis. More importantly, constitutive activation of tyrosine kinase signalling results in the activation of Warburg Effect by the constant phosphorylation of many glycolytic enzymes such as the LDHA, the phosphoglycerate mutase 1 (PGAM1), and the pyruvate kinase M2 isoform (PKM2) (Wiese and Hitosugi, 2018).

1.2.1.2. Alternative energy sources and mitochondrial respiration

Enhanced glycolysis with the Warburg Effect is one way for cancer cells to generate energy. However, several other catabolic reactions are also reprogrammed in cancer to fuel tumour cells with the appropriate energy and sustain survival and proliferation. It is now well established that cancer cells obtain part of their cellular energy from the oxidation of glucose, glutamine and from other nutrients that produce the precursors to initiate the TCA cycle and OXPHOS (Kim, 2018; Koppenol et al., 2011). Glutamine is the most abundant non-essential amino acid in the human body (Vinnars et al., 1975). After entering the cells through the SLC1A5 (or ASCT2) transporter, glutamine is converted to glutamate (GLU) and ammonia (NH_4+) by glutaminases (GLS1/2). Then, the glutamate dehydrogenase enzyme (GDH) metabolise glutamate to α-ketoglutarate (αKG), which is

channelled into the TCA cycle and produce NADH and $FADH_2$, the intermediates for ATP production from the ETC (**Figure 1.3**). This anaplerosis of the TCA cycle, or else, the replenishing of TCA cycle intermediates from glutamine catabolism (also known as glutaminolysis), is one of the major metabolic reprogramming events in several cancers (McKeehan, 1982).

Furthermore, recent studies suggest that the production of abundant cytosolic NADH can be used as an electron source for the ETC from cancer cells (Kang et al., 2016; Lee et al., 2016). NADH is generated in the cytoplasm as a by-product from several catabolic reactions and enters into mitochondria where OXPHOS occurs through the malate-aspartate shuttle (MAS). MAS is a multi-step procedure, in which the high energy electrons from the NADH are using malate as a “vehicle” to enter mitochondria. In TCA cycle, malate is converted to oxaloacetate, giving rise to NADH and thus ATP through the ETC (Greenhouse and Lehnlnger, 1976).

Another theory relies on a symbiotic model in tumours in which some cancer cells secrete lactate (that derives from the consumption of glucose) for neighbouring cells to consume (Faubert et al., 2017; Sonveaux et al., 2008). Lactate enters the cells through the monocarboxylate transporters (MCT1/4) and it is converted back to pyruvate by the LDHA enzyme. Pyruvate can then generate acetyl-CoA for fatty acids synthesis or fuels the TCA cycle. An alternative source that contributes in cytosolic acetyl-CoA pool is the catabolism of acetate. It is found that several cancer tissues upregulate the enzyme acetyl-CoA synthetase 2

(ACSS2) under hypoxia or nutrient-limiting conditions (Kamphorst et al., 2014; Mashimo et al., 2014). The ACSS2 synthesise acetyl-CoA from the ligation of acetate and CoA to support biomass production or to produce energy (**Figure 1.3**) (Schug et al., 2015).

Finally, fatty acids oxidation (FAO; or β-oxidation) also serves as an energy source in different cancer types. Fatty acids are required for membrane synthesis in cells and thus are necessary for cell growth and proliferation. They can be synthesised *de novo* from acetyl-CoA, imported from extracellular matrix, or by the degradation of intracellular lipid droplets, an autophagic process that is called lipophagy (Singh et al., 2009). Elevated uptake of fatty acids in cancer cells is used not only to maintain lipid homeostasis and prevent lipotoxicity, but also through FAO provide an extra source of ATP during conditions of metabolic stress (Koundouros and Poulogiannis, 2020). The FAO pathway is a cycle that generates NADH and $FADH_2$ by removing two carbons from fatty acids in each round. As mentioned before, NADH and $FADH_2$ enter in ETC to produce ATP (Carracedo et al., 2013). Recent evidence demonstrate that activation of FAO is essential for cancer cell survival and proliferation, particularly in haematological cancers (Monti et al., 2005; Samudio et al., 2010).

1.2.2. Biosynthetic reprogramming

The ability of cells to uptake nutrients not only covers the cellular energy demands but also provides the metabolic precursors for biosynthesis of proteins, lipids and nucleotides. Cells are using amino acids to build proteins, acetyl-CoA for lipids, and purines and pyrimidines for nucleotides, which are the building blocks for the nucleic acids (DNA and RNA). Multitudinous anabolic reactions are involved in the *de novo* synthesis of these essential molecules to support cell growth and proliferation. Glucose, besides a major energy supplier is also a carbon provider for biosynthesis. Along glycolysis, glucose is metabolised in several metabolic precursors used by other branching biosynthetic pathways (**Figure 1.3**). The glucose 6-phosphate (G6P), an intermediate metabolite in the steps of glycolysis, is utilised from the pentose phosphate pathway (PPP) for the formation of ribose 5-phosphate and ultimately the synthesis of nucleotides. During this process NADPH is generated and it is used for reductive biosynthesis reactions or to prevent cells from oxidative stress (Wamelink et al., 2008). The 3-phosphoglycerate (3PG) also derives from glycolysis and it is used for the *de novo* production of serine, a non-essential amino acid. Following, serine can convert to glycine by the serine hydromethylotransferases (SHMT1 and SHMT2), in a reaction that contributes the most in the one-carbon pool for nucleotide synthesis and methylation (Labuschagne et al., 2014).

In cancer cells, the part of pyruvate that is not converted to lactate, passes in the mitochondria and in the form of acetyl-CoA, fuels the TCA cycle both for

bioenergetic and other anabolic purposes. The TCA cycle generates citrate, αKG, succinate, fumarate, malate, and oxaloacetate (**Figure 1.3**). The isocitrate dehydrogenase (IDH) enzyme is responsible for converting citrate to αKG. Cancer cells with mutations in IDH genes, disrupt the normal enzymatic activity and result in the production of a “new” metabolite that is called 2-hydroxyglutarate (2HG) (Dang et al., 2009). The 2HG together with succinate and fumarate are characterised as onco-metabolites and they can regulate the epigenome by inhibiting histone or DNA demethylases (Nowicki and Gottlieb, 2015). Alternatively, acetyl-CoA and citrate can export from mitochondria to support fatty acids and cholesterol biosynthesis (Kato et al., 2018). Moreover, glutaminolysis is also a major contributor in multiple biosynthetic pathways in cancer. Metabolism of glutamine enriches cell’s nitrogen pools for the synthesis of purines and pyrimidines, to detoxify ammonia, and to activate mTOR signalling (Bott et al., 2019). In mitochondria, the anaplerosis of the TCA cycle via glutaminolysis produces malate that generates pyruvate and NAPDH, oxaloacetate (which can be converted to aspartate), and asparagine (via the aspartate transaminases GOT1 and GOT2) (Vazquez et al., 2016), and citrate (for lipid synthesis) (**Figure 1.3**) (Kato et al., 2018).

The rapid proliferation of cancer cells and the mitochondrial metabolism enhance the intracellular levels of reactive oxygen species (ROS) (Murphy, 2009). The production of ROS such as hydrogen peroxide (H_2O_2), a major by-product of mitochondrial oxidative phosphorylation, quickly oxidise nucleotides, proteins and

lipids leading the cell to apoptosis. To avoid this, cancer cells increase their antioxidant capacity through glutathione (GSH) oxidation-reduction and generation of NADPH from PPP and one carbon metabolism. GSH is synthesized from cysteine, glutamate, and glycine. The enzymes glutathione peroxidases (GPXs) can detoxify H_2O_2 to H_2O by generating glutathione disulphide (GSSG), which interacts with NADPH back to glutathione in a reaction catalysed by glutathione reductase (GR) (**Figure 1.3**) (Cox et al., 2009). Therefore, ROS detoxification through the reduction of GSSG back to GSH coupled to NADPH oxidation requires a constant supply of NADPH. As stated previously, various pathways contribute to NADPH production from $NADP^+$. The main source of NADPH comes from glucose at the first step of PPP, by glucose-6-phosphate dehydrogenase (G6PDH). Other sources can be serine via one carbon metabolism or the malate dehydrogenase enzyme, which oxidizes malate to pyruvate, while $NADP^+$ is reduced to NADPH. Similar to malate dehydrogenase, the isocitrate dehydrogenase (IDH) and the glutamate dehydrogenase (GDH) also generate NADPH. Besides ROS detoxification, NADPH contributes in many biosynthetic reactions and anabolic pathways, such as cholesterol synthesis, fatty acid synthesis, ascorbic acid synthesis or steroid synthesis.

Overall, it is becoming apparent that the role of mitochondria in malignant cells is altered to serve more as a producer of the anabolic intermediates for biosynthesis and redox homeostasis than a power generator of ATP.

1.2.3. Nutrients' acquisition

As the tumour expands, it creates areas with insufficient nutrients supply due to the limited vascularity. Cancer cells in these areas maintain survival by enabling autophagy (Guo et al., 2011; Yang et al., 2011). This mechanism is also used by normal cells to degrade and recycle the malfunctioning organelles via recruitment of the autophagosome and their fusion in the lysosomes, providing precursors to support the bioenergetic and biosynthetic cellular needs (Settembre and Ballabio, 2014). However, this process is not sufficient to maintain proliferation. For this purpose, cancer cells are using alternative mechanisms to obtain the necessary nutrients from their environment. As stated previously, small molecules such as glucose or glutamine are imported through the upregulation of membrane transporters. On the other hand, larger molecules such as proteins are recovered via endocytic mechanisms involving the micropinocytosis, the phagocytosis, and the entosis (Commisso et al., 2013; Krajcovic et al., 2013). These mechanisms are using the cellular membrane to engulf large amounts of nutrients or even cells and transfers them inside the cell. In the cell the lysosomes are responsible to break them down into their building blocks and fuel cell's metabolism (**Figure 1.3**). The oncogenic Ras proteins and the mTORC1 signalling pathways are associated with the regulation of these mechanisms, particularly in micropinocytosis. However, the exact mechanism behind this regulation remains poorly understood (Commisso et al., 2013; Kamphorst et al., 2015; Palm et al., 2015). Remarkably, even under limited nutrients delivery conditions, tumours utilise both intracellular and

extracellular macromolecules, to gain the advantage and sustain survival and proliferation.

1.2.4. Tumour microenvironmental factors

Tumours consist of diverse and heterogenic populations of cancer cells, creating areas with poor angiogenesis and blood supply as they expand. As a result, limited oxygenation occurs, thus generating a hypoxic environment, particularly in the core of the tumour. Consequently, most cancer cells survive in a hypoxic environment in between 1% to 2% oxygen concentration, whilst most normal tissues require a range between 4.6% to 9.5% (Muz et al., 2015). Consequently, several biosynthetic reactions that are using the molecular oxygen as an electron acceptor are suppressed in tumour cells. The metabolic response system in hypoxia activates HIF-1A and triggers the Warburg effect by inducing the expression of GLUTs and several other glycolytic enzymes, such as the pyruvate dehydrogenase kinase (PDK) (causing the phosphorylation and inactivation of the PDH) (Lu et al., 2002; Papandreou et al., 2006). The high production and secretion of lactate, as a result of elevated glycolysis, increases the levels of this metabolite extracellularly and creates an acidic microenvironment that impacts the extracellular matrix (ECM) and nearby cells. However, high levels of lactate and hypoxia promote immunosuppression by decreasing the activation and function of several immune cells, such as the dendritic and T cells (Fischer et

al., 2007; Gottfried et al., 2006). To this end, the release of glutamate from cancer cells, as a response to elevated glutaminolysis, is also found to regulate T-cell activation (Pacheco et al., 2006). Additionally, lactate efforts to sustain angiogenesis by stabilizing the HIF-1A expression and stimulating the expression of the vascular endothelial growth factor (VEGF), a pro-angiogenic molecule to promote vascularisation, in the neighbour endothelial cells (Sonveaux et al., 2012; Tang et al., 2004). The acidic niche in tumours is also enhanced by the release of intracellular H^+ and HCO_3^- derived from the catalysis of CO_2 (generated from PPP pathway) with H_2O to produce H_2CO_3 (Swietach et al., 2007). Among the carbonic anhydrases (CAs) that are responsible for this reaction, the CAIX isoform is elevated in several cancers and it is associated with hypoxia and tumour invasion (İlie et al., 2010; Yang et al., 2015). Altogether, several tumour microenvironment factors are contributing in the tumour niche, which also influence the metabolism of the surrounding cells to promote tumour survival, progression and metastasis.

1.2.5. Targeting cancer metabolism

The metabolic reprogramming of cancer cells is an emerging field for cancer therapy. The most promising therapeutic interventions on metabolic enzymes that are currently being accessed for their effectiveness and toxicity in clinical trials are highlighted in **Figure 1.3**. Research in cancer metabolism aims to find and target unique metabolic elements of tumours, such as the production of lactate, to

increase the specificity of a new therapeutic approach. To this direction, the AT-101 (Gossypol) molecule, which is an LDHA inhibitor, has demonstrated little efficiency in clinical trials to date (Fiveash et al., 2009; Sacco et al., 2014; Schelman et al., 2015). Another target is the MCT1 transporter, that is regulated by c-Myc and can facilitate both import and export of lactate. This transporter is inhibited by the AZD3965 agent, which is currently in phase I of clinical trials for Non-Hodgkin lymphoma (NHL), with an estimated completion date of May 2021 (NCT01791595). Pre-clinical studies have demonstrated the anti-tumour activity of the AZD3965 *in vitro* and *in vivo* in xenograft models (Curtis et al., 2017; Noble et al., 2017). However, Beloueche-Babari et al. have shown that inhibition with AZD3965 increases the mitochondrial metabolism in cancer cells and they suggested the combined use of AZD3965 with metformin or with the mitochondrial pyruvate carrier inhibitor UK5099 (PF-1005023), as a more effective cancer-targeted therapy (Beloueche-Babari et al., 2017). Metformin is a commonly used drug in type 2 diabetes, acting by reducing hepatic gluconeogenesis and regulating insulin levels in the blood stream. Interestingly, numerous studies highlighted the antitumorigenic effects of metformin (Dowling et al., 2012; Evans et al., 2005; Storozhuk et al., 2013). Beside regulating the circulating glucose and insulin levels, it was proven that metformin decreases the mitochondrial ATP production by inhibiting the mitochondrial ETC complex I (Owen et al., 2000; Wheaton et al., 2014). Currently, several clinical trials are evaluating the combined use of metformin with conventional cancer treatment as an improved therapeutic strategy (Saraei et al., 2019).

The bioenergetic profile of cancer cells, mainly supported from elevated glycolysis and glutaminolysis, is another popular target. In human cells, fluxes in the glycolytic pathway are regulated mostly by three regulatory enzymes (hexokinase, phosphofructokinase, and pyruvate kinase), as a response to intracellular and extracellular signals. The first step of glycolysis is to convert the intracellular glucose to G6P by utilizing the hexokinase (HK2) enzyme. In HIF1-A and c-Myc driven cancer cells, several glycolytic enzymes including the HK2 are upregulated to promote glycolysis (Kim et al., 2007). HK2 is inhibited by a glucoses mimetic, the 2-deoxy-D-glucose (2DG). Although, it is known since 1958 that the 2DG molecule is able to reduce the number of leukemic cells in patients, this molecule failed to proceed further than phase 1 of clinical trials (Landau et al., 1958; Stein et al., 2010). This is because the 2DG is antagonizing glucose, which has a concentration of 60-fold more (7 to 10 mg/ml) than the concentration (0.116 mg/ml) of the maximum tolerated dose of 2DG (Raez et al., 2013). Still, the use of 2DG alone or in combination with other cancer treatments, appears to provide a clinical benefit in cancers with elevated glycolysis. Other important enzymes involved in the regulation of glycolysis are the phosphofructokinase and the pyruvate kinase. Phosphorylation of these enzymes, as a response to intracellular signals, can drive either glycolysis or gluconeogenesis.

To target glutaminolysis, inhibitors for the GLSs enzymes have been developed and tested in glutamine-addicted cancers. The bis-2-(5-phenylacetamido-1,2,4-thiadiazol-2-yl)ethyl sulphide (BIPTES) and the CB-839

(Telaglenastat) are two promising molecules that are inhibiting the action of the GLS1 protein. Studies with the CB-839 molecule have shown that this agent can reduce tumour cell growth in patient derived xenografts and it is now being assessed in phase I clinical trials for advance/metastatic solid tumours and haematological malignancies (Gross et al., 2014; NCT02071888; NCT03875313). Similarly, the BIPTES molecule is also found to suppress tumour growth in IDH1 mutant cancer cells and in xenografts (Le et al., 2012; Seltzer et al., 2010; Xiang et al., 2015). Cancers with IDH mutations are also being investigated with inhibitors that aim to suppress the production of 2HG oncometabolite by blocking the mutant IDH enzymes (Golub et al., 2019). Recently, the IDH1 inhibitor AT-120 (Ivosidenib) and the IDH2 inhibitor AG-221 (Enasidenib) have been FDA-approved for treatment of patients with relapsed/refractory acute myeloid leukaemia (AML) (Dhillon, 2018; Dugan and Pollyea, 2018). Additionally, preliminary results from a phase I clinical study with another inhibitor the AG-881 (Vorasidenib, targets both mutated IDH1 or IDH2 enzyme) demonstrate 2HG suppression with >90% reduction in patients with glioma tumours (NCT02481154). Despite the significant progress of the above pharmaceutical inhibitors, targeting cancer metabolisms requires a much better understanding of the involved metabolic and signalling pathways to avoid cytotoxicity in the normal proliferative tissues and increase the survival rate of cancer patients.

1.3 Haematological malignancies

1.3.1 Overview and classification

Haematological malignancies or blood cancers are tumours derived from the haematopoietic system. This system contains organs and tissues, including the bone marrow, spleen, thymus and lymph nodes, involved in the production of blood and lymph cellular components. Blood contains red blood cells (erythrocytes), white blood cells (leukocytes), platelets (thrombocytes) and plasma, and it is responsible for tissue oxygenation, immunity and haemostasis (wound healing). Lymph consists of interstitial fluid (or tissue fluid) with high number of lymphocytes (types of white blood cells), such as natural killer (NK) cells, B-cells and T-cells. Most of the blood and lymph cells are generated from haematopoietic stem cells in the bone marrow, a spongy tissue in the centre of bones. In the bone marrow, haematopoietic stem cells reproduce themselves and differentiate into myeloid or lymphoid precursor cells in a process known as haematopoiesis (Jagannathan-Bogdan and Zon, 2013). Several haematopoietic growth factors (such as interleukins) regulate the differentiation of these myeloid and lymphoid precursors into any/all forms of mature blood cells. Myeloid precursors develop to erythrocytes, granulocytes and platelets, while lymphoid precursors mostly generate NK, B-cells and T-cells (**Figure 1.4**).

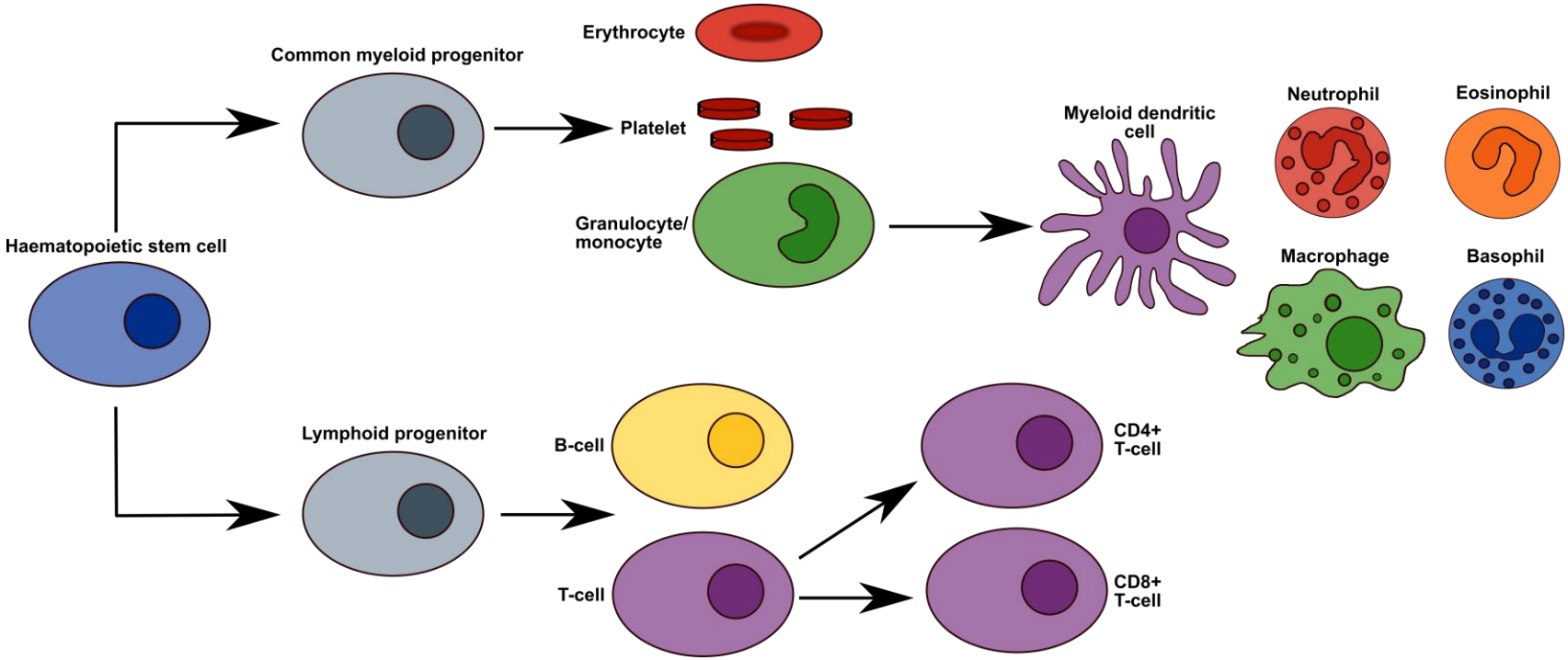


Figure 1.4 Haematopoiesis. The formation of all forms of mature blood cells derived from pluripotent haematopoietic stem cells that originate in the bone marrow (Jagannathan-Bogdan and Zon, 2013).

In general, haematological malignancies refers to cancers of white blood cells and they are divided into leukaemia, lymphoma and myeloma. Leukaemia and myeloma are cancers that derive mainly from blood cells in the bone marrow, while lymphoma are cancers of lymphocytes in the lymphatic system. According to the National Cancer Registration and Analysis Service, there were 33,719 cases and 12,767 deaths of haematological cancers reported within 2017 in England alone (NCRAS, 2017). Although new developments in therapeutics have decreased the morality rates of haematological cancers, the incidence rates for most of these diseases are rising or remain the same (NCIN report, 2014). The Haematological Malignancy Research Network expects that 44,160 new cases occur in the UK each year (HMRN, 2019).

The classification of haematological malignancies is a complex procedure because of the great biological and clinical diversity that these tumours present. In the recent International Classification of Diseases 11th revision (ICD-11, version 04/2019) (WHO, 2019), haematological cancers are subcategorised as “Neoplasms of haematopoietic and lymphoid tissues” under the broad group of “Neoplasms”. Within this thesis, emphasis is given to diseases that belong to the “Mature B-cell neoplasms” (BL, DLBCL and CLL) category as highlighted in **Figure 1.5**.

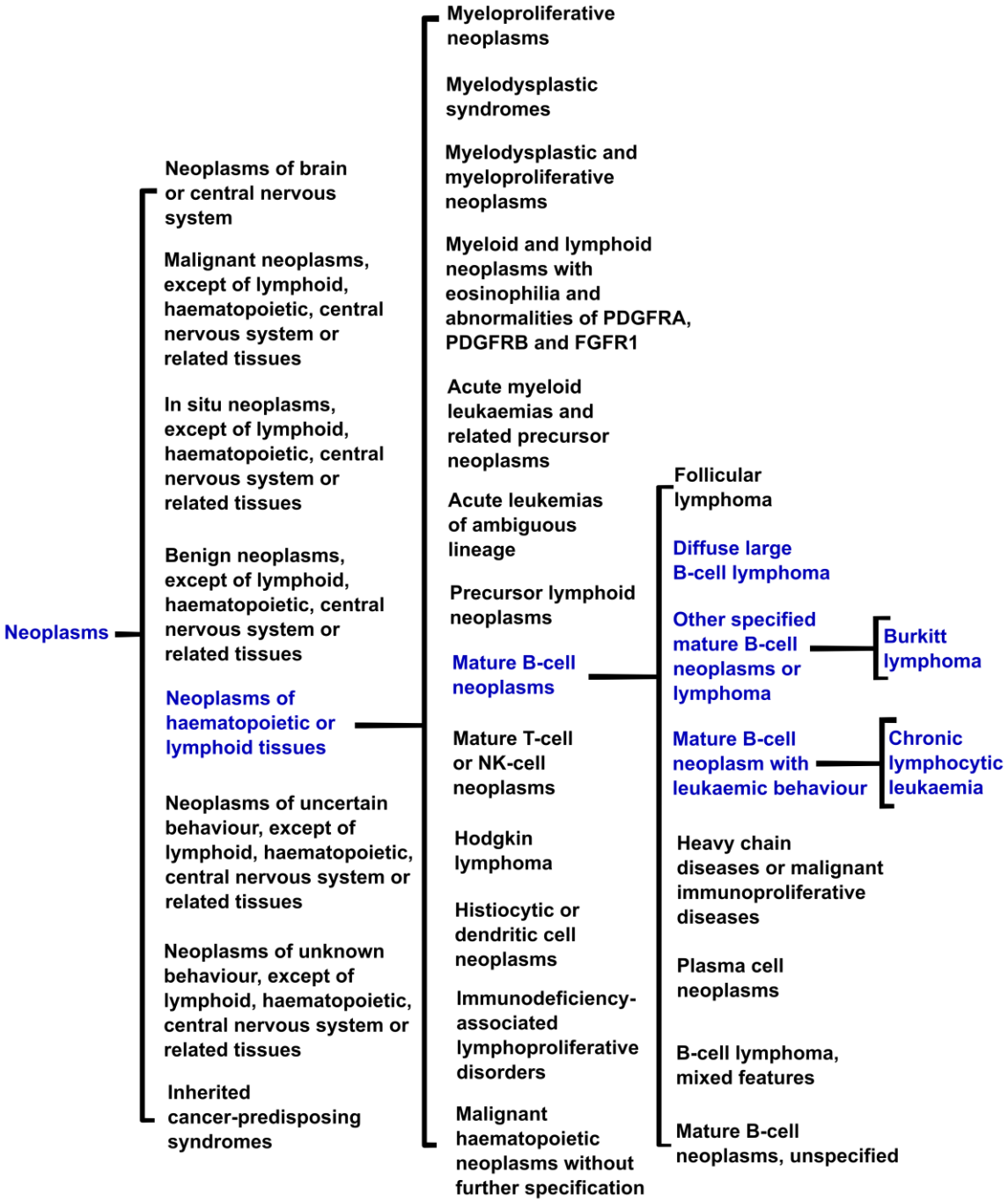


Figure 1.5. Haematological cancers classification. Disease groups are separated on the basis of the International Classification of Diseases 11th revision (WHO, 2019). Fonts with blue color indicate the disease groups that are examined in this thesis.

1.3.2 Mature B-cell neoplasms

Mature B-cell neoplasms are a broad category of haematological malignancies that rise from the clonal expansion of mature B-cells in the blood, bone marrow and secondary lymphoid organs, such as the spleen or the lymph nodes. B-cells are white blood cells with a pivotal role in the defence mechanism of the human body against diseases (Alberts et al., 2002). They are responsible for the secretion of high affinity antibodies, which recognise pathogens and trigger more advanced defence actions, known as adaptive immune responses. Naïve B-cells are produced in the bone marrow from lymphoid progenitor cells that derive from the haematopoietic stem cells. They move to secondary lymphoid organs, where their selection and differentiation occur in lymphoid follicle sites, known as the germinal centres (GCs). In GCs, B-cells development, selection and rapid proliferation involves multitudinous and complex procedures, such as somatic hypermutation and class-switch recombination (Mesin et al., 2016). This results in the production of antigen-secreting plasma cells and memory B-cells that exit the GCs to support adaptive immunity. Any disruptions in DNA damage and cell transformation checkpoints during B-cell maturation can cause lymphomagenesis and generate mature B-cell neoplasms.

The majority of mature B-cell neoplasms are lymphomas, such as NHL that represent the larger group of lymphoid neoplasms. According to the Global Cancer Observatory, in 2018 NHL were responsible for 509,590 new cancer cases worldwide, making it the leading cause of haematological cancers (WHO, 2018).

NHL consist of several subtypes with different morphologic, immunophenotypic, genetic and clinical characteristics. Based on the progress of the disease, they can be separated as indolent or aggressive lymphomas. The indolent lymphomas, such as the Follicular lymphomas (FL) or the Marginal Zone lymphomas, show few symptoms and evolve slowly. Inversely, the aggressive NHL like Diffuse large B-cell lymphoma (DLBCL) or the Burkitt lymphoma (BL), tend to spread quickly with serious symptoms and rapidly become fatal if remain untreated (Armitage et al., 2017). Another common characteristic between FL, BL and DLBCL is that all of them originate from GCs malignant B cells and therefore these lymphomas are also known as GC-derived B-cell lymphomas (Mlynarczyk et al., 2019). GC-derived B-cell lymphomas are highly diverse tumours with a broad spectrum of genomic, epigenetic and metabolic profiles. This thesis investigates the aggressive NHL that are also GC-derived B-cell lymphomas and hence it describes works on BL and DLBCL cases.

Although NHL represent the majority of haematological malignancies, leukaemia are still the most lethal, being accountable for 309,006 deaths in 2018 globally (WHO, 2018). Lymphoid leukaemia, such as the Chronic Lymphocytic Leukaemia (CLL) or the Hairy Cell Leukaemia, also rise during the maturation procedure of B-cells, thus sharing common characteristics during malignant transformation with other mature B-cell neoplasms. Herein, lymphoid leukaemia tumours were also investigated by examining CLL cases.

1.3.2.1 Burkitt Lymphoma

Burkitt Lymphomas (BL) are highly aggressive NHL that derive from the GCs development and most commonly occur in children. Three main types of BL exist: endemic, sporadic and immunodeficiency-associated. BL are generally characterised by a monotonous infiltrate of medium-sized blastic lymphoid cells that display round nuclei with clumped chromatin and multiple nucleoli. These tumour cells are characterised by high proliferation rate (95% or higher with Ki-65 staining) and high rate of cell death or apoptosis leading to a morphological pattern termed ‘starry sky’ (Rosenwald and Ott, 2008). The BL cells immunophenotypic profile shows similarities with GC cells being positive for CD20, CD10, BCL6 and negative for Mum-1, CD44, CD138 and BCL2 antibodies (Schmitz et al., 2014). Positivity for Epstein-Barr virus (EBV) infection is a hallmark of BL, as it is found in 98% of the endemic and 20% of sporadic BL cases (Dave et al., 2006). At the genetic level, several translocations involving the oncogenic transcription factor c-Myc (translocation t(8;14)(q24;q32)) are common to all subtypes of BL (Bellan et al., 2009; Gerbitz et al., 1999). c-Myc influences several cell functions such as cell cycle, DNA damage, protein synthesis and metabolism. Additionally, genome sequence studies have linked BL with cells in the dark zone of GCs by identifying mutation in the transcription factor 3 (TCF3) and in ID3 gene (but not in DLBCL) (Schmitz et al., 2012; the ICGC MMML-Seq Project, 2012).

1.3.2.2 Diffuse Large B-cell Lymphoma

One in two NHL cases in the UK is caused by a Diffuse large B-cell lymphoma (DLBCL) tumour (HMRN, 2019). Most DLBCL cases are curable with chemotherapy (such as CHOP) or combined chemotherapy with immunotherapy (such as R-CHOP) (Armitage et al., 2017). The diagnosis of DLBCL occurs by the presence of large neoplastic B-cells comprising centroblastic, immunoblastic, T-cell/histiocyte-rich and anaplastic morphological variants (Liu and Barta, 2019). Studies have shown that 5 to 15% of DLBCL cases are positive for EBV infection, while others have associated 5 to 10% of DLBCL cases with translocations in c-Myc (Castillo et al., 2016; Rosenwald and Ott, 2008). Many DLBCL cases with c-Myc translocations and either BCL2 or BCL6 mutations present a more aggressive clinical behaviour. High expression of BCL2 protein, usually due to translocation t(14;18), can inhibit apoptosis, giving a survival advantage to affected B-cells. In addition, BCL6 acts as a transcription repressor, protecting the cell from apoptosis (Rosenthal and Younes, 2017). Gene expression profiles (GEP) studies have classified DLBCL into two molecular subtypes: the germinal centre like group (GCB) and the activated B-cell like group (ABC), reflecting the derivation of B-cells based on their cell-of-origin when first oncogenesis occurred. Genetic alterations, molecular signalling pathways, and different clinical outcomes are associated with these subtypes (Campo et al., 2011). However, 10-15% of DLBCL cases do not classify with GEP into these two subgroups (Swerdlow et al., 2016). The GCB subgroup shows similar GEP with normal GC B-cells and is associated with a good

clinical outcome (Rosenwald et al., 2002). Moreover, several genetic alterations have been related with this subgroup. The most common chromosomal translocations in GCB-DLBCL cases involve the BCL2 and c-Myc translocations (Lenz et al., 2008; Rosenwald et al., 2002), whilst BCL6 translocations and inactivation of acetyltransferases genes CREBBP and EP300 can be found in both subgroups (Pasqualucci et al., 2011). Somatic mutations in histone-modifying genes and mutations of the EZH2 methyltransferase are also observed in GCB (Morin et al., 2010, 2011). In contrast, the more severe ABC DLBCL subtype of DLBCL seems to derive from post germinal centre B-cells that are arrested during plasmacytic differentiation. Among the most common genetic alterations in ABC group, are genetic defects in B-cell antigen receptor (BCR) and mutations in genes (TNFAIP3, CARD11, CD79B and MYD88) that enhance activation of the NF-kB pathway (Pasqualucci et al., 2011).

1.3.2.3 Chronic Lymphocytic Leukaemia

Chronic Lymphocytic Leukaemia (CLL) is the most common type of leukaemia in the Western world with around 3.800 new cases and 990 deaths in the UK every year (Cancer Research UK, 2015). The disease is characterised by the clonal proliferation of small mature-appearing B-cells (more than $5x10^9$/L) and mostly affects people over the age of 60 (Watson et al., 2008). At an early stage, the disease is usually asymptomatic, and it can be detected by a routine full (or complete) blood count test. This is a common blood test which evaluates the types

and quantities of cells in patients' blood, gives an indication about patients' general health and detects eventual signs of health problems. For example, a very high white blood cells counts can be a sign of leukaemia, like CLL, and can help early diagnosis of the disease. However, more specific cancer blood tests need to be applied at a second stage to fully characterise the type of cancer through detection of specific cancer biomarkers. In later stages, more severe symptoms such as lymphadenopathy, cytopenia, hepatomegaly or splenomegaly occur (Hallek et al., 2008). The time from diagnosis to disease progression can vary from months to decades.

CLL as a disease presents a wide clinical and biological heterogeneity. The clinical status varies from common indolent or progressive cases to rare regress cases. The disease prognosis is associated with the expression of CD38 or ZAP70 proteins, and with the mutational status of immunoglobulin heavy chain (IgHV) genes (Boonstra et al., 2006; Orchard et al., 2004; Stevenson et al., 2011). The genomic landscape in CLL includes several chromosomal aberrations (e.g. deletions in 13q, 17p, 11q or trisomy 12), recurrent mutations and somatic copy number variations that are affecting genes including ATM, TP53, NOTCH1, MYD88, SF3B1, FBXW7, POT1, CHD2, RPS15, IKZF3, ZNF292, ZMYM3, ARID1A and PTPN11 (Hallek, 2017; Puente et al., 2011; Stankovic and Skowronska, 2014). Survival of CLL cells putatively relies on resistance in apoptosis. Microenvironment signals and B-cell receptor signalling block pro-apoptotic factors or stimulate anti-apoptotic factors of BCL-2 and the IAP family

proteins, such as BCL-2, MCL-1, and BCL-XL (Billard, 2014).

Although, remarkable progress has been made to reveal the molecular and cellular mechanisms of CLL, the cell-of-origin status of CLL remains controversial. Seifert *et al.* utilized GEP in a study, suggesting that the IgHV mutated CLL cases are associated with the GC B-cells, while the unmutated CLL cases are associated with naïve B-cells (Seifert et al., 2012). Moreover, 2 to 8% of CLL cases tend to transform into DLBCL, known as “Richter’s syndrome” (Parikh et al., 2013). Similar to DLBCL, the main therapeutic strategy for CLL are pathway inhibitors (PIs), such as inhibitors of Bruton tyrosine kinase (BTK), phosphatidylinositol 3 kinase (PI3K) and BCL2 (Dreger et al., 2018), and CHOP-R for Richter’s syndrome.

1.4 Molecular Omics

Ample revolutionary innovations in both biomedicine and informatics have set the foundations for Computational Biology that combines the knowledge and the technologies of these two scientific disciplines. Utilizing Computational Biology methods can assist researchers in understanding biological systems and fighting diseases, such as cancer. Scientists are now using advanced computational and mathematical approaches to analyse, integrate and interpret large biological datasets that derive from Omics technologies such as genomics, transcriptomics, proteomics and metabolomics. Genomics is the systematic study of the whole

deoxyribonucleic acid (DNA) of an organism, the genome, whereas the transcriptomics is the study of the complete ribonucleic acid (RNA), the transcriptome. Similarly, proteomics investigates the total number of proteins in cells, tissues and biofluids, while metabolomics focusses on the metabolome, which is the sum of small molecules called metabolites. Today these are the most prevalent Omics studied, however this field is constantly expanding with new areas of biomedical sciences entering the field and gaining more attention such as epigenomics, fluxomics, microbiomics and drugomics. Today's challenge for Computational Biology is to utilise, and develop when necessary, the most sophisticated bioinformatics tools and biostatistics methods to integrate all the available Omics (or else multi-Omics) datasets and build a holistic picture of the biological mechanisms under investigation. This integrative multi-Omics approach is now sometimes referred to as integromics or panomics (Manzoni et al., 2018). Nevertheless, most Omics datasets are still generated independently rather than as an integrated concept and they require multi-disciplinary expertise for the analysis. Consequently, several issues are affecting the analysis, such as incomplete sampling across the datasets, missing features within the samples, and different types of experimental noise and error. Moreover, each datatype contains thousands or even up to a million features which comes with challenges to extract the most important biological information out of such high-dimensional data. For instance, biological and technical variation can contribute to unrelated features which antagonise (or dominate) the important features in high-dimensional space (Ronan et al., 2016). Then, multi-Omics integration of such data introduces an

extra layer of large variation that needs to be considered. Therefore, the integration of multi-Omics datasets is a complex and challenging task (discussed further in section 5.5). The emphasis of this thesis is on the integration of transcriptomics and metabolomics data derived from B-cell neoplasms with a focus on cancer metabolic reprogramming.

1.4.1 Transcriptomics

The transcriptome consists of categories of coding and non-coding RNA. The coding RNA derives from the transcription of genomic DNA (coding regions in genes) to messenger RNA (mRNA) and then to proteins. This represents only 1~4% of total RNA in a eukaryotic cell. The rest (>95%) is the non-coding RNA, with the most abundant forms being ribosomal RNA (rRNA) and transfer RNA (tRNA). Several other RNAs with catalytic functions, such as the small nuclear RNA (snRNA), the microRNA, the long non-coding (lncRNA) and the small interfering RNA (siRNA) are also members of the non-coding RNA category (Pevsner, 2015).

In recent decades, the need to identify the expression of genes, usually by comparing two or more conditions (i.e. disease vs healthy or control vs drug), has developed gene expression profile (GEP) studies, which in most cases are measuring the cytoplasmic mRNA transcript levels under a defined condition. Most GEP studies are employing the high-throughput technologies of RNA microarrays

or next generation sequencing (NGS) methods, such as RNA sequencing (RNAseq) and single-cell RNA sequencing (scRNAseq) to measure transcriptomic profiles. These technologies provide a broad picture of gene expression by extracting the total RNA from samples and converting it to complementary DNA (cDNA) before subsequent sequencing. In RNA microarrays, the cDNA is hybridized to a collection of probes (biochip) that are specific for a defined number of genes, whilst the NGS methods are based on cDNA fragmentation and library preparation, followed by sequencing and alignment to a reference genome (Pevsner, 2015). Although an RNA microarrays approach is a robust and economic option, it is limited to measuring only the expression of genes of a pre-designed biochip that is used. On the contrary, the RNAseq technology can provide a more comprehensive picture of RNA expression, allowing the identification of new transcript isoforms or even gene fusion events (Manzoni et al., 2018). Single-cell transcriptomic sequencing provides the advantage of also uncovering the cell-cell heterogeneity of expression between tissues or even cell types. Thus, the NGS methods have become established as the most preferential strategy on the generation and analysis of transcriptomics data for the latest GEP studies. All the transcriptomic profiles analysed in this thesis were generated with the RNAseq technology.

1.4.2 Metabolomics

As mentioned in the previous section, metabolomics is the systematic study of the metabolome in cells, tissues, biofluids or even in geo-climatic environments. The metabolome is defined as the complete set of metabolites, which are molecules that are the end products of enzymatic chemical reactions. Measuring the metabolome reveals the biochemical activity of an organism and assist to understand the effect of several environmental factors. Metabolomics is now increasingly used in biomedicine to identify prognostic biomarkers, discover new drug targets or even predict treatment responses (Alonso et al., 2015). Currently, the most common technologies to measure the metabolic profile of a biological sample are nuclear magnetic resonance (NMR) and mass spectroscopy (MS). NMR is a highly reproducible spectroscopy, which detects the electromagnetic signal arising from the spin of certain atomic nuclei (^{1}H, ^{13}C and ^{31}P) during or after a radiofrequency pulse inside a strong magnetic field. Most metabolomics studies are using the one-dimensional NMR (1D-NMR) to identify compounds and quantify their concentrations, while 2D-NMR is mainly used to obtain additional information of the structural variation of metabolites (Alonso et al., 2015; Dumas, 2012). In contrast, MS is a technique that is based on the ionisation of the biological sample, which is coupled to a chromatographic separation of metabolites in either liquid (LC-MS) or gas (GC-MS) chromatographic columns. MS gives the advantage of studying hundreds of known metabolites in biofluids or tissue samples. Metabolomics studies are using both these technologies in untargeted and

targeted approaches with the untargeted studies to aim in the discovery or generation of a hypothesis and the targeted studies on testing these specific hypothesis (Misra et al., 2019). Despite the major advances in these high-throughput technologies, at the current time only 15% of all 300,000 known metabolites can be identified, which restricts the information acquired compared to other Omics technologies. However, studying the metabolome allows deeper insights into the bio-molecular mechanisms behind disease pathogenesis and reveals new vulnerabilities for novel therapeutic strategies. To achieve this, metabolomic data from NMR were analysed in chapter 3 and data from LC-MS in chapter 4.

1.4.3 Transcriptomics-metabolomics integration

Metabolites are linked with enzymes into biochemical reactions which form biochemical pathways and create metabolic networks. These networks are regulated by various genetic and signalling interactions, as a response to several environmental factors. To analyse and understand these interactions, several web-based, GUI, or command-line computational tools have been developed based on statistical, machine learning, pathway-based or even Genome Scale Metabolic Modelling approaches. These tools provide well established workflows for pre-processing, visualization and integrative analysis of Omics data (**Table 1.1**). Following the success of several other integrative studies (Brial et al., 2019; Cazier

et al., 2012; Dumas et al., 2016), the integration of transcriptomics with metabolomics datasets unveils the role of metabolic malfunctions in human diseases and suggests novel therapeutic approaches. However, the integration between transcriptomics and metabolomics is quite complex. This is because there is no direct association between transcript and metabolite, due to post-transcriptional modifications and protein expression (Auslander et al., 2017). Overall, linking known metabolites with gene expression via their common metabolic reactions and pathways is a powerful tool to understand the role of metabolism in tumour progression and identify new metabolic drug targets.

Table 1.1. Selection of tools for Omics integration with metabolomics data.

Tools	Approach	Interface	Programming languages	Reference
MetaboAnalyst	Statistical/Pathway-based	Web-based	Java, R	(Xia et al., 2009)
integrOmics	Statistical	Command line	R	(Lê Cao et al., 2009)
PathVisio	Statistical	GUI	Java	(Kutmon et al., 2015)
IMPaLA	Statistical	Web-based	Python, SOAP/WSDL	(Kamburov et al., 2011)
MetScape	Statistical	GUI/Cytoscape app)	Java	(Gao et al., 2010)
MassTRIX	Statistical	Web-based	Perl	(Wägele et al., 2012)
COVAIN	Statistical	Command line	MATLAB, C	(Sun & Wolfram, 2012)
MeltDB	Statistical	Web-based	Perl	(Neuweger et al., 2008)
MetaMapp	Statistical	Command line	Javascript, R	(Barupal et al., 2012)
PiMP	Statistical	Web-based	Python	(Gloaguen et al., 2017)
MarVis-Pathway	Statistical	GUI	MATLAB	(Kaever et al., 2015)
Metabox	Statistical	Command line	R	(Wanichthanarak et al., 2017)
INIT	GSMM	Command line	MATLAB	(Agren et al., 2012)
iMAT	GSMM	Command line	MATLAB	(Zur et al. 2010)
GIMME	GSMM	Command line	MATLAB	(Becker & Palsson, 2008)
MetDisease	Pathway-based	GUI/Cytoscape app	Java	(Duren et al., 2014)
rMTA	GSMM	Command line	MATLAB	(Valcárcel et al., 2019a)
MetFlow	Statistical	Web-based	Java	(Shen and Zhu, 2019)
3Omics	Statistical	Web-based	Perl, PHP	(Kuo et al., 2013)
SyNDI	Statistical	GUI/Cytoscape app	Java	(Lindfors et al., 2018)
MetaBridge	Statistical	Web-based, Command line	Javascript, R	(Hinshaw et al., 2018)
Pathway Commons	Pathway-based	Web-based	Java	(Rodchenkov et al., 2019)
mQTL.NMR	Statistical	Command line	R	(Hedjazi et al., 2015)
OmicsNet	Pathway-based	Web-based	Javascript	(Zhou and Xia, 2018a)
MetaboSignal	Pathway-based	Command line	R	(Rodriguez-Martinez et al., 2017)
mixOmics/DIABLO	Statistical/Machine-Learning	Command line	R	(Rohart et al., 2017)
MapMan	Pathway-based	GUI	Java	(Thimm et al., 2004)

1.4.3.1 Genome scale metabolic modelling

Advances in the field of Systems Biology has enabled the development of in silico models to reconstruct the metabolism of different species, known as Genome Scale Metabolic Models (GSMMs). The use of GSMMs is now one of the most common approaches to simulate metabolism in industrial or medical research. These computational models are built by gene-protein-reaction (GPR) associations to create a stoichiometric matrix of metabolites and mass-balanced metabolic reactions for the whole metabolic network of an organism (Gu et al., 2019). Linear programming with flux balance analysis (FBA) is utilizing GSMMs with kinetic data for constrain-based modelling, in order to calculate the metabolic fluxes by maximizing a specific cellular process, such as the biomass reaction (Kauffman et al., 2003). Apart from the traditional FBA, GSMMs are also integrated with omics data to reconstruct metabolism in a condition of interest (Machado and Herrgård, 2014). Due to the widespread use of NGS datasets, most of the current computational tools (INIT, GIMMIE, iMAT) can now integrate transcriptomic data with GSMMs. This novelty, together with the development of human GSMMs, such as the Recon series (Duarte et al., 2007), allowed the reconstruction of condition-specific GSMMs to investigate metabolic alterations in various cancers or viral infections (Aller et al., 2018; Asgari et al., 2018; Bidkhori et al., 2018). More specifically, the Metabolic Transformation Algorithm (MTA), a computational method based on condition-specific GSMMs, has been developed and validated across numerous published perturbation experiments (Yizhak et al., 2013a). MTA

was firstly applied in cellular ageing to predict lifespan-extending genes in yeast (BY4741 strains) or to predict metabolic drug targets in human muscle tissue that can transform it back to its young state (Yizhak et al., 2013a). Besides ageing, MTA predicted metabolic drug targets for Alzheimer's disease (Stempler et al., 2014a) and colorectal cancer (Auslander et al., 2017a). Here, we have applied a robust version of MTA, named as rMTA (Valcárcel et al., 2019b), to predict metabolic vulnerabilities in CLL. As an example of this utility, rMTA was successfully applied with publicly available gene expression data from prostate cancer to highlight the regulatory role of PGC1α gene in tumour progression (Valcárcel et al., 2019b).

1.4.3.2 Pathway or network - based integration approaches

There are a wide range of statistical tools that provide normalization, statistical analysis and integration of metabolomics data. A common approach to integrate metabolomics with transcriptomics data is the pathway based-integration method. This approach is using the existing biological knowledge stored within online databases such as the Kyoto Encyclopaedia of Genes and Genomes (KEGG), to map metabolites and genes on duly annotated metabolic pathways. Most of the available tools are using either over representations analysis (ORA) or enrichment analysis to map metabolites in metabolic pathways. The enrichment methods are mostly based on gene set enrichment analysis (GSEA) (Subramanian et al., 2005), a quite popular method to analyse genomic data. GSEA is a

computational method where a set of genes is preferentially associated with known gene sets (e.g. a specific signalling pathway) and statistically significant differences between two biological conditions can be computed. More advanced tools are combining the expression of genes, proteins and metabolites with topological features in order to calculate the expression of significant biological pathways (Rodriguez-Martinez et al., 2016; Simillion et al., 2017; Zhou and Xia, 2018). Visualization of these significant pathways are usually demonstrated with networks, graphically represented with metabolites as nodes and the edges as the reactions involved. These graphs are highly informative to researchers to interpret transcriptional, translational and post-translational modifications that affect metabolism in a study condition. Pathway based integration approaches were used in this thesis to investigate key metabolic differences in NHL lymphomas.

1.4.3.3 Machine-learning approaches for multi-Omics data integration

The rapidly growing and accessible computational power enhanced the development and implementation of machine learning (ML) algorithms in a wide range of applications. Most of these algorithms rely on the utilization of the train/validate/test strategy to build a model that learns from the data and gives solutions to a particular problem. ML can handle classification and regression problems in a supervised, semi-supervised and unsupervised learning approach. Supervised learning depends on the collection of labelled samples to train a model which can then predict a label for a novel input sample. In contrast, unsupervised

learning approaches use mostly unlabelled samples to commonly perform dimensionality reduction or clustering (Burkov, 2019). To deal with large datasets or else big data, ML approaches also incorporate artificial neural network architectures, known as deep learning (DL). These DL algorithms can utilise both labelled and unlabelled samples (semi-supervised learning) to identify predictable relationships and interaction in diverse forms of data (Bagherzadeh and Asil, 2019).

Both ML and DL methods are employed in computational biological studies, especially for Omics analysis and integration. However, the small sample size of biological experiments or the low signal-to-noise in Omics data restrict their application usually to data visualization, or to build more interpretable models with dimensionality reduction techniques such as the Principal Component Analysis (PCA). Despite these limitations, an increasing number of studies are taking advantage of the unique opportunities of ML and DL approaches to integrate multi-Omics datasets for patient classification (Alakwaa et al., 2018), drug sensitivity modelling (Ali and Aittokallio, 2019), and biomarker discovery (Grapov et al., 2018), making great strides towards precision medicine.

Herein, both supervised and unsupervised machine learning approaches were applied for single Omics analysis (chapter 3) and for multi-Omics data integration analysis (chapter 4).

1.5 Aims

The overall aim of this work is to uncover novel biological insights into the metabolic reprogramming of haematological cancers and identify metabolic vulnerabilities for future therapeutic studies. This is achieved by integrating transcriptomic with metabolomic datasets from cancer patients or cell lines, derived mostly from mature B-cell neoplasms such as Burkitt Lymphoma, DLBCL and CLL. This category of haematological cancers includes both leukaemia and lymphomas and it is responsible for the majority of neo-diagnosed cases of haematological cancers worldwide (WHO, 2018). All the datasets presented in this thesis have been acquired to explain hypotheses from different collaborative projects. Therefore, comparisons within or between the datasets are driven by the availability of the datasets, the need to answer several biological questions, and to highlight the advantages and limitations of Omics integration strategies. Hence, the power of multi-Omics data integration approach is highlighted here using different methodologies to accomplish the additional following aims:

1) ***Gene expression analysis and Genome Scale Metabolic modelling predict metabolic vulnerabilities in CLL (chapter 2).*** The transcriptomic profile of CLL patients with different clinical outcome is used to investigate cancer metabolism in this type of mature B-cell neoplasms with leukaemic characteristics.

Our aim is to identify key differences in the expression of metabolic genes between spontaneous regressed and non-regressed CLL cases. Finally, integration of transcriptomic data with GSMMs aims to identify metabolic genes that can act as potential metabolic drug targets in CLL for future functional studies.

2) ***Metabolic features and pathways underpin the Germinal Centre-derived B-cell lymphomagenesis (chapter 3).*** GC-derived B-cell lymphomas comprise a range of multi-factorial diseases, our aim is to understand the metabolic regulation in aggressive NHL that derived from GC development – focusing on endemic Burkitt lymphoma and germinal centre –like subtype of DLBCL. This work not only explores the key metabolome regulators in GC-derived B-cell lymphomas but it is also aiming to highlight potential metabolic drug targets for these lymphomas.

3) ***Explore the transcriptomic and metabolic diversity of cancers cell lines (chapter 4).*** Cancer metabolism is now one of the hallmarks of cancer, however the metabolic profile for each cancer type demonstrates a broad variety. Our aim is to integrate publicly available transcriptomic and metabolomic datasets from the Cancer Cell Line Encyclopaedia (CCLE) database to explore cancer's metabolic diversity and identify key associations between genes and metabolites that separate haematological cancers from the other types.

CHAPTER 2

METABOLIC MODELLING INTEGRATION TO REVEAL METABOLIC VULNERABILITIES IN CLL

2.1 Introduction

In CLL, most of the patients experience an indolent or a progressive status of the disease. In those patients an elevated number of CLL cells in the blood tends either to remain stable or slowly increase in time. Surprisingly, in less than 2% of CLL cases, the disease spontaneously regresses in the absence of any treatment (Giudice et al., 2009; Thomas et al., 2002). These spontaneous regression CLL cases are usually CD38, ZAP70 negative and they have mutated IgHV genes (Giudice et al., 2009). Our work highlighted the genetic and microenvironmental factors that underpin the clonal attrition in spontaneous CLL regression machinery, revealing the important role of metabolism as well. Results from unsupervised hierarchical clustering analysis of the entire RNAseq dataset, which also included age-matched untreated progressive CLL case, IgHV unmutated CLL cases and healthy controls, revealed that most of the spontaneously regressed CLL segregated into a distinct cluster adjacent to the indolent CLL cluster (**Figure 2.1**). This is consistent with spontaneously regressed CLL cases having a unique transcriptomic profile that bears the closest resemblance to indolent CLL cases and is different from healthy controls (**Figure 2.1**). More importantly, RNAseq analysis of spontaneously regressed CLL cases highlighted downregulation of several metabolic genes (such as c-Myc gene) and other metabolic pathways that indicated a low metabolic, quiescent state (Kwok et al., 2019). In addition, previous studies have also associated remissions in CLL with increased mitochondrial mass

and mitochondrial ROS production (Carew et al., 2004; Jones et al., 2018). Although new therapeutic approaches are targeting metabolism in CLL (Adekola et al., 2015; Galicia-Vázquez and Aloyz, 2019), the mechanisms of CLL metabolic reprograming remain poorly understood.

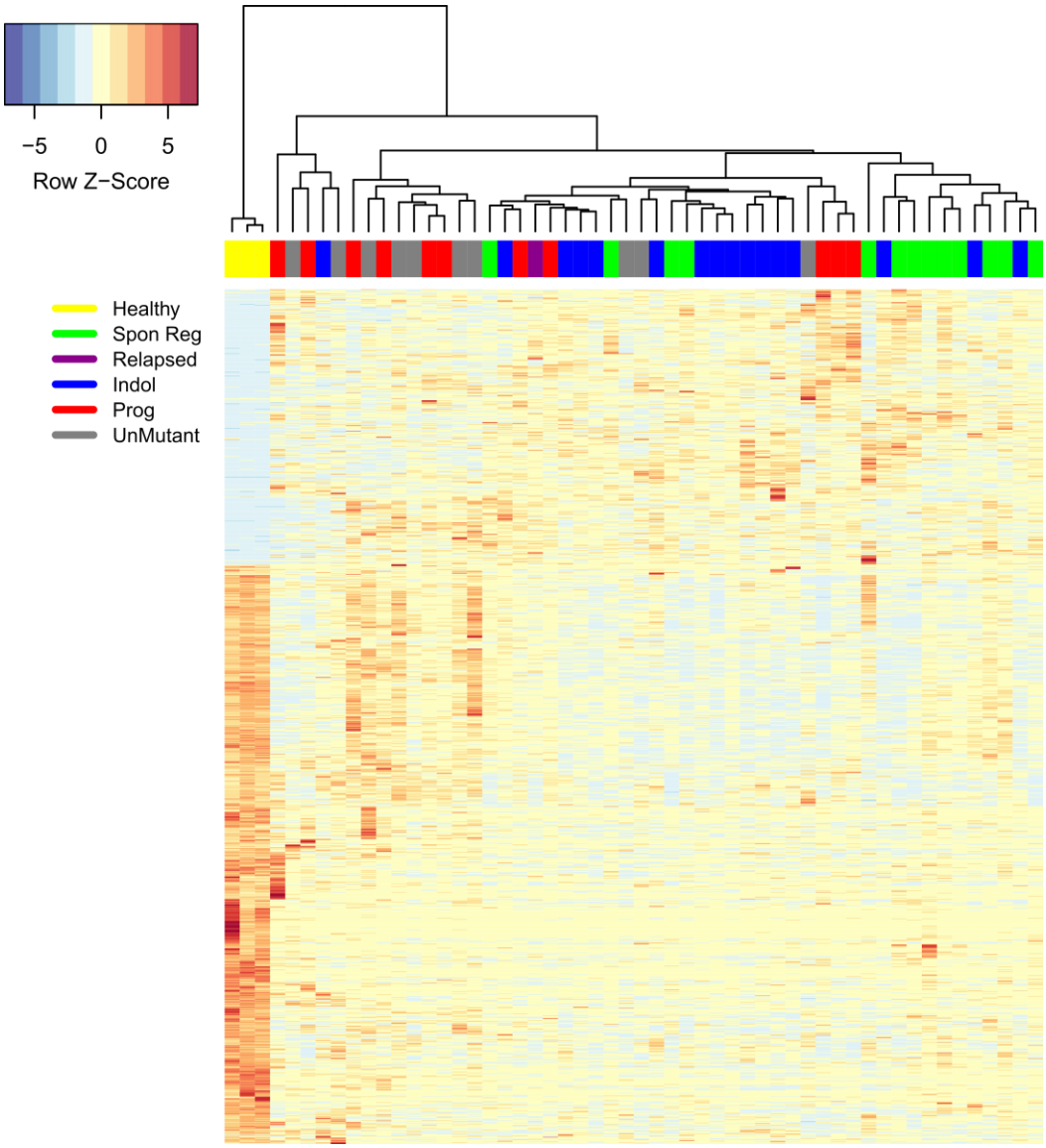


Figure 2.1. Gene expression profile of CLL cases and healthy donors. Dendrogram in hierarchical clustering analysis was produced with Ward.D2 method and distance: 1 – Spearman‘s rank correlation. Heatmap represents expression values of 39,297 genes converted to a Z-score scale along the rows for case comparisons between: 3 healthy controls (Healthy, yellow), 13 spontaneously regressed CLL (Spon Reg, green), 1 relapsed CLL (Relapsed, purple), 16 indolent CLL (Indol, blue), 11 progressive CLL (Prog, red) and 10 IgHV unmutated CLL cases (UnMutant, grey) (Kwok et al., 2019).

In this chapter we have used RNAseq data to interrogate the role of metabolism in groups of CLL patients having different clinical outcomes. We looked for differences in expression of metabolic genes and pathways between spontaneous regression and non-regression CLL cases. Finally, we integrated these results with Genome Scale Metabolic Modelling (GSMM), by utilizing independently the Robust Metabolic Transformation Algorithm (rMTA) and the genetic Minimal Cut Sets (gMCSs) computational approaches to predict metabolic genes as vulnerabilities in CLL (**Figure 2.2**).

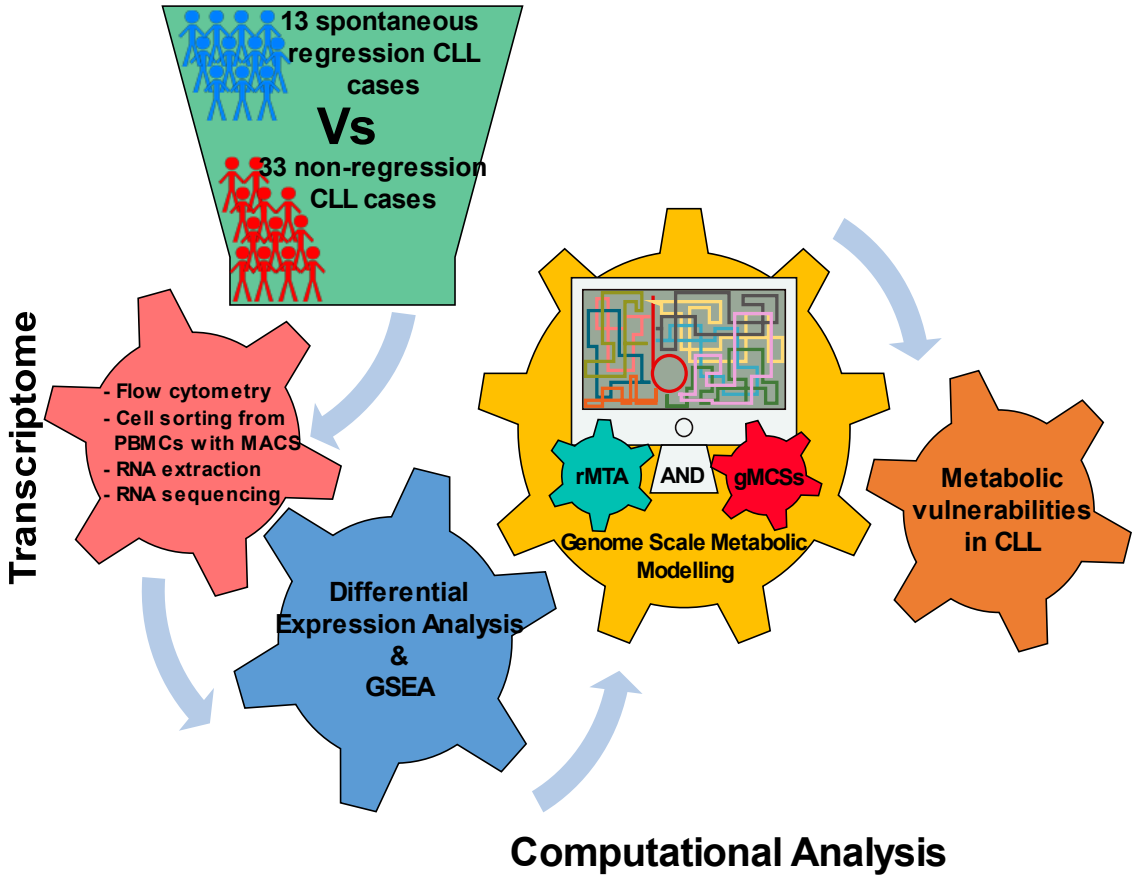


Figure 2.2. The study flowchart. Schematic representation illustrating the steps of the analysis in CLL cases.

2.2 Materials and methods

2.2.1. Transcriptomic data from CLL patients

All the analyses described in this chapter were performed with data that derived from primary CLL samples. The transcriptomic dataset was obtained from our collaborators (Institute of Cancer and Genomic Sciences, University of Birmingham) and derived from the largest cohort of spontaneous CLL regression cases worldwide. According to our initial study, data derived from patients with untreated CLL who attended 4 haemato-oncology centres in the United Kingdom between 2010 and 2016 (records of 1425 CLL patients were reviewed). The design, selection and data generation of the CLL cohort was performed by Dr Marwan Kwok and is fully described in his recent publication (Kwok et al., 2019). For this study, he identified subjects with complete spontaneous CLL regression on the basis of a sustained reduction in absolute lymphocyte count (ALC) to below 4 x 10^9/L, with complete resolution of CLL-related symptoms, anaemia (<100g/L), thrombocytopenia (<100 x 10^9/L) and clinically detectable adenopathy that may be present at diagnosis. He also identified subjects with partial spontaneous regression based on sustained reduction of lymphocytosis by ≥50% from the highest level, with regressing nodal disease. Individuals with a potential explanation for disease regression were excluded. These include patients with

concurrent infections or second malignancies and those receiving myelosuppressive or immunosuppressive therapies, including systemic corticosteroids, for any indication immediately preceding or coinciding with the onset of CLL regression. Subjects diagnosed with a second malignancy following the onset of CLL regression were not excluded but were categorized separately if the subsequent malignancy was diagnosed within 5 years of the onset of CLL regression. For comparison purposes, untreated CLL cases with indolent disease were recruited locally, progressive cases were sourced from multicentre trials and three healthy controls were recruited. Indolent CLL was defined as Binet stage A disease with a lymphocyte doubling time of ≥2 years monitored over ≥5 years. Peripheral blood samples were obtained after the clinically characterization of the regression state with written consent from participating subjects and with prior institutional ethical approval. For progressive cases, a sample obtained immediately before treatment was used. The dataset is also publicly available at the Sequence Read Archive (SRA) with Bioproject accession no. PRJNA535508; URL: https://www.ncbi.nlm.nih.gov/bioproject/ PRJNA535508 (Kwok et al., 2019). From the initial cohort we have selected the data from 46 CLL cases with 13 cases categorised as spontaneously regressed and 33 cases as non-regressed (Appendix 1). These 33 non-regression cases include 16 indolent IgHV mutated cases, 8 progressive IgHV mutated and 9 IgVH unmutated cases. In CLL, both progressive IgHV mutated and unmutated cases are related to unfavourable clinical prognosis compared to the other CLL groups. All spontaneous regression cases were IgHV mutated.

All immunophenotyping analyses were performed by Dr Marwan Kwok on fresh blood. For immunophenotyping, red blood cells contained within the sample were lysed using ammonium chloride prior to antibody incubation. Antibody staining was carried out for 30 minutes at 4°C on cells resuspended in FACSFlow (BD Biosciences) with 2% bovine serum albumin. For antibody panels where intracellular markers were included, Intrasure and FACSLyse reagents (BD Biosciences) were used to fix, lyse and permeabilize the cells prior to incubation with intracellular antibodies.

Following acquisition of a minimum of 200,000 events per sample using FACSCanto or LSR II flow cytometers (BD Biosciences), the mononuclear cell population was gated on the FSC vs SSC plot and doublets were excluded. For CLL immunophenotyping, events were displayed on a CD19 vs CD20 plot. CD20 provides the single best discriminator between the CLL cells and normal B cells. A gate was applied to include all CD19+ CD20-low CLL cells but exclude normal B cells that would be CD19+ CD20-high. For T cell immunophenotyping, events were displayed on a CD3 vs SSC plot, and a third gate was applied on the CD3+ T cell population, with further gating to differentiate between CD3+ CD4+ and CD3+ CD8+ T cells. The gated singlet CLL or T cell population was then analysed for the expression of various markers. Biological controls were used to determine the setting of gates which demarcate the positive vs negative populations. A CLL phenotype of a residual monoclonal B lymphocyte population (CD19, CD5, CD23 and CD43 positive, CD20, CD79b and CD81 weak, CD10 negative and Igl/λ-

restricted) was identified by multiparameter flow cytometry in the regressed CLL cases. To distinguish from normal B cell population, the CLL population was quantified on the basis that normal B cells have high expression levels of CD20, CD79b and CD81 and low expression levels of CD43, CLL cells typically have high expression levels of CD43 and low expression levels of CD20, CD79b and CD81. Thus, it was determined that residual CLL cells accounted for a median of 92.5% of B cells (range 71.6-99.8%) at the time of regression.

For cell sorting, mononuclear cells were isolated from peripheral blood (PBMCs) by Dr Marwan Kwok using density gradient centrifugation with Lymphoprep solution (Axis-Shield). The isolation of CD19+ CD5+ CLL cells from PBMCs was carried out using a two-step magnetic-activated cell sorting (MACS) process. This involves first isolating CD19+ B lymphocytes by depleting all other cell types using the human B-CLL cell isolation kit (Miltenyl Biotec). The sorted CD19+ B lymphocyte population was then enriched for CD19+ CD5+ CLL cells by positive sorting for CD5+ cells using human CD5-biotin antibody (clone UCHT2; Miltenyl Biotec) and anti-biotin microbeads (Miltenyl Biotec). The sorted cell fraction was confirmed to be >95% CD19+ CD5+ by flow cytometry prior to DNA and RNA extraction as well as their use in the telomerase and β-galactosidase assays.

RNA was extracted by Dr Marwan Kwok from sorted CLL cells using the RNeasy Mini or Micro kit (Qiagen) respectively according to the manufacturer's instructions. Nucleic acid samples were quantified and their purity confirmed using

a NanoDrop spectrophotometer (Thermo Fisher Scientific). For samples prepared for RNA sequencing, the nucleic acid concentration was further verified by a Qubit 2.0 fluorometer (Thermo Fisher Scientific).

Qubit-quantified RNA from sorted CLL cells was quality assessed using TapeStation 2200 (Agilent Technologies), with an RNA integrity number (RIN) of ≥7 being considered acceptable. Library preparation was performed using the TruSeq Stranded mRNA Library Prep Kit for NeoPrep (Illumina), with 16 RNA samples being pooled into a single library. In brief, RNA purification beads were added to 50 ng of each RNA sample, which were subsequently heated in a Veriti thermal cycler at 65°C for 5 minutes. Each sample was then loaded onto a NeoPrep library card, alongside their corresponding index adaptors and other reagents required for library preparation. Subsequent cDNA synthesis, A-tailing, adaptor ligation, hybridization, enrichment, PCR amplification, library quantification and pooling steps were automated upon loading of the library card onto the NeoPrep system (Illumina). Transcriptome sequencing (RNAseq) was performed on the prepared cDNA libraries using NextSeq 500/550 High Output Kit v2 (Illumina). Altogether, 16 RNA samples were sequenced within a single flow cell, allowing an average of 25 million reads per sample.

2.2.2. Transcriptomic analysis for CLL dataset

The RNAseq data for the CLL study were analysed with the *Kallisto-Sleuth* computational approach. First, quality control of the paired-end RNAseq read counts was performed for every sample using *FastQC 0.11.7* software (Andrews, 2010). Diagnostic plots generated via FastQC were examined for per base sequence distribution, GC%, per sequence quality distribution and vector or adapter contamination. RNAseq data with both R1 and R2 read-pair were selected for further analysis based on: the per base sequence quality score with median for any base was ≥25; the averaged quality score per read was ≥20 (this equates to a 1% error rate); and the absence of adapter contamination. Then, we used *Kallisto 0.43.0* (Bray et al., 2016) to pseudo-align reads to the human reference genome GRCh38 cDNA index and quantify the transcripts abundances for every sample. Next, differential expression analysis (DEA) was performed with *Sleuth 0.30.0* R package (Pimentel et al., 2017), comparing the spontaneous regressed CLL cases to the non-regressed CLL cases. Significantly altered gene expression was identified with the "Wald" parametric statistical test to perform DEA. The test calculates "beta" values to demonstrate the gene expression under each condition. False discovery rate (FDR) was calculated to correct for multiple comparisons problem with the Benjamini-Hochberg method, using a threshold of 10% (q values < 0.1). Heatmaps were generated using the $\log_2$TPM+1 normalised values with *gplot v3.01.1* R package (Warnes et al., 2019). Hierarchical clustering was

performed with the internal *hclust* R function, using the “Ward.D2” method with distance: 1 – Spearman’s rank correlation. Finally, rather than focusing only on significant genes, we have used gene set enrichment analysis with the *SetRank v1.1.0* R software package (Simillion et al., 2017) to identify statistically significant pathways and we built gene set networks and interactome maps with *Cytoscape v3.7.2* (Shannon et al., 2003).

2.2.3. Genome Scale Metabolic Modelling approaches

2.2.3.1. robust Metabolic Transformation Algorithm (rMTA)

We applied the rMTA algorithm (Valcárcel et al., 2019) from COBRA Toolbox v3.0 MATLAB software (Heirendt et al., 2017) to integrate genome scale metabolic models with the expression profile of non-regression CLL cases as a “source” metabolic state and the spontaneous regression cases as a “target” metabolic state. The human metabolic network *Recon 2.v04* (Thiele et al., 2013) was used as a starting genome scale metabolic reconstruction. First, a mean flux distribution for the source state was computed from 2000 integration Metabolic Analysis (iMAT). In the mean flux distribution, a closed loop formed by reactions r0170 and r0575, both of which being catalysed by FDFT1, was identified. Because such loop could lead to overestimating the importance of this gene, this was corrected by constraining *Recon 2.v04* to make r0575 irreversible and repeating

the sampling step. Such inconsistencies are mostly related with gaps in reactions or metabolites (Orth and Palsson, 2010; Ponce-de-Leon et al., 2015) and they are one of the vast limitations in the use of rMTA. Additional limitations are fully described in section 5.2 of chapter 5. Then, the results from DEA between non-regression and spontaneous regression cases (false discovery rate at 10%) were used as an input to rMTA with the alpha parameter set to 0.99 value. Finally, perturbations were simulated for the 33 genes that were significantly upregulated in non-regression CLL cases (**Table 2.2**) and an rTS score was assigned to each gene based on the ability to transform the source to the target state.

2.2.3.2. genetic Minimal Cut Sets (gMCSs)

Genetic minimal cut sets were calculated for every significant upregulated gene in non-regression CLL cases. We used the gMCSs (Apaolaza et al., 2017) function from COBRA Toolbox v3.0 MATLAB software (Heirendt et al., 2017) to define a set of genes, whose removal would block proliferation (biomass production) in the *Recon 2.v04* model. Next, the differential expression analysis results were mapped in every gMCSs, selecting as the most important ones those where only a single gene is having higher expression in non-regression CLL cases with the rest being lowly expressed (Apaolaza et al., 2017).

2.3. Results

2.3.1. Differentially expressed genes

Our first goal in this study was to investigate the transcriptomic profile of 13 spontaneous regressed and 33 non-regressed CLL cases (**Figure 2.2**, Appendix 1) using the RNA-seq data from the largest cohort study of spontaneous regression cases (Kwok et al., 2019). Transcripts abundances were quantified for 17,051 genes and then DEA with Wald-test identified 530 differentially expressed genes (208 upregulated; 322 downregulated; q value < 0.1) between the two CLL groups (**Figure 2.3A**). Furthermore, unsupervised hierarchical clustering separated most of spontaneous regression cases from the non-regression ones, which suggests a distinct expression profile between them (**Figure 2.3A**). Moreover, Dr. Marwan Kwok highlighted specific phenotypic features in spontaneous regressed CLL cases compared to non-regressed CLL cases. Immunophenotypic analysis revealed low or absence of CLL proliferation or a lack of recently proliferated cells, as evidenced by low Ki-67 and high CXCR4 expression in spontaneous regressed CLL cases (Calissano et al.; Coelho et al., 2013; Kwok et al., 2019). Another feature of spontaneous regressed CLL cases was the reduced CD49d and ROR1 expression, and increased CD95/FasR expression (Kwok et al., 2019). Together these findings highlight that spontaneous CLL regression status presents a unique transcriptome profile with distinct phenotypic features.

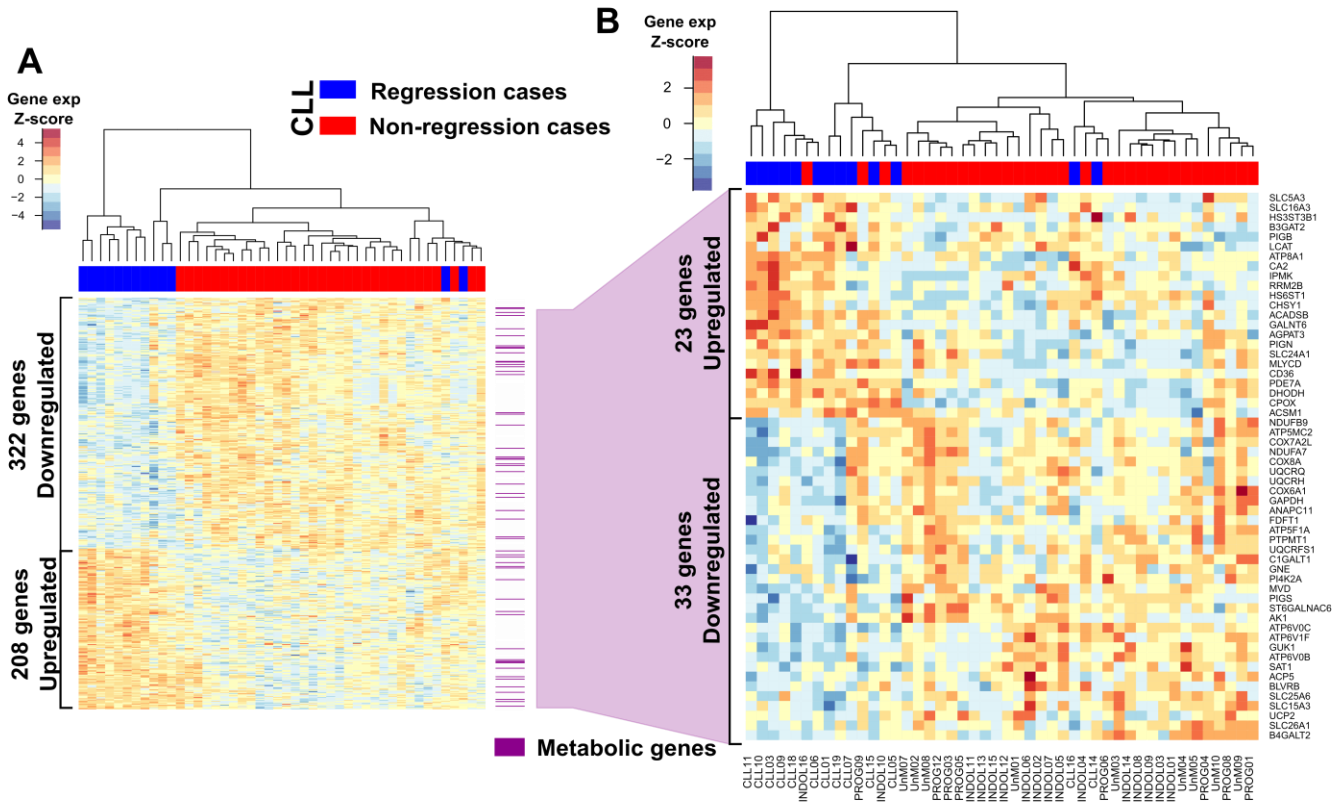


Figure 2.3. Differential expression analysis for RNAseq data from CLL patients. A) Heatmap representing the 530 differentially expressed genes between spontaneous regression (blue colour) and non-regression CLL cases (red colour) with FDR at 10% (q value < 0.1). Rows with purple colour indicate the 56 metabolic genes identified from *Recon 2.v04* model. **B)** Expression of these 56 differentially expressed metabolic genes in *Recon 2.v04*. Gene expression values have been converted to a Z-score scale along the rows for case comparisons. Dendrogram in hierarchical clustering analysis was produced with Ward.D2 method and distance: 1 – Spearman‘s rank correlation.

However, two of the spontaneous regressed CLL cases (CLL14 and CLL16) clustered between the non-regressed CLL cases **(Figure 2.3A)**. These two cases also presented a sustained reduction of lymphocytosis by 50% from the peak level with regressing nodal disease. This evidence resulted in their characterization as *partial* spontaneous disease regression cases, which separated them from other complete spontaneous regressed CLL cases. Furthermore, resemblance of their expression profile with non-regressed CLL cases may suggests the possibility of these cases to progress in the future. In addition, SNP array analysis performed by Dr Marwan Kowk and our collaborators associated these two cases with deletion in 13q14.2-q14.3 region and the loss of the microRNA miR15a/16-1 (Kwok et al., 2019). Deletions of the 13q14 region are a frequent event in CLL and the loss of miR15a/16-1 cluster is suggested to be associated with BCL-2 overexpression and CLL cell proliferation (Calin et al., 2008; Cimmino et al., 2005; Klein et al., 2010).

To assess the role of metabolism we examined the expression of the 2140 metabolic genes that are present in *Recon 2.v04* metabolic reconstruction. We identified 56 metabolic genes as significantly altered, with 33 genes showed higher expression in non-regressed CLL cases and 23 genes elevated in spontaneous regressed cases (**Figure 2.3B**, **Table 2.1**). Importantly, we observed that most of these upregulated genes in non-regression cases have a key role in the electron transport chain (NDUFB9, NDUFA7, ATP5AF1, COX6A1, COX8A, COX7A2L, UQCRQ, UQCRFS1 and UQCRH). Similarly, the mitochondrial inner membrane transporter SLC25A6, which is responsible for exporting ATP to the cytosol, was

enriched in non-regressed cases. Transcriptomic differences in these metabolic genes may indicate a highly active role of energy metabolism in non-regressed cases compare to regressed ones. To clarify this issue, we next performed gene set enrichment analysis (GSEA).

Table 2.1. Statistically significant metabolic genes that are included in *Recon 2.v04* model as calculated by DEA.

Gene symbols	Ensembl ID	Entrez ID	pval	qval	beta value
CA2	ENSG00000104267	760	0.00007	0.013	1.26
HS3ST3B1	ENSG00000125430	9953	0.00002	0.006	0.90
IPMK	ENSG00000151151	253430	0.00014	0.019	0.82
CD36	ENSG00000135218	948	0.00114	0.057	0.79
ACSM1	ENSG00000166743	116285	0.00118	0.058	0.73
ATP8A1	ENSG00000124406	10396	0.00063	0.040	0.67
CPOX	ENSG00000080819	1371	0.00073	0.042	0.64
GALNT6	ENSG00000139629	11226	0.00150	0.067	0.62
LCAT	ENSG00000213398	3931	0.00081	0.045	0.59
SLC16A3	ENSG00000141526	9123	0.00002	0.006	0.59
CHSY1	ENSG00000131873	22856	0.00312	0.094	0.56
PDE7A	ENSG00000205268	5150	0.00003	0.008	0.56
SLC5A3	ENSG00000198743	6526	0.00334	0.097	0.53
HS6ST1	ENSG00000136720	9394	0.00146	0.066	0.49
B3GAT2	ENSG00000112309	135152	0.00173	0.072	0.48
SLC24A1	ENSG00000074621	9187	0.00183	0.075	0.48
DHODH	ENSG00000102967	1723	0.00193	0.076	0.46
RRM2B	ENSG00000048392	50484	0.00143	0.066	0.42
ACADSB	ENSG00000196177	36	0.00275	0.090	0.41
MLYCD	ENSG00000103150	23417	0.00178	0.074	0.39
AGPAT3	ENSG00000160216	56894	0.00288	0.091	0.33
PIGB	ENSG00000069943	9488	0.00311	0.094	0.31
PIGN	ENSG00000197563	23556	0.00289	0.091	0.30
GNE	ENSG00000159921	10020	0.00315	0.094	-0.30
ATP5MC2	ENSG00000135390	517	0.00242	0.085	-0.30
UQCRH	ENSG00000173660	7388	0.00145	0.066	-0.32
COX6A1	ENSG00000111775	1337	0.00148	0.066	-0.32
NDUFA7	ENSG00000267855	4701	0.00303	0.093	-0.33
PIGS	ENSG00000087111	94005	0.00211	0.079	-0.34
GUK1	ENSG00000143774	2987	0.00066	0.040	-0.34
ANAPC11	ENSG00000141552	51529	0.00322	0.096	-0.35
COX7A2L	ENSG00000115944	9167	0.00045	0.032	-0.36
ATP5F1A	ENSG00000152234	498	0.00031	0.026	-0.36
C1GALT1	ENSG00000106392	56913	0.00124	0.060	-0.36
PTPMT1	ENSG00000110536	114971	0.00078	0.044	-0.36
SAT1	ENSG00000130066	6303	0.00197	0.077	-0.36
FDFT1	ENSG00000079459	2222	0.00026	0.024	-0.37
SLC25A6	ENSG00000169100	293	0.00017	0.020	-0.37
UQCRFS1	ENSG00000169021	7386	0.00019	0.020	-0.38
COX8A	ENSG00000176340	1351	0.00012	0.016	-0.39
UQCRQ	ENSG00000164405	27089	0.00061	0.039	-0.39
ATP6V0C	ENSG00000185883	527	0.00023	0.023	-0.39
ST6GALNAC6	ENSG00000160408	30815	0.00003	0.008	-0.41
MVD	ENSG00000167508	4597	0.00312	0.094	-0.45
NDUFB9	ENSG00000147684	4715	0.00001	0.005	-0.45
ACP5	ENSG00000102575	54	0.00262	0.088	-0.47
AK1	ENSG00000106992	203	0.00218	0.081	-0.49
SLC15A3	ENSG00000110446	51296	0.00008	0.014	-0.49
ATP6V0B	ENSG00000117410	533	0.00005	0.010	-0.49
GAPDH	ENSG00000111640	2597	0.00031	0.027	-0.50
UCP2	ENSG00000175567	7351	0.00110	0.056	-0.50
BLVRB	ENSG00000090013	645	0.00302	0.093	-0.51
ATP6V1F	ENSG00000128524	9296	0.00006	0.012	-0.52
PI4K2A	ENSG00000155252	55361	0.00002	0.006	-0.52
SLC26A1	ENSG00000145217	10861	0.00092	0.048	-0.57
B4GALT2	ENSG00000117411	8704	0.00180	0.074	-0.84

2.3.2. Gene set enrichment analysis in CLL dataset

To further determine the role of the most essential metabolic pathways, we used an advanced method of GSEA with *SetRank* by utilizing 979 annotated gene sets from the KEGG database (Kanehisa et al., 2019). These gene sets represent the majority of biological processes and pathways in the cell. Results from GSEA identified 31 significantly enriched gene sets (SetRank parameters thresholds: setPCutoff = 0.01 and fdrCutoff = 0.05) (Simillion et al., 2017) with the most significant gene sets being: the ribosome in eukaryotes, the transcriptional mis-regulation in cancer and the oxidative phosphorylation (OXPHOS) pathway (**Figure 2.4**, Appendix 2). These findings are in accordance with a previous study that has also associated spontaneous regression CLL cases with ribosomal genes, signal transduction and transcription regulators (Giudice et al., 2009). Moreover, despite that OXPHOS pathway was under investigation in CLL (Bruno et al., 2015; Galicia-Vázquez and Aloyz, 2019), its role in CLL spontaneous regression remains unclear.

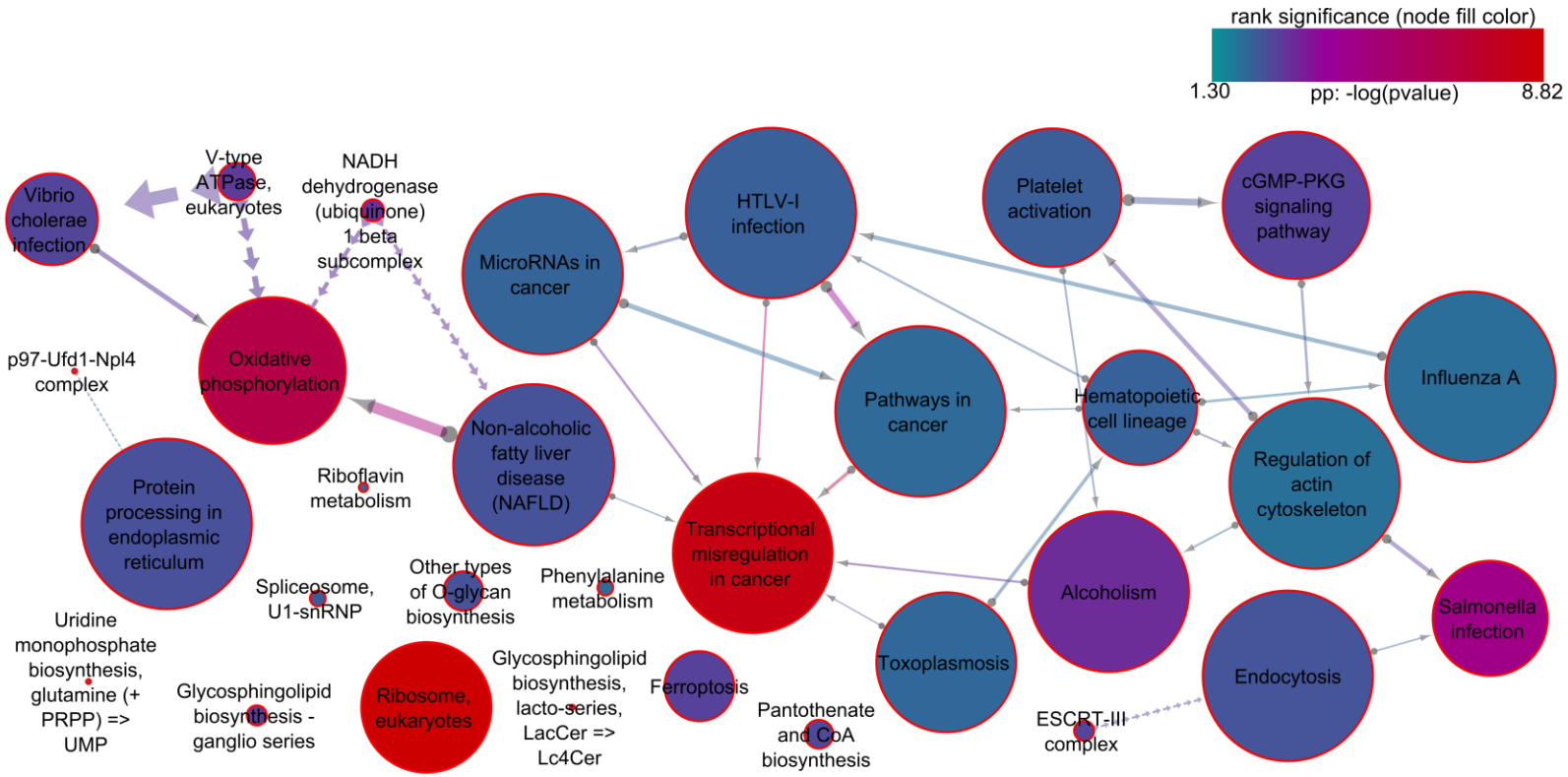


Figure 2.4 . Gene set network for significant pathways from GSEA. The network highlights the intersections between the 31 significant altered gene sets from GSEA with SetRank. The node fill colour reflects the corrected p-value, going from blue to red with decreasing p-value (increasing significance, pp denotes the negative logarithm of the p-value). The node border colour reflects the SetRank p-value using the same colouring coding. Edge thickness reflects the size of the interactions between two gene sets. The edge arrows point from least significant gene set to more significant one.

To further explore the genes involved in OXPHOS and their interactions, we constructed a gene interaction network with interactions from the STRING database (Szklarczyk et al., 2019). Interestingly, we observed that the most important nodes in the network represent genes that are significantly upregulated in non-regressed CLL cases (**Figure 2.5**). Overall, our findings suggested that non-regressed CLL cells depend more on mitochondrial metabolism compared to spontaneous regressed cases.

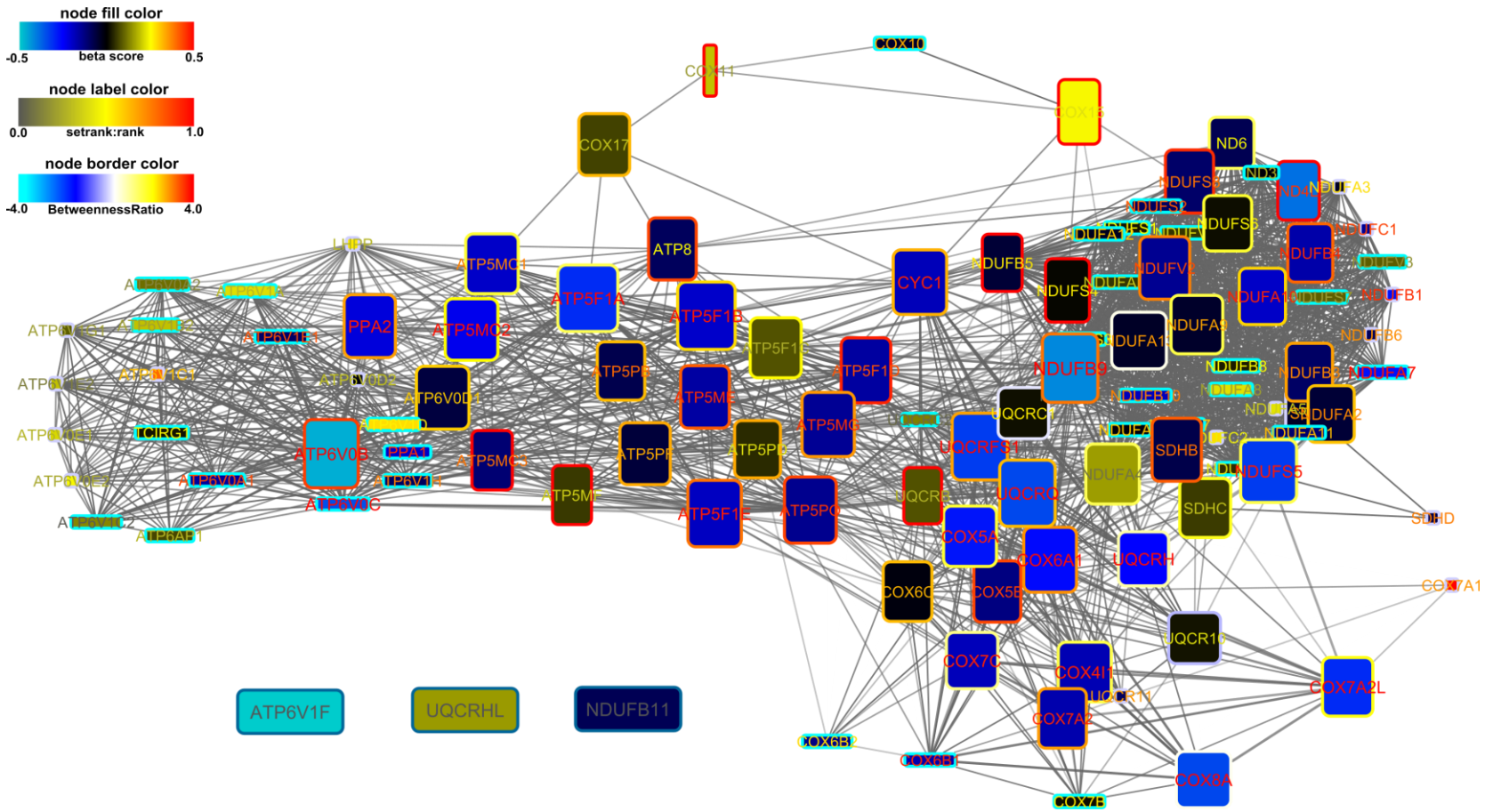


Figure 2.5. Gene interactome network for OXPHOS pathway from *SetRank*. The node colour reflects the beta scores from DEA of spontaneous regressed vs non-regressed CLL cases; the size of the node labels reflects the significance of difference in expression; node label colour reflects the gene rank when sorted by p-value (rank of 1.0 means the lowest p-value). The node border colour reflects the log-ratio between the local and global betweenness.

2.3.3. Genome Scale Metabolic Modelling results

An extension of our work was to integrate the CLL transcriptomic profiles with the *Recon 2.v04* Genome-Scale Metabolic Model to predict metabolic genes as metabolic vulnerabilities. We applied independently two different computational approaches: the rMTA and the gMCSs. First, we employed the rMTA to identify metabolic gene knockouts that revert metabolism of a given metabolic state (source state) to another (target state) (Auslander et al., 2017; Stempler et al., 2014; Yizhak et al., 2013). We defined as a source metabolic state the expression status of non-regressed CLL cases and as a target state the status of spontaneous regressed cases. The algorithm calculated the robust transformation score (rTS) for every significantly upregulated gene in non-regressed CLL cases, which indicates the ability of gene perturbation to alter/transform metabolism closer to the regression metabolic state. The highest scoring gene was the SLC26A1, which encodes a sulfate anion transporter (rTS=7.07, **Table 2.2**). The GUK1 gene showed the second highest score (rTS=2.7, **Table 2.2**) followed by the SLC25A6 gene (rTS=1.9, **Table 2.2**) and GNE gene (rTS=0.7, **Table 2.2**). Thus, according to rMTA prediction those are the best putative targets for inhibition to revert the non-regressed CLL cases to the status of spontaneous regressed cases. Prior of any future inhibition study, it is important to confirm and validate for any changes of these genes at mRNA and protein level using RT-PCR and Western Blot analysis.

Table 2.2. rMTA results for upregulated genes in non-regression CLL cases. A unified score (rTS) was calculated, considering a best-case (bTS), a worst-case (wTS) scenario and a Minimization of Metabolic Adjustment score (mTS) to achieve robustness.

Gene symbols	Entrez ID	Description	bTS	mTS	wTS	rTS
SLC26A1	10861	solute carrier family 26 member 1	0.173	0.234	-0.129	7.080
GUK1	2987	guanylate kinase 1	0.127	0.111	-0.116	2.701
SLC25A6	293	solute carrier family 25 member 6	0.100	0.101	-0.096	1.970
GNE	10020	glucosamine (UDP-N-acetyl)-2-epimerase/N-acetylmannosamine kinase	0.081	0.057	-0.048	0.735
PTPMT1	114971	protein tyrosine phosphatase, mitochondrial 1	0.047	0.026	-0.016	0.165
MVD	4597	mevalonate diphosphate decarboxylase	0.038	0.025	-0.024	0.155
COX6A1	1337	cytochrome c oxidase subunit 6A1	-0.003	0.005	0.003	0.005
COX8A	1351	cytochrome c oxidase subunit 8A	-0.003	0.005	0.003	0.005
COX7A2L	9167	cytochrome c oxidase subunit 7A2 like	-0.003	0.005	0.003	0.005
UQCRQ	27089	ubiquinol-cytochrome c reductase complex III subunit VII	0.003	0.004	-0.002	0.002
UQCRFS1	7386	ubiquinol-cytochrome c reductase, Rieske iron-sulfur polypeptide 1	0.003	0.004	-0.002	0.002
UQCRH	7388	ubiquinol-cytochrome c reductase hinge protein	0.003	0.004	-0.002	0.002
ST6GALNAC6	30815	ST6 N-acetylgalactosaminide alpha-2,6-sialyltransferase 6	0.003	-0.001	0.073	-0.001
ANAPC11	51529	anaphase promoting complex subunit 11	0.001	-0.003	0.075	-0.003
FDFT1	2222	farnesyl-diphosphate farnesyltransferase 1	0.001	-0.003	0.075	-0.003
C1GALT1	56913	core 1 synthase, glycoprotein-N-acetylgalactosamine 3-beta-galactosyltransferase 1	0.001	-0.003	0.075	-0.003
PIGS	94005	phosphatidylinositol glycan anchor biosynthesis class S	0.001	-0.003	0.075	-0.003
NDUFA7	4701	NADH:ubiquinone oxidoreductase subunit A7	0.001	-0.003	0.076	-0.003
NDUFB9	4715	NADH:ubiquinone oxidoreductase subunit B9	0.001	-0.003	0.076	-0.003
ATP5A1	498	ATP synthase, H+ transporting, mitochondrial F1 complex, alpha subunit 1, cardiac muscle [Source:HGNC Symbol;Acc:HGNC:823]	-0.482	-0.003	0.490	-0.003
ATP6V0C	527	ATPase H+ transporting V0 subunit c	0.010	-0.004	0.028	-0.004
ATP6V0B	533	ATPase H+ transporting V0 subunit b	0.010	-0.004	0.028	-0.004
ATP6V1F	9296	ATPase H+ transporting V1 subunit F	0.010	-0.004	0.028	-0.004

Furthermore, we utilised independently the computational method of gMCSs to identify synthetic lethality metabolic genes by targeting proliferation in *Recon 2.v04* model. This algorithm predicted metabolic candidates by calculating minimal cut sets for all the upregulated genes in non-regressed cases. We identified minimal cuts sets for the FDF1, PIGS, AK1, PTPMT1, GADPH, ATP5G2, GUK1, ATP5A1, GNE, MVD and SLC25A6 genes. **Table 2.3** presents the 49 most important gMCSs out of a total 29034 gMCSs. The selection criteria were a maximum number of 20 genes in each cut set and up to two upregulated genes in non-regressed cases (beside the gene of interest), while all the other genes are downregulated or lowly expressed. Both AK1 and GUK1 genes seemed to be promising candidates for single or combined targeted strategies, since they both identified with several important cut sets (**Table 2.3**). In addition, gMCSs results indicated that proliferation can be blocked either if the FDFT1 or the MVD gene is simultaneously knocked out with the PLPP3 and the SGPP1 genes (**Table 2.3**). Similar effect is predicted for a combined targeting of the PTPMT1 gene with the CTPS2 and the UCK1 genes. Lastly, combined inhibition of the GNE enzyme and the equilibrative nucleoside transporters (SLC29A1 and SLC29A2) will possibly have a synthetic lethality effect in non-regressed CLL cells.

Overall, findings derived from these two independent GSMM approaches are complementary, suggesting a number of genes as potential metabolic vulnerabilities in CLL. In brief, rMTA unveiled putative genes associated to non-regressed CLL phenotype and gMCSs analysis highlighted genes that can be targeted to kill CLL cells displaying the non-regressed CLL phenotype.

Table 2.3. The 49 most important gMCSs calculated for significant upregulated genes in non-regressed CLL cases. Colour scale represents the beta values from DEA results with red colour indicating genes that show higher gene expression in non-regression CLL cases and blue colour higher gene expression in spontaneous regression CLL cases.

gMCSs	gene_1	gene_2	gene_3	gene_4	gene_5	gene_6	gene_7	gene_8	gene_9	gene_10	gene_11	gene_12	gene_13	gene_14	gene_15	gene_16	gene_17	gene_18	gene_19
FDFT1	CRAT	FASN	FDFT1	HMGCL	PLPP1	PLPP2	PLPP3	SGPL1	SLC27A5	SLC27A4	SLC27A3	COQ2	HMGCLL1	MCCC1	SGPP1	SGPP2	SLC27A1		
FDFT1	AUH	CRAT	FASN	FDFT1	HMGCL	PLPP1	PLPP2	PLPP3	SGPL1	COQ7	SLC27A5	SLC27A4	SLC27A3	HMGCLL1	SGPP1	SGPP2	SLC27A1		
AK1	AK1	AK2	AK4	TXNRD1	AK5	RRM2B	AK7	LOC100507855											
AK1	AK1	AK2	AK4	RRM2	TXNRD1	AK5	AK7	LOC100507855											
AK1	AK1	AK2	AK4	RRM2	AK5	RRM2B	AK7	LOC100507855											
AK1	AK1	AK2	AK4	RRM1	TXNRD1	AK5	AK7	LOC100507855											
AK1	AK1	AK2	AK4	RRM1	AK5	RRM2B	AK7	LOC100507855											
AK1	AK1	AK2	AK4	DGUOK	TXNRD1	AK5	CMPK1	AK7	LOC100507855										
AK1	AK1	AK2	AK4	DCK	DCTD	DGUOK	TXNRD1	AK5	AK7	LOC100507855									
AK1	AK1	AK2	AK4	DTYMK	GUK1	TXNRD1	AK5	CMPK1	AK7	LOC100507855									
AK1	AK1	AK2	AK4	DCK	DCTD	DTYMK	GUK1	TXNRD1	AK5	AK7	LOC100507855								
AK1	AK1	AK2	AK4	SLC29A2	PNP	TXNRD1	SLC28A2	AK5	CMPK1	SLC28A3	AK7	LOC100507855							
AK1	AK1	AK2	AK4	DCTD	DGUOK	SLC29A2	TXNRD1	SLC28A1	AK5	SLC28A3	AK7	LOC100507855							
AK1	AK1	AK2	AK4	DGUOK	SLC29A2	TXNRD1	UPP1	SLC28A1	AK5	SLC28A3	AK7	UPP2	LOC100507855						
AK1	AK1	AK2	AK4	DCK	DCTD	SLC29A2	PNP	TXNRD1	SLC28A2	AK5	SLC28A3	AK7	LOC100507855						
AK1	AK1	AK2	AK4	DPYD	TYMP	GUK1	SLC29A2	UPP1	SLC28A2	SLC28A1	AK5	SLC28A3	RBKS	AK7	UPP2	LOC100507855			
AK1	AK1	AK2	AK4	DPYD	TYMP	GUK1	SLC29A2	UPP1	SLC28A2	SLC28A1	AK5	PGM2	SLC28A3	AK7	UPP2	LOC100507855			
AK1	AK1	AK2	AK4	DPYD	DPYS	GUK1	SLC29A2	UPP1	SLC28A2	SLC28A1	AK5	SLC28A3	RBKS	AK7	UPP2	LOC100507855			
AK1	AK1	AK2	AK4	DPYD	DPYS	GUK1	SLC29A2	UPP1	SLC28A2	SLC28A1	AK5	PGM2	SLC28A3	AK7	UPP2	LOC100507855			
AK1	AK1	AK2	AK4	DCTD	NME1	NME3	PKLR	TK1	TK2	TXNRD1	NME6	AK5	NME7	CMPK1	AK7	LOC100507855			
AK1	AK1	AK2	AK4	DCK	NME1	NME3	PKLR	TK1	TK2	TXNRD1	NME6	AK5	NME7	CMPK1	AK7	LOC100507855			
AK1	ACAA1	AK1	AK2	ABCD1	AMPD1	AMPD2	AMPD3	FASN	PNP	SLC27A5	NT5C2	AK5	NT5C	NT5C3A	NT5C1A	NT5C1B	AK7	SLC27A1	
AK1	ACAA1	AK1	AK2	ABCD1	AMPD1	AMPD2	AMPD3	ACSL1	FASN	PNP	SLC27A5	NT5C2	AK5	NT5C	NT5C3A	NT5C1A	NT5C1B	AK7	
AK1	ACOX1	AK1	AK2	ABCD1	AMPD1	AMPD2	AMPD3	ABCC2	ACSL1	FASN	PNP	NT5C2	AK5	NT5C	NT5C3A	NT5C1A	NT5C1B	AK7	
AK1	ACAA1	AK1	AK2	ABCD1	AMPD1	AMPD2	AMPD3	ACSL1	FASN	SLC27A5	NT5C2	QPRT	AK5	NT5C	NT5C3A	NT5C1A	NT5C1B	NAPRT	AK7
PTPMT1	CTPS1	RRM1	UCK2	CDIPT	CTPS2	UCK1	PTPMT1												
PTPMT1	CTPS1	UCK2	CDIPT	RRM2B	CTPS2	UCK1	PTPMT1												
PTPMT1	CTPS1	RRM2	UCK2	CDIPT	CTPS2	UCK1	PTPMT1												
GUK1	GUK1	SLC25A19																	
GUK1	GUK1	RRM2																	
GUK1	GUK1	RRM2B																	
GUK1	DGUOK	GUK1																	
GUK1	GUK1	RRM1																	
GUK1	GUK1	SLC29A2	PNP	SLC28A2	SLC28A3														
GUK1	DCTD	DTYMK	GUK1	SLC29A2	PNP	TXNRD1	SLC28A1	SLC28A3											
GUK1	DTYMK	GUK1	SLC29A2	PNP	TXNRD1	SLC28A1	AK5	CMPK1	SLC28A3										
GUK1	AK1	AK2	AK4	DTYMK	GUK1	TXNRD1	AK5	CMPK1	AK7	LOC100507855									
GUK1	DTYMK	DUT	GUK1	SLC29A2	ITPA	PNP	TXNRD1	SLC28A1	SLC28A3	UPP2									
GUK1	GUK1	SLC29A2	NME2	NME3	PNP	PKLR	TK1	TK2	NME6	NME7	CMPK1	UPP2							
GUK1	GUK1	SLC29A2	NME1	NME3	PNP	PKLR	TK1	TK2	NME6	NME7	CMPK1	UPP2							
GUK1	DPYD	GUK1	SLC29A2	NME1	NME3	PKLR	TK1	TK2	UPP1	NME6	NME7	CMPK1	UPP2						
GUK1	DPYD	GUK1	SLC29A2	NME2	NME3	PKLR	TK1	TK2	UPP1	NME6	NME7	CMPK1	UPP2						
GUK1	DPYD	GUK1	SLC29A2	NME2	NME3	PNP	PGM1	PKLR	TK1	TK2	NME6	NME7	CMPK1	PGM2					
GUK1	DPYD	GUK1	SLC29A2	NME1	NME3	PNP	PGM1	PKLR	TK1	TK2	NME6	NME7	CMPK1	PGM2					
GNE	CDA	SLC29A1	GALNS	GALT	SLC29A2	TXNRD1	UGP2	UMPS	SLC28A2	SLC28A1	GNE	RRM2B	A4GNT	CMPK1	AICDA	SLC28A3			
GNE	CDA	SLC29A1	GALT	SLC29A2	NAGA	RRM1	TXNRD1	UGP2	UMPS	SLC28A2	SLC28A1	GNE	A4GNT	CMPK1	AICDA	SLC28A3			
MVD	CRAT	FASN	HMGCL	MVD	PLPP1	PLPP2	PLPP3	SGPL1	SLC27A5	SLC27A4	SLC27A3	HMGCLL1	MCCC1	SGPP1	SGPP2	SLC27A1			
MVD	CRAT	FASN	HMGCL	MVD	SLC22A5	PLPP1	PLPP2	PLPP3	SGPL1	SLC27A5	SLC27A4	SLC27A3	HMGCLL1	SGPP1	SGPP2	SLC27A1			
MVD	AUH	CRAT	FASN	HMGCL	MVD	PLPP1	PLPP2	PLPP3	SGPL1	SLC27A5	SLC27A4	SLC27A3	HMGCLL1	SGPP1	SGPP2	SLC27A1			

2.4. Discussion

We started our investigation with an effort to explore metabolism in CLL and identify metabolic vulnerabilities by comparing spontaneous regressed CLL cases with non-regressed cases. As shown in **Figure 2.1**, the transcriptomic profile of CLL cases was distinct from those of healthy donors, which indicates that data derived mostly from CLL cells. This is consistent with flow cytometry data whereby CLL cells are distinguished based on high expression of levels of CD43 and low expression levels of CD20, CD79b and CD81, whereas normal B-cells present high expression levels of CD20, CD79b and CD81. Although, normal B-cells and CLL cells are both CD19 positive, the expression of CD20 provides the single best discriminator between the CLL cells and normal B-cells as previously highlighted by Rawstron et al. (Rawstron et al., 2016). Despite our effort to isolate pure CLL cells, the acquired RNAseq data may also contain RNA from normal B-cells. This limitation could be overcome in future studies with the generations of single cell RNAseq data.

Results from DEA and GSEA showed that non-regressed CLL cases have a differential reliance on OXPHOS compared to spontaneous regressed cases. These findings are in accordance with recent studies that target OXPHOS in CLL using metformin (NCT01750567), a mitochondrial Complex 1 inhibitor (Bruno et al., 2015; Galicia-Vázquez and Aloyz, 2018). Another study has also reported that leukemic stem cells (LSC) depend on OXPHOS and amino acids catabolism to

support proliferation (Jones et al., 2018). Nevertheless, we highlighted here upregulation of OXPHOS pathway associated to the non-regressed CLL cases, which has not previously been reported. However, further experiments should be performed to prove functional upregulation of OXPHOS in non-regressed CLL cells. Results from our collaborators Kwok et al. suggest a model in which clonal anergy, reduced CLL trafficking and possibly underpin as a mechanism spontaneous regression in CLL (Kwok et al., 2019). It will be of importance to investigate the role of mitochondrial metabolism in CLL clonal anergy and its contribution in the activation process of CLL cells in relapsed CLL cases. Glycolysis and OXPHOS activity can be quantified with the Seahorse extracellular flux analyser (Agilent, CA, USA) by measuring the extracellular acidification rate (ECAR) for glycolysis, and the oxygen consumption rate (OCR) for OXPHOS. Complementary, enzymatic assays activities can be measured in cellular extracts from CLL cells to determine glycolytic and mitochondrial metabolism. In addition, it will be of interest to examine if mitochondrial fusion and/or mitochondrial mass is associated with upregulation of OXPHOS in CLL cells, as it was reported in other cell types (Wai and Langer, 2016; Youle and van der Bliek, 2012). Mitochondrial fusion in CLL cells can be assessed with immunofluorescence microscopy using antibodies against manganese superoxide dismutase to visualize mitochondria.

Analysing gene expression is mostly limited to identification of transcriptional alterations. Our goal was to go beyond these alterations and try to elucidate the metabolic changes in CLL by integrating with GSMM. Our results

demonstrated that rMTA can be applied to identify genes that are putative key drivers to revert non-regressed CLL phenotype and highlighted the SLC26A1 gene, which encodes the human sulfate anion transporter 1 (or SAT1), as the most promising metabolic vulnerability, which opens a new perspective on CLL therapy to be further explored and validated experimentally. Although, SAT1 is a key regulator of both oxalate and sulfate homeostasis (Dawson et al., 2010), its role in CLL metabolism remains undetermined. Therefore, it is important to confirm, firstly, for any differences in mRNA level of the SLC26A1 gene using RT-PCR analysis. Protein levels of the sulfate anion transporter 1 can also be compared with Western Blot analysis, while protein concentrations can be determined with assays such as the Bio-Rad protein assay (BioRad, Hercules, CA, USA). Further transcriptomic analysis comparing RNAseq data from CLL cases and healthy donors will highlight for any differences in the expression of SLC26A1 gene in normal B-cells and CLL cells. Following knock-out experiments of this gene either in CLL cell lines or in vivo models, such as with the CRISPR-Cas method (Ishibashi et al., 2020), can reveal important aspects of its metabolic role in CLL. It is noteworthy that when initially applying rMTA, we identified the need for curation of reaction catalysed by FDFT1 as the actual implementation in *Recon 2.v04* generate a closed loop formed by the reactions r0170 and r0575 (both of which are catalysed by FDFT1). We corrected such loop by constraining *Recon 2.v04* to make r0575 irreversible and we verified that the problem was solved. Such inconsistencies are well known in Genome Scale Metabolic Models and they are mostly related with gaps in reactions or metabolites (Orth and Palsson, 2010; Ponce-de-Leon et al., 2015).

In parallel, the second approach of gMCSs proposed the combined inhibition of FDFT1, PLPP3 and SGPP1 genes in synthetic lethality–strategy. In particular relevance to the present study, cholesterol metabolism has been shown to contribute to chemotherapy resistance in several other cancers (Benakanakere et al., 2014; Montero et al., 2008; Storch et al., 2007). A previous study has also showed enhanced chemoimmuno-sensitivity in MEC-2 CLL cells when targeting squalene synthase with YM-5360 or TAK-475 inhibitors (Benakanakere et al., 2014). More importantly, both GSMM approaches examine changes in mRNA levels and infer changes in protein and metabolome level. However, the expression and the activity of metabolic enzymes is affected by post-translational and metabolic regulations. As previously mentioned, the mRNA and protein levels for gMCSs finding should be validated with RT-PCR and Western Blot analysis.

Moreover, our results allow us to hypothesise that non-regressed CLL cases depend on mitochondrial respiration, particularly on OXPHOS, and cholesterol metabolism to support cell growth and proliferation, which seems to enhance the aggressive status of the disease. Further experimental functional and drug target studies, such as those described in the previous paragraphs, using either CLL cell lines or primary samples will be needed to investigate the contribution of mitochondrial metabolism in CLL and validate the therapeutic contribution of these GSMM putative targets.

CHAPTER 3

PATHWAY INTEGRATION TO CHARACTERISE METABOLIC VARIATIONS IN GC-DERIVED LYMPHOMAS

3.1. Introduction

Germinal centres (GCs) are the sites of lymph nodes where B-cells undergo proliferation and selection on the basis of affinity antigen (Basso and Dalla-Favera, 2015). Firstly, B-cells enter and pass from the structure of GCs known as the dark zone, where immunoglobulin somatic hypermutation and proliferation occur. Next, they transfer to another region of GCs the so-called light zone, where processes such as B-cells activation and selection (based on their affinity for antigen) is taking place. Then either B-cells exit the GCs as plasma or memory B-cells, or they recycle back to the dark zone (**Figure 3.1**). The transit of B-cells from the GCs requires a series of signalling and transcriptional events and any disruption during these procedures can lead to malignant transformation. Thus, lymph nodes are also the histological structures where most mature B-cell lymphomas arise. As mentioned before, Burkitt Lymphoma (BL) and Diffuse Large B-cell Lymphoma (DLBCL) are both GCs mature aggressive B-cell Non Hodgkin Lymphomas (NHL) that tend to spread quickly with serious symptoms (WHO, 2014). BL cases derive from malignant B-cells of the dark zone and they are grouped in sporadic, endemic and HIV-associated cases. The main characteristics in BL cells are the c-Myc translocations and the EBV positivity, specifically in the endemic cases (Schmitz et al., 2014). DLBCL is categorised in the germinal centres like (GCB) and the activated B-cell like (ABC) groups (Alizadeh et al., 2000). GCB-DLBCL cases originate from the light zone malignant B-cells, whilst ABC-DLBCL cases contain

malignant arrested B-cells in the plasmablast stage (or else “immature plasma B-cells”). Although, BL and DLBCL are genetically, phenotypically and clinically distinct, the available treatment methods are mostly based on chemotherapy, radiotherapy and autologous stem cell transplantation. Occasionally, GC-derived B-cells cases present morphological, immunophenotypic and cytogenetic intermediate features between BL and DLBCL, making these cases difficult to classify in diagnosis level. In addition, gene expression profiling studies have also illustrated a common transcriptomic profile between BL and DLBCL (Campo et al., 2011). Furthermore, several GCB-DLBCL cases present c-Myc translocations similar to those that characterise the aggressive BL cases. The absence of any biomarker or specific therapeutic target in GC-derived lymphomas justifies the need to investigate further the molecular differences at the transcriptional or metabolic level.

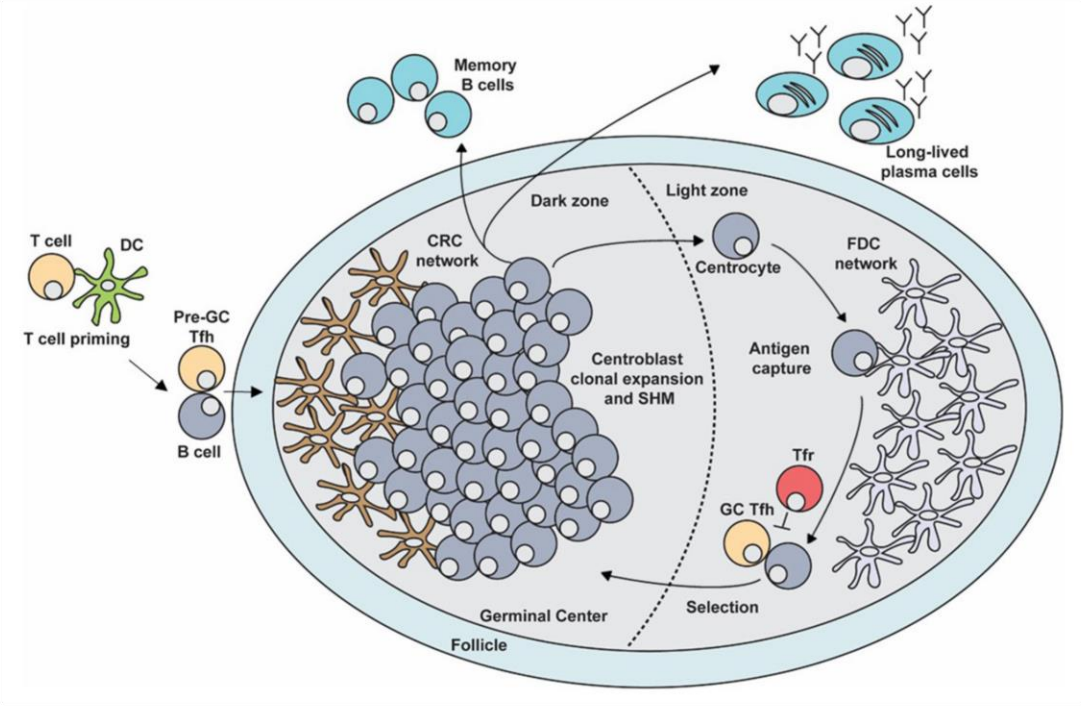


Figure 3.1. The germinal centres (GCs) structure and response. The maturation steps occurring in B-cells during their transit through the GCs (Stebegg et al., 2018).

Cancer as mentioned previously is beginning to be recognised as a metabolic disorder, suggesting new metabolic molecular targets that reprogram metabolism and enhance carcinogenesis. Alterations along glycolysis and other metabolic pathways such as the reductive metabolism of glutamine are now considered essential for malignant transformation (Dong et al., 2017). Significant progress has also been made in identifying the role of metabolism in different stages of GCs lymphomagenesis. A previous metaboproteomics study has demonstrated downregulation of glycolysis and pyruvate metabolism, while one carbon metabolism was upregulated in BL compared to DLBCL (Schwarzfischer et al., 2017). Deregulated expression of c-Myc has been associated with the upregulation of glutamine catabolism especially in BL (Le et al., 2012; Wise et al., 2008). These are only a few examples of metabolic alterations that accelerate lymphomagenesis in these types of lymphomas.

Given the significance of metabolism in GCs lymphomagenesis and driven by Mrs Zuhal Eraslan's complementary work and hypothesis on the role of serine in BL, we sought to better define the fundamental aspects of these complex metabolic regulations in GC-derived lymphomas and try to identify new molecules as potential metabolic targets. In this chapter, we performed transcriptomic analysis in publicly available RNAseq NHL datasets generated from primary tumours. A pathway-based computational approach was then followed to integrate transcriptomics (RNAseq data) with metabolomics (untargeted 1D ^{1}H-NMR data) profiles from our in-house NHL cell lines to explore interactions between metabolic

genes that co-expressed with metabolites.

3.2. Materials and methods

3.2.1. NHL transcriptome sequencing profiles

Publicly available RNAseq data from BL (Abate et al., 2015) and DLBCL (Teater et al., 2018) primary tumours were retrieved from the Sequence Read Archive (SRA) database (Leinonen et al., 2011). The accession numbers for BL dataset is SRP062178 and for DLBCL is SRP100105. The raw data for BL cases is available online (https://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP062178) and for DLBCL (https://trace.ncbi.nlm.nih.gov.Traces/sra/?study=SRP100105). Raw data from BL cases was generated from formalin-fixed and paraffin-embedded (FFPE) samples to investigate the mutational and viral landscape of endemic BL. All BL cases were consistent with BL diagnosis: t(8;14)-positive CD20+, CD10+ BCL-6+, Ki67>98% and BCL-2- (Abate et al., 2015), however the extend of non-tumour cell contamination or tumour purity in the samples was not reported. The authors validated the RNAseq results from BL cases using two distinct series of cases of which matched normal controls were available. The data from DLBCL cases were generated from frozen solid tissue sections with tumour purity above 80-90% based on histological observation to investigate epigenetic

heterogeneity in DLBCL (Teater et al., 2018). Tumour purity is a confounding factor in transcriptomic analysis because tissue samples represent a mixture of RNA transcripts of tumour and non-tumour cells (Aran et al., 2015). So far, studies have used deconvolution methods to estimate the relative proportion between tumour-infiltrating lymphocytes and other immune cells in transcriptomic data (Li et al., 2016a; Newman et al., 2015). However, it remains unclear how to estimate the impact of tumour purity on gene expression or which gene pairs are associated with purity adjustment. In addition, comparison of data derived from FFPE samples versus data from frozen samples have an impact on the results of the transcriptomic analysis. It is well known that FFPE tissues are partially degraded, resulting in fragmentation of RNA transcripts and low quality of RNA data (Chung et al., 2008). However, the absence of publicly available dataset derived from fresh/frozen tissues of endemic BL cases, which are mostly detected in Africa, lead us to the use of data from FFPE samples. Furthermore, several studies have reported consistent and robust results for combined analyses across FFPE/frozen samples and platforms (Bossel Ben-Moshe et al., 2018; Newton et al., 2020; Turnbull et al., 2020). Here, we analysed datasets where both studies have generated paired-end RNAseq data, using the Illumina HiSeq2000 platform (Illumina, San Diego, USA). In total, 19 endemic BL cases and 12 GCB-DLBCL cases were used further for gene expression analysis (Appendix 3).

3.2.2. In-house cell lines transcriptome profiles

RNAseq data from in-house NHL cell lines were generated using 4 BL (BL-31, Ezema, SAV and Glor) and 4 DLBCL (Farage, SUDHL-4, SUDHL-5, SUDHL-6) cell lines to perform integrative analysis (Appendix 5). Cell lines arise from long-standing work in Dr. Farhat Khanim's group (University of Birmingham) under whose supervision the data was generated. All the cell lines were purchased from DSMZ (Braunschweig, Germany) and cultured in RPMI 1640 media (Gibco-Invitrogen Ltd, Paisley, U.K.) with 10% FBS (FBS, Gibco-Invitrogen) supplemented with penicillin (100 U/ml) and streptomycin (100µg/ml). The cultures were routinely passaged every 2 days to maintain exponential phase by Mrs Zuhal Eraslan. Cells were authenticated regularly to control variation with NorthGene service for STR profiling. Mycoplasma test was performed with DAPI stain (Sigma Aldrich).

Two biological replicates from each cell line were cultured at 37^{o}C with 5% CO_2 to generate the RNAseq data. The TruSeq Stranded mRNA Sample prep kit (IIlumina, San Diego, USA) was used for library preparation. Following, cDNA synthesis, hybridization, PCR amplification and library quantification were performed by our partners. The prepared cDNA libraries were sequenced (RNAseq) by the Theragen Etex (Theragen Co Ltd, Suwon, Korea) using the Illumina HiSeq2500 platform (IIlumina, San Diego, USA). All the work presented in

this section was undertaken by Mrs Zuhal Eraslan and supervised by Dr. Farhat Khanim and Prof. Ulrich Günther. Lastly, the raw data were analysed by the author of this thesis as explained in section 3.2.4. Transcriptomic data from NHL cell lines were analysed independently by comparing data from BL and DLBCL cell lines, while no comparison between transcriptomic data from cell lines and primary tumours was performed in this study.

3.2.3. In-house cell lines metabolomic signatures

Metabolomic data were generated by Mrs Zuhal Eraslan from the same exponentially growing in-house cell lines (4 BL and 4 DLBCL) under the same media conditions (RPMI1640 and 10% FBS) that were used for the RNAseq experiment. Six technical replicates per cell line were taken from each flask, where every cell line was cultured separately, to measure the intracellular metabolites with NMR spectroscopy. Data derived from replicates were analysed independently and they are presented in section 3.3.4 of this chapter. Technical replicates for the NMR experiments were taken at a different time point than those for the RNAseq experiment and not pooled together. A total number of $5x10^7$ cells for each replicate, were used to perform cell extraction. Cell suspensions were centrifuged in falcon tubes at 1500 rpm for 5 minutes at 21°C. 10 ml of the supernatant was stored for media analysis and the remaining supernatant was disposed of. Cell pellets were then washed once with 1 ml of pre-warmed PBS and transferred to Eppendorf tubes. Supernatants were discarded after centrifugation at 14000 rpm for 20 seconds. After this, 400 µl of HPLC grade methanol were rapidly added. Cell pellets were resuspended in methanol on dry ice and vortexed for 10 seconds before storing at −80 °C until extraction. For the extraction, cell pellets in methanol were transferred into the Wheaton™ clear glass sample vials (MERK). 325 µl of distilled HPLC grade H2O and 400 µl of chloroform, pre-chilled on wet ice, were added. Samples were vortexed for 40 seconds and then incubated on the bench for 5 minutes. After centrifugation at 4000 rpm, for 10

minutes at 4°C, polar and nonpolar samples were transferred to Eppendorf tubes and the Wheaton™ clear glass sample vials, respectively. Polar samples were dried using a vacuum dryer.

All polar extracts were dissolved in 50 µl of 100 mM sodium phosphate buffer (pH 7.0) prepared with 90%H2O/10% D2O or 100% D2O (99.9% pure; GOSS Scientific Instruments Ltd.), 3 mM sodium azide (NaN3) and 0.5 mM timethylsilyl-propanoic acid (TMSP, Cambridge Isotope Laboratories) as a chemical shift reference. Samples were sonicated for 15 minutes and vortexed for 10 seconds. 35 ml of sample was transferred into 1.7 mm NMR tubes. Samples that were dissolved in sodium phosphate buffer containing 90% H2O/10% D2O were transferred into 1.7 mm NMR tubes using the GILSON sample preparation unit, while samples that were prepared with 100% D2O buffer were manually transferred. For preparation of media samples for NMR, 162 µl of the previously saved media was resuspended in 18 µl of metabolomics buffer made from 1 M phosphate buffer (pH 7.0) prepared with D2O, 0.5 mM TMSP and 3 mM NaN3. Then the samples were transferred to 3.0 mm NMR tubes. All samples were kept at 4°C until measurement.

All 1D 1H-NOESY spectra for cell extracts were acquired at 300 K using a Bruker 600 MHz spectrometer, equipped with a 1.7-mm TCI probe and a cooled Bruker SampleJet autosampler. The 1H 1D spectra were obtained using the 1D NOESY pulse sequence (noesygppr1d) with water pre-saturation. Key parameters were as follows: spectral width 7183.9 Hz; complex points, TD 32768; interscan delay, d1=4 seconds; NOE mixing time, d8=10 milliseconds; number of transient,

ns = 128; steady state transient, ds = 8. Total acquisition time was 14 minutes. For the 1H 1D spectra of media samples, Bruker 600 76 MHz spectrometer with a 5-mm TXO cryogenic probe with a cooled Bruker SampleJet autosampler was used. The standard Bruker pulse sequence noesygppr1d was used to obtain 1H 1D spectra. The key parameters used were as follows: spectral width: 7183,9 Hz; TD=32768; d1=5 seconds; NOE mixing time, d8=10 milliseconds; ns=32; ds=8. Total acquisition time was 5 minutes. All spectra were measured once.

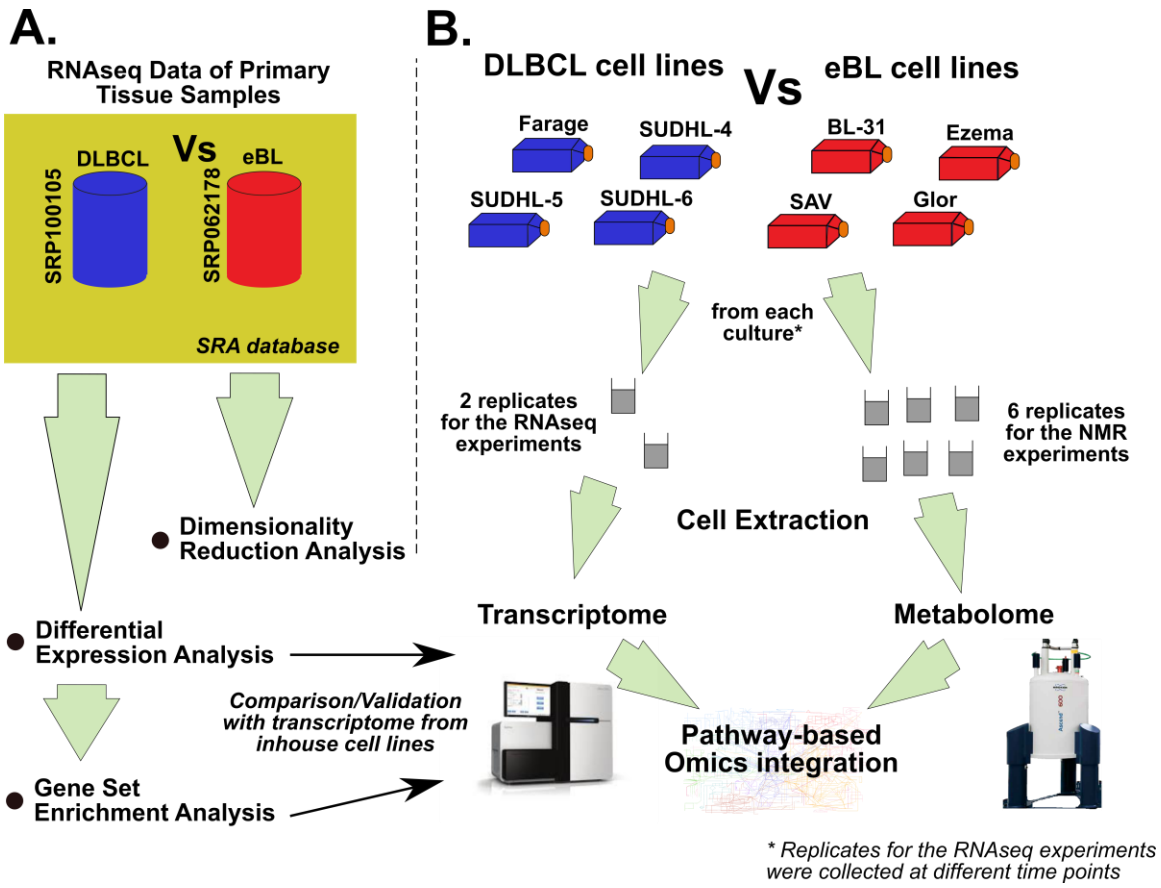


Figure 3.2. Flow diagram of the analyses and the experimental procedures in chapter 3. (A) Transcriptomic analysis performed with SRA datasets derived from primary tissue samples. **(B)** Pathway-based Omics integration analysis performed with Omics datasets from inhouse cell lines.

3.2.4. Transcriptome sequencing data analysis

The transcriptomic datasets were analysed with the *Kallisto-Sleuth* computational workflow (Yalamanchili et al., 2017). For the public datasets, the raw RNAseq data were downloaded in *sra* format from the SRA database and they were converted to *fastq* format with the *SRA Toolkit 2.9.2* (Leinonen et al., 2011). Quality control metrics were obtained with the *FastQC 0.11.7* software (Andrews, 2010); reads were aligned to the GRCh38 human reference genome cDNA index (Ensembl rel.99) and counted to quantify for transcripts abundances with the *Kallisto 0.43.0* software (Bray et al., 2016). Gene-level differential expression analysis was performed with the *Sleuth 0.30.0* R statistical package (Pimentel et al., 2017), comparing BL to DLBCL cases. Differentially expressed genes (DEGs) were calculated with the Wald statistical test, correcting for multiple comparisons with the Benjamini-Hochberg method using a false discovery rate (FDR) threshold of 1% (q values < 0.01). Ensembl gene transcripts were annotated with Entrez IDs, official gene symbols and KEGG enzymes with the *BioMart 2.40.3* R statistical package (Durinck et al., 2009). Transcripts per million (TPM) expression values were calculated to normalise for sequencing depth and gene length (Li et al., 2010). Log_2TPM+1 values were used in the unsupervised method Principal Component Analysis (PCA) with the *PCAtools 1.0.0* R statistical package (Blighe and Lun, 2019) to identify clusters and outliers within the data. Furthermore, a list of 2,552 metabolic enzymes was retrieved from the global KEGG metabolic network for human (map01100 KEGG pathway: https://www.genome.jp/dbget-

bin/get_linkdb?-t+enzyme+path: map01100) (Kanehisa et al., 2019) to study metabolic genes in PCA. The log_2TPM+1 expression values were also used in heatmaps generation and in hierarchical clustering with the Ward method and distance: 1 – Spearman's rank correlation. Finally, Gene Set Enrichment Analysis (GSEA) was performed with the *fgsea 1.10.0* R statistical package (Korotkevich et al., 2016) to study a collection of hallmark gene sets from Molecular Signature Database (Subramanian et al., 2005), using as a significance level the FDR threshold of 5% (q values <0.05).

3.2.5. Metabolome NMR data analysis

The NMR concept design, data acquisition and analysis presented in this section came jointly from Mrs Zuhal Eraslan, Dr. Farhat Khanim and Prof. Ulrich Günther. All 1D ^{1}H NMR spectra were measured once and they were manually phase corrected and chemical shift referenced to TMSP at δ 0.00 ppm, and they were aligned on the TMSP signal using *MetaboLab* (Ludwig and Günther, 2011), a MatLab version R2017a (MathWorks, Massachusetts, USA) based program. MetaboLab was also used to pre-processed all the NMR spectra by performing Fourier Transformation and baseline correction. The free induction decay (FID) signal was zero filled to 32768 points once and Fourier transformed using an exponential line broadening of 0.3 Hz. Additionally, the region between 4.5 to 5.15 ppm was deleted with the same tool for water suppression. The spectra were

scaled to a probabilistic quotient normal normalization (PQN-scaling). Then, segmental alignment (using icoshift tool) was performed in order to align several metabolite. A total number of 21 Metabolites were identified manually by using the *Chenomx NMR Suit 7.6* software (Chenomx Inc., Edmonton, Canada). Metabolite-intensity data from the 1D ^{1}H NMR spectra were generated for further statistical and integrative analysis. The intensity of metabolites was determined by semi-manual integration (ITN tool) in MetaboLab within MatLab. Metabolite intensities are normalised according to cell number as follows: normalised signal intensity value=signal intensity x normalization factor, with normalisation factor= $\frac{1X10^{6}}{\text{cell density}}$.

Data were transformed by the author with generalised logarithm (log) and row-wise normalised with quantile normalization to make features more comparable between the two diseases. The normalised data were analysed using the metabolomic data processing server *MetaboAnalyst 4.0* (Chong et al., 2018). First, hierarchical cluster analysis was performed with the *hclust* R function using the Ward's clustering method and Euclidean distances to determine clusters between the data. Following, univariate analysis tested for changes in metabolites intensities that are significant to discriminate the two conditions. Normality was tested with the Shapiro-Wilk test, which compares whether the sample distribution of the data deviates from a normal distribution. We assumed normality for those metabolites that had p-value > 0.05, which implies that the distribution of the data was not significantly different from the normal distribution (Appendix_6). P-values

for univariate analysis were calculated with the t-test for metabolites with normal distribution or the non-parametric Wilcoxon Mann Whitney for metabolites with non-normal distribution of the data. Additionally, fold change was calculated to detect which metabolites are increased or decreased in each condition. The goal of fold change is to compare the absolute values of changes between two group means. Because log transformation significantly changes the absolute values, fold change was calculated as the ratios between the two groups means using data before log transformation. However, fold changes were calculated in this study to examine changes (either up or down) in metabolites between the two conditions. Therefore, no threshold value was applied to highlight significant metabolites with the fold change. As a result, metabolites with very low fold changes close to value 1, which is the minimum value for fold change indicating no change, are reported in Appendix 6. Finally, to adjust the p-values for multiple testing corrections in univariate analysis, the FDR values were determined for each metabolite with Benjamini-Hochberg approach (Appendix 6). Metabolites with an FDR threshold of 5% (q values < 0.05) were considered statistically significant and selected for pathway analysis.

3.2.6. Pathway-based Omics integration

A pathway-based integration approach was implemented using the “Joint-Pathway Analysis” module from the *MetaboAnalyst 4.0* toolbox to map and

visualise both metabolomic and transcriptomic data. This module provides integration using both pathway topological analysis combined with enrichment analysis. Both tables with significant altered metabolites and common significant genes (primary and cell lines transcriptomic data) together with their expression values (fold changes or beta values, Appendix 6 and 7) were uploaded and matched to the information gathered from KEGG, HMDB and STITCH databases for metabolites and gene annotations. Data were mapped to 31 metabolic pathways from KEGG database that include both metabolic genes and metabolites. Moreover, an integration analysis for both 180 metabolic and regulatory pathways was also examined (Appendix 8). Pathway topology analysis evaluated the importance of a molecule based on its position within a pathway by measuring the Degree centrality, which is the number of links that connect to a node and calculating pathway impact values. Enrichment analysis was also performed with the integration method of combine queries, in which genes and metabolites are pooled into a single query. Finally, metabolic pathway results were visualised in the KEGG global metabolic network to explore interaction between metabolomic and gene expression data between BL and DLBCL tumours.

3.3. Results

3.3.1. Dimensionality reduction in NHL primary tumours

Our investigation began with the analysis of the transcriptome profile of 19 endemic BL and 12 GCB-DLBCL cases, by examining publicly available RNAseq datasets from the SRA database as described in the materials and methods section of this chapter. Firstly, 39,320 transcripts abundances were quantified for 17,048 genes in both datasets. The TPM expression values were used with PCA for dimensionality reduction to explore any transcriptomic associations between the two diseases. PCA aims to define directions that explain the maximum variance in a dataset by summarizing the data into much fewer variables called scores and weighting profiles of the original variables, which called loadings. The results from PCA highlights that cumulatively the first 20 principal components (PCs) represent more than 90% of the explained variation (**Figure 3.3A**). The first component (PC1) which explains 37.65% of the variation, is the optimal component to segregate the two cancers (**Figure 3.3B**), as tested by performing a t-test between BL and DLBCL cases (p value = 4,96e-13, q value = 5e-13). Unfortunately, the available clinical characteristics for BL and DLBCL cases was not informative enough to clearly explain the separation in PC2 or PC3 (Appendix 4 and 9). Together, the PC1 and PC2 separated the transcriptomic profile of BL

from that of DLBCL cases, suggesting that the two diseases are transcriptionally distinct. However, the BL60, BL15 and BL22 cases, which derive from jaw and neck BL tumours, presented an intermediate expression profile between BL and DLBCL cases. Both these three cases were from male patients with stage C cancer, similar to 8 other BL cases. All BL cases were EBV positive and HIV negative except BL15, which status did not record (NR) for these viruses. Moreover, BL15 was lost in the follow up, while BL22 had a complete response to treatment and no relapse in contrast to BL60, which did not respond to treatment. Similar to the BL60 case, BL20 and BL35 did not respond to treatment and 8 other cases were lost in follow up. A common clinical characteristic that potentially is associated with the intermediate profile of all these three BL cases was their positivity in human cytomegalovirus (CMV) (Appendix 4). Although, endemic Burkitt lymphoma is strongly associated with EBV infection, the synergistic role of CMV and other human herpes viruses in the development of lymphoma still remains unclear. For feature selection, we explored the loading values of every gene in the PCA. Out of the total 17048 genes, the top 3 genes that were most responsible for variation along PC1 were KRT13, PSPHP1 and FCRL5. The KRT13 gene, which was significantly upregulated in BL cases (beta = 6.12, q value = 3.87E-05), encodes a structural intermediate filament protein responsible for the maintenance of the integrity of epithelial cells. Although KRT13 demonstrates a diverse expression profile in cancer, several studies have associated KRT13 with regulatory roles in cancer invasion, migration, and metastasis (König et al., 2013; Li et al., 2016c; Sihto et al., 2011). More specifically, Hamakawa et al. reported an

increased gene expression of KRT13 in micrometastases in the lymph nodes of cervical cancer (Hamakawa et al., 2000). In our study, it is unclear whether higher gene expression of KRT13 was related to metastatic or aggressive BL cases. Like the KRT13 gene, PSPHP1 was significantly upregulated in BL cases (beta = 5.19, q value = 2.59E-27). PSPHP1 is a pseudogene, a non-functional gene, which possibly arises from the phosphoserine phosphate (PSPH) gene. PSPH is involved in serine biosynthesis **(Figure 3.5C)**, which was considered upregulated in BL cases, as explained next in section 3.4. By contrast, the FCRL5 gene was upregulated in DLBCL cases (beta = -3.29, q value = 1.11E-08). The FCRL5 gene express a member of the FC receptor like family that previously found upregulated in post-GC cells of the marginal zone and was suggested that is responsible for B-cell development and regulation of human immune system (Polson, 2006). Now, given our interest on metabolism, we examined the metabolic genes retrieved from the KEGG database (as mentioned in the methods section). The loading values showed that the PYCR1, PHGDH, PSAT1, IDO1 and CHIT1 were the most important metabolic genes driving the variation along PC1 (**Figure 3.3C**). Along PC2, the CDA gene of the pyrimidine salvage pathway, which is essential for DNA/RNA synthesis, was the most extreme among metabolic genes that drives the variation (**Figure 3.3C**).

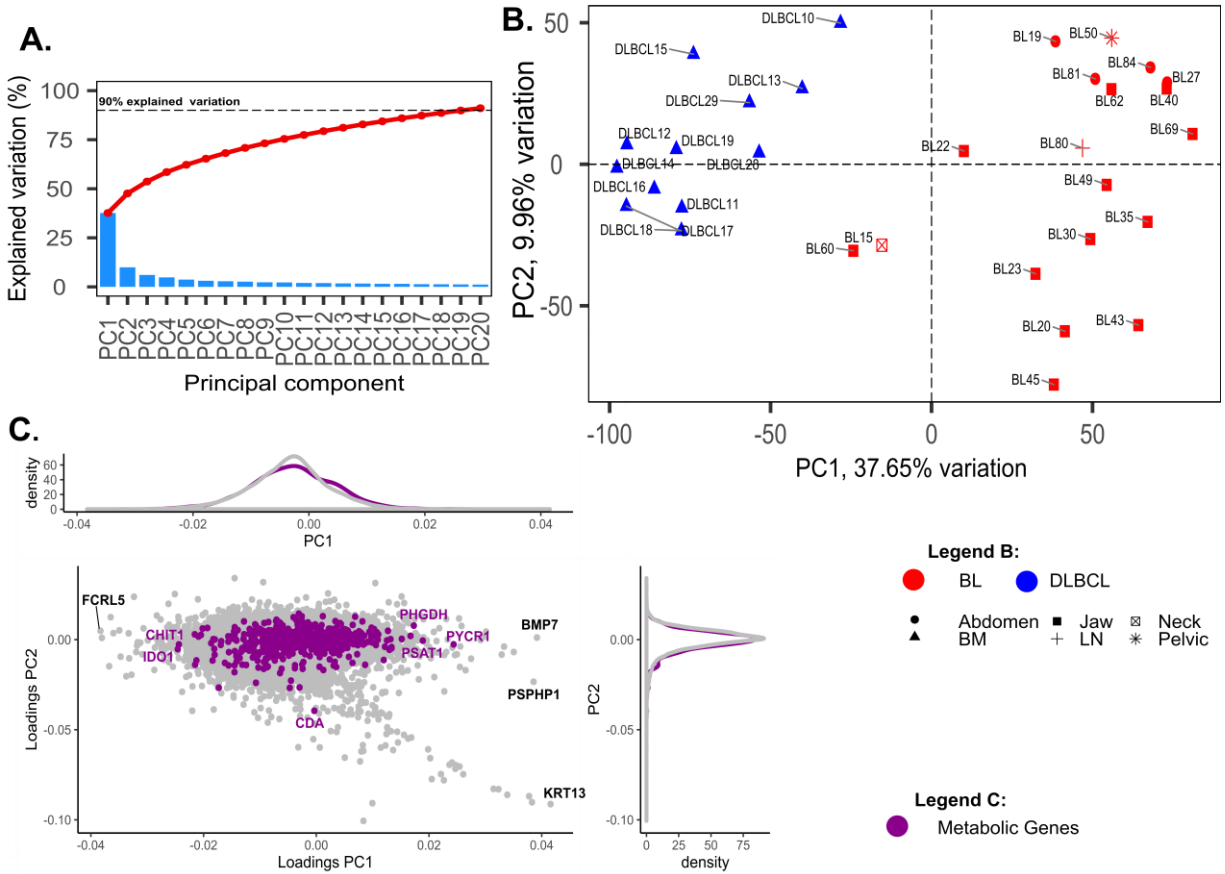


Figure 3.3. Principal component analysis performed on transcriptome profile of primary tumours. (A) The first 20 principal components (PCs) account for more than 90% of explained variation. **(B)** Scatter plot with the first and second principal components are contributing to 37.65% and 9.96% of the total explained variation, respectively. The red colour circles represent the BL cases, and the blue colour circles the DLBCL cases. The marker shapes represent the origin of isolated malignant B-cells: abdomen (circle), bone marrow (BM, triangle), jaw (square), lymph nodes (LN, cross), neck (square X) and pelvic (star). **(C)** PC1 and PC2 loadings values for every gene, highlighting the most important genes that drive the variation along PC1 and PC2. The metabolic genes in scatterplot and density plots are demonstrated with magenta colour.

3.3.2. Differential expression analysis

To further clarify the role of metabolic genes in GCs-derived lymphomas we performed differential gene expression analysis with the Wald-test, which calculates the β-coefficient on every gene (beta values). Our analysis identified 6475 significantly altered genes with 3892 upregulated and 2583 downregulated (q value <0.01) in BL compared to DLBCL cases (**Figure 3.4A**). Like in the previous section, we focused on metabolic genes with 182 genes presenting significantly (q value <0.01) higher expression in BL compared to 82 genes, which were significantly overexpressed in DLBCL. To test whether the relative proportions of metabolic and other genes were the same in two diseases, a contingency table was created. The two-sided Fisher's exact test was applied and revealed that there was a significant difference (p value = 0.0025, odds ratio = 0.66) in BL/DLBCL between metabolic and other significant genes (**Figure 3.4B**). These findings highlighted the key role of metabolic reprogramming in GCs derived lymphomas.

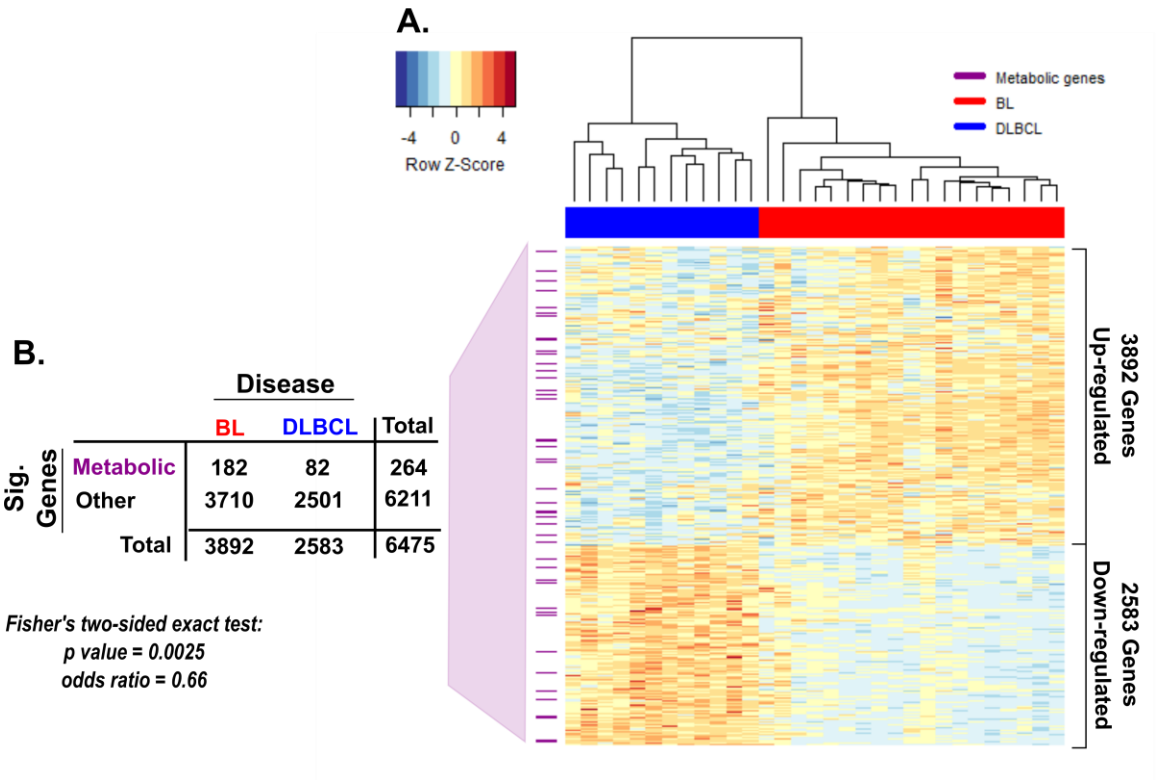


Figure 3.4. Differentially expressed genes between BL and DLBCL cases. (A) The 6475 statistically significant altered genes (FDR <0.01) from differential expression analysis are visualised in a heatmap. Gene expression values have been converted to a Z-score scale along the rows for case comparisons. Dendrogram in hierarchical clustering analysis was produced with Ward method and distance 1- Spearman's rank correlation. **(B)** Contingency table used for two-sided Fisher's Exact test to compare the relative proportions of significant genes between the BL and DLBCL.

3.3.3. Gene set enrichment analysis

To access the role of biological processes and pathways, we performed gene set enrichment analysis (GSEA) using the hallmark gene sets from Molecular Signature Database. The analysis was performed by ranking all the genes based on their beta values from differential expression analysis and identifying the rank positions of all the members of the gene set in the ranked data. Next, an enrichment score was calculated to represent the difference of the observed rankings with an assuming random rank distribution; and normalised to the mean enrichment of random sample of the same size, known as normalised enrichment score (NES). Taking into consideration these NES values with an FDR threshold at 5% (q value <0.05) we identified 30 significant gene sets with 16 gene sets upregulated and 14 downregulated in BL compared to DLBCL (**Figure 3.5A**). GSEA results for BL cases showed upregulation of gene sets related to metabolism (e.g. Glycolysis, Oxidative Phosphorylation and MYC targets) relative to DLBCL cases. Importantly, we observed that the mTORC1 signalling pathway, which is associated with PHGDH and PSAT1 genes, was also upregulated in BL. Moreover, the CDA gene which showed higher expression in BL cases was associated with upregulation of the DNA repair mechanism in BL.

In contrast, the DLBCL cases have demonstrated upregulation of several cellular signalling pathways such as the NF-kB, the JAK-STAT and the KRAS, as a response to inflammation, TNF and interferon gamma (INFγ) signals (**Figure 3.5A**). More specifically, the IDO1 metabolic gene, which presented higher

expression in DLBCL compare to BL, belongs to the interferon gamma response gene set. Together, GSEA findings suggested that DLBCL cases likely alter their metabolism as a result to inflammation and other extracellular signals.

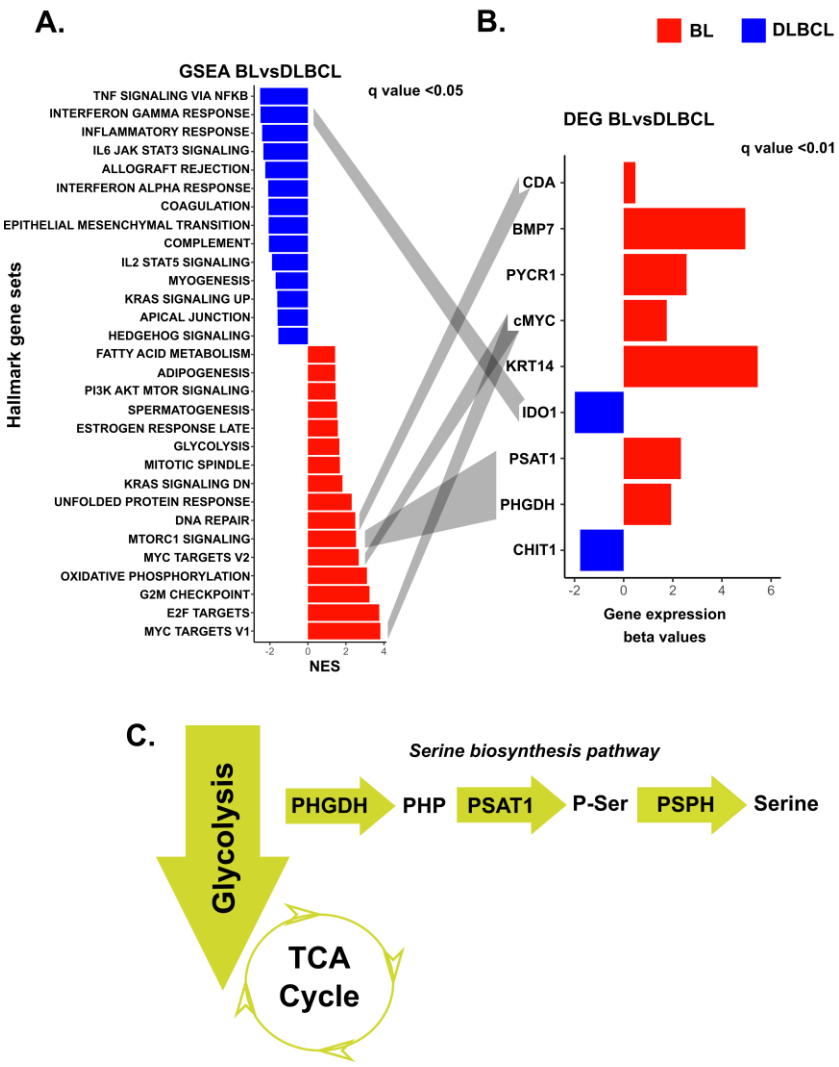


Figure 3.5. Altered genes and pathways between BL and DLBCL. (A) GSEA results identified 30 significant altered gene sets with an FDR threshold of 5%. **(B)** Significant altered differentially expressed genes (DEG) with an FDR threshold of 1%, selected from PCA. **(C)** Schematic representation of the genes involved in the serine pathway.

3.3.4. Statistical analysis of NMR metabolomic data

Untargeted 1D ^{1}H NMR metabolomic data were generated by Mrs Zuhal Eraslan using cell extraction samples from 4 BL and 4 DLBCL cell lines. Metabolites peak intensities were generated as described in the material and methods section for exploratory data analysis between the two diseases. Firstly, hierarchical clustering analysis was utilised to determine clusters between the samples. The Ward's linkage clustering algorithm, which minimises the sum of squares of any two clusters, clearly demonstrated an unsupervised separation between BL and DLBCL (**Figure 3.6A**). To explore further which metabolites are significantly altered and discriminate the two lymphomas, I preformed univariate analysis. Normally distributed metabolites were analysed with t-test, while others with the nonparametric Wilcoxon Mann Whitney test to compare BL and DLBCL cell lines. P-values were calculated for each metabolite and corrected for multiple testing, setting the FDR threshold at 5% (q values < 0.05). Nine metabolites were considered as significantly altered (Appendix 6). The metabolites L-Alanine, phosphorylcholine (CHoP), glycine, myoinositol and D-Glutamic acid were up-regulated in BL, while taurine, creatine, UDP-N-acetyl-D-glucose and L-Glutamine were down-regulated (**Figure 3.6B**). These significantly altered metabolites were used next as an input list for integrative pathway analysis.

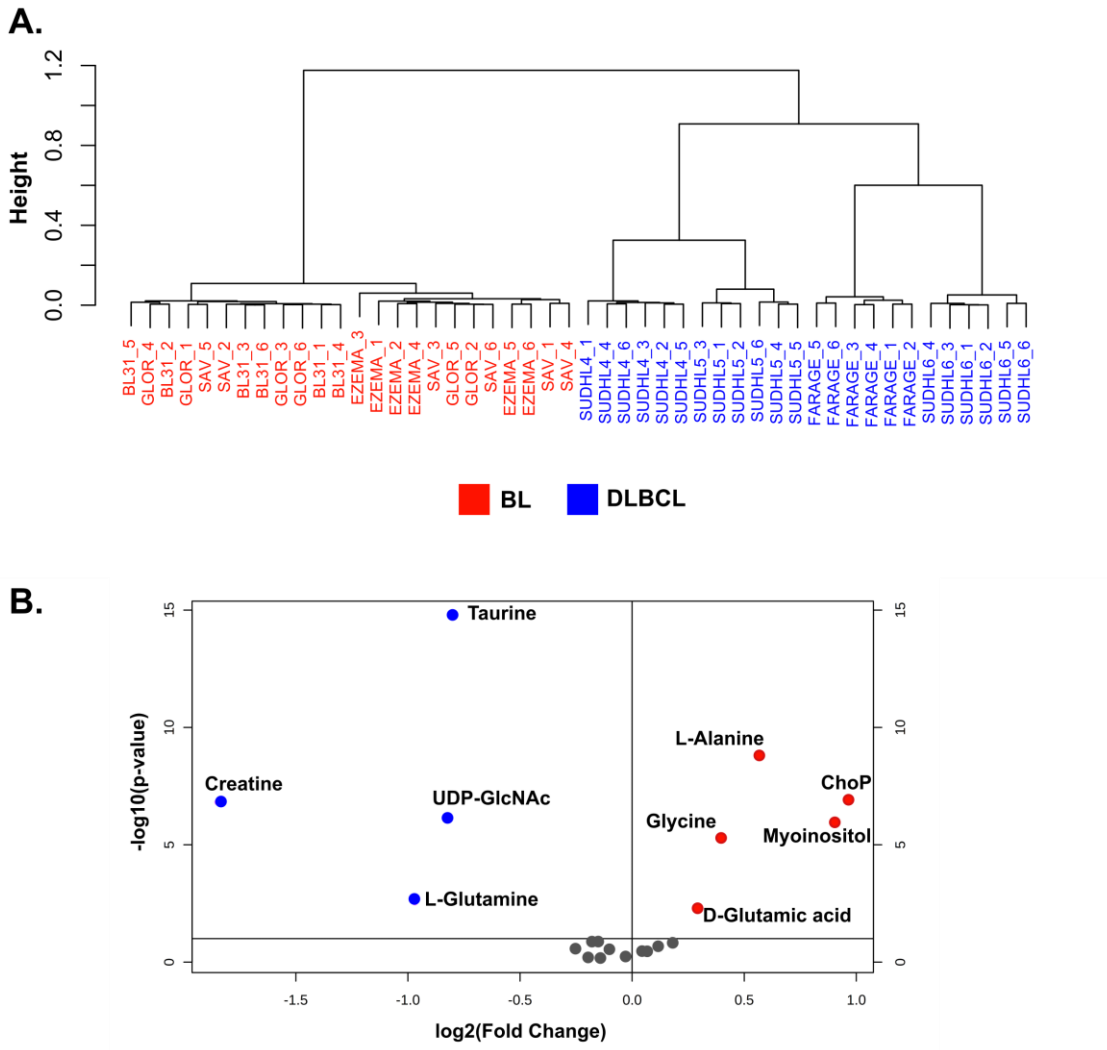


Figure 3.6. Hierarchical clustering and univariate analysis with metabolomic data. (A) Hierarchical clustering analysis was applied with Ward clustering method and euclidean distance measuring to determine clusters between the data. Results were visualised in the form of a dendrogram. **(B)** Significant metabolites selected by univariate analysis with FDR threshold of 5% are presented in the volcano plot. Both fold changes and p-values are log transformed.

3.3.5. Integrative analysis between BL and DLBCL

The NMR metabolomic data were used for integrative pathway analysis together with the transcriptomic data that we retrieved from the same NHL cell lines. RNAseq data were generated and analysed using the same pipeline as for the RNAseq data from primary tumours. Differential expression analysis identified 39226 transcripts corresponding to 14,421 genes of which 335 genes were statistically significant altered (q values < 0.1). We identified 113 of these genes shared common expression with the 6,475 significant differentially expressed genes from the primary tumours, (61 upregulated and 52 downregulated in BL compared to DLBCL) (Appendix 7).

We looked from a pathway perspective to integrate both lists with common significantly altered genes and metabolites to explore any interactions in the network between metabolism and gene expression. Enrichment and topology analysis revealed 31 metabolic pathways from KEGG database (Appendix 8). The first 6 metabolic pathways were nominally significant (p-values < 0.05) (**Figure 3.8A**). However, correcting for multiple testing with an FDR threshold of 10% none of them could be considered significantly altered. Still, these pathways together with the next 4 metabolic pathways (to contain the pathway of interest *glycine, serine and threonine* metabolism) were visualised in the KEGG global metabolic network to explore any associations (**Figure 3.7**). The *alanine, aspartate and glutamine pathway* had the lowest p-value (p-value = 0.0013) and contained the most mapped features (4 genes and metabolites out of 61) compared to the other

metabolic pathways. The non-essential amino acid L-alanine (fold change = 1.03, q value < 0.001) together with the metabolic genes ALDH5A1 (beta = 1.61, q value = 0.022) and RIMKLB (beta = 4.48, q values < 0.001) were upregulated in BL, while glutamine (fold change = -1.05, q value < 0.0014) was downregulated (**Figure 3.8B**). Moreover, integrative analysis with KEGG metabolic and regulatory pathways also identified the *alanine, aspartate and glutamate metabolism* as the most significant pathway (p-value = 0.0004, q value = 0.107) among the 180 KEGG pathways (Appendix 10). Other metabolic genes that mapped together with metabolites were the PCCA (beta = -2.03, q values = 0.0017) in *glyoxylate and dicarboxylate metabolism*; and the MBOAT2 (beta = 2.84, q values < 0.001) and LPIN1 (beta = -0.99, q values < 0.073) in *glycerophospholipid metabolism* (**Figure 3.8A**). Pathways that mapped only significant genes or metabolites were not examined any further in this study.

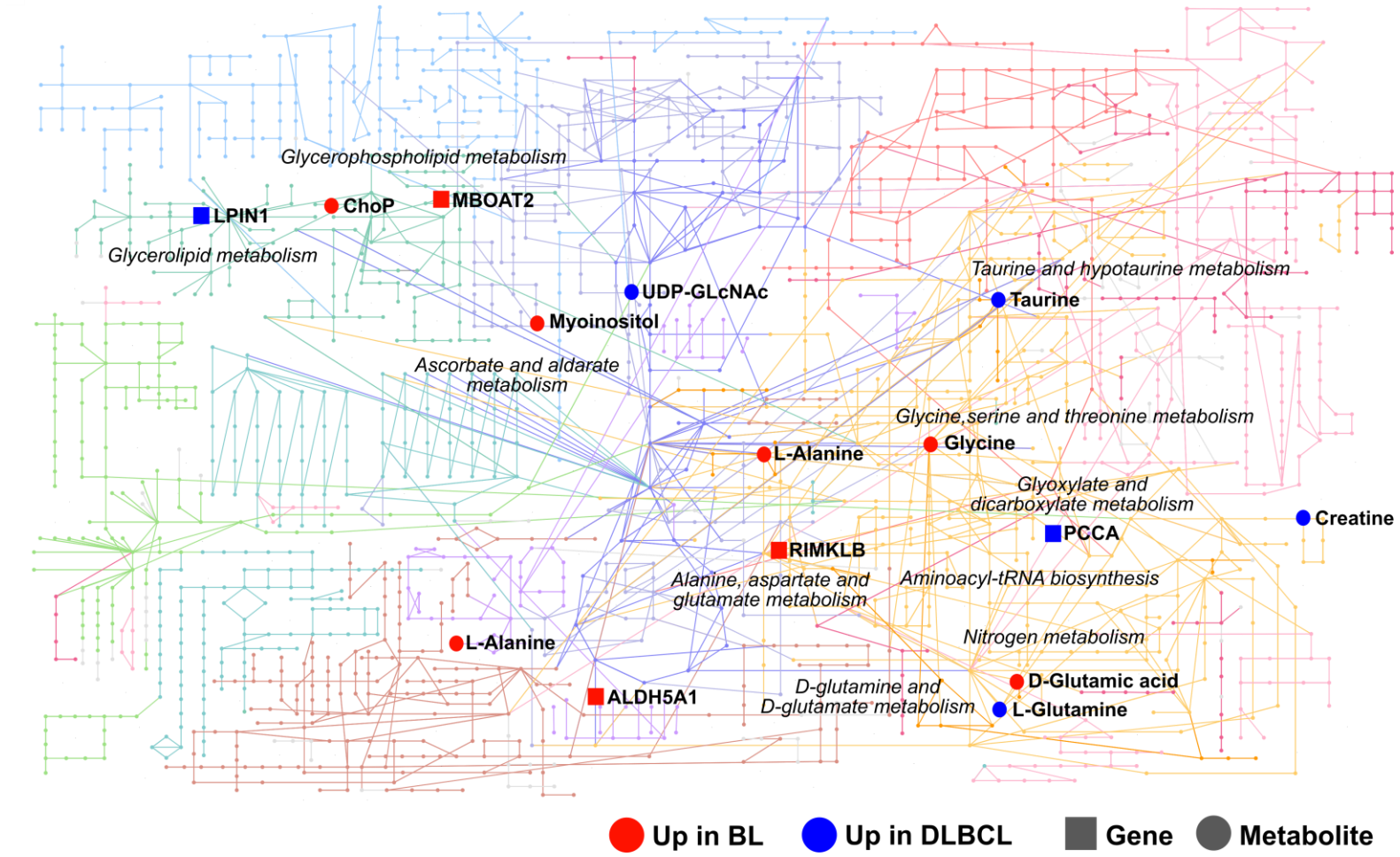


Figure 3.7. Pathway based integration analysis with metabolomic and transcriptomics data between BL and DLBCL cell lines. Integration pathway results were mapped in the KEGG global metabolic network.

A.

Top 10 metabolic pathways	Total	Expected	Hits	p-value	-log(p-value)	q value	Impact	Features
Alanine, aspartate and glutamate metabolism	61	0.50	4	0.0013	6.67	0.107	0.233	L-Alanine, L-Glutamine, ALDH5A1, RIMKLB
D-Glutamine and D-glutamate metabolism	10	0.08	2	0.0028	5.89	0.116	0.333	D-Glutamic acid, L-Glutamine
Glyoxylate and dicarboxylate metabolism	56	0.46	3	0.0099	4.62	0.277	0.109	Glycine, L-Glutamine, PCCA
Aminoacyl-tRNA biosynthesis	74	0.61	3	0.0211	3.86	0.442	0.041	L-Glutamine, Glycine, L-Alanine
Glycerophospholipid metabolism	86	0.70	3	0.0312	3.47	0.452	0.129	ChoP,MBOAT2, LPIN1
Glycerolipid metabolism	35	0.29	2	0.0323	3.43	0.452	0.147	MBOAT2, LPIN1
Nitrogen metabolism	10	0.08	1	0.0792	2.54	0.950	0.111	L-Glutamine
Ascorbate and aldarate metabolism	13	0.11	1	0.1017	2.29	0.967	0.083	Myoinositol
Glycine, serine and threonine metabolism	68	0.56	2	0.1054	2.25	0.967	0.284	Glycine, Creatine
Taurine and hypotaurine metabolism	16	0.13	1	0.1237	2.09	0.967	0.267	Taurine

B.

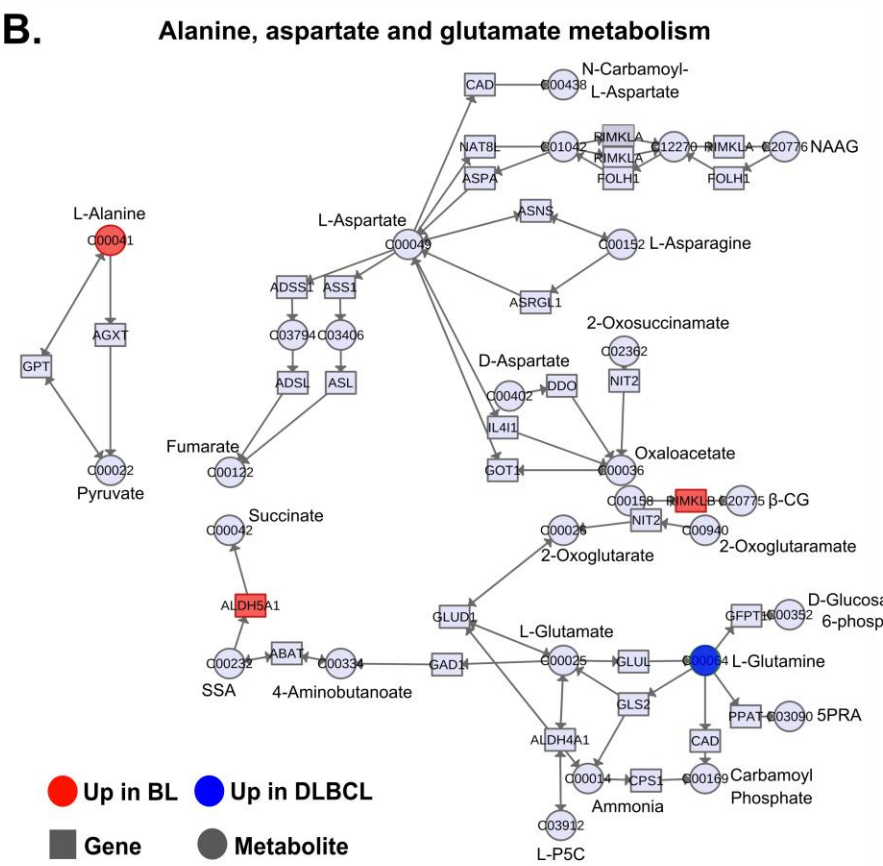


Figure 3.8. Details of integration pathway results. (A) Top 10 matched pathways from integration pathway analysis with p-values from the pathway enrichment analysis and pathway impact values from pathway topology analysis. **(B)** The integration network of the *alanine, aspartate and glutamate metabolism* pathway, as retrieved from MetaboAnalyst module.

3.4. Discussion

Metabolic research began to unveil the key regulatory roles in GCs lymphomagenesis and other haematological cancers. To date, there has been no therapeutic metabolic targets or metabolic biomarkers widely available for GCs-derived lymphomas. To bridge this gap, we started by exploring the transcriptomic profile of 19 endemic BL and 12 GCB-DLBCL primary tumours for any transcriptomic differences in these complex metabolic regulations. Dimensionality reduction results with PCA suggested a transcriptomic distinction (t-test for PC1: p-value = 4,96e-13, q value = 5e-13) between BL and DLBCL tumours (**Figure 3.3B**). These findings are in line with previous gene expression profile studies (Dave et al., 2006; Schmitz et al., 2012). However, we additionally report genes related to cytoskeleton structure (KRT13) and to Fc receptors (FCRL5) as the even more important ones compared to the other studies to drive this separation. Although metabolic genes did not seem to contribute the most in the separation of the two diseases, the PYCR1 in proline metabolism and genes involved in biosynthesis of serine (PHGDH and PSAT1) are the most extreme metabolic genes to drive the variation along the first principal component which explains 37.65% of the variance. Nevertheless, differential expression analysis and two-sided Fisher exact test revealed significant difference (p value = 0.0025, odds ratio = 0.66) in BL/DLBCL between the number of significant metabolic genes and the rest of regulatory significant genes (**Figure 3.4B**). These findings highlight the

necessity of understanding and monitoring metabolic properties in these two lymphomas. We also showed that the previously reported serine genes PHGDH and PSAT1 had a significant higher expression in BL compared to DLBCL cases, which was not observed in transcriptomic data from our in-house NHL cell lines. More emphasis was given in the results from transcriptomic analysis of data from primary tumours rather than those from cell lines. Although, immortalized cell lines are a valuable in vitro model for cancer research, several inherent limitations are related from their use and might have affected our analysis. Most important of them are the possible misidentification of cell lines, the cross contamination, and the genotypic instability that impacts cells phenotype (Mirabelli et al., 2019). Immortalized cell lines are artificial systems that have adapted/changed many biological processes to sustain proliferation in vitro compared to the initial tumour from which they are derived from. Nevertheless, transcriptomic data from cell lines represent RNA transcripts from pure tumour cells, while primary tumour samples contain a mixture of RNA transcripts between tumour and non-tumour cells (as previously explained in section 3.2.1). The extend of non-tumour cell contamination is another limitation that affects our transcriptomic analysis, especially for BL primary tumours of which tumour purity was not reported. it remains unclear, the impact of tumour purity on gene expression or which gene pairs are associated with purity adjustment. In BL cases the upregulation of metabolic genes in the de novo biosynthesis of serine is potentially important, since a) serine has a major role in tumour cell growth and proliferation as a precursor for protein and amino acid biosynthesis; b) serine contributes one-

carbon units to folate cycle and thus is essential for synthesis of nucleic acids and NADPH regeneration (Fan et al., 2014; Yang and Vousden, 2016); and c) recent evidence points to support antioxidant activity via the production of glutathione to protect from ROS (Mehrmohamadi et al., 2014; Wang et al., 2017b; Ye et al., 2014). Altogether, we propose that elevated expression of serine genes combining with findings from Mrs Zuhal Eraslan work can contribute to key functional properties of metabolic regulation of GCs lymphomagenesis in BL. However, additional investigation is needed using knockout mouse models for PHGDH to reveal key properties of the serine metabolism in these types of lymphomas.

To examine this concept in more detail I performed GSEA with the aim to elucidate the role of metabolic and signalling gene-sets/pathways. As expected, BL cases showed upregulation in gene sets relative to c-Myc gene (MYC targets V1 and MYC targets V2). More importantly though, this showed upregulation in BL of the mTORC1 signalling pathway, which is known to control cell growth and metabolism (Laplante and Sabatini, 2012; Valvezan and Manning, 2019). Of particular relevance to the previous results, mTORC1 regulates the serine/glycine de novo synthesis via gene expression of glycolysis and serine biosynthesis pathway (Wang et al., 2017b). Furthermore, GSEA results suggested that the DLBCL cases upregulated cellular signalling pathways such as the NFkB, the JAK-STAT and the KRAS pathways, possibly as a response to inflammation, TNF and INFγ signals. Most of these pathways are associated with metabolic readjustments in cancer cells facilitated by the tumour microenvironment (Habtetsion et al., 2018;

Londhe et al., 2018; Son et al., 2013). Moreover, higher expression of the IDO1 gene, which is involved in interferon gamma response pathway, possibly protects DLBCL tumours from immune response. Evidence suggested that the IDO1 gene is responsible for the depletion of the essential amino acid tryptophan in kynurenine pathway, leading to immunosuppression (van Baren and Van den Eynde, 2015). More importantly, the IDO1 enzyme plays a key role in cancer immunosurveillance. When the first malignant cells arise, activated dendritic cells (DCs) can secret low levels of the IDO1 enzyme and inhibit tumour growth by depleting tryptophan from the tumour microenvironment. In the phase when tumour growth escapes from the control of immune system, tumour cells produce high levels of IDO1 (Hornyák et al., 2018). The depletion of tryptophan and the accumulation of kynurenine, caused by the IDO1 enzyme, lead to immunosuppression and immunological tolerance by inhibiting effector T-cell and NK cell functions and stimulating regulatory T-cells (Zhao et al., 2012). IDO1 also regulates the activation of myeloid-derived suppressor cells (MDSCs), which supress the activity of antitumour effector T-cells function (Holmgaard et al., 2015). Increased kynurenine levels activate the aryl hydrocarbon receptor (AhR) that regulates DCs from immunogenic to tolerogenic (Mellor et al., 2002). Godin-Ethier et al. also showed that elevated expression of IDO1 gene in different cancer types is associated with unfavourable clinical outcome (Godin-Ethier et al., 2011). Overall, the transcriptomic data recapitulate that the BL cases possessed a distinct gene expression profile compared to DLBCL, suggestive of altered function of metabolism, such as the serine metabolic genes mediated by the mTORC1

pathway. On the opposite, DLBCL cases were more dependant to extracellular signals either from cytokines (INFγ response) or inflammations that triggers activation of cellular signalling pathways to support tumour resistance and proliferation (Chen et al., 2018).

Going beyond the transcriptional differences, multi-Omics integration at the pathway level was employed to explore any interactions between genes and metabolites in gene metabolic pathways. In analysing the untargeted 1D ^{1}H NMR data from the NHL cell lines (from Dr. Farhat Khanim's lab, prepared by Mrs Zuhal Eraslan) we were able to assign and eventually examine 21 metabolites (Appendix 6). Although, this small number of assigned metabolites restricted the total network interactions, we were still able to extract some useful observations contributing to the generation of a new experimental hypothesis. Results from these metabolites of the NMR data showed an unsupervised separation between the two diseases (**Figure 3.6A**) similar to the one observed with the transcriptomic data. Moreover, we found that BL cell lines upregulated non-essential amino acids, possibly synthesised from glycolytic intermediates, such as the L-alanine (FDR < 0.001, Fold Change = 1.03) and the glycine (FDR < 0.001, Fold Change = 1.02). Cells can either import non-essential amino acids from microenvironment or they can synthesise them *de novo*. Glycine uses serine as a precursor for biosynthesis and thus elevated glycine in BL may indicate upregulation of serine biosynthesis. As mentioned previously, upregulation of serine biosynthetic metabolic genes was observed in BL primary tumours but not in BL cell lines. Of relevance to these

findings, elevated serine biosynthesis possibly increased glycolysis and suppressed OXPHOS (Samanta and Semenza, 2016). In contrast, we observed that DLBCL cell lines depended more on L-Glutamine (FDR = 0.0014, Fold Change = -1.05) and taurine (FDR < 0.001, Fold Change = -1.04) to increase cell growth and modulate inflammatory pathways (Cluntun et al., 2017; Hensley et al., 2013; Sartori et al., 2018).

The integration analysis for both metabolomic and transcriptomic data highlighted the *alanine, aspartate and glutamine metabolism* as the most significant pathway (p-value = 0.0013) with the most mapped features. However, no direct interactions between significant genes and metabolites were identified in this network (**Figure 3.8B**). Surprisingly, the RIMKLB gene which encodes the enzyme β-citrylglutamate synthase B and catalyses the synthesis of β-citryl-L-glutamate (BCG) metabolite, was upregulated (beta = 4.48, q values < 0.001) in BL. The BCG is a dipeptide mostly detected in central nervous system (CNS) and in testis, with studies to suggest that act as an iron (Fe) and copper (Cu) chelator (Hamada-Kanazawa et al., 2010; Narahara et al., 2010). Both these two metals are essential for rapid cellular proliferation and chelators related to them are compounds responsible to bind, regulate and detoxify the cells from these metals. To date, chelators are under investigation as potential anti-tumour targets for cancer therapy (Fryknäs et al., 2016; Gaur et al., 2018; Lee et al., 2016; Lui et al., 2015). Similar expression to the RIMKLB gene in BL showed the acetaldehyde dehydrogenase (ALDH5A1) gene (beta = 1.61, q value = 0.022), which encodes

the succinate-semialdehyde dehydrogenase enzyme to produce succinate from succinate semialdehyde and NADH. In general, high levels of aldehydes dehydrogenases enzymes (ALDHs) were found in several cancers (Dollé et al., 2012; Kahlert et al., 2012; Kang et al., 2016; Tomita et al., 2016). Although, ALDHs detoxify the cells from aldehyde substrates, increasing evidence revealed their key role in mitochondrial redox homeostasis regulating the NAD+/NADH ratio (Missihoun et al., 2018; Wang et al., 2017a). In BL cells the reduction of NAD^+ to NADH partially from the activity of the ALDH5A1 is likely to be associated with the increased expression of the PYCR1 gene (beta = 2.56, q values < 0.001). As previously mentioned, the PYCR1 gene which was found as the most essential metabolic gene in PCA, encodes a mitochondrial NADH-oxidising enzyme in biosynthesis of proline. This finding is important as it has been recently suggested that upregulation of PYCR1 in cancer cell lines lower the $NADH/NAD^+$ which retains the TCA cycle activity when ETC flux is limiting (Hollinshead et al., 2018). Thus, we suggested that inhibition of the PYRC1 or the ALDH5A1 enzyme may act to disturb the mitochondrial redox homeostasis and consequently make these BL tumours more vulnerable to current available treatments.

CHAPTER 4

MULTI-OMICS DATA INTEGRATION FOR CANCER CELL LINES WITH MACHINE LEARNING

4.1. Introduction

So far, we have described multi-Omics data integration to highlight novel metabolic genes and pathways related to CLL and Non-Hodgkin lymphomas. Our next step was to determine which features (genes and/or metabolites) have a predictive value to classify new cases into haematological cancers. However, such classification models cannot be constructed from the methods used so far, namely Genome Scale Metabolic Modelling nor pathway-based approaches. Nowadays, Machine Learning (ML) algorithms have the ability to get trained (or "learn") from the data and build models that are able to accurately assign new cases into a specified class.

We have also seen that transcriptomics and metabolomics high-throughput technologies generate distinct data that capture and explain the biological information at different stages of the transition from the genome to the phenotype. The integration and analysis of such Omics modalities is not only affected by the inherently different biological nature of the data, but it is also influenced by any limitations of each Omics platform that generate the data (Leek et al., 2010). An additional complexity is that Omics modalities measure millions or tens of thousands of features (as known as variables, or dimensions) per sample and thus these features have to be combined and analysed in high dimensional spaces. The expression *"curse of dimensionality"*, was introduced to describe the pitfalls of analysing data in high dimensional space (Bellman, 2010). Working with such data

is demanding for most statistical significance approaches, since these try to extract statistical inference from a large number of supposedly independent variables when some of these may not be independent (multicollinearity). Furthermore, it is also challenging for data visualization, as humans are able to interpret up to three dimensions plots. Most ML algorithms can extract the most informative variables from high dimensional Omics data that are not always statistically associated with the phenotype. Thus, ML approaches are now widely used with Omics datatypes to perform dimensionality reduction, feature selection or construct classification/regression models.

The application of metabolomics in cancer research significantly contributed to the understanding of cancer development and progression. However, to date, there has been no systematic metabolomic profiling for primary cancer and matched normal samples, similar to “The Cancer Genome Atlas Program” (TCGA, https://www.cancer.gov/tcga) for Genomics, which has characterized genomic, epigenomic, transcriptomic and proteomic data for over 80,000 primary cancer samples. The absence of such systematic effort to characterize the metabolic profile of primary samples, together with difficulties associated with identifying metabolites from NMR and MS spectra, restricts most metabolic studies to use cancers cell lines as a model to investigate cancer metabolism. In cancer research, several studies have already used machine learning approaches to integrate Omics data from cancer cell lines and construct computational models for biomarker discovery or for drug response predictive models (Garnett et al., 2012;

Geeleher et al., 2014; Stransky et al., 2015). Cancer cell lines have served as a valuable tool for in vitro modelling since 1951, when the first cancer cell line established from patient's tumour cells that were capable to grow in vitro for prolonged periods (Scherer et al., 1953). Today, thousands of human and animal cell lines have been formed with Omics profile data acquired and stored in well-organised databases. One of the most significant efforts to collect the profile for 1457 cell lines is the Cancer Cell Line Encyclopaedia (CCLE) project (Barretina et al., 2012). The CCLE provides public access to datasets on gene expression, mutation, miRNA, copy number variations and drug sensitivity for most cell lines and thus it is an ideal source of well-structured and complete datasets for multi-Omics integration studies. So far, studies have explored CCLE datasets with a diversity of machine learning algorithms, such as elastic net regression (Jang et al., 2013), random forest (Berlow et al., 2014), support vector machine (Dong et al., 2015) and dual layer network (Zhang et al., 2015). In addition, deep learning approaches were also used with these datasets to integrate Omics data with drug sensitivity measurements (Ding et al., 2018; Li et al., 2019b; Zhao et al., 2019). Most of these studies depend on multi-Omics integration to build prediction models for drug response, however none of them has used machine learning approaches to explore metabolism. A recent study has released the metabolic profile of 928 CCLE cell lines, which contains measurements for 225 metabolites with liquid chromatography-mass spectrometry (LC-MS) (Li et al., 2019a). Li et al. have effectively applied the statistical method of linear regression analysis to identify associations between genetic (mutation, copy number variation and RNAseq data)

and metabolic features.

These high-quality, comprehensive metabolomic data can highlight the metabolic diversity between haematological and other cancer types and reveal association dependencies across the cell lines, which can lead to the discovery of new anti-cancer metabolomic targets. Therefore, we integrated transcriptomic and metabolomic data from CCLE using ML approaches to explore associations between gene expression and metabolites that are related to haematopoietic cancer cell lines. Taking into consideration the impact of environmental factors in cell's metabolic response, we have selected only 427 cell lines which grow under the same media condition. We started by performing dimensionality reduction in both unique and integrated Omics datasets with popular unsupervised ML approaches to explore the heterogeneity of cancer cell lines. After revealing the distinct transcriptomic and metabolic profiles of haematopoietic cell lines, we utilised a supervised learning approach to integrate Omics data and construct a classification model that identifies highly correlated or co-expressed genes and metabolites associated with haematopoietic cell lines.

4.2. Materials and methods

4.2.1. Omics datasets

Publicly available transcriptomic and metabolomic data were retrieved from the CCLE database (https://portals.broadinstitute.org/ccle/data). Out of the 1,457 cancer related cell lines in the database, 891 cell lines were chosen for having both transcriptomic (1019 cell lines) and metabolomic (928 cell lines) profiles. From those a total number of 427 cell lines were selected based on their ability to grow under common culture media conditions (RPMI1640 plus 10% foetal bovine serum (FBS)). 90 cell lines were classified as haematopoietic cell lines (leukaemia, lymphomas, and multiple myeloma) and the rest as other tumour cell lines belonging to 15 different cancer types, including lung, skin and colon cancers to name a few (Appendix 11). Despite transcriptomic profiles from microarrays and RNAseq data were available, we retrieved only the RNAseq data due to higher sensitivity and specificity of the method (Zhao et al., 2014). Then, the corresponding transcripts per million (TPM) values for 52,173 genes were quantile normalised with the *affy 1.66* R statistical package and log transformed. Similarly, LC-MS metabolomic data with intensities for 124 polar and 101 lipid metabolites were used for the integration analysis with the RNAseq data. The curated LC-MS data (Li et al., 2019a) were used here, meaning that these data had already been normalised by median, log transformed and scaled. For scaling raw metabolomic

data across samples, pooled 20 samples composed of mixed metabolites from 11 cell lines (NCIH446, DMS79, NCIH460, DMS53, NCIH69, HCC1954, CAMA1, KYSE180, NMCG1, UACC257 and AU565) were used as a reference. The peak area for each metabolite in each sample was standardized by computing the ratio between the value observed in the sample and the value observed in the “nearest neighbour” pooled sample. These ratios were then multiplied by the mean value of all reference samples for each analyte to obtain standardized peak areas. To concatenate the Omics data for the integration analysis both datasets were organised in matrices with the rows being the cell lines samples (in the same order and number of rows) and the columns the features (either genes or metabolites) measured from each individual Omics technology. Concatenation of the two Omics datasets by columns was done in Python 3.7.4 scripting language using the *np.concatenation* function *of numpy* library.

4.2.2. Dimensionality reduction with machine learning

We applied unsupervised learning to reduce dimensions and explore each Omics datatype separately as well as in an integrative multi-Omics fashion (**Figure 4.1A**). For comparison reasons, three commonly used dimensionality reduction techniques were employed: Principal Component Analysis (PCA) (Tipping and Bishop, 1999), t-Distributed Stochastic Neighbor Embedding (tSNE) (van der Maaten and Hinton, 2008) and Uniform Manifold Approximation and Projection

(UMAP) (McInnes et al., 2018). These analyses were performed in Python 3.7.4 scripting language. Multi-Omics integration of the data itself was achieved by concatenating the RNAseq with the LC-MS matrices into a big matrix. Then, the linear dimension reduction method of PCA was first implemented with the *sklearn.decomposition.PCA* class from the python library *scikit-learn 0.21.3*. Similarly, the class *sklearn.decomposition.tSNE* was used for the non-linear method tSNE and the python library *umap 0.4* for the UMAP method, which is also a non-linear dimension reduction approach.

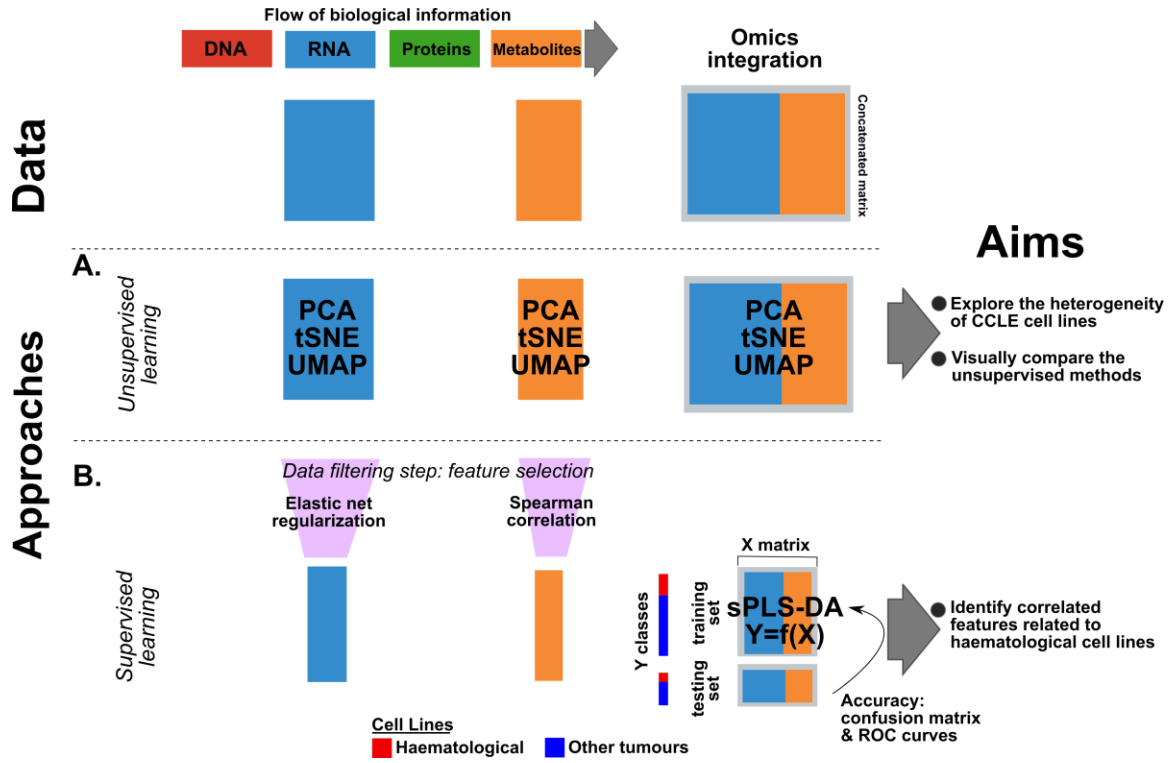


Figure 4.1. Overview of Omics integration analysis with CCLE datasets. Two independent approaches (unsupervised and supervised learning) were applied to explore each Omics datatype separately and in an integrative multi-Omics concept.

4.2.3. Supervised Omics integration with sPLS-DA

4.2.3.1. Principle of sPLS-DA

So far, we have used PCA as a method to reduce data's dimensions but also retain the most vital information. We can achieve this with PCA by projection into linear combination of new variables, known as latent variables or components. Similar to PCA, the Partial Least Squares Discriminant analysis (PLS-DA), which is based on the Partial Least Squares (PLS) regression method, also projects the data into PLS-components. However, instead of maximizing the variance of components like PCA does, PLS-DA maximises the covariance between PLS-components from two datasets (Wold et al., 2001). Thus, PLS-DA is often used in a supervised manner by combining quantitative with qualitative matrices. To define PLS-components, PLS-DA calculates coefficients for each data variable, known as loading vectors. Loading vectors demonstrate the importance of each variable in PLS-components and they can be used for feature selection and classification. An extension of PLS-DA is the sparse PLS-DA (sPLS-DA) which introduce the LASSO penalties (or else L1 regularization) (Lê Cao et al., 2011) to loading vectors. It enables the generation of a sparse model (a simpler model that shrinks loading vectors defining the PLS-components to zero) that selects simultaneous features from Omics datasets discriminating for classes (Singh et al., 2016). Here, we have used the sPLS-DA method to integrate Omics data and extract the most important features that discriminate haematopoietic cell lines from the other tumour cell lines.

4.2.3.2. Pre- select essential features from Omics datasets.

Prior to the use of sPLS-DA, each Omics dataset needs to be filtered in order to select the most informative features and consequently reduce the variables building a less complex model, which is more interpretable and computationally efficient during the next step of tuning parameters (Rohart et al., 2017; Singh et al., 2016). Dealing with the high dimensional RNAseq data, the elastic net regularization method (a combination of L1 and L2 regularization) from *glmnet 4.0* R statistical package (Simon et al., 2011) was applied to select 81 genes as the most essential variables/features for the integration analysis (Appendix 12). Furthermore, univariate feature selection was applied to the LC-MS metabolomic data with the R function *cor.test* for the non-parametric Spearman correlation method. Correcting for multiple testing with False Discovery Rate, a cut-off of 0.05 (FDR < 0.05) was used to select 188 metabolites as the most significant related to haematopoietic cell lines (Appendix 13).

4.2.3.3. Construction process of the predictive multi-Omics model

Omics data were concatenate into a matrix X and used as an input to a typical machine learning setup: Y=f(X). The Y vector is the class label for haematopoietic or other cancer cell line and the function f() is the classification rule learnt from the sparse PLS-DA algorithm (**Figure 4.1B**). The DIABLO method from the *mixOmics 6.10.9* R package (Rohart et al., 2017; Singh et al., 2016) was employed to

implement sPLS-DA as previously described. DIABLO is partly based on the Generalised Canonical Correlation Analysis (Tenenhaus et al., 2014) to multiple match datasets. Firstly, the initial dataset of 427 cell lines was split into a training set (60 haematopoietic and 224 other tumours cell lines) and a testing set (30 haematopoietic cell lines and 114 other tumours cell lines). A five-fold cross-validation with 50 repeats was performed without variable selection to assess the overall performance of the model and select the best hyperparameters. Hyperparameters are parameters set to configure the structure of the model prior to the training. PLS-DA requires an optimal number of PLS components and an optimal number of features to be extracted from the loading vectors of each PLS-components. The classification error rates with both Mahalanobis and maximum classification distances were computed to select the number of optimal components for the final model as suggested by the tool authors (Rohart et al., 2017; Singh et al., 2019 and their tutorial: https://mixomicsteam.github.io/Bookdown/plsda.html#tuning:sPLSDA). The Balanced error rate (BER) represents the average proportion of wrongly classified samples, weighted by the total number of samples in each class and thus is suitable for our study since the number of haematopoietic cell lines is much lower than the other tumour cell lines. Once again, we run a five-fold cross validation repeated 50 times under the assumption of a strong correlation between features and cancer types to extract the average number of features on each component across all folds and repeats as the optimal number of features to be extracted from the final model. After selecting the hyperparameters, the final and tuned sparse PLS-DA model was constructed. The

LASSO algorithm selected the most essential features from the loading vectors of the selected components in each Omics data. Results were displayed with a relevance network, which represents the correlation structure between genes and metabolites. Finally, the data testing set was utilised to assess the classification performance of the final PLS-DA model with a confusion matrix and the area under the ROC (Receiver Operating Characteristic) curve method.

4.3. Results

4.3.1. Reducing dimension in Omics datasets

We first employed the most common unsupervised algorithms for dimensionality reduction to visualise any interesting aspects in single and integrated Omics data. We started with the PCA approach as fully presented in chapter 3. Results from PCA showed the first principal component (PC1), which contains most of the data variation in RNAseq (13%), LC-MS (24%) and integrated Omics data (16%), to clearly separate the haematopoietic cell lines from the other types (**Figure 4.2A1-A3**). However, PCA performs poorly when integrating unbalanced scaled datasets, where the features of one dataset are measured on a different scale (or range of values) than the features of the other. As a consequence, the datasets do not contribute equally to the model fitting, where the

higher values dominate in PCA over the other dataset, such as in our case where the RNAseq data dominates over the LC-MS data. One method to overcome this limitation is to re-scale the distributions by using the ratio of the distance of each value from the minimum value in each dataset to the range of values in each dataset: (x – min) / (max – min). This feature scaling approach fits the distribution of each dataset on a scale between 0 to 1, however such an approach was not applied in our analysis. Thus, performing PCA for integrative Omics datasets serves mostly for illustration purposes. Due to PCA limitation in preserving any non-linear structure of the data and tendency to select samples with large pairwise differences to maximise the variance, the tSNE approach was also applied here as an alternative to PCA. As expected, tSNE achieved clearer separation and more dense groups between different types of cell lines compared to the PCA method. This is because tSNE is primarily designed to preserve small pairwise differences by bringing together the neighbouring samples. In other words, tSNE preserved the local structure of the data which resulted in clearer groups and visualised better in 2D-plots the heterogeneity of single and integrated Omics data (**Figure 4.2B1-B3**). More specifically, both in RNAseq and integrated Omics datasets tSNE not only separated the haematopoietic cell lines, like PCA did, but also distinguished groups among leukaemia, lymphomas and multiple myelomas cell lines (**Figure 4.2B1, B3**). However, tSNE is designed to preserve the distance within rather than between the groups, known as the global structure of the data. To tackle this issue, the recent dimension reduction method of UMAP was also applied. UMAP not only captures the local structure of the data similar to tSNE, but it preserves non-linear

distances on a global scale (Becht et al., 2019; McInnes et al., 2018). Here, UMAP was used for exploratory data analysis and it was also compared visually against PCA and tSNE. Consequently, plotting UMAP components showed even more distinct and tight groups between cancer types of the cell lines compared to previous methods (**Figure 4.2C1-C3**). Furthermore, UMAP strongly highlighted that haematopoietic cancer cell lines were dissimilar from the other types both in transcriptome and metabolome level.

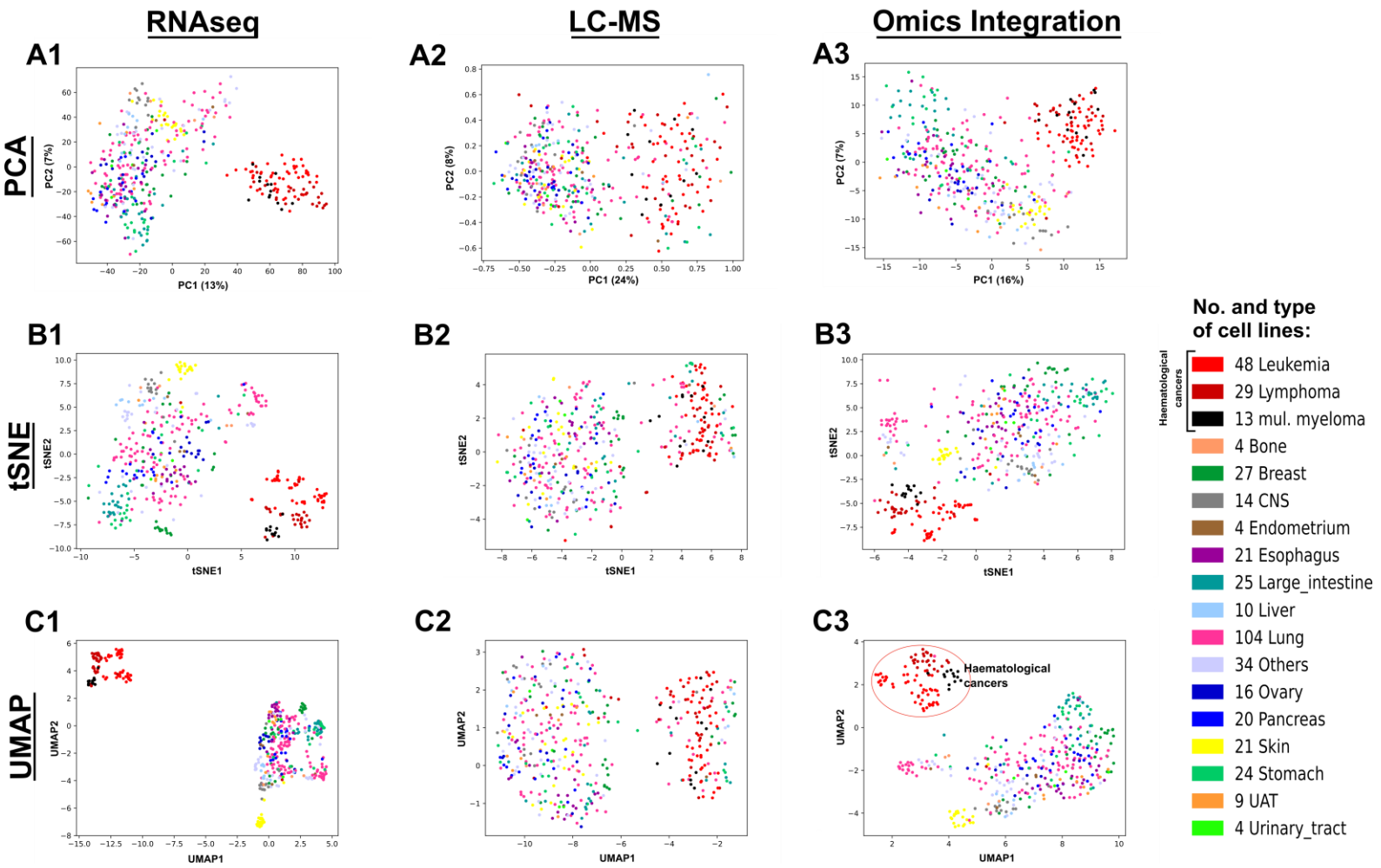


Figure 4.2. Dimensionality reduction in CCLE Omics datasets. Dimensionality reduction applied with a series of unsupervised algorithms: PCA, tSNE and UMAP for single and integrated Omics data.

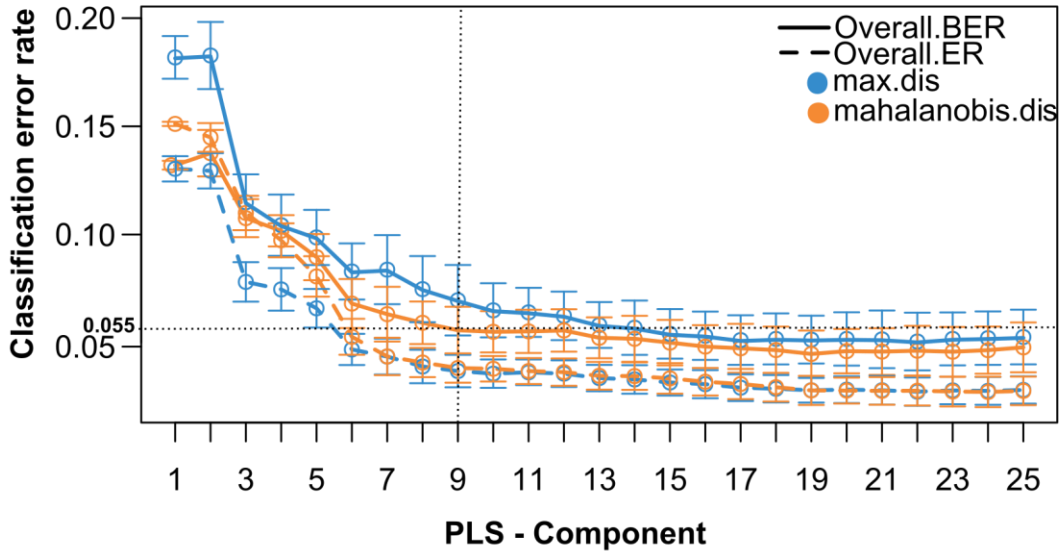
4.3.2. Supervised analysis for Omics integration

We have previously shown that haematopoietic cell lines possess a distinct transcriptomic and metabolic profiles compared to other types of cancer cell lines. We then turned to apply multi-Omics integration using a supervised approach, as reported in previous studies (Gromski et al., 2015; Koenig et al., 2018; Lê Cao et al., 2011; Singh et al., 2019), to identify highly correlated genes and metabolites related to haematological cancer cell lines. As a pre-processing step, feature selection was applied to each Omics datatype individually to address the issues of curse of dimensionality and multicollinearity. Feature selection significantly reduced the variance from the high-dimensional RNAseq data by selecting the 81 most essentially expressed genes to be used for the integration analysis. Similarly, the non-parametric Spearman correlation test was used with the LC-MS data to evaluate the relationship between each individual metabolite against the types of cancer cell lines and identified 188 metabolites related to the haematopoietic cell lines.

The reduced RNAseq and LC-MS datasets were concatenated into a matrix X and used as an input to the sPLS-DA method together with a class vector Y representing the cancer types (**Figure 4.1B**). Due to the relatively low number of samples, input data were split into a training set (284 cell lines) to build the PLS-DA classification model and a testing set (143 cell lines) to evaluate the classification performance of this model. Tuning hyperparameters for the final model with cross validation the algorithm predicted the first 9 PLS-components

(n=9) as the optimal number of components for the final model. As shown in **Figure 4.3A**, both classification distances for BER seems to reach a plateau/low error rate (BER=0.055 for Mahalanobis distance) at the 9th PLS-component to achieve good performance for the model. Furthermore, the table in **Figure 4.3B** summarises the average number of features across all folds and repeats for each pair of selected PLS-component to be used downstream in the final model.

A.



B.

Average no. of features per PLS-component

Data	Comp1	Comp2	Comp3	Comp4	Comp5	Comp6	Comp7	Comp8	Comp9
RNAseq	5	8	10	9	10	7	8	10	10
LC-MS	6	2	1	1	1	1	1	1	1

Figure 4.3. Tuning hyperparameters. (A) The overall and Balanced classification error rate were calculated with Mahalanobis and maximum classification distances from a five-fold cross-validation with 50 repeats to select the optimal number of PLS- components. **(B)** Similarly, a five-fold cross validation with 50 repeats calculated the average number of featured per PLS-component for each datatype to be extracted from the final model.

Selecting these hyperparameters, we run the final sPLS-DA to identify strong relationships between genes and metabolites based on constrains with the PLS-components (**Figure 4.1**). sPLS-DA generates a pair of components each associated to each Omics dataset. Each individual Omics datatype was examined to assess for any similarities between the sample cell lines in the reduced dimensional space spanned by the first two PLS-components. As illustrated with other unsupervised methods (**Figure 4.2**), sPLS-DA also discriminated the haematopoietic cell lines from the other cancer cell lines on the 1st PLS-component both in RNAseq and LC-MS data (**Figure 4.4A**). Both RNAseq and LC-MS features with the highest loading scores for the 1st PLS-components drive the separation between haematopoietic cell lines and other tumour cell lines. The selected genes for the 1st PLS-component based on their loading scores in the RNAseq data were: AC004687.1, WAS, IKZF1, CD53 and GMFG. Likewise, the loading scores for the 1st PLS-component from the LC-MS data highlighted the metabolites: xanthine, creatine, C20:4 CE, oxalate, C16:0CE and phosphocreatine; as the most important metabolites to maximise the correlation between Omics and the separation between classes (**Figure 4.4B**). Interestingly, all the selected genes for the 1st PLS-component were highly expressed in haematological cancers. By observing the metabolites, only oxalate was upregulated in haematological cancers, while all the other metabolites for the 1st PLS-component of the LC-MS data were highly abundant in other cancer types. Interestingly, two cholesterol esters (CEs) the C20:4CE (also among the top 3 metabolites with Spearman correlation, Appendix

13) and the C16:0CE were among the most important metabolites related to solid tumours based on the loadings scores of the 1st PLS-component.

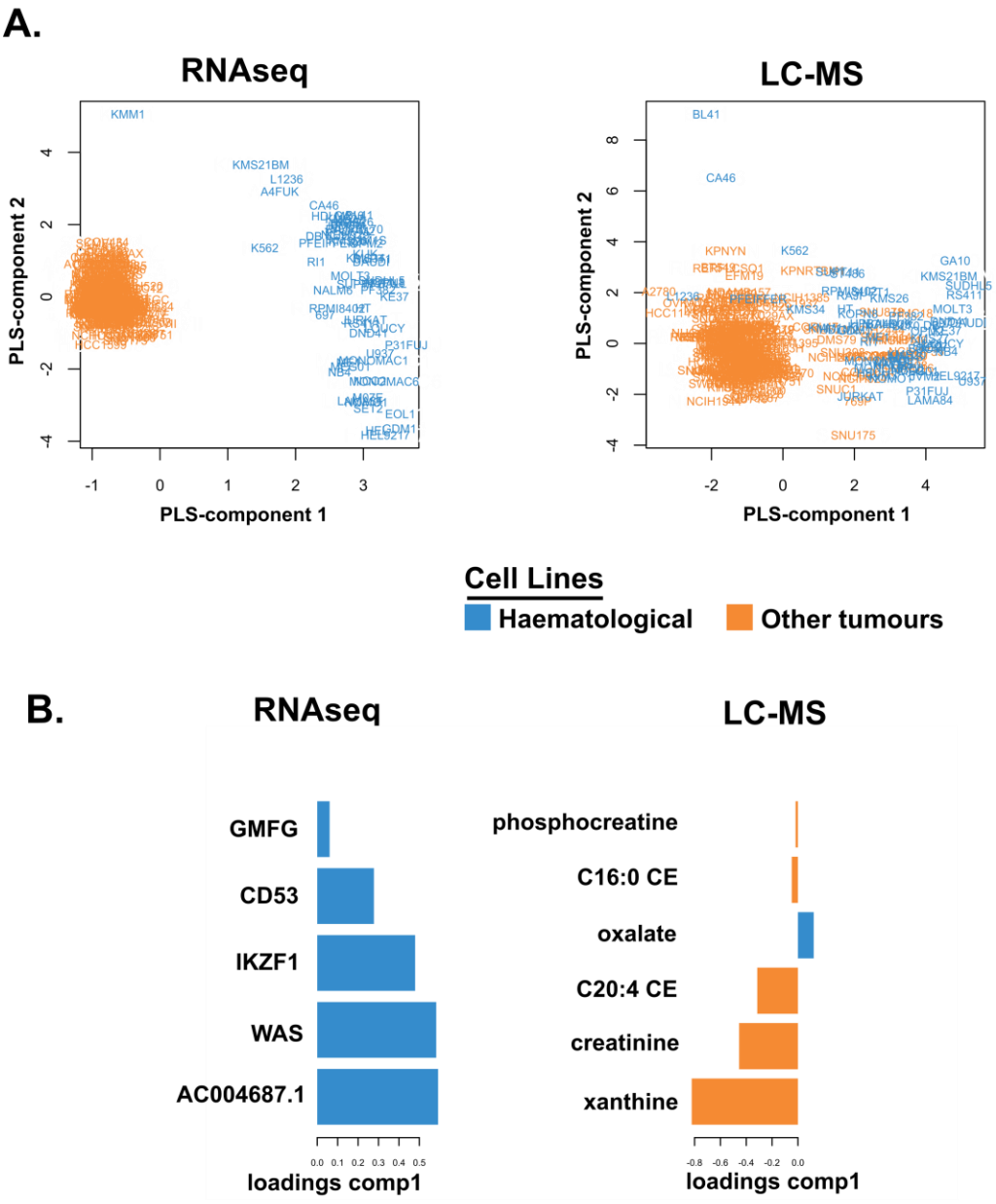


Figure 4.4. Omics integration at PLS-components level. (A) Cell lines are plotted for the first two PLS-components in different Omics datasets. **(B)** Expression of features as presented from the loading scores of the firsts PLS-components.

Moreover, the final sPLS-DA model extracted the most informative features from all the 9 PLS-components that we set as optimal prior to modelling. The algorithm computes similarity scores (González et al., 2012) that are analogous to Pearson correlation coefficients to represent correlations within features of different Omics. Here, we examined the 66 most essential correlations between the extracted genes and metabolites using a correlation cut-off of $r > |0.7|$ as previously used in other studies (Rohart et al., 2017; Singh et al., 2016), and we presented them together with the average expression value of each feature (**Figure 4.5A**). In general, we observe that 24 genes with higher expression in haematopoietic cell lines were negatively correlated with 8 metabolites (Appendix 12). The AMOTL2 gene, which showed higher expression in other tumour cell lines, was positively correlated with xanthine and negatively correlated with C36:1PC lipid metabolite. Furthermore, the xanthine, creatinine and betaine, which presented lower levels in haematopoietic cell lines, were positively correlated with 14 genes that were also downregulated in haematopoietic cell lines (Appendix 14).

Figure 4.5A is quite informative about correlation within features and between Omics, however it is still hard to interpret associations between features with more than one pair of correlations. An alternative representation of the same

results is with a relevance network (**Figure 4.5B**), which visualises correlations of the extracted features in a pairwise fashion. To simplify the network and highlight the strongest associations between genes and metabolites we generated a second relevance network setting the correlation cut-off at r >|0.8| (**Figure 4.5C**). We identified strong negative correlations among the genes WAS, AC004687.1, CD53, IKZF1, RHOH, DOCK2 and GMFG with creatinine and xanthine metabolites. Strong positive correlations were calculated between xanthine and the FKBP9 and TJP1 genes separately. The TJP1 gene was also strongly positive correlated with the creatinine metabolite. Similar strong correlation was observed between the lipid metabolite C46:2TAG and the AL161932.3 and OR4C7P pseudogenes.

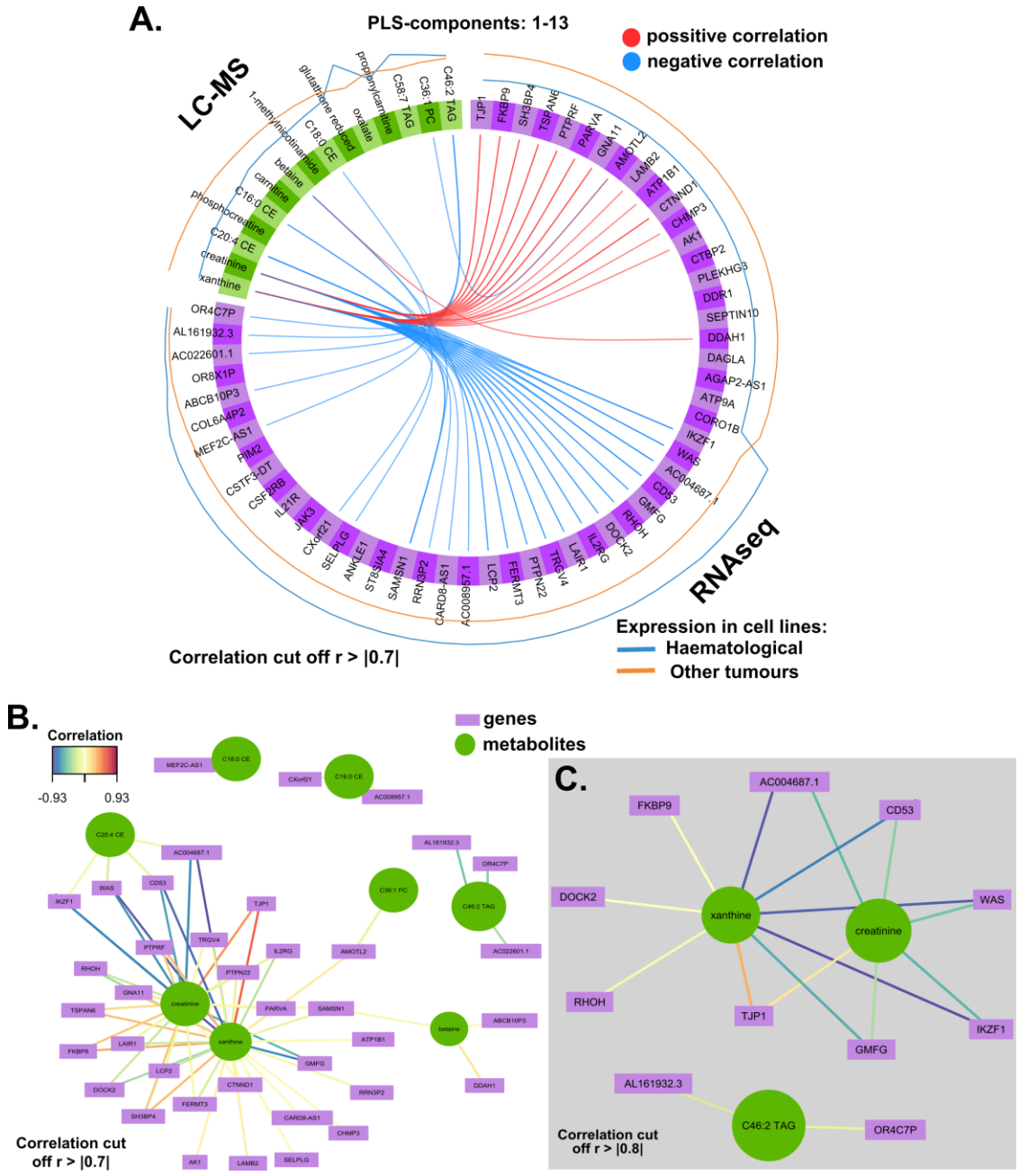


Figure 4.5. Correlations among the most informative features. (A) Important correlations (r>0.7) between all the extracted features from the pairs of the 13 PLS-components. **(B)** Relevance network depicting correlated (r>0.7) genes and metabolites. **(C)** Relevance network presenting stronger correlations (r>0.8).

After training the final sPLS-DA model, we used the testing set of the 143 cell lines to independently evaluate the classification performance of the final model by predicting which of these cell lines derived from haematological cancers. The confusion matrix outlined in **Figure 4.6A** summarises the number of cell line in each class as predicted from the 1st PLS-component of the final sPLS-DA model. Out of the 30 haematopoietic cell lines in the testing set, the model correctly classified 27 (27 true positives), none were incorrectly classified (0 false negative) and for 3 cell lines (WSUDLCL2, KMS34, L1236) the classes were not predicted (NA) by the model. Similarly, out of the 113 cell lines that derive from other tumours 98 were correctly classified (98 true negatives), again none of them were incorrectly classified (0 false negatives) and 15 cell lines could not be determined (15 NA: NCIH2291, HCC1187, NCIH146, NCIH82, NCIH508, GSS, SNU283, NCIH2347, CADOES1, NCIH1930, SNU878, SNU1214, CORL24, NCIH2444 and HCC2218). Overall, the final model generalises well as it showed a good classification performance with the testing set of the data achieving an accuracy (true positives + true negatives / total number of cell line samples) of 87% and a BER of 11%.

Finally, to assess the predictive performance of the extracted genes and metabolites from the 1st PLS-component in each Omics, the area under the ROC curve (AUC) was calculated individually. Both the selected genes and metabolites seemed to be good classifiers (AUR = 0.99 and AUR=0.91, **Figure 4.6B,C**) for predicting haematological cancers.

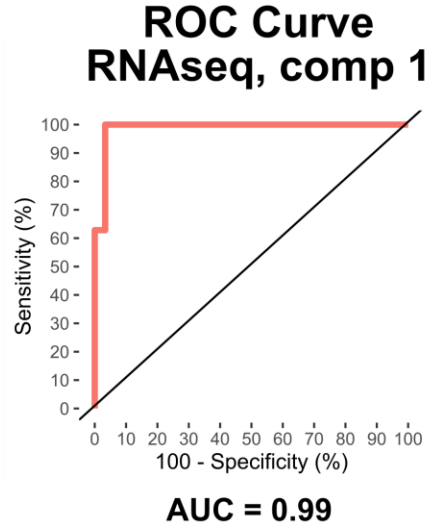
A.

Confusion matrix

Class	predicted as Haematological	predicted as Other tumour	predicted as NA
Haematological	27	0	3
Other tumour	0	98	15

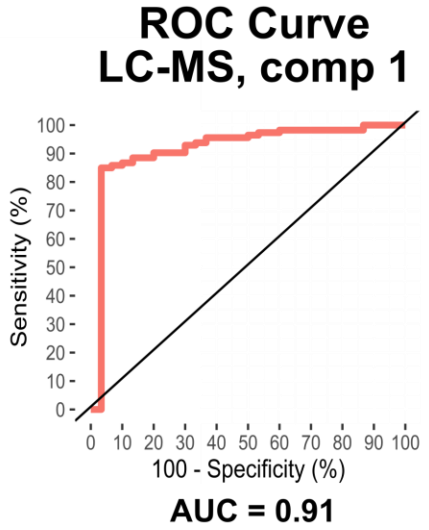
Accuracy = 87% and Balanced Error Rate = 11%

B.



AUC = 0.99

C.



AUC = 0.91

Haematological vs other cell lines

Figure 4.6. Prediction performance of the final PLS-DA model. (A) A confusion matrix compares the real with the predicted classes in the testing set. **(B)** and **(C)** the area under the ROC curve (AUC) was computed to assess the prediction performance of the extracted features for the 1st PLS-component in each Omics.

4.4. Discussion

In previous chapters we showed that a critical feature of CLL and NHL cells is alterations along metabolism to sustain lymphomagenesis. However, it did not address how different these metabolic alterations were compared to other tumours. To investigate the metabolic diversity between cancer types, we analysed the transcriptomic and metabolic profiles from cancer cell lines retrieved from the CCLE database. High-dimensional RNAseq and LC-MS data were explored by analysing them separately and integrating them together. Firstly, we reduced dimensions to address for data complexity and highlight the main source of data's variation. We applied a series of well-established unsupervised machine learning methods for dimensionality reduction and explore the linear (with PCA) or non-linear combination (with tSNE or UMAP) in each datatype, by preserving either the local (with tSNE or UMAP) or global structure (with PCA or UMAP) of the data. We have avoided considering one method better than the other, since mapping a high-dimensional data into low dimensions significantly reduces with each method differently the whole structure of the data. There is always a trade-off of with different methods having different advantage/drawbacks, therefore we have decided to explore and visualize the data with all common approaches. Confirming observations of the initial study (Li et al., 2019a), all our methods revealed that haematopoietic cell lines were the major source of variation in both RNAseq and LC-MS datasets (**Figure 4.2**), suggesting that haematological cancers are

transcriptionally and metabolically distinct from the other cancer types. In line with literature, the majority of cancer cells tend to form solid tumours mostly affected by the tumour microenvironment (Paolicchi et al., 2016; Pires et al., 2012; Zhao et al., 2018). On the contrary, haematological malignant cells derive from the highly dynamic environments of the bone marrow or the lymph nodes and most of them are circulating in the circulatory system (Gharbaran et al., 2014; Mulder et al., 2019; Pedersen et al., 2013; Scott and Gascoyne, 2014). Moreover, despite contrasting opinions on whether tSNE or UMAP is the most preferable method to preserve the global structure of the data (Becht et al., 2019; Kobak and Linderman, 2019), here both methods clearly separated groups of distinct cancer types, separating even different types of haematological cancers (**Figure 4.2B,C**). However, visualizing UMAP results in 2D scatter plots were more easily interpreted than tSNE results, since UMAP achieved more distinct groups compared to tSNE, even for the integrative datasets. Therefore, UMAP is suggested here to better segregate clusters both in biological and integrated datasets. This observation is in accordance with Kobak and Linderman findings, where UMAP produced denser and more compact clusters than t-SNE, with more white space in between (Kobak and Linderman, 2021). They have computed Pearson correlation between pairwise Euclidean distances in three different datasets to quantify preservation of global structure, like Becht et al. (Becht et al., 2019). This quantification was used as a default metric to measure the distance between two points without considering correlated variables, such in our case between LC-MS and RNAseq data, and therefore was not computed in our analysis. Further machine learning analyses

are needed in combination with functional experiments to highlight the mathematical and algorithmic difference of all methods and prove which one is more informative for the biological data. Although, such comparison is quite challenging to be measured, the Mahalanobi's distances can be used in future studies as a metric to quantify preservation of global structure in correlated datasets. In general, we observed that dimensionality reduction with integrated multi-Omics data, independently of the method applied, seemed to be more informative visually than analysing the LC-MS data alone, but not from the RNAseq.

After illustrating the transcriptomic and metabolic heterogeneity of cancer cell lines, we focused on the identification of key associations between the most essential genes and metabolites that were responsible to discriminate the haematopoietic cell lines from the other tumours. Hence, we applied the supervised method of sPLS-DA, which selects and calculates correlations between the most important features by maximizing the separation between classes in the reduced dimensional PLS space. We identified a pattern of genes and metabolites (**Figure 4.4B**) discriminating haematopoietic cancer cell lines from the other tumours on the 1st PLS-component. By examining the expression of the selected genes, we observed that all of them showed higher expression in the haematopoietic cell lines. Most of these genes (WAS, CD53, IKZF1 and GMFG) are proven to predominantly be normally expressed in haematopoietic cells (Greenberg et al., 2020; Marke et al., 2018; Shi et al., 2006; Sun et al., 2019). An interesting finding is the elevated expression of the AC004687.1 transcript in

haematopoietic cell lines. Although, expression in haematopoietic cell lines has not previously been reported, the regulatory role of this novel lncRNA is undefined. Next, focusing on the selected metabolites, oxalate was elevated in haematopoietic cell lines, while all the rest of the metabolites were more abundant in other tumours. Oxalate is a metabolic end product, primarily of hepatic metabolism, which is excreted into the urine (Greger et al., 1978). Oxalate has an important role regulating the homeostasis of divalent ions, which are structural and functional co-factors for many biochemical interactions. Evidence in immune cells revealed signalling functions, as secondary messengers, for divalent ions, such as Ca^{2+}, the Mg^{2+} and the ZN^{2+} (Chaigne-Delalande and Lenardo, 2014; Kaltenberg et al., 2010; Li et al., 2011). In addition, Castellaro et al. have associated oxalate with carcinogenic effects in breast epithelial cells (Castellaro et al., 2015). However, questions still remain whether oxalate, as an ion, or calcium oxalate, which is more abundant in the human body, is responsible for inducing breast cancer. Taken together, it will be of importance to examine the role of oxalate, as a second messenger, to promote cancer in haematopoietic/immune cells. As a start, treatment with both calcium oxalate and potassium oxalate (highly soluble) separately, can be compared by measuring proliferation (with cell proliferation assays) in haematopoietic cell lines. In vivo experiments in mouse models are also required to demonstrate oxalate's capacity to induce haematopoietic cancers in such models. In contrast, higher levels of creatinine and phosphocreatine found in several other tumours, especially in oesophagus cells, indicated a potential dependency of these cell lines on creatine metabolism to fuel their energy

demands. Both these metabolites are derived from creatine, which is well known to be a high energy metabolite for fast production of ATP (Wyss and Kaddurah-Daouk, 2000). Similarly, xanthine accumulated mostly in endometrial, ovarian and other cancer cell lines and presented lower levels in haematopoietic cell lines. Xanthine is involved in purine catabolism and it is metabolised to uric acid by the xanthine oxidoreductase enzyme. High levels of this enzyme are normally expressed in tissues, where cell lines with abundant levels of xanthine derive from (liver, breast, colon and kidney) (Battelli et al., 2016). Most importantly, both xanthine and creatinine were strongly associated with genes related to cellular cytoskeleton, presenting negative correlations with actin remodelling genes (WAS, GMFG and DOCK2) and positive correlations with cell adhesion genes (TJP1, AMOTL2 and CTNND1). Moreover, two cholesterol esters (CEs), the C20:4CE and the C16:0CE were elevated in solid tumours compared to haematopoietic cell lines. CEs are formed by the esterification of cholesterol with long chain fatty acids linked to a hydroxyl group, as a mean either to store cholesterol intracellularly or to transport cholesterol trough the blood stream (Tosi and Tugnoli, 2005). Cholesterol is a critical component of the plasma membrane and intracellular levels of cholesterol are regulated by several metabolic processes, whose equilibrium is altered in cancer (as previously discussed in section 2.4 of chapter 2). Here, our observation of elevated CEs in solid tumours is in accordance with several studies that have reported increased levels of CEs in breast cancer (de Gonzalo-Calvo et al., 2015), pancreatic cancer (Li et al., 2016b) and prostate cancer (Lee et al., 2018; Yue et al., 2014). All of them suggest targeting cholesterol esterification as

a strategy to supress tumour growth. Overall, all these findings generate several questions for future studies to investigate the regulatory role of metabolism in cancer cells cytoskeletal rearrangements. As a first step, validation of gene expression results is required with RT-PCR analysis and examination for protein expression with Western Blot Analysis, as previously described in section 2.4. It will be of importance to investigate primary tumour samples for similar association dependencies in transcriptome and metabolome level among cancer types. Although, transcriptomic data are available in TCGA, the unavailability of metabolomic profiling data for primary tumour samples restricts such investigation. Finally, despite the fact that this pattern of features presented a good classification performance (accuracy 87%) with the current datasets (testing set), additional studies are also needed to demonstrate their predictive ability in primary tumours and clarify their role as potential biomarkers.

CHAPTER 5

CONCLUSION

5.1. Multi-Omics integration in haematological cancers

This thesis presents work undertaken to highlight aspects on the expanding field of cancer metabolic reprogramming in haematological cancers. It is driven by the computational approach on multi-Omics data integration with practical application to CLL (chapter 2), GC-derived lymphomas (chapter 3), and cancer cell lines (chapter 4). Finally, we unveil the power of integrative methodologies based on Genome Scale Metabolic Modelling (GSMM), pathway level, and Machine Learning approaches to identify novel biological insights and predict metabolic vulnerabilities in these cancers.

5.2. Omics integration with GSMM in CLL.

CLL is a disease with a wide clinical and biological heterogeneity that putatively relies on resistance of malignant CLL cells in apoptosis. The clinical outcome of CLL patients ranges from progressive CLL cases, where the malignant cells vastly proliferate and do not die; to indolent CLL cases, where the malignant cells are in a quiescent state; or rare spontaneous regression cases, where the high number of malignant cells decrease over time. Although, extracellular signals and BCR signalling is highly responsible for the malignant transformation and the

transition of CLL cells across these phenotypes, in chapter 2 we revealed with Differential Expression Analysis and Metabolic modelling the contribution of metabolic reprogramming in CLL towards these regulations. Results from differential expression analysis showed deregulated expression of metabolic genes and pathways between spontaneous regression cases versus non-regression CLL cases. Non-regressed CLL cells presented a differential reliance on oxidative phosphorylation and mitochondrial respiration compared to spontaneous regressed cells. Going beyond gene expression results, we simulated the metabolic fluxes by integrating the CLL transcriptome profiles with GSMMs using two independent computational approaches. The **rMTA** method highlighted the sulfate anion transporter SLC26A1 as highly responsible to transform the non-regression metabolic flux state to a regression state (see 2.3.3). Additionally, the second method of **gMCSs** identified several genes (AK1, GUK1, FDFT1, MVD, PTPMT1 and GNE) in CLL as potential metabolic targets. The **gMCSs** method selects elevated genes which belong to a minimum subset with lowly expressed genes, whose simultaneous knockout blocks biomass production. Although, both approaches have predicted several genes as ~~a~~ metabolic vulnerabilities in CLL, each one provides unique opportunities to investigate metabolism. In this context, the main advantage of using the **rMTA** approach, is the metabolic reconstruction and comparison of two different metabolic states (regression versus non-regression state). Instead, the **gMCSs** approach gives the opportunity to extract subset of genes related to a given metabolic task, such as the biomass production in our study to explore synthetic lethality in CLL.

These findings highlighted the key role of metabolic reprogramming and suggested the possibility of targeting several metabolic genes in CLL. Further experiments and functional analyses, discussed in more details in section 2.4, will be required to validate these metabolic targets. However, several limitations affected the results of this study. Firstly, we examined only alterations in gene expression profiles, and we inferred that any changes in mRNA levels may resemble changes in protein and metabolite levels. However, several post-translational modifications and other metabolic regulations appear to affect the expression and activity of metabolic enzymes, which need to be further explored in future studies. Therefore, it is important to acquire proteomics and metabolomics datasets and integrate them with the transcriptomic to better understand metabolic alterations in CLL. Secondly, additional limitations are inherent to the integration analysis with GSMMs. As described in paragraph 1.4.3.1, GSMMs are computational models constructed based on gene-protein-reaction (GPR) associations and thus are limited in the established scientific knowledge of their time. Therefore, any GSMM approach is unable to detect and examine any new interactions or associations between metabolic features that are not present in the initial model. Therefore, both rMTA and gMCSs analyses should be repeated in the future using the current or a new CLL dataset with the updated GSMM model of human metabolism to explore novel interactions that will be included in the updated model. Moreover, the next step will be to prove our predictions and validate any changes in the mRNA and protein expression level, such as the expression of the sulfate anion transporter SLC26A1. To achieve this, functional

experiments with the RT-PCR and the Western blot technology, highlighted also in section 2.4, are required to be performed in malignant cells from the same CLL patients and be compared with cells from normal hematopoietic tissues from donors. However, such additional experimental step would add a complexity to track and recruit the same patients again. Although, we have used here the largest cohort of spontaneous regressed CLL cases for our analysis, a study with a larger cohort of CLL cases could provide a more powerful validation.

Despite these limitations, we believe that by understanding the metabolic reprogramming in indolent, progressive and regression CLL status, novel metabolic targets will emerge for new therapeutic interventions, such as those described in subsection 1.2.5.

5.3. Pathway based integration in GC-derived lymphomas.

Non-Hodgkin Lymphoma (NHL) represent most of the mature B-cell neoplasms with the aggressive subtypes being rapidly fatal if remain untreated. Both Burkitt Lymphomas (BL) and Diffuse Large B-cell Lymphomas (DLBCL) are aggressive NHL that originate from germinal centres (GCs) malignant B-cells and display a broad spectrum of genomic, epigenetic and metabolic profiles. In chapter 3 we highlighted metabolic properties in NHL related to the germinal centres' development. Although, a q-value threshold of 5% is commonly used in transcriptomic analyses, a stricter threshold (less than 1%) was selected for the differential expression analysis in chapter 3 and a weaker threshold (less than 10%) in chapter 2. By comparing two different cancers types of the same cell origin, such as BL and DLBCL, results in a relatively large number of differentially expressed genes. Therefore, a threshold of 1% was applied to minimize false positive results and try to capture the most significant transcriptomic differences between the two cancers. On the contrary, the less strict threshold of 10% was applied in chapter 2, where different conditions of the CLL diseases were analysed. Thus, we decided to investigate more differentially expressed genes with even less significant differences between the CLL conditions. More specifically, in chapter 3 we suggested that BL cells rely more on the mTORC1/serine/glycine axis probably to enhance the cells' antioxidant ability and support their growth and proliferation.

DLBCL cases on the other hand, seemed to be more dependent on extracellular signals to alter their metabolism and escape immune response. Furthermore, to explore any interactions between genes and metabolites in gene metabolic pathways, integration analysis at the pathway level was applied with Omics datasets from cell lines. Pathway-based integration analysis provides a useful visual interface to explore features not only in metabolic pathways, like GSMM methods do, but also investigate signalling and drug related biological pathways. As explained in subsection 3.3.5, results from integrative analysis suggested that BL cells may depend more on non-essential amino acids to support and maintain the mitochondrial redox homeostasis compared to DLBCL cells. In line with this current work, Mrs Zuhal Eraslan (Dr. Farhat Khanim and Prof. Ulrich Günther) findings highlighted the serine production and uptake in NHL cell lines. More specifically, Mrs Zuhal Eraslan implemented a tracer-based approach to examine the role of serine by the presence of asparagine. Next, inhibition of the serine PHGDH enzyme was performed in combination with asparaginase (ASNase) to examine the impact of this intervention in cell viability assays. Preliminary results indicate that BL cell lines are more sensitive in this combined inhibition compared to DLBCL. Together, these findings set an example of the numerous opportunities that our work in NHL provides.

Similarly, to the CLL study, several limitations influenced the investigation in NHL. To begin with, the small number of available in-house NHL cell lines restricted the power of the study and limited to the comparison of only the endemic

BL subtype versus the GCB-DLBCL. A study with a larger number of NHL cell lines can provide more accurate results and confirm (or contradict to) our findings in other subtypes of BL and DLBCL. Secondly, any observed transcriptomic and metabolomic alterations in cell lines do not necessarily represent in primary tumours. Results from PCA analysis of RNAseq data from primary tumours and cell lines, clearly highlighted that cell lines have separated from primary tumours in PC1 (**Figure 5.1**). This indicates differences in transcriptomic profile between cell lines and primary tumours. It is well known now that cell lines have adapted their phenotype to sustain survival and proliferation in vitro. As described in section 3.4, BL primary tumours presented elevated expression of serine genes, which was not observed in BL cell lines. To emphasize these differences between primary tumours and inhouse cell lines in PCA, we have also included data from BL and DLBCL cell lines from the CCLE dataset used in chapter 4. Inhouse cell lines, in **Figure 5.1**, clustered closer to CCLE cell lines, which implies a more similar transcriptomic profile. However, despite the transcriptomic differences between cell lines and primary tumours, the use of cell lines gave us the opportunity to reach an adequate number of cells for the NMR experiments and acquire metabolomic data, which was not the case for primary tumours.

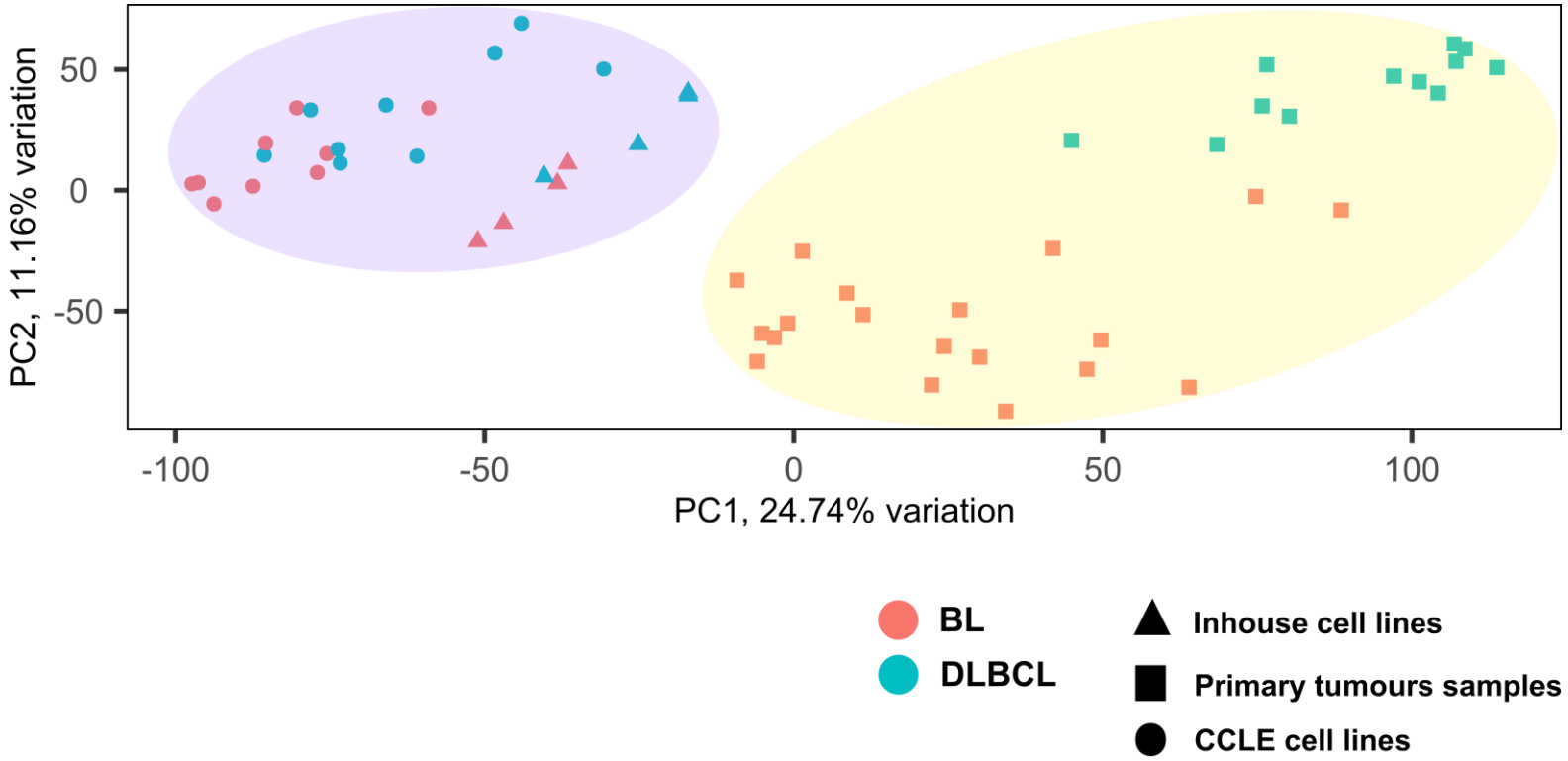


Figure 5.1 Unsupervised principal component analysis with transcriptomic data from BL/DLBCL cell lines and primary tumours. The first and second principal components are contributing to 24.74% and 11.16% of the total explained variation, respectively. The marker shapes represent inhouse cell lines (triangle) and primary tumours (square) from chapter 3 and CCLE cell lines (circle) from chapter 4. The purple circle includes only cell lines samples, while the yellow circle only primary tumours samples.

Furthermore, another limitation lies in the small number of metabolites that were identified manually with untargeted 1D ^{1}H NMR. A bias in selection occurs because some metabolites are easier to identify than others. This bias affects pathway-based integration analysis, by causing pathways, which includes the measured metabolites, to be always significantly over-represented and enriched. An alternative is either to deconvolute NMR peaks and automatically assign a possibly larger number of metabolites or perform univariate analysis to select the most significant NMR peaks. A final limitation is that the acquired metabolic profiles of the cell lines were measured at a steady-state condition, meaning that we assumed that the examined metabolites reached an equilibrium. A study with time courses can provide the opportunity to examine metabolic fluxes and understand network dynamics by using GSMM and Flux Balanced Analysis.

Altogether, our analysis combined with pathway-based multi-Omics integration highlighted useful insights into metabolic reprogramming of NHL. Several new hypotheses emerged from our findings, which can be experimentally tested, as recommended in section 3.4, to identify biomarkers or metabolic targets either in BL or DLBCL. Though, the limitations and biases of this study should be also considered prior to any experimental testing.

5.4. Machine Learning for multi-Omics data integration in cancer cell lines

Reprogramming cellular metabolism is now considered a hallmark of cancer, however the metabolic profile of each cancer type is quite distinct. In chapter 4 we explored cancer's transcriptomic and metabolic heterogeneity in cell lines datasets from the Cancer Cell Line Encyclopaedia (CCLE) database. In addition, we identified key associations between genes and metabolites that separated haematological cancers from the other tumours. To achieve this, Machine Learning approaches analysed and combined Omics data by constructing classification models that were not able to be constructed with the previous applied integrative methods. As stated in section 4.1, most Machine Learning methods provide the advantage to "learn" from the data and highlight informative associations that are not necessarily statistically linked with the phenotype. Results from dimensionality reduction methods: **PCA**, **tSNE**, and **UMAP** revealed the heterogenous transcriptomic and metabolomic profiles of haematopoietic cell lines compared to other tumours (subsection 4.3.1). Next, the **sPLS-DA** method identified gene expression changes related to cellular cytoskeleton and cell adhesion with deregulated metabolites, such as the xanthine and the creatinine (subsection 4.3.2). Thus, it will be of importance to assess in future studies, such as those described in section 4.4, the role of these metabolic features, highlighted by our integrative analysis, into the regulation of cancer cellular metabolism.

Again, several limitations should be considered in this study. Firstly, the use of public datasets was a convenient approach for our analysis, but many concerns can be raised about the quality of the data. For example, LC-MS data were acquired from only one biological replicate for most of the cell lines samples. This increases the likelihood for several error types and biases. Despite the quality control checked by the authors (Li et al., 2019), a study with more replicates can provide better statistical power and minimise errors. Furthermore, as highlighted previously, cell lines may have altered many biological functions compared to the primary tumours where they derived from. Thus, it will be of interest to examine data from primary cells for any of the alterations observed here. Similar to the NMR experiments, the assigned metabolites in the LC-MS spectra represent only a few biochemical pathways and not the full range of cellular metabolic processes. Therefore, a study on a larger number of metabolites can reveal information on more metabolic pathways. Moreover, this study is limited in performing transcriptomic and metabolomic data integration. Nowadays, there are available additional Omics datasets in CCLE, such as DNA methylation, miRNA expression or CNV data, that can be integrated with the same methodology. Such investigation is necessary since it will link metabolism with even more genetic and epigenetic features. Nevertheless, this integration can be complex since the additional Omics datasets introduce more technical and biological limitations.

In summary, multi-Omics data integration with Machine Learning approaches highlighted the highly dynamic transcriptomic and metabolomic

variations between haematopoietic cancers cells and other tumours. Regardless of the limitations, our findings generated several interesting questions which provide new hypotheses for future work.

5.5. Challenges in multi-Omics data integration strategy

Integration of multi-Omics data is a challenging task and a ‘golden standard’ approach still does not exist. As stated in section 1.4, Computational Biology aims to provide a holistic picture of the biological mechanisms under study by integrating multi-Omics datasets with the most sophisticated bioinformatics tools and biostatistics methods. Therefore, various computational methodologies were applied here to integrate and analyse transcriptomic with metabolomic profiles based on the research questions and the sample availability in each study.

So far, many methods have been proposed to integrate Omics datasets measured either on the same samples or on independent samples from different studies. Most of these methods have relied on statistical, pathway-based, Machine Learning, and Metabolic modelling approaches. Despite the advantages of each unique approach, common challenges need to be addressed for a successful multi-Omics integration analysis. Firstly, in most integrative studies the combined Omics datasets are still generated independently rather than an integrated concept

(unified extraction). Consequently, several issues are affecting the analysis, such as incomplete sampling across the datasets, missing features within the samples, and different types of experimental noises and errors. Thus, it is important to consider the experimental design and the sample collection strategy before the integration analysis to achieve robustness and reproducibility. A second challenge is handling and processing of the individual datasets successfully. This requires biological knowledge and experience on data cleaning, annotation, filtering, and data normalization. These steps are quite diverse in each data type and strongly related with the high-throughput platforms from which the data were generated. In addition, further issues emerge at the integration level. Each data type deals with unique biases, which can further mitigate when data combined and should be taken into account. Furthermore, each platform literally generates Big Data with thousands or even up to millions of features which comes with its own practical challenges of data handling. Moreover, as already described in section 4.1, dealing with such high-dimensional data is challenging to extract the most important biological information related to the examined phenotype. For instance, biological and technical variation can contribute to unrelated features which antagonise (or dominate) the important ones in high-dimensional space (Ronan et al., 2016). Consequently, interpreting the results from an integration analysis of such data is even more complex since each data type introduces an extra layer of large variation. To tackle this, dimensionality reduction methodologies are usually applied, such as demonstrated in chapters 3 and 4. Truly, applying multi-Omics data integration requires a strong biological knowledge of the system under study

combined with computational and analytical skills.

Despite all these challenges that need addressing, multi-Omics data integration with its extensive applications allows the identification of key features of the dynamic biological networks that are usually non-obvious in individual Omics data analysis. Therefore, it becomes an asset when working closely with computational biologists to identify the molecular relationships that associate genetic, epigenetic, and metabolic variation with the phenotype. By revealing these relationships scientists are able to better understand malignancies and develop novel therapeutic strategies for treatment.

REFERENCES

Abate, F., Ambrosio, M.R., Mundo, L., Laginestra, M.A., Fuligni, F., Rossi, M., Zairis, S., Gazaneo, S., De Falco, G., Lazzi, S., et al. (2015). Distinct Viral and Mutational Spectrum of Endemic Burkitt Lymphoma. PLoS Pathog *11*, e1005158.

Adekola, K.U.A., Aydemir, S.D., Ma, S., Zhou, Z., Rosen, S.T., and Shanmugam, M. (2015). Investigating and Targeting Chronic Lymphocytic Leukemia Metabolism with the HIV Protease Inhibitor Ritonavir and Metformin. Leuk Lymphoma *56*, 450–459.

Agren, R., Bordel, S., Mardinoglu, A., Pornputtapong, N., Nookaew, I., and Nielsen, J. (2012). Reconstruction of Genome-Scale Active Metabolic Networks for 69 Human Cell Types and 16 Cancer Types Using INIT. PLoS Comput Biol *8*, e1002518.

Alakwaa, F.M., Chaudhary, K., and Garmire, L.X. (2018). Deep Learning Accurately Predicts Estrogen Receptor Status in Breast Cancer Metabolomics Data. J. Proteome Res. *17*, 337–347.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). Molecular Biology of the Cell (New York: Garland Science).

Ali, M., and Aittokallio, T. (2019). Machine learning and feature selection for drug response prediction in precision oncology applications. Biophys Rev *11*, 31–39.

Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature *403*, 503–511.

Aller, S., Scott, A., Sarkar-Tyson, M., and Soyer, O.S. (2018). Integrated human-virus metabolic stoichiometric modelling predicts host-based antiviral targets against Chikungunya, Dengue and Zika viruses. J. R. Soc. Interface *15*, 20180125.

Alonso, A., Marsal, S., and JuliÃ, A. (2015). Analytical Methods in Untargeted Metabolomics: State of the Art in 2015. Front. Bioeng. Biotechnol. *3*.

Andrews, S. (2010). FastQC A Quality Control tool for High Throughput Sequence Data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Apaolaza, I., San José-Eneriz, E., Tobalina, L., Miranda, E., Garate, L., Agirre, X., Prósper, F., and Planes, F.J. (2017). An in-silico approach to predict and exploit synthetic lethality in cancer metabolism. Nature Communications *8*.

Aran, D., Sirota, M., and Butte, A.J. (2015). Systematic pan-cancer analysis of tumour purity. Nat Commun *6*, 8971.

Armitage, J.O., Gascoyne, R.D., Lunning, M.A., and Cavalli, F. (2017). Non-Hodgkin lymphoma. The Lancet *390*, 298–310.

Asgari, Y., Khosravi, P., Zabihinpour, Z., and Habibi, M. (2018). Exploring candidate biomarkers for lung and prostate cancers using gene expression and flux variability analysis. Integr. Biol. *10*, 113–120.

Auslander, N., Cunningham, C.E., Toosi, B.M., McEwen, E.J., Yizhak, K., Vizeacoumar, F.S., Parameswaran, S., Gonen, N., Freywald, T., Bhanumathy, K.K., et al. (2017). An integrated computational and experimental study uncovers FUT 9 as a metabolic driver of colorectal cancer. Mol Syst Biol *13*, 956.

Bagherzadeh, J., and Asil, H. (2019). A review of various semi-supervised learning models with a deep learning and memory approach. Iran J Comput Sci *2*, 65–80.

van Baren, N., and Van den Eynde, B.J. (2015). Tryptophan-Degrading Enzymes in Tumoral Immune Resistance. Front. Immunol. *6*.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehár, J., Kryukov, G.V., Sonkin, D., et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature *483*, 603–607.

Barupal, D.K., Haldiya, P.K., Wohlgemuth, G., Kind, T., Kothari, S.L., Pinkerton, K.E., and Fiehn, O. (2012). MetaMapp: mapping and visualizing metabolomic data by integrating information from biochemical pathways and chemical and mass spectral similarity. BMC Bioinformatics *13*, 99.

Basso, K., and Dalla-Favera, R. (2015). Germinal centres and B cell lymphomagenesis. Nat Rev Immunol *15*, 172–184.

Battelli, M.G., Polito, L., Bortolotti, M., and Bolognesi, A. (2016). Xanthine oxidoreductase in cancer: more than a differentiation marker. Cancer Med *5*, 546–557.

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I.W.H., Ng, L.G., Ginhoux, F., and Newell, E.W. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol *37*, 38–44.

Becker, S.A., and Palsson, B.O. (2008). Context-Specific Metabolic Networks Are Consistent with Experiments. PLoS Comput Biol *4*, e1000082.

Bellan, C., Stefano, L., Giulia, D.F., Rogena, E.A., and Lorenzo, L. (2009). Burkitt lymphoma versus diffuse large B-cell lymphoma: a practical approach. Hematol. Oncol. n/a-n/a.

Bellman, R. (2010). Dynamic programming (Princeton, N.J.: Princeton University Press).

Beloueche-Babari, M., Wantuch, S., Casals Galobart, T., Koniordou, M., Parkes, H.G., Arunan, V., Chung, Y.-L., Eykyn, T.R., Smith, P.D., and Leach, M.O. (2017). MCT1 Inhibitor AZD3965 Increases Mitochondrial Metabolism, Facilitating Combination Therapy and Noninvasive Magnetic Resonance Spectroscopy. Cancer Res *77*, 5913–5924.

Benakanakere, I., Johnson, T., Sleightholm, R., Villeda, V., Arya, M., Bobba, R., Freter, C., and Huang, C. (2014). Targeting cholesterol synthesis increases chemoimmuno-sensitivity in chronic lymphocytic leukemia cells. Exp Hematol Oncol *3*, 24.

Berlow, N., Haider, S., Wan, Q., Geltzeiler, M., Davis, L.E., Keller, C., and Pal, R. (2014). An Integrated Approach to Anti-Cancer Drug Sensitivity Prediction. IEEE/ACM Trans. Comput. Biol. and Bioinf. *11*, 995–1008.

Bidkhori, G., Benfeitas, R., Elmas, E., Kararoudi, M.N., Arif, M., Uhlen, M., Nielsen, J., and Mardinoglu, A. (2018). Metabolic Network-Based Identification and Prioritization of Anticancer Targets Based on Expression Data in Hepatocellular Carcinoma. Front. Physiol. *9*, 916.

Billard, C. (2014). Apoptosis inducers in chronic lymphocytic leukemia. Oncotarget *5*.

Blighe, K., and Lun, A. (2019). PCAtools: everything Principal Component Analysis (bioconductor).

Boonstra, J.G., van Lom, K., Langerak, A.W., Graveland, W.J., Valk, P.J.M., Kraan, J., van 't Veer, M.B., and Gratama, J.W. (2006). CD38 as a prognostic factor in B cell chronic lymphocytic leukaemia (B-CLL): Comparison of three approaches to analyze its expression. Cytometry *70B*, 136–141.

Bossel Ben-Moshe, N., Gilad, S., Perry, G., Benjamin, S., Balint-Lahat, N., Pavlovsky, A., Halperin, S., Markus, B., Yosepovich, A., Barshack, I., et al. (2018). mRNA-seq whole transcriptome profiling of fresh frozen versus archived fixed tissues. BMC Genomics *19*, 419.

Bott, A., Maimouni, S., and Zong, W.-X. (2019). The Pleiotropic Effects of Glutamine Metabolism in Cancer. Cancers *11*, 770.

Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol *34*, 525–527.

Brial, F., Le Lay, A., Hedjazi, L., Tsang, T., Fearnside, J.F., Otto, G.W., Alzaid, F., Wilder, S.P., Venteclef, N., Cazier, J.-B., et al. (2019). Systems Genetics of Hepatic Metabolome Reveals Octopamine as a Target for Non-Alcoholic Fatty Liver Disease Treatment. Sci Rep *9*, 3656.

Bruno, S., Ledda, B., Tenca, C., Ravera, S., Orengo, A.M., Mazzarello, A.N., Pesenti, E., Casciaro, S., Racchi, O., Ghiotto, F., et al. (2015). Metformin inhibits cell cycle progression of B-cell chronic lymphocytic leukemia cells. Oncotarget *6*, 22624–22640.

Burkov, A. (2019). The hundred-page machine learning book (Andriy Burkov).

Calin, G.A., Cimmino, A., Fabbri, M., Ferracin, M., Wojcik, S.E., Shimizu, M., Taccioli, C., Zanesi, N., Garzon, R., Aqeilan, R.I., et al. (2008). MiR-15a and miR-16-1 cluster functions in human leukemia. PNAS *105*, 5166–5171.

Calissano, C., Damle, R.N., Marsilio, S., Yan, X.-J., Yancopoulos, S., Hayes, G., Emson, C., Murphy, E.J., Hellerstein, M.K., Sison, C., et al. Intraclonal Complexity in Chronic Lymphocytic Leukemia: Fractions Enriched in Recently Born/Divided and Older/Quiescent Cells. 9.

Campo, E., Swerdlow, S.H., Harris, N.L., Pileri, S., Stein, H., and Jaffe, E.S. (2011). The 2008 WHO classification of lymphoid neoplasms and beyond: evolving concepts and practical applications. Blood *117*, 5019–5032.

Cancer Research UK (2015). Chronic lymphocytic leukaemia (CLL) statistics. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/leukaemia-cll

Carew, J.S., Nawrocki, S.T., Xu, R.H., Dunner, K., McConkey, D.J., Wierda, W.G., Keating, M.J., and Huang, P. (2004). Increased mitochondrial biogenesis in primary leukemia cells: the role of endogenous nitric oxide and impact on sensitivity to fludarabine. Leukemia *18*, 1934.

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

Castellaro, A.M., Tonda, A., Cejas, H.H., Ferreyra, H., Caputto, B.L., Pucci, O.A., and Gil, G.A. (2015). Oxalate induces breast cancer. BMC Cancer *15*, 761.

Castillo, J.J., Beltran, B.E., Miranda, R.N., Young, K.H., Chavez, J.C., and Sotomayor, E.M. (2016). EBV-positive diffuse large B-cell lymphoma of the elderly: 2016 update on diagnosis, risk-stratification, and management: EBV+ DLBCL 2016 update. Am. J. Hematol. *91*, 529–537.

Cazier, J.-B., Kaisaki, P.J., Argoud, K., Blaise, B.J., Veselkov, K., Ebbels, T.M.D., Tsang, T., Wang, Y., Bihoreau, M.-T., Mitchell, S.C., et al. (2012). Untargeted Metabolome Quantitative Trait Locus Mapping Associates Variation in Urine Glycerate to Mutant Glycerate Kinase. J. Proteome Res. *11*, 631–642.

Chaigne-Delalande, B., and Lenardo, M.J. (2014). Divalent cation signaling in immune cells. Trends in Immunology *35*, 332–344.

Cham, C.M., and Gajewski, T.F. (2005). Glucose Availability Regulates IFN-γ Production and p70S6 Kinase Activation in CD8 $^{+}$ Effector T Cells. J Immunol *174*, 4670–4677.

Chen, L., Deng, H., Cui, H., Fang, J., Zuo, Z., Deng, J., Li, Y., Wang, X., and Zhao, L. (2018). Inflammatory responses and inflammation-associated diseases in organs. Oncotarget *9*.

Chong, J., Soufan, O., Li, C., Caraus, I., Li, S., Bourque, G., Wishart, D.S., and Xia, J. (2018). MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. Nucleic Acids Research *46*, W486–W494.

Chung, J.-Y., Braunschweig, T., Williams, R., Guerrero, N., Hoffmann, K.M., Kwon, M., Song, Y.K., Libutti, S.K., and Hewitt, S.M. (2008). Factors in Tissue Handling and Processing That Impact RNA Obtained From Formalin-fixed, Paraffin-embedded Tissue. J Histochem Cytochem. *56*, 1033–1042.

Cimmino, A., Calin, G.A., Fabbri, M., Iorio, M.V., Ferracin, M., Shimizu, M., Wojcik, S.E., Aqeilan, R.I., Zupo, S., Dono, M., et al. (2005). miR-15 and miR-16 induce apoptosis by targeting BCL2. Proceedings of the National Academy of Sciences *102*, 13944–13949.

Cluntun, A.A., Lukey, M.J., Cerione, R.A., and Locasale, J.W. (2017). Glutamine Metabolism in Cancer: Understanding the Heterogeneity. Trends in Cancer *3*, 169–180.

Coelho, V., Krysov, S., Steele, A., Sanchez Hidalgo, M., Johnson, P.W., Chana, P.S., Packham, G., Stevenson, F.K., and Forconi, F. (2013). Identification in CLL of circulating intraclonal subgroups with varying B-cell receptor expression and function. Blood *122*, 2664–2672.

Commisso, C., Davidson, S.M., Soydaner-Azeloglu, R.G., Parker, S.J., Kamphorst, J.J., Hackett, S., Grabocka, E., Nofal, M., Drebin, J.A., Thompson, C.B., et al.

(2013). Macropinocytosis of protein is an amino acid supply route in Ras-transformed cells. Nature *497*, 633–637.

Cox, A.G., Winterbourn, C.C., and Hampton, M.B. (2009). Mitochondrial peroxiredoxin involvement in antioxidant defence and redox signalling. Biochem. J. *425*, 313–325.

Curtis, N.J., Mooney, L., Hopcroft, L., Michopoulos, F., Whalley, N., Zhong, H., Murray, C., Logie, A., Revill, M., Byth, K.F., et al. (2017). Pre-clinical pharmacology of AZD3965, a selective inhibitor of MCT1: DLBCL, NHL and Burkitt’s lymphoma anti-tumor activity. Oncotarget *8*.

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

Dave, S.S., Fu, K., Wright, G.W., Lam, L.T., Kluin, P., Boerma, E.-J., Greiner, T.C., Weisenburger, D.D., Rosenwald, A., Ott, G., et al. (2006). Molecular Diagnosis of Burkitt's Lymphoma. N Engl J Med *354*, 2431–2442.

Dawson, P.A., Russell, C.S., Lee, S., McLeay, S.C., van Dongen, J.M., Cowley, D.M., Clarke, L.A., and Markovich, D. (2010). Urolithiasis and hepatotoxicity are linked to the anion transporter Sat1 in mice. J. Clin. Invest. *120*, 706–712.

Dhillon, S. (2018). Ivosidenib: First Global Approval. Drugs *78*, 1509–1516.

Ding, M.Q., Chen, L., Cooper, G.F., Young, J.D., and Lu, X. (2018). Precision Oncology beyond Targeted Therapy: Combining Omics Data with Machine Learning Matches the Majority of Cancer Cells to Effective Therapeutics. Mol Cancer Res *16*, 269–278.

Dollé, L., Best, J., Empsen, C., Mei, J., Van Rossen, E., Roelandt, P., Snykers, S., Najimi, M., Al Battah, F., Theise, N.D., et al. (2012). Successful isolation of liver progenitor cells by aldehyde dehydrogenase activity in naïve mice. Hepatology *55*, 540–552.

Dong, W., Keibler, M.A., and Stephanopoulos, G. (2017). Review of metabolic pathways activated in cancer cells as determined through isotopic labeling and network analysis. Metabolic Engineering *43*, 113–124.

Dong, Z., Zhang, N., Li, C., Wang, H., Fang, Y., Wang, J., and Zheng, X. (2015). Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection. BMC Cancer *15*, 489.

Dowling, R.J.O., Niraula, S., Stambolic, V., and Goodwin, P.J. (2012). Metformin in cancer: translational challenges. Journal of Molecular Endocrinology *48*, R31–R43.

Dreger, P., Ghia, P., Schetelig, J., van Gelder, M., Kimby, E., Michallet, M., Moreno, C., Robak, T., Stilgenbauer, S., and Montserrat, E. (2018). High-risk chronic lymphocytic leukemia in the era of pathway inhibitors: integrating molecular and cellular therapies. Blood *132*, 892–902.

Duarte, N.C., Becker, S.A., Jamshidi, N., Thiele, I., Mo, M.L., Vo, T.D., Srivas, R., and Palsson, B.O. (2007). Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proceedings of the National Academy of Sciences *104*, 1777–1782.

Dugan, J., and Pollyea, D. (2018). Enasidenib for the treatment of acute myeloid leukemia. Expert Review of Clinical Pharmacology *11*, 755–760.

Duhaylongsod, F.G., Lowe, V.J., Patz, E.F., Vaughn, A.L., Coleman, R.E., and Wolfe, W.G. (1995). Detection of primary and recurrent lung cancer by means of F-18 fluorodeoxyglucose positron emission tomography (FDG PET). The Journal of Thoracic and Cardiovascular Surgery *110*, 130–140.

Dumas, M.-E. (2012). Metabolome 2.0: quantitative genetics and network biology of metabolic phenotypes. Mol. BioSyst. *8*, 2494.

Dumas, M.-E., Domange, C., Calderari, S., Martínez, A.R., Ayala, R., Wilder, S.P., Suárez-Zamorano, N., Collins, S.C., Wallis, R.H., Gu, Q., et al. (2016). Topological analysis of metabolic networks integrating co-segregating transcriptomes and metabolomes in type 2 diabetic rat congenic series. Genome Med *8*, 101.

Duren, W., Weymouth, T., Hull, T., Omenn, G.S., Athey, B., Burant, C., and Karnovsky, A. (2014). MetDisease—connecting metabolites to diseases via literature. Bioinformatics *30*, 2239–2241.

Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc *4*, 1184–1191.

Evans, J.M.M., Donnelly, L.A., Emslie-Smith, A.M., Alessi, D.R., and Morris, A.D. (2005). Metformin and reduced risk of cancer in diabetic patients. BMJ *330*, 1304–1305.

Fan, J., Ye, J., Kamphorst, J.J., Shlomi, T., Thompson, C.B., and Rabinowitz, J.D. (2014). Quantitative flux analysis reveals folate-dependent NADPH production. Nature *510*, 298–302.

Faubert, B., Li, K.Y., Cai, L., Hensley, C.T., Kim, J., Zacharias, L.G., Yang, C., Do, Q.N., Doucette, S., Burguete, D., et al. (2017). Lactate Metabolism in Human Lung Tumors. Cell *171*, 358-371.e9.

Fearon, E.R., and Vogelstein, B. (1990). A genetic model for colorectal tumorigenesis. Cell *61*, 759–767.

Fischer, K., Hoffmann, P., Voelkl, S., Meidenbauer, N., Ammer, J., Edinger, M., Gottfried, E., Schwarz, S., Rothe, G., Hoves, S., et al. (2007). Inhibitory effect of tumor cell–derived lactic acid on human T cells. Blood *109*, 3812–3819.

Fiveash, J.B., Chowdhary, S.A., Peereboom, D., Mikkelsen, T., Nabors, L.B., Lesser, G.J., Rosenfeld, M.R., Ye, X., and Grossman, S.A. (2009). NABTT-0702: A phase II study of R-(-)-gossypol (AT-101) in recurrent glioblastoma multiforme (GBM). JCO *27*, 2010–2010.

Fryknäs, M., Zhang, X., Bremberg, U., Senkowski, W., Olofsson, M.H., Brandt, P., Persson, I., D'Arcy, P., Gullbo, J., Nygren, P., et al. (2016). Iron chelators target both proliferating and quiescent cancer cells. Sci Rep *6*, 38343.

Galicia-Vázquez, G., and Aloyz, R. (2018). Ibrutinib Resistance Is Reduced by an Inhibitor of Fatty Acid Oxidation in Primary CLL Lymphocytes. Front Oncol *8*.

Galicia-Vázquez, G., and Aloyz, R. (2019). Metabolic rewiring beyond Warburg in chronic lymphocytic leukemia: How much do we actually know? Critical Reviews in Oncology/Hematology *134*, 65–70.

Gao, J., Tarcea, V.G., Karnovsky, A., Mirel, B.R., Weymouth, T.E., Beecher, C.W., Cavalcoli, J.D., Athey, B.D., Omenn, G.S., Burant, C.F., et al. (2010). Metscape: a Cytoscape plug-in for visualizing and interpreting metabolomic data in the context of human metabolic networks. Bioinformatics *26*, 971–973.

Garnett, M.J., Edelman, E.J., Heidorn, S.J., Greenman, C.D., Dastur, A., Lau, K.W., Greninger, P., Thompson, I.R., Luo, X., Soares, J., et al. (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature *483*, 570–575.

Gaur, K., Vázquez-Salgado, A., Duran-Camacho, G., Dominguez-Martinez, I., Benjamín-Rivera, J., Fernández-Vega, L., Carmona Sarabia, L., Cruz García, A., Pérez-Deliz, F., Méndez Román, J., et al. (2018). Iron and Copper Intracellular Chelation as an Anticancer Drug Strategy. Inorganics *6*, 126.

Geeleher, P., Cox, N.J., and Huang, R. (2014). Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. Genome Biol *15*, R47.

Gerbitz, A., Mautner, J., Geltinger, C., Hörtnagel, K., Christoph, B., Asenbauer, H., Klobeck, G., Polack, A., and Bornkamm, G.W. (1999). Deregulation of the proto-oncogene c-myc through t(8;22) translocation in Burkitt's lymphoma. Oncogene *18*, 1745–1753.

Gharbaran, R., Park, J., Kim, C., Goy, A., and Suh, K.S. (2014). Circulating tumor cells in Hodgkin's lymphoma—A review of the spread of HL tumor cells or their putative precursors by lymphatic and hematogenous means, and their prognostic significance. Critical Reviews in Oncology/Hematology *89*, 404–417.

Giudice, I.D., Chiaretti, S., Tavolaro, S., Propris, M.S.D., Maggio, R., Mancini, F., Peragine, N., Santangelo, S., Marinelli, M., Mauro, F.R., et al. (2009). Spontaneous regression of chronic lymphocytic leukemia: clinical and biologic features of 9 cases. Blood *114*, 638–646.

Gloaguen, Y., Morton, F., Daly, R., Gurden, R., Rogers, S., Wandy, J., Wilson, D., Barrett, M., and Burgess, K. (2017). PiMP my metabolome: an integrated, web-based tool for LC-MS metabolomics data. Bioinformatics *33*, 4007–4009.

Godin-Ethier, J., Hanafi, L.-A., Piccirillo, C.A., and Lapointe, R. (2011). Indoleamine 2,3-Dioxygenase Expression in Human Cancers: Clinical and Immunologic Perspectives. Clinical Cancer Research *17*, 6985–6991.

Golub, D., Iyengar, N., Dogra, S., Wong, T., Bready, D., Tang, K., Modrek, A.S., and Placantonakis, D.G. (2019). Mutant Isocitrate Dehydrogenase Inhibitors as Targeted Cancer Therapeutics. Front. Oncol. *9*, 417.

González, I., Le Cao, K.-A., Davis, M., and Dejean, S. (2012). Visualising associations between paired 'omics' data sets. BioData Mining *5*.

de Gonzalo-Calvo, D., López-Vilaró, L., Nasarre, L., Perez-Olabarria, M., Vázquez, T., Escuin, D., Badimon, L., Barnadas, A., Lerma, E., and Llorente-Cortés, V. (2015). Intratumor cholesteryl ester accumulation is associated with human breast cancer proliferation and aggressive potential: a molecular and clinicopathological study. BMC Cancer *15*, 460.

Gottfried, E., Kunz-Schughart, L.A., Ebner, S., Mueller-Klieser, W., Hoves, S., Andreesen, R., Mackensen, A., and Kreutz, M. (2006). Tumor-derived lactic acid modulates dendritic cell activation and antigen expression. Blood *107*, 2013–2021.

Grapov, D., Fahrmann, J., Wanichthanarak, K., and Khoomrung, S. (2018). Rise of Deep Learning for Genomic, Proteomic, and Metabolomic Data Integration in Precision Medicine. OMICS: A Journal of Integrative Biology *22*, 630–636.

Greenberg, Z.J., Monlish, D.A., Bartnett, R.L., Yang, Y., Shen, G., Li, W., Bednarski, J.J., and Schuettpelz, L.G. (2020). The Tetraspanin CD53 Regulates Early B Cell Development by Promoting IL-7R Signaling. J.I. *204*, 58–67.

Greenhouse, W.V.V., and Lehnlnger, A.L. (1976). Occurrence of the Malate-Aspartate Shuttle in Various Tumor Types. 6.

Greger, R., Lang, F., Oberleithner, H., and Deetjen, P. (1978). Handling of oxalate by the rat kidney. Pflugers Arch. *374*, 243–248.

Gromski, P.S., Muhamadali, H., Ellis, D.I., Xu, Y., Correa, E., Turner, M.L., and Goodacre, R. (2015). A tutorial review: Metabolomics and partial least squares-discriminant analysis – a marriage of convenience or a shotgun wedding. Analytica Chimica Acta *879*, 10–23.

Gross, M.I., Demo, S.D., Dennison, J.B., Chen, L., Chernov-Rogan, T., Goyal, B., Janes, J.R., Laidig, G.J., Lewis, E.R., Li, J., et al. (2014). Antitumor Activity of the Glutaminase Inhibitor CB-839 in Triple-Negative Breast Cancer. Molecular Cancer Therapeutics *13*, 890–901.

Gu, C., Kim, G.B., Kim, W.J., Kim, H.U., and Lee, S.Y. (2019). Current status and applications of genome-scale metabolic models. Genome Biol *20*, 121.

Guo, J.Y., Chen, H.-Y., Mathew, R., Fan, J., Strohecker, A.M., Karsli-Uzunbas, G., Kamphorst, J.J., Chen, G., Lemons, J.M.S., Karantza, V., et al. (2011). Activated Ras requires autophagy to maintain oxidative metabolism and tumorigenesis. Genes & Development *25*, 460–470.

Gwangwa, M.V., Joubert, A.M., and Visagie, M.H. (2018). Crosstalk between the Warburg effect, redox regulation and autophagy induction in tumourigenesis. Cell Mol Biol Lett *23*, 20.

Hallek, M. (2017). Chronic lymphocytic leukemia: 2017 update on diagnosis, risk stratification, and treatment: HALLEK. Am J Hematol *92*, 946–965.

Hallek, M., Cheson, B.D., Catovsky, D., Caligaris-Cappio, F., Dighiero, G., Döhner, H., Hillmen, P., Keating, M.J., Montserrat, E., Rai, K.R., et al. (2008). Guidelines for the diagnosis and treatment of chronic lymphocytic leukemia: a report from the International Workshop on Chronic Lymphocytic Leukemia updating the National Cancer Institute–Working Group 1996 guidelines. Blood *111*, 5446–5456.

Hamada-Kanazawa, M., Kouda, M., Odani, A., Matsuyama, K., Kanazawa, K., Hasegawa, T., Narahara, M., and Miyake, M. (2010). b-Citryl-L-glutamate Is an Endogenous Iron Chelator That Occurs. *33*, 9.

Hamakawa, H., Fukuzumi, M., Bao, Y., Sumida, T., Kayahara, H., Onishi, A., and Sogawa, K. (2000). Keratin mRNA for detecting micrometastasis in cervical lymph nodes of oral cancer. Cancer Letters *160*, 115–123.

Hanahan, D., and Weinberg, R.A. (2000). The Hallmarks of Cancer. Cell *100*, 57–70.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of Cancer: The Next Generation. Cell *144*, 646–674.

Hedjazi, L., Gauguier, D., Zalloua, P.A., Nicholson, J.K., Dumas, M.-E., and Cazier, J.-B. (2015). mQTL.NMR: An Integrated Suite for Genetic Mapping of Quantitative Variations of ^{1} H NMR-Based Metabolic Profiles. Anal. Chem. *87*, 4377–4384.

Heirendt, L., Arreckx, S., Pfau, T., Mendoza, S.N., Richelle, A., Heinken, A., Haraldsdóttir, H.S., Wachowiak, J., Keating, S.M., Vlasov, V., et al. (2017). Creation and analysis of biochemical constraint-based models: the COBRA Toolbox v3.0. ArXiv:1710.04038 [q-Bio].

Hensley, C.T., Wasti, A.T., and DeBerardinis, R.J. (2013). Glutamine and cancer: cell biology, physiology, and clinical opportunities. J. Clin. Invest. *123*, 3678–3684.

Hinshaw, S.J., H Y Lee, A., Gill, E.E., and E W Hancock, R. (2018). MetaBridge: enabling network-based integrative analysis via direct protein interactors of metabolites. Bioinformatics *34*, 3225–3227.

HMRN, 2019 HMRN - QuickStats. https://www.hmrn.org/statistics/quick

Hollinshead, K.E.R., Munford, H., Eales, K.L., Bardella, C., Li, C., Escribano-Gonzalez, C., Thakker, A., Nonnenmacher, Y., Kluckova, K., Jeeves, M., et al. (2018). Oncogenic IDH1 Mutations Promote Enhanced Proline Synthesis through PYCR1 to Support the Maintenance of Mitochondrial Redox Homeostasis. Cell Reports *22*, 3107–3114.

Holmgaard, R.B., Zamarin, D., Li, Y., Gasmi, B., Munn, D.H., Allison, J.P., Merghoub, T., and Wolchok, J.D. (2015). Tumor-Expressed IDO Recruits and Activates MDSCs in a Treg-Dependent Manner. Cell Reports *13*, 412–424.

Hornyák, L., Dobos, N., Koncz, G., Karányi, Z., Páll, D., Szabó, Z., Halmos, G., and Székvölgyi, L. (2018). The Role of Indoleamine-2,3-Dioxygenase in Cancer Development, Diagnostics, and Therapy. Front. Immunol. *9*, 151.

İlie, M., Mazure, N.M., Hofman, V., Ammadi, R.E., Ortholan, C., Bonnetaud, C., Havet, K., Venissac, N., Mograbi, B., Mouroux, J., et al. (2010). High levels of carbonic anhydrase IX in tumour tissue and plasma are biomarkers of poor prognostic in patients with non-small cell lung cancer. Br J Cancer *102*, 1627–1635.

Ishibashi, A., Saga, K., Hisatomi, Y., Li, Y., Kaneda, Y., and Nimura, K. (2020). A simple method using CRISPR-Cas9 to knock-out genes in murine cancerous cell lines. Sci Rep *10*, 22345.

Jagannathan-Bogdan, M., and Zon, L.I. (2013). Hematopoiesis. Development *140*, 2463–2467.

Jang, I.S., Neto, E.C., Guinney, J., Friend, S.H., and Margolin, A.A. (2013). SYSTEMATIC ASSESSMENT OF ANALYTICAL METHODS FOR DRUG SENSITIVITY PREDICTION FROM CANCER CELL LINE DATA. In Biocomputing 2014, (Kohala Coast, Hawaii, USA: WORLD SCIENTIFIC), pp. 63–74.

Jones, C.L., Stevens, B.M., D'Alessandro, A., Reisz, J.A., Culp-Hill, R., Nemkov, T., Pei, S., Khan, N., Adane, B., Ye, H., et al. (2018). Inhibition of Amino Acid Metabolism Selectively Targets Human Leukemia Stem Cells. Cancer Cell *34*, 724-740.e4.

Kaever, A., Landesfeind, M., Feussner, K., Mosblech, A., Heilmann, I., Morgenstern, B., Feussner, I., and Meinicke, P. (2015). MarVis-Pathway: integrative and exploratory pathway analysis of non-targeted metabolomics data. Metabolomics *11*, 764–777.

Kahlert, C., Gaitzsch, E., Steinert, G., Mogler, C., Herpel, E., Hoffmeister, M., Jansen, L., Benner, A., Brenner, H., Chang-Claude, J., et al. (2012). Expression Analysis of Aldehyde Dehydrogenase 1A1 (ALDH1A1) in Colon and Rectal Cancer in Association with Prognosis and Response to Chemotherapy. Ann Surg Oncol *19*, 4193–4201.

Kaltenberg, J., Plum, L.M., Ober-Blöbaum, J.L., Hönscheid, A., Rink, L., and Haase, H. (2010). Zinc signals promote IL-2-dependent proliferation of T cells. Eur. J. Immunol. *40*, 1496–1503.

Kamburov, A., Cavill, R., Ebbels, T.M.D., Herwig, R., and Keun, H.C. (2011). Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. Bioinformatics *27*, 2917–2918.

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

Kamphorst, J.J., Nofal, M., Commisso, C., Hackett, S.R., Lu, W., Grabocka, E., Vander Heiden, M.G., Miller, G., Drebin, J.A., Bar-Sagi, D., et al. (2015). Human Pancreatic Cancer Tumors Are Nutrient Poor and Tumor Cells Actively Scavenge Extracellular Protein. Cancer Research *75*, 544–553.

Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K., and Tanabe, M. (2019a). New approach for understanding genome variations in KEGG. Nucleic Acids Res *47*, D590–D595.

Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K., and Tanabe, M. (2019b). New approach for understanding genome variations in KEGG. Nucleic Acids Research *47*, D590–D595.

Kang, J.H., Lee, S.-H., Hong, D., Lee, J.-S., Ahn, H.-S., Ahn, J.-H., Seong, T.W., Lee, C.-H., Jang, H., Hong, K.M., et al. (2016a). Aldehyde dehydrogenase is used by cancer cells for energy metabolism. Exp Mol Med *48*, e272–e272.

Kang, J.H., Lee, S.-H., Hong, D., Lee, J.-S., Ahn, H.-S., Ahn, J.-H., Seong, T.W., Lee, C.-H., Jang, H., Hong, K.M., et al. (2016b). Aldehyde dehydrogenase is used by cancer cells for energy metabolism. Exp Mol Med *48*, e272–e272.

Kato, Y., Maeda, T., Suzuki, A., and Baba, Y. (2018). Cancer metabolism: New insights into classic characteristics. Japanese Dental Science Review *54*, 8–21.

Kauffman, K.J., Prakash, P., and Edwards, J.S. (2003). Advances in flux balance analysis. Current Opinion in Biotechnology *14*, 491–496.

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

Kim, J. -w., Zeller, K.I., Wang, Y., Jegga, A.G., Aronow, B.J., O'Donnell, K.A., and Dang, C.V. (2004). Evaluation of Myc E-Box Phylogenetic Footprints in Glycolytic Genes by Chromatin Immunoprecipitation Assays. Molecular and Cellular Biology *24*, 5923–5936.

Kim, J. -w., Gao, P., Liu, Y.-C., Semenza, G.L., and Dang, C.V. (2007). Hypoxia-Inducible Factor 1 and Dysregulated c-Myc Cooperatively Induce Vascular Endothelial Growth Factor and Metabolic Switches Hexokinase 2 and Pyruvate Dehydrogenase Kinase 1. Molecular and Cellular Biology *27*, 7381–7393.

Klein, U., Lia, M., Crespo, M., Siegel, R., Shen, Q., Mo, T., Ambesi-Impiombato, A., Califano, A., Migliazza, A., Bhagat, G., et al. (2010). The DLEU2/miR-15a/16-1 Cluster Controls B Cell Proliferation and Its Deletion Leads to Chronic Lymphocytic Leukemia. Cancer Cell *17*, 28–40.

Knudson, A.G. (1971). Mutation and Cancer: Statistical Study of Retinoblastoma. Proceedings of the National Academy of Sciences *68*, 820–823.

Kobak, D., and Linderman, G.C. (2019). UMAP does not preserve global structure any better than t-SNE when using the same initialization (Bioinformatics).

Kobak, D., and Linderman, G.C. (2021). Initialization is critical for preserving global data structure in both t-SNE and UMAP. Nat Biotechnol *39*, 156–157.

Koenig, A.M., Karabatsiakis, A., Stoll, T., Wilker, S., Hennessy, T., Hill, M.M., and Kolassa, I.-T. (2018). Serum profile changes in postpartum women with a history of childhood maltreatment: a combined metabolite and lipid fingerprinting study. Sci Rep *8*, 3468.

König, K., Meder, L., Kröger, C., Diehl, L., Florin, A., Rommerscheidt-Fuss, U., Kahl, P., Wardelmann, E., Magin, T.M., Buettner, R., et al. (2013). Loss of the Keratin Cytoskeleton Is Not Sufficient to Induce Epithelial Mesenchymal Transition in a Novel KRAS Driven Sporadic Lung Cancer Mouse Model. PLoS ONE *8*, e57996.

Koppenol, W.H., Bounds, P.L., and Dang, C.V. (2011). Otto Warburg's contributions to current concepts of cancer metabolism. Nat Rev Cancer *11*, 325–337.

Korotkevich, G., Sukhov, V., and Sergushichev, A. (2016). Fast gene set enrichment analysis (Bioinformatics).

Koundouros, N., and Poulogiannis, G. (2020). Reprogramming of fatty acid metabolism in cancer. Br J Cancer *122*, 4–22.

Krajcovic, M., Krishna, S., Akkari, L., Joyce, J.A., and Overholtzer, M. (2013). mTOR regulates phagosome and entotic vacuole fission. MBoC *24*, 3736–3745.

Kuo, T.-C., Tian, T.-F., and Tseng, Y. (2013). 3Omics: a web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data. BMC Syst Biol *7*, 64.

Kutmon, M., van Iersel, M.P., Bohler, A., Kelder, T., Nunes, N., Pico, A.R., and Evelo, C.T. (2015). PathVisio 3: An Extendable Pathway Analysis Toolbox. PLoS Comput Biol *11*, e1004085.

Kwok, M., Oldreive, C., Rawstron, A.C., Goel, A., Papatzikas, G., Jones, R.E., Drennan, S., Agathanggelou, A., Sharma-Oates, A., Evans, P., et al. (2019). Integrative analysis of spontaneous CLL regression highlights genetic and microenvironmental interdependency in CLL. Blood blood.2019001262.

Labuschagne, C.F., van den Broek, N.J.F., Mackay, G.M., Vousden, K.H., and Maddocks, O.D.K. (2014). Serine, but Not Glycine, Supports One-Carbon Metabolism and Proliferation of Cancer Cells. Cell Reports *7*, 1248–1258.

Landau, B.R., Laszlo, J., Stengle, J., and Burk, D. (1958). Certain metabolic and pharmacologic effects in cancer patients given infusions of 2-deoxy-D-glucose. J. Natl. Cancer Inst. *21*, 485–494.

Laplante, M., and Sabatini, D.M. (2012). mTOR Signaling in Growth Control and Disease. Cell *149*, 274–293.

Le, A., Lane, A.N., Hamaker, M., Bose, S., Gouw, A., Barbi, J., Tsukamoto, T., Rojas, C.J., Slusher, B.S., Zhang, H., et al. (2012). Glucose-Independent Glutamine Metabolism via TCA Cycling for Proliferation and Survival in B Cells. Cell Metabolism *15*, 110–121.

Lê Cao, K.-A., González, I., and Déjean, S. (2009). integrOmics: an R package to unravel relationships between two omics datasets. Bioinformatics *25*, 2855–2856.

Lê Cao, K.-A., Boitard, S., and Besse, P. (2011). Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. BMC Bioinformatics *12*, 253.

Lee, H.J., Li, J., Vickman, R.E., Li, J., Liu, R., Durkes, A.C., Elzey, B.D., Yue, S., Liu, X., Ratliff, T.L., et al. (2018). Cholesterol Esterification Inhibition Suppresses Prostate Cancer Metastasis by Impairing the Wnt/β-catenin Pathway. Mol Cancer Res *16*, 974–985.

Lee, J.-C., Chiang, K.-C., Feng, T.-H., Chen, Y.-J., Chuang, S.-T., Tsui, K.-H., Chung, L.-C., and Juang, H.-H. (2016a). The Iron Chelator, Dp44mT, Effectively Inhibits Human Oral Squamous Cell Carcinoma Cell Growth in Vitro and in Vivo. IJMS *17*, 1435.

Lee, J.-S., Kang, J.H., Lee, S.-H., Hong, D., Son, J., Hong, K.M., Song, J., and Kim, S.-Y. (2016b). Dual targeting of glutaminase 1 and thymidylate synthase elicits death synergistically in NSCLC. Cell Death Dis *7*, e2511–e2511.

Leek, J.T., Scharpf, R.B., Bravo, H.C., Simcha, D., Langmead, B., Johnson, W.E., Geman, D., Baggerly, K., and Irizarry, R.A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet *11*, 733–739.

Leinonen, R., Sugawara, H., Shumway, M., and on behalf of the International Nucleotide Sequence Database Collaboration (2011). The Sequence Read Archive. Nucleic Acids Research *39*, D19–D21.

Lenz, G., Wright, G.W., Emre, N.C.T., Kohlhammer, H., Dave, S.S., Davis, R.E., Carty, S., Lam, L.T., Shaffer, A.L., Xiao, W., et al. (2008). Molecular subtypes of diffuse large B-cell lymphoma arise by distinct genetic pathways. Proceedings of the National Academy of Sciences *105*, 13520–13525.

Li, B., Ruotti, V., Stewart, R.M., Thomson, J.A., and Dewey, C.N. (2010). RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics *26*, 493–500.

Li, B., Severson, E., Pignon, J.-C., Zhao, H., Li, T., Novak, J., Jiang, P., Shen, H., Aster, J.C., Rodig, S., et al. (2016a). Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. Genome Biol *17*, 174.

Li, F.-Y., Chaigne-Delalande, B., Kanellopoulou, C., Davis, J.C., Matthews, H.F., Douek, D.C., Cohen, J.I., Uzel, G., Su, H.C., and Lenardo, M.J. (2011). Second messenger role for Mg2+ revealed by human T-cell immunodeficiency. Nature *475*, 471–476.

Li, H., Ning, S., Ghandi, M., Kryukov, G.V., Gopal, S., Deik, A., Souza, A., Pierce, K., Keskula, P., Hernandez, D., et al. (2019a). The landscape of cancer cell line metabolism. Nat Med *25*, 850–860.

Li, J., Gu, D., Lee, S.S.-Y., Song, B., Bandyopadhyay, S., Chen, S., Konieczny, S.F., Ratliff, T.L., Liu, X., Xie, J., et al. (2016b). Abrogating cholesterol esterification suppresses growth and metastasis of pancreatic cancer. Oncogene *35*, 6378–6388.

Li, M., Wang, Y., Zheng, R., Shi, X., li, yaohang, Wu, F., and Wang, J. (2019b). DeepDSC: A Deep Learning Method to Predict Drug Sensitivity of Cancer Cell Lines. IEEE/ACM Trans. Comput. Biol. and Bioinf. 1–1.

Li, Q., Yin, L., Jones, L.W., Chu, G.C.-Y., Wu, J.B.-Y., Huang, J.-M., Li, Q., You, S., Kim, J., Lu, Y.-T., et al. (2016c). Keratin 13 expression reprograms bone and brain metastases of human prostate cancer cells. Oncotarget *7*, 84645–84657.

Lindfors, E., van Dam, J.C.J., Lam, C.M.C., Zondervan, N.A., Martins dos Santos, V.A.P., and Suarez-Diez, M. (2018). SyNDI: synchronous network data integration framework. BMC Bioinformatics *19*, 403.

Liu, Y., and Barta, S.K. (2019). Diffuse large B-cell lymphoma: 2019 update on diagnosis, risk stratification, and treatment. Am J Hematol *94*, 604–616.

Lu, H., Forbes, R.A., and Verma, A. (2002). Hypoxia-inducible Factor 1 Activation by Aerobic Glycolysis Implicates the Warburg Effect in Carcinogenesis. J. Biol. Chem. *277*, 23111–23115.

Ludwig, C., and Günther, U.L. (2011). MetaboLab - advanced NMR data processing and analysis for metabolomics. BMC Bioinformatics *12*, 366.

Lui, G.Y.L., Kovacevic, Z., Richardson, V., Merlot, A.M., Kalinowski, D.S., and Richardson, D.R. (2015). Targeting cancer by binding iron: Dissecting cellular signaling pathways. Oncotarget *6*.

Lunt, S.Y., and Vander Heiden, M.G. (2011). Aerobic glycolysis: meeting the metabolic requirements of cell proliferation. Annu. Rev. Cell Dev. Biol. *27*, 441–464.

van der Maaten, L., and Hinton, G. (2008). Visualizing Data using t-SNE. Journal of Machine Learning Research 1 *9*, 2579--2605.

Machado, D., and Herrgård, M. (2014). Systematic Evaluation of Methods for Integration of Transcriptomic Data into Constraint-Based Models of Metabolism. PLoS Comput Biol *10*, e1003580.

Manzoni, C., Kia, D.A., Vandrovcova, J., Hardy, J., Wood, N.W., Lewis, P.A., and Ferrari, R. (2018). Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. Briefings in Bioinformatics *19*, 286–302.

Marke, R., van Leeuwen, F.N., and Scheijen, B. (2018). The many faces of IKZF1 in B-cell precursor acute lymphoblastic leukemia. Haematologica *103*, 565–574.

Mashimo, T., Pichumani, K., Vemireddy, V., Hatanpaa, K.J., Singh, D.K., Sirasanagandla, S., Nannepaga, S., Piccirillo, S.G., Kovacs, Z., Foong, C., et al. (2014). Acetate Is a Bioenergetic Substrate for Human Glioblastoma and Brain Metastases. Cell *159*, 1603–1614.

McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. ArXiv:1802.03426 [Cs, Stat].

McKeehan, W.L. (1982). Glycolysis, glutaminolysis and cell proliferation. Cell Biol. Int. Rep. *6*, 635–650.

Mehrmohamadi, M., Liu, X., Shestov, A.A., and Locasale, J.W. (2014). Characterization of the Usage of the Serine Metabolic Network in Human Cancer. Cell Reports *9*, 1507–1519.

Mellor, A.L., Keskin, D.B., Johnson, T., Chandler, P., and Munn, D.H. (2002). Cells Expressing Indoleamine 2,3-Dioxygenase Inhibit T Cell Responses. J Immunol *168*, 3771–3776.

Mesin, L., Ersching, J., and Victora, G.D. (2016). Germinal Center B Cell Dynamics. Immunity *45*, 471–482.

Mirabelli, Coppola, and Salvatore (2019). Cancer Cell Lines Are Useful Model Systems for Medical Research. Cancers *11*, 1098.

Misra, B.B., Langefeld, C., Olivier, M., and Cox, L.A. (2019). Integrated omics: tools, advances and future approaches. Journal of Molecular Endocrinology R21–R45.

Missihoun, T.D., Kotchoni, S.O., and Bartels, D. (2018). Aldehyde Dehydrogenases Function in the Homeostasis of Pyridine Nucleotides in Arabidopsis thaliana. Sci Rep *8*, 2936.

Mlynarczyk, C., Fontán, L., and Melnick, A. (2019). Germinal center-derived lymphomas: The darkest side of humoral immunity. Immunol Rev *288*, 214–239.

Monteiro, A., and Waizbort, R. (2007). The accidental cancer geneticist: Hilário de Gouvêa and hereditary retinoblastoma: Cancer Biology & Therapy: Vol 6, No 5. Cancer Biology and Theraphy *6*, 811–813.

Montero, J., Morales, A., Llacuna, L., Lluis, J.M., Terrones, O., Basañez, G., Antonsson, B., Prieto, J., García-Ruiz, C., Colell, A., et al. (2008). Mitochondrial Cholesterol Contributes to Chemotherapy Resistance in Hepatocellular Carcinoma. Cancer Res *68*, 5246–5256.

Monti, S., Savage, K.J., Kutok, J.L., Feuerhake, F., Kurtin, P., Mihm, M., Wu, B., Pasqualucci, L., Neuberg, D., Aguiar, R.C.T., et al. (2005). Molecular profiling of diffuse large B-cell lymphoma identifies robust subtypes including one characterized by host inflammatory response. Blood *105*, 1851–1861.

Morin, R.D., Johnson, N.A., Severson, T.M., Mungall, A.J., An, J., Goya, R., Paul, J.E., Boyle, M., Woolcock, B.W., Kuchenbauer, F., et al. (2010). Somatic mutations altering EZH2 (Tyr641) in follicular and diffuse large B-cell lymphomas of germinal-center origin. Nat Genet *42*, 181–185.

Morin, R.D., Mendez-Lago, M., Mungall, A.J., Goya, R., Mungall, K.L., Corbett, R.D., Johnson, N.A., Severson, T.M., Chiu, R., Field, M., et al. (2011). Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. Nature *476*, 298–303.

Mukherjee, S. (2011). The emperor of all maladies (London: Fourth Estate).

Mulder, Wahlin, Österborg, and Palma (2019). Targeting the Immune Microenvironment in Lymphomas of B-Cell Origin: From Biology to Clinical Application. Cancers *11*, 915.

Murphy, M.P. (2009). How mitochondria produce reactive oxygen species. Biochemical Journal *417*, 1–13.

Muz, B., de la Puente, P., Azab, F., and Azab, A.K. (2015). The role of hypoxia in cancer progression, angiogenesis, metastasis, and resistance to therapy. HP 83.

Narahara, M., Hamada-Kanazawa, M., Kouda, M., Odani, A., and Miyake, M. (2010). Superoxide Scavenging and Xanthine Oxidase Inhibiting Activities of Copper–b-Citryl-L-glutamate Complex. *33*, 6.

NCIN report (2014). Trends in incidence and outcome for haematological cancers in England: 2001-2010.

NCRAS, 2017 Haematological cancers. http://www.ncin.org.uk/cancer_type_and_topic_specific_work/cancer_type_specific_work/haematological_cancers/

NCT01791595 A Phase I Trial of AZD3965 in Patients With Advanced Cancer - Tabular View - https://www.clinicaltrials.gov/ct2/show/NCT01791595

NCT02071888 Study of the Glutaminase Inhibitor CB-839 in Hematological Tumors - Full Text View - https://clinicaltrials.gov/ct2/show/NCT02071888

NCT02481154 Study of Orally Administered AG-881 in Patients With Advanced Solid Tumors, Including Gliomas, With an IDH1 and/or IDH2 Mutation - Full Text View - https://clinicaltrials.gov/ct2/show/NCT02481154

NCT03875313 Study of CB-839 (Telaglenastat) in Combination With Talazoparib in Patients With Solid Tumors - https://clinicaltrials.gov/ct2/show/NCT03875313

Neuweger, H., Albaum, S.P., Dondrup, M., Persicke, M., Watt, T., Niehaus, K., Stoye, J., and Goesmann, A. (2008). MeltDB: a software platform for the analysis and integration of metabolomics experiment data. Bioinformatics *24*, 2726–2732.

Newman, A.M., Liu, C.L., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., Hoang, C.D., Diehn, M., and Alizadeh, A.A. (2015). Robust enumeration of cell subsets from tissue expression profiles. Nat Methods *12*, 453–457.

Newton, Y., Sedgewick, A.J., Cisneros, L., Golovato, J., Johnson, M., Szeto, C.W., Rabizadeh, S., Sanborn, J.Z., Benz, S.C., and Vaske, C. (2020). Large scale, robust, and accurate whole transcriptome profiling from clinical formalin-fixed paraffin-embedded samples. Sci Rep *10*, 17597.

Noble, R.A., Bell, N., Blair, H., Sikka, A., Thomas, H., Phillips, N., Nakjang, S., Miwa, S., Crossland, R., Rand, V., et al. (2017). Inhibition of monocarboxyate transporter 1 by AZD3965 as a novel therapeutic approach for diffuse large B-cell lymphoma and Burkitt lymphoma. Haematologica *102*, 1247–1257.

Nowicki, S., and Gottlieb, E. (2015). Oncometabolites: tailoring our genes. FEBS J *282*, 2796–2805.

Orchard, J.A., Ibbotson, R.E., Davis, Z., Wiestner, A., Rosenwald, A., Thomas, P.W., Hamblin, T.J., Staudt, L.M., and Oscier, D.G. (2004). ZAP-70 expression and prognosis in chronic lymphocytic leukaemia. The Lancet *363*, 105–111.

Orth, J.D., and Palsson, B.Ø. (2010). Systematizing the generation of missing metabolic knowledge. Biotechnol. Bioeng. *107*, 403–412.

Owen, M.R., Doran, E., and Halestrap, A.P. (2000). Evidence that metformin exerts its anti-diabetic effects through inhibition of complex 1 of the mitochondrial respiratory chain. 8.

Pacheco, R., Oliva, H., Martinez-Navío, J.M., Climent, N., Ciruela, F., Gatell, J.M., Gallart, T., Mallol, J., Lluis, C., and Franco, R. (2006). Glutamate Released by Dendritic Cells as a Novel Modulator of T Cell Activation. J Immunol *177*, 6695–6704.

Palm, W., Park, Y., Wright, K., Pavlova, N.N., Tuveson, D.A., and Thompson, C.B. (2015). The Utilization of Extracellular Proteins as Nutrients Is Suppressed by mTORC1. Cell *162*, 259–270.

Paolicchi, E., Gemignani, F., Krstic-Demonacos, M., Dedhar, S., Mutti, L., and Landi, S. (2016). Targeting hypoxic response for cancer therapy. Oncotarget *7*, 13464–13478.

Papandreou, I., Cairns, R.A., Fontana, L., Lim, A.L., and Denko, N.C. (2006). HIF-1 mediates adaptation to hypoxia by actively downregulating mitochondrial oxygen consumption. Cell Metabolism *3*, 187–197.

Parikh, S.A., Rabe, K.G., Call, T.G., Zent, C.S., Habermann, T.M., Ding, W., Leis, J.F., Schwager, S.M., Hanson, C.A., Macon, W.R., et al. (2013). Diffuse large B-cell lymphoma (Richter syndrome) in patients with chronic lymphocytic leukaemia (CLL): a cohort study of newly diagnosed patients. Br J Haematol *162*, 774–782.

Pasqualucci, L., Trifonov, V., Fabbri, G., Ma, J., Rossi, D., Chiarenza, A., Wells, V.A., Grunn, A., Messina, M., Elliot, O., et al. (2011). Analysis of the coding genome of diffuse large B-cell lymphoma. Nat Genet *43*, 830–837.

Pedersen, E.A., Shiozawa, Y., Mishra, A., and Taichman, R.S. (2013). Structure and function of the solid tumor niche. Front Biosci 22.

Pevsner, J. (2015). Bioinformatics and Functional Genomics (USA: Wiley Balckwell).

Pimentel, H., Bray, N.L., Puente, S., Melsted, P., and Pachter, L. (2017). Differential analysis of RNA-seq incorporating quantification uncertainty. Nature Methods *14*, 687–690.

Pires, I.M., Olcina, M.M., Anbalagan, S., Pollard, J.R., Reaper, P.M., Charlton, P.A., McKenna, W.G., and Hammond, E.M. (2012). Targeting radiation-resistant hypoxic tumour cells through ATR inhibition. Br J Cancer *107*, 291–299.

Polson, A.G. (2006). Expression pattern of the human FcRH/IRTA receptors in normal tissue and in B-chronic lymphocytic leukemia. International Immunology *18*, 1363–1373.

Ponce-de-Leon, M., Calle-Espinosa, J., Peretó, J., and Montero, F. (2015). Consistency Analysis of Genome-Scale Models of Bacterial Metabolism: A Metamodel Approach. PLoS ONE *10*, e0143626.

Puente, X.S., Pinyol, M., Quesada, V., Conde, L., Ordóñez, G.R., Villamor, N., Escaramis, G., Jares, P., Beà, S., González-Díaz, M., et al. (2011). Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. Nature *475*, 101–105.

Raez, L.E., Papadopoulos, K., Ricart, A.D., Chiorean, E.G., DiPaola, R.S., Stein, M.N., Rocha Lima, C.M., Schlesselman, J.J., Tolba, K., Langmuir, V.K., et al. (2013). A phase I dose-escalation trial of 2-deoxy-d-glucose alone or combined with docetaxel in patients with advanced solid tumors. Cancer Chemother Pharmacol *71*, 523–530.

Rawstron, A.C., Fazi, C., Agathangelidis, A., on behalf of ERIC (European Research Initiative on CLL), Villamor, N., Letestu, R., Nomdedeu, J., Palacio, C., Stehlikova, O., Kreuzer, K.-A., et al. (2016). A complementary role of multiparameter flow cytometry and high-throughput sequencing for minimal residual disease detection in chronic lymphocytic leukemia: an European Research Initiative on CLL study. Leukemia *30*, 929–936.

Rodchenkov, I., Babur, O., Luna, A., Aksoy, B.A., Wong, J.V., Fong, D., Franz, M., Siper, M.C., Cheung, M., Wrana, M., et al. (2019). Pathway Commons 2019 Update: integration, analysis and exploration of pathway data. Nucleic Acids Research gkz946.

Rodriguez-Martinez, A., Ayala, R., Posma, J.M., Neves, A.L., Gauguier, D., Nicholson, J.K., and Dumas, M.-E. (2016). MetaboSignal: a network-based approach for topological analysis of metabotype regulation *via* metabolic and signaling pathways. Bioinformatics btw697.

Rohart, F., Gautier, B., Singh, A., and Lê Cao, K.-A. (2017). mixOmics: An R package for 'omics feature selection and multiple data integration. PLoS Comput Biol *13*, e1005752.

Ronan, T., Qi, Z., and Naegle, K.M. (2016). Avoiding common pitfalls when clustering biological data. Science Signaling *9*, re6–re6.

Rosenthal, A., and Younes, A. (2017). High grade B-cell lymphoma with rearrangements of MYC and BCL2 and/or BCL6: Double hit and triple hit lymphomas and double expressing lymphoma. Blood Reviews *31*, 37–42.

Rosenwald, A., and Ott, G. (2008). Burkitt lymphoma versus diffuse large B-cell lymphoma. Annals of Oncology *19*, iv67–iv69.

Rosenwald, A., Wright, G., Chan, W.C., Connors, J.M., Campo, E., Fisher, R.I., Gascoyne, R.D., Muller-Hermelink, H.K., Smeland, E.B., Giltnane, J.M., et al. (2002). The Use of Molecular Profiling to Predict Survival after Chemotherapy for Diffuse Large-B-Cell Lymphoma. N Engl J Med *346*, 1937–1947.

Sacco, A., Nor, J., Belile, E., Sukari, A., Chepeha, D., Bradford, C., Eisbruch, A., Wolf, G., Urba, S., and Worden, F. (2014). Phase 2 Trial of AT-101 in Combination With Docetaxel for Recurrent, Locally Advanced, or Metastatic Head-and-Neck Squamous Cell Carcinoma (HNSCC). International Journal of Radiation Oncology*Biology*Physics *88*, 507.

Samanta, D., and Semenza, G.L. (2016). Serine Synthesis Helps Hypoxic Cancer Stem Cells Regulate Redox. Cancer Res *76*, 6458–6462.

Samudio, I., Harmancey, R., Fiegl, M., Kantarjian, H., Konopleva, M., Korchin, B., Kaluarachchi, K., Bornmann, W., Duvvuri, S., Taegtmeyer, H., et al. (2010). Pharmacologic inhibition of fatty acid oxidation sensitizes human leukemia cells to apoptosis induction. J. Clin. Invest. *120*, 142–156.

Saraei, P., Asadi, I., Kakar, M.A., and Moradi-Kor, N. (2019). The beneficial effects of metformin on cancer prevention and therapy: a comprehensive review of recent advances. CMAR *Volume 11*, 3295–3313.

Sartori, T., Galvão dos Santos, G., Nogueira-Pedro, A., Makiyama, E., Rogero, M.M., Borelli, P., and Fock, R.A. (2018). Effects of glutamine, taurine and their association on inflammatory pathway markers in macrophages. Inflammopharmacol *26*, 829–838.

Schelman, W.R., Mohammed, T.A., Traynor, A.M., Kolesar, J.M., Marnocha, R.M., Eickhoff, J., Keppen, M., Alberti, D.B., Takebe, N., and Liu, G. (2015). A Phase I study of AT-101 with Cisplatin and Etoposide in patients with advanced solid tumors with an Expanded Cohort in Extensive-Stage Small Cell Lung Cancer. 16.

Scherer, W.F., Syverton, J.T., and Gey, G.O. (1953). STUDIES ON THE PROPAGATION IN VITRO OF POLIOMYELITIS VIRUSES. J Exp Med *97*, 695–710.

Schmitz, R., Young, R.M., Ceribelli, M., Jhavar, S., Xiao, W., Zhang, M., Wright, G., Shaffer, A.L., Hodson, D.J., Buras, E., et al. (2012). Burkitt lymphoma

pathogenesis and therapeutic targets from structural and functional genomics. Nature *490*, 116–120.

Schmitz, R., Ceribelli, M., Pittaluga, S., Wright, G., and Staudt, L.M. (2014). Oncogenic Mechanisms in Burkitt Lymphoma. Cold Spring Harbor Perspectives in Medicine *4*, a014282–a014282.

Schug, Z.T., Peck, B., Jones, D.T., Zhang, Q., Grosskurth, S., Alam, I.S., Goodwin, L.M., Smethurst, E., Mason, S., Blyth, K., et al. (2015). Acetyl-CoA Synthetase 2 Promotes Acetate Utilization and Maintains Cancer Cell Growth under Metabolic Stress. Cancer Cell *27*, 57–71.

Schwarzfischer, P., Reinders, J., Dettmer, K., Kleo, K., Dimitrova, L., Hummel, M., Feist, M., Kube, D., Szczepanowski, M., Klapper, W., et al. (2017). Comprehensive Metaboproteomics of Burkitt's and Diffuse Large B-Cell Lymphoma Cell Lines and Primary Tumor Tissues Reveals Distinct Differences in Pyruvate Content and Metabolism. Journal of Proteome Research *16*, 1105–1120.

Scott, D.W., and Gascoyne, R.D. (2014). The tumour microenvironment in B cell lymphomas. Nat Rev Cancer *14*, 517–534.

Seifert, M., Sellmann, L., Bloehdorn, J., Wein, F., Stilgenbauer, S., Dürig, J., and Küppers, R. (2012). Cellular origin and pathophysiology of chronic lymphocytic leukemia. The Journal of Experimental Medicine *209*, 2183–2198.

Selak, M.A., Armour, S.M., MacKenzie, E.D., Boulahbel, H., Watson, D.G., Mansfield, K.D., Pan, Y., Simon, M.C., Thompson, C.B., and Gottlieb, E. (2005). Succinate links TCA cycle dysfunction to oncogenesis by inhibiting HIF-α prolyl hydroxylase. Cancer Cell *7*, 77–85.

Seltzer, M.J., Bennett, B.D., Joshi, A.D., Gao, P., Thomas, A.G., Ferraris, D.V., Tsukamoto, T., Rojas, C.J., Slusher, B.S., Rabinowitz, J.D., et al. (2010). Inhibition of Glutaminase Preferentially Slows Growth of Glioma Cells with Mutant IDH1. Cancer Research *70*, 8981–8987.

Settembre, C., and Ballabio, A. (2014). Lysosome: regulator of lipid degradation pathways. Trends in Cell Biology *24*, 743–750.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res. *13*, 2498–2504.

Shen, X., and Zhu, Z.-J. (2019). MetFlow: an interactive and integrated workflow for metabolomics data cleaning and differential metabolite discovery. Bioinformatics *35*, 2870–2872.

Shi, Y., Chen, L., Liotta, L.A., Wan, H.-H., and Rodgers, G.P. (2006). Glia Maturation Factor Gamma (GMFG): A Cytokine-Responsive Protein During Hematopoietic Lineage Development and Its Functional Genomics Analysis. Genomics, Proteomics & Bioinformatics *4*, 145–155.

Sihto, H., Lundin, J., Lundin, M., Lehtimäki, T., Ristimäki, A., Holli, K., Sailas, L., Kataja, V., Turpeenniemi-Hujanen, T., Isola, J., et al. (2011). Breast cancer biological subtypes and protein expression predict for the preferential distant metastasis sites: a nationwide cohort study. Breast Cancer Res *13*, R87.

Simillion, C., Liechti, R., Lischer, H.E.L., Ioannidis, V., and Bruggmann, R. (2017). Avoiding the pitfalls of gene set enrichment analysis with SetRank. BMC Bioinformatics *18*, 151.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. J. Stat. Soft. *39*.

Singh, A., Shannon, C.P., Gautier, B., Rohart, F., Vacher, M., Tebbutt, S.J., and Lê Cao, K.-A. (2016). DIABLO: from multi-omics assays to biomarker discovery, an integrative approach (Bioinformatics).

Singh, A., Shannon, C.P., Gautier, B., Rohart, F., Vacher, M., Tebbutt, S.J., and Lê Cao, K.-A. (2019). DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. Bioinformatics *35*, 3055–3062.

Singh, R., Kaushik, S., Wang, Y., Xiang, Y., Novak, I., Komatsu, M., Tanaka, K., Cuervo, A.M., and Czaja, M.J. (2009). Autophagy regulates lipid metabolism. Nature *458*, 1131–1135.

Sonveaux, P., Végran, F., Schroeder, T., Wergin, M.C., Verrax, J., Rabbani, Z.N., De Saedeleer, C.J., Kennedy, K.M., Diepart, C., Jordan, B.F., et al. (2008). Targeting lactate-fueled respiration selectively kills hypoxic tumor cells in mice. J. Clin. Invest. JCI36843.

Sonveaux, P., Copetti, T., De Saedeleer, C.J., Végran, F., Verrax, J., Kennedy, K.M., Moon, E.J., Dhup, S., Danhier, P., Frérart, F., et al. (2012). Targeting the Lactate Transporter MCT1 in Endothelial Cells Inhibits Lactate-Induced HIF-1 Activation and Tumor Angiogenesis. PLoS ONE *7*, e33418.

Stankovic, T., and Skowronska, A. (2014). The role of *ATM* mutations and 11q deletions in disease progression in chronic lymphocytic leukemia. Leukemia & Lymphoma *55*, 1227–1239.

Stebegg, M., Kumar, S.D., Silva-Cayetano, A., Fonseca, V.R., Linterman, M.A., and Graca, L. (2018). Regulation of the Germinal Center Response. Front. Immunol. *9*, 2469.

Stein, M., Lin, H., Jeyamohan, C., Dvorzhinski, D., Gounder, M., Bray, K., Eddy, S., Goodin, S., White, E., and DiPaola, R.S. (2010). Targeting tumor metabolism with 2-deoxyglucose in patients with castrate-resistant prostate cancer and advanced malignancies. Prostate *70*, 1388–1394.

Stempler, S., Yizhak, K., and Ruppin, E. (2014). Integrating Transcriptomics with Metabolic Modeling Predicts Biomarkers and Drug Targets for Alzheimer's Disease. PLoS ONE *9*, e105383.

Stevenson, F.K., Krysov, S., Davies, A.J., Steele, A.J., and Packham, G. (2011). B-cell receptor signaling in chronic lymphocytic leukemia. Blood *118*, 4313–4320.

Storch, C.H., Ehehalt, R., Haefeli, W.E., and Weiss, J. (2007). Localization of the Human Breast Cancer Resistance Protein (BCRP/ABCG2) in Lipid Rafts/Caveolae and Modulation of Its Activity by Cholesterol in Vitro. J Pharmacol Exp Ther *323*, 257–264.

Storozhuk, Y., Hopmans, S.N., Sanli, T., Barron, C., Tsiani, E., Cutz, J.-C., Pond, G., Wright, J., Singh, G., and Tsakiridis, T. (2013). Metformin inhibits growth and enhances radiation response of non-small cell lung cancer (NSCLC) through ATM and AMPK. Br J Cancer *108*, 2021–2032.

Stransky, N., Ghandi, M., Kryukov, G.V., Garraway, L.A., and Amzallag, A. (2015). Pharmacogenomic agreement between two cancer cell line data sets. Nature *528*, 84–87.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences *102*, 15545–15550.

Sun, X., and Wolfram, W. (2012). COVAIN: a toolbox for uni- and multivariate statistics, time-series and correlation network analysis and inverse estimation of the differential Jacobian from metabolomics covariance data. Metabolomics 81–93.

Sun, X., Wei, Y., Lee, P.P., Ren, B., and Liu, C. (2019). The role of WASp in T cells and B cells. Cellular Immunology *341*, 103919.

Swerdlow, S.H., Campo, E., Pileri, S.A., Harris, N.L., Stein, H., Siebert, R., Advani, R., Ghielmini, M., Salles, G.A., Zelenetz, A.D., et al. (2016). The 2016 revision of The World Health Organization classification of lymphoid neoplasms. Blood *127*, 2375–2390.

Swietach, P., Vaughan-Jones, R.D., and Harris, A.L. (2007). Regulation of tumor pH and the role of carbonic anhydrase 9. Cancer Metastasis Rev *26*, 299–310.

Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P., et al. (2019). STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res *47*, D607–D613.

Tang, N., Wang, L., Esko, J., Giordano, F.J., Huang, Y., Gerber, H.-P., Ferrara, N., and Johnson, R.S. (2004). Loss of HIF-1␣ in endothelial cells disrupts a hypoxia-driven VEGF autocrine loop necessary for tumorigenesis. CANCER CELL 11.

Teater, M., Dominguez, P.M., Redmond, D., Chen, Z., Ennishi, D., Scott, D.W., Cimmino, L., Ghione, P., Chaudhuri, J., Gascoyne, R.D., et al. (2018). AICDA drives epigenetic heterogeneity and accelerates germinal center-derived lymphomagenesis. Nat Commun *9*, 222.

Tenenhaus, A., Philippe, C., Guillemot, V., Le Cao, K.-A., Grill, J., and Frouin, V. (2014). Variable selection for generalized canonical correlation analysis. Biostatistics *15*, 569–583.

The ICGC MMML-Seq Project (2012). Recurrent mutation of the ID3 gene in Burkitt lymphoma identified by integrated genome, exome and transcriptome sequencing. Nat Genet *44*, 1316–1320.

Thimm, O., Bläsing, O., Gibon, Y., Nagel, A., Meyer, S., Krüger, P., Selbig, J., Müller, L.A., Rhee, S.Y., and Stitt, M. (2004). mapman: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. The Plant Journal *37*, 914–939.

Thomas, R., Ribeiro, I., Shepherd, P., Johnson, P., Cook, M., Lakhani, A., Kaczmarski, R., Carrington, P., and Catovsky, D. (2002). Spontaneous clinical regression in chronic lymphocytic leukaemia. British Journal of Haematology *116*, 341–345.

Thompson, S.J., Thompson, S.E.M., and Cazier, J.-B. (2019). CaStLeS (Compute and Storage for the Life Sciences): a collection of compute and storage resources for supporting research at the University of Birmingham.

Tipping, M.E., and Bishop, C.M. (1999). Mixtures of Probabilistic Principal Component Analysers. 30.

Tomita, H., Tanaka, K., Tanaka, T., and Hara, A. (2016). Aldehyde dehydrogenase 1A1 in stem cells and cancer. Oncotarget *7*.

Tosi, M.R., and Tugnoli, V. (2005). Cholesteryl esters in malignancy. Clinica Chimica Acta *359*, 27–45.

Turnbull, A.K., Selli, C., Martinez-Perez, C., Fernando, A., Renshaw, L., Keys, J., Figueroa, J.D., He, X., Tanioka, M., Munro, A.F., et al. (2020). Unlocking the transcriptomic potential of formalin-fixed paraffin embedded clinical tissues: comparison of gene expression profiling approaches. BMC Bioinformatics *21*, 30.

Valcárcel, L.V., Torrano, V., Tobalina, L., Carracedo, A., and Planes, F.J. (2019). rMTA: robust metabolic transformation analysis. Bioinformatics *35*, 4350–4355.

Valvezan, A.J., and Manning, B.D. (2019). Molecular logic of mTORC1 signalling as a metabolic rheostat. Nat Metab *1*, 321–333.

Vander Heiden, M.G., and DeBerardinis, R.J. (2017). Understanding the Intersections between Metabolism and Cancer Biology. Cell *168*, 657–669.

Vazquez, A., Kamphorst, J.J., Markert, E.K., Schug, Z.T., Tardito, S., and Gottlieb, E. (2016). Cancer metabolism at a glance. J Cell Sci *129*, 3367–3373.

Vinnars, E., Bergström, J., and Fürst, P. (1975). Influence of the Postoperative State on the Intracellular Free Amino Acids in Human Muscle Tissue: Annals of Surgery *182*, 665–671.

Wägele, B., Witting, M., Schmitt-Kopplin, P., and Suhre, K. (2012). MassTRIX Reloaded: Combined Analysis and Visualization of Transcriptome and Metabolome Data. PLoS ONE *7*, e39860.

Wai, T., and Langer, T. (2016). Mitochondrial Dynamics and Metabolic Regulation. Trends in Endocrinology & Metabolism *27*, 105–117.

Walenta, S., Wetterling, M., Lehrke, M., Schwickert, G., Sundfør, K., Rofstad, E.K., and Mueller-Klieser, W. High Lactate Levels Predict Likelihood of Metastases, Tumor Recurrence, and Restricted Patient Survival in Human Cervical Cancers. 7.

Wamelink, M.M.C., Struys, E.A., and Jakobs, C. (2008). The biochemistry, metabolism and inherited defects of the pentose phosphate pathway: a review. J. Inherit. Metab. Dis. *31*, 703–717.

Wang, B., Chen, X., Wang, Z., Xiong, W., Xu, T., Zhao, X., Cao, Y., Guo, Y., Li, L., Chen, S., et al. (2017a). Aldehyde dehydrogenase 1A1 increases NADH levels and promotes tumor growth via glutathione/dihydrolipoic acid-dependent NAD^+ reduction. Oncotarget *8*.

Wang, D., Wu, L., Cao, Y., Yang, L., Liu, W., E, X., Ji, G., and Bi, Z. (2017b). A novel mechanism of mTORC1-mediated serine/glycine metabolism in osteosarcoma development. Cellular Signalling *29*, 107–114.

Wanichthanarak, K., Fan, S., Grapov, D., Barupal, D.K., and Fiehn, O. (2017). Metabox: A Toolbox for Metabolomic Data Analysis, Interpretation and Integrative Exploration. PLoS ONE *12*, e0171046.

Warburg, O. (1924). Über den Stoffwechsel der Carcinomzelle. Naturwissenschaften *12*, 1131–1137.

Warburg, O. (1956). On the origin of cancer cells. Science *123*, 309–314.

Warburg, O., Wind, F., and Negelein, E. (1927). The metabolism of tumors in the body. J Gen Physiol *8*, 519–530.

Warnes, G., Bolker, B., Bonebakker, L., Gentleman, R., Huber, W., Liaw, A., Lumley, T., Maechler, M., Mangusson, A., Moeller, S., et al. (2019). gplots: Various R Programming Tools for Plotting Data.

Watson, L., Wyld, P., and Catovsky, D. (2008). Disease burden of chronic lymphocytic leukaemia within the European Union. European Journal of Haematology *81*, 253–258.

Wheaton, W.W., Weinberg, S.E., Hamanaka, R.B., Soberanes, S., Sullivan, L.B., Anso, E., Glasauer, A., Dufour, E., Mutlu, G.M., Budigner, G.S., et al. (2014). Metformin inhibits mitochondrial complex I of cancer cells to reduce tumorigenesis. ELife *3*, e02242.

WHO, Stewart B, Stewart, B.W., Wild, C., International Agency for Research on Cancer, and World Health Organization (2014). World cancer report 2014.

WHO, 2018 Cancer Today. https://gco.iarc.fr/today/data/factsheets/cancers/34-Non-hodgkin-lymphoma-fact-sheet.pdf

WHO, 2019 WHO | International Classification of Diseases, 11th Revision (ICD-11). https://www.who.int/classifications/icd/en/

Wiese, E.K., and Hitosugi, T. (2018). Tyrosine Kinase Signaling in Cancer Metabolism: PKM2 Paradox in the Warburg Effect. Front. Cell Dev. Biol. *6*, 79.

Wise, D.R., DeBerardinis, R.J., Mancuso, A., Sayed, N., Zhang, X.-Y., Pfeiffer, H.K., Nissim, I., Daikhin, E., Yudkoff, M., McMahon, S.B., et al. (2008). Myc regulates a transcriptional program that stimulates mitochondrial glutaminolysis and leads to glutamine addiction. Proceedings of the National Academy of Sciences *105*, 18782–18787.

Wold, S., Sjöström, M., and Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems *58*, 109–130.

World Health Organization (2018). World health statistics 2018: monitoring health for the SDGs.

Wyss, M., and Kaddurah-Daouk, R. (2000). Creatine and Creatinine Metabolism. *80*, 107.

Xia, J., Psychogios, N., Young, N., and Wishart, D.S. (2009). MetaboAnalyst: a web server for metabolomic data analysis and interpretation. Nucleic Acids Research *37*, W652–W660.

Xiang, Y., Stine, Z.E., Xia, J., Lu, Y., O'Connor, R.S., Altman, B.J., Hsieh, A.L., Gouw, A.M., Thomas, A.G., Gao, P., et al. (2015). Targeted inhibition of tumor-specific glutaminase diminishes cell-autonomous tumorigenesis. J. Clin. Invest. *125*, 2293–2306.

Yalamanchili, H.K., Wan, Y., and Liu, Z. (2017). Data Analysis Pipeline for RNA-seq Experiments: From Differential Expression to Cryptic Splicing. Current Protocols in Bioinformatics *59*.

Yang, M., and Vousden, K.H. (2016). Serine and one-carbon metabolism in cancer. Nat Rev Cancer *16*, 650–662.

Yang, J.-S., Lin, C.-W., Chuang, C.-Y., Su, S.-C., Lin, S.-H., and Yang, S.-F. (2015). Carbonic anhydrase IX overexpression regulates the migration and progression in oral squamous cell carcinoma. Tumor Biol. *36*, 9517–9524.

Yang, S., Wang, X., Contino, G., Liesa, M., Sahin, E., Ying, H., Bause, A., Li, Y., Stommel, J.M., Dell'Antonio, G., et al. (2011). Pancreatic cancers require autophagy for tumor growth. Genes & Development *25*, 717–729.

Ye, J., Fan, J., Venneti, S., Wan, Y.-W., Pawel, B.R., Zhang, J., Finley, L.W.S., Lu, C., Lindsten, T., Cross, J.R., et al. (2014). Serine Catabolism Regulates Mitochondrial Redox Control during Hypoxia. Cancer Discovery *4*, 1406–1417.

Yizhak, K., Gabay, O., Cohen, H., and Ruppin, E. (2013). Model-based identification of drug targets that revert disrupted metabolism and its application to ageing. Nat Commun *4*, 2632.

Youle, R.J., and van der Bliek, A.M. (2012). Mitochondrial Fission, Fusion, and Stress. Science *337*, 1062–1065.

Yue, S., Li, J., Lee, S.-Y., Lee, H.J., Shao, T., Song, B., Cheng, L., Masterson, T.A., Liu, X., Ratliff, T.L., et al. (2014). Cholesteryl Ester Accumulation Induced by

PTEN Loss and PI3K/AKT Activation Underlies Human Prostate Cancer Aggressiveness. Cell Metabolism *19*, 393–406.

Zhang, N., Wang, H., Fang, Y., Wang, J., Zheng, X., and Liu, X.S. (2015). Predicting Anticancer Drug Responses Using a Dual-Layer Integrated Cell Line-Drug Network Model. PLoS Comput Biol *11*, e1004498.

Zhao, Q., Kuang, D.-M., Wu, Y., Xiao, X., Li, X.-F., Li, T.-J., and Zheng, L. (2012). Activated CD69 $^{+}$ T Cells Foster Immune Privilege by Regulating IDO Expression in Tumor-Associated Macrophages. J.I. *188*, 1117–1124.

Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K., and Liu, X. (2014). Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. PLoS ONE *9*, e78644.

Zhao, Y., Dong, Q., Li, J., Zhang, K., Qin, J., Zhao, J., Sun, Q., Wang, Z., Wartmann, T., Jauch, K.W., et al. (2018). Targeting cancer stem cells and their niche: perspectives for future therapeutic targets and strategies. Seminars in Cancer Biology *53*, 139–155.

Zhao, Z., Li, K., Toumazou, C., and Kalofonou, M. (2019). A computational model for anti-cancer drug sensitivity prediction. In 2019 IEEE Biomedical Circuits and Systems Conference (BioCAS), (Nara, Japan: IEEE), pp. 1–4.

Zhou, G., and Xia, J. (2018). OmicsNet: a web-based tool for creation and visual analysis of biological networks in 3D space. Nucleic Acids Research *46*, W514–W522.

Zur, H., Ruppin, E., and Shlomi, T. (2010). iMAT: an integrative metabolic analysis tool. Bioinformatics *46*, 3140-3142.

APPENDICES

Appendix 1 (Chapter 2, section 2.2.1): The RNAseq data from CLL primary samples. Table presents clinical features of the CLL cases of which the data were generated. Results from the alignment of processed reads from each sample with the GRCh38 human reference genome cDNA index are also highlighted in this table.

Cases	Group	Clinical status	Processed reads	Pseudoaligned reads	% mapped reads	averaged frag. length
CLL01	Regression	Complete spontaneous regression	30943705	21616435	69.86	171.26
CLL03	Regression	Complete spontaneous regression	36527963	28327583	77.55	173.17
CLL05	Regression	Complete spontaneous regression	30048891	22305966	74.23	172.45
CLL06	Regression	Complete spontaneous regression	35622729	25881842	72.66	182.16
CLL07	Regression	Complete spontaneous regression	26148321	21073875	80.59	161.67
CLL10	Regression	Complete spontaneous regression	30565307	18970445	62.07	169.97
CLL11	Regression	Complete spontaneous regression	30785248	21912374	71.18	178.07
CLL14	Regression	Partial spontaneous regression	38669608	30232432	78.18	172.46
CLL15	Regression	Partial spontaneous regression	28716128	20901237	72.79	164.80
CLL16	Regression	Partial spontaneous regression	25791319	19992820	77.52	175.42
CLL18	Regression	Complete spontaneous regression	37505477	26462638	70.56	172.57
CLL19	Regression	Partial spontaneous regression	29291895	18831836	64.29	172.21
INDOL01	Non-regression	Indolent IGHV mutated	35945164	27992348	77.88	179.06
INDOL02	Non-regression	Indolent IGHV mutated	38980529	29335691	75.26	176.37
INDOL03	Non-regression	Indolent IGHV mutated	28932077	21854524	75.54	175.59
INDOL04	Non-regression	Indolent IGHV mutated	31063099	25015139	80.53	168.70
INDOL05	Non-regression	Indolent IGHV mutated	30254604	21232739	70.18	161.39
INDOL06	Non-regression	Indolent IGHV mutated	28638109	22118156	77.23	169.37
INDOL07	Non-regression	Indolent IGHV mutated	35244033	28152754	79.88	169.66
INDOL08	Non-regression	Indolent IGHV mutated	34771283	27464527	78.99	175.53
INDOL09	Non-regression	Indolent IGHV mutated	29340625	23269717	79.31	169.10
INDOL10	Non-regression	Indolent IGHV mutated	54384370	37590872	69.12	185.69
INDOL11	Non-regression	Indolent IGHV mutated	35535158	26407816	74.31	168.06
INDOL12	Non-regression	Indolent IGHV mutated	40883910	31997134	78.26	189.45
INDOL13	Non-regression	Indolent IGHV mutated	26186150	19677182	75.14	164.03

Cases	Group	Clinical status	Processed reads	Pseudoaligned reads	% mapped reads	averaged frag. length
INDOL14	Non-regression	Indolent IGHV mutated	36596740	29440931	80.45	172.17
INDOL15	Non-regression	Indolent IGHV mutated	31950052	22500620	70.42	176.24
INDOL16	Non-regression	Indolent IGHV mutated	36801963	28818446	78.31	177.83
PROG01	Non-regression	Progressive IGHV mutated	28950500	22171926	76.59	176.21
PROG03	Non-regression	Progressive IGHV mutated	33482176	26009568	77.68	182.24
PROG04	Non-regression	Progressive IGHV mutated	29758319	20270659	68.12	180.07
PROG05	Non-regression	Progressive IGHV mutated	30731201	23549922	76.63	188.91
PROG06	Non-regression	Progressive IGHV mutated	28859177	23110222	80.08	183.10
PROG08	Non-regression	Progressive IGHV mutated	30355668	24018284	79.12	176.14
PROG09	Non-regression	Progressive IGHV mutated	27190300	20165190	74.16	188.51
PROG12	Non-regression	Progressive IGHV mutated	32741757	25978940	79.34	182.88
UnM01	Non-regression	IGHV Unmutated	32780244	25729318	78.49	167.73
UnM02	Non-regression	IGHV Unmutated	27610284	17571877	63.64	181.08
UnM03	Non-regression	IGHV Unmutated	34836801	27907967	80.11	177.48
UnM04	Non-regression	IGHV Unmutated	38010111	29553844	77.75	180.41
UnM05	Non-regression	IGHV Unmutated	29048322	23478370	80.83	179.56
UnM07	Non-regression	IGHV Unmutated	34293057	24142725	70.40	181.51
UnM08	Non-regression	IGHV Unmutated	31351804	23997920	76.54	188.28
UnM09	Non-regression	IGHV Unmutated	27978904	20713517	74.03	173.89
UnM10	Non-regression	IGHV Unmutated	36790532	25819707	70.18	181.24

Appendix 2 (Chapter 2, section 2.3.2): Statistically significant gene sets from GSEA with SetRank (parameters thresholds: setPCutoff = 0.01 and fdrCutoff = 0.05). Table presents details of the KEGG disease gene sets, such as the KEGG name, the description, and the size of gene set. The setRank value, the associated p-value and adjusted p-value are also presented in this table. The pp denotes for the negative logarithm of the p-value.

KEGG name	Description	size	setRank	p-value SetRank	corrected p-value	adjusted p-value	pp
M00177	Ribosome, eukaryotes	89	0.019	1	5.84E-35	7.01E-34	34.23
hsa05202	Transcriptional misregulation in cancer	129	0.148	0.26237	6.37E-09	7.01E-08	8.20
hsa00190	Oxidative phosphorylation	111	0.044	1	2.88E-07	7.01E-08	6.54
hsa05132	Salmonella infection	68	0.051	1	2.84E-05	7.01E-08	4.55
hsa05034	Alcoholism	130	0.049	1	0.0006998	7.01E-08	3.16
M00147	NADH dehydrogenase (ubiquinone) 1 beta subcomplex	13	0.019	1	0.0009362	7.01E-08	3.03
M00160	V-type ATPase, eukaryotes	20	0.019	1	0.001251	7.01E-08	2.90
hsa00604	Glycosphingolipid biosynthesis - ganglio series	12	0.019	1	0.0014319	0.014319	2.84
hsa04216	Ferroptosis	35	0.019	1	0.0016068	0.014461	2.79
hsa04022	cGMP-PKG signaling pathway	112	0.034	1	0.0017403	7.01E-08	2.76
M00412	ESCRT-III complex	12	0.019	1	0.0018315	7.01E-08	2.74
hsa05110	Vibrio cholerae infection	45	0.019	1	0.0019041	7.01E-08	2.72
hsa00770	Pantothenate and CoA biosynthesis	16	0.019	1	0.0021073	0.016858	2.68
hsa04141	Protein processing in endoplasmic reticulum	156	0.019	1	0.0029436	0.020605	2.53
hsa00514	Other types of O-glycan biosynthesis	21	0.019	1	0.003034	0.020605	2.52
hsa04932	Non-alcoholic fatty liver disease (NAFLD)	131	0.019	1	0.0032068	7.01E-08	2.49
hsa04144	Endocytosis	231	0.019	1	0.0033851	7.01E-08	2.47
M00351	Spliceosome, U1-snRNP	10	0.019	1	0.0046049	0.023024	2.34
hsa04611	Platelet activation	100	0.034	1	0.0048242	7.01E-08	2.32
hsa05166	HTLV-I infection	225	0.046	1	0.0053498	7.01E-08	2.27
M00051	Uridine monophosphate biosynthesis, glutamine (+ PRPP) => UMP	3	0.019	1	0.0054388	0.023024	2.26
M00070	Glycosphingolipid biosynthesis, lacto-series, LacCer => Lc4Cer	4	0.019	1	0.0054427	0.023024	2.26
hsa04640	Hematopoietic cell lineage	67	0.027	1	0.0056124	7.01E-08	2.25

KEGG name	Description	size	setRank	p-value SetRank	corrected p-value	adjusted p-value	pp
M00400	p97-Ufd1-Npl4 complex	3	0.019	1	0.0058753	0.020605	2.23
hsa05206	MicroRNAs in cancer	130	0.032	1	0.0062573	7.01E-08	2.20
hsa00740	Riboflavin metabolism	7	0.019	1	0.0072449	0.023024	2.14
hsa05145	Toxoplasmosis	102	0.019	1	0.007551	7.01E-08	2.12
hsa00360	Phenylalanine metabolism	10	0.019	1	0.0076842	0.023024	2.11
hsa05200	Pathways in cancer	305	0.052	1	0.0076936	7.01E-08	2.11
hsa05164	Influenza A	146	0.025	1	0.0093294	7.01E-08	2.03
hsa04810	Regulation of actin cytoskeleton	169	0.054	1	0.0097915	7.01E-08	2.01

Appendix 3 (Chapter 3, section 3.2.1): NHL primary tumours RNAseq data. Table presents clinical features of the NHL primary tumours of which the data were generated. Results from the alignment of processed reads from each sample with the GRCh38 human reference genome cDNA index are also highlighted in this table.

Sample ID	SRA	GEO	condition	processed reads	aligned reads	% mapped reads
DLBCL10	SRR6033228	GSM2782656	GCB-DLBCL	28824118	22989325	79.76
DLBCL11	SRR6033229	GSM2782657	GCB-DLBCL	28411571	23726207	83.51
DLBCL12	SRR6033230	GSM2782658	GCB-DLBCL	73331478	57769946	78.78
DLBCL13	SRR6033231	GSM2782659	GCB-DLBCL	30515523	25094935	82.24
DLBCL14	SRR6033232	GSM2782660	GCB-DLBCL	49357055	35099019	71.11
DLBCL15	SRR6033233	GSM2782661	GCB-DLBCL	77697531	58539509	75.34
DLBCL16	SRR6033234	GSM2782662	GCB-DLBCL	35639754	28928846	81.17
DLBCL17	SRR6033235	GSM2782663	GCB-DLBCL	39714656	31439509	79.16
DLBCL18	SRR6033236	GSM2782664	GCB-DLBCL	48996527	41076006	83.83
DLBCL19	SRR6033237	GSM2782665	GCB-DLBCL	44856565	34153244	76.14
DLBCL28	SRR6033246	GSM2782674	GCB-DLBCL	38789171	32080797	82.71
DLBCL29	SRR6033247	GSM2782675	GCB-DLBCL	33675412	27137265	80.58
BL30	SRR2149954	PRJNA292327	endemic BL	25130506	21839237	86.90
BL84	SRR2149952	PRJNA292327	endemic BL	50426015	41484253	82.27
BL81	SRR2149951	PRJNA292327	endemic BL	40113517	34667346	86.42
BL80	SRR2149950	PRJNA292327	endemic BL	50426015	41484253	82.27
BL62	SRR2149948	PRJNA292327	endemic BL	46828468	40844565	87.22
BL60	SRR2149947	PRJNA292327	endemic BL	42232435	36136846	85.57
BL50	SRR2149946	PRJNA292327	endemic BL	50070622	45594231	91.06
BL49	SRR2149945	PRJNA292327	endemic BL	32488327	29058864	89.44
BL45	SRR2149943	PRJNA292327	endemic BL	27181627	21996400	80.92
BL43	SRR2149942	PRJNA292327	endemic BL	35583467	31548435	88.66
BL40	SRR2149940	PRJNA292327	endemic BL	50697844	45567388	89.88
BL35	SRR2149938	PRJNA292327	endemic BL	52731358	40169925	76.18
BL27	SRR2149937	PRJNA292327	endemic BL	41949475	36930032	88.03
BL23	SRR2149936	PRJNA292327	endemic BL	36708639	31707933	86.38
BL22	SRR2149935	PRJNA292327	endemic BL	62743675	53280821	84.92
BL20	SRR2149897	PRJNA292327	endemic BL	45746442	40939102	89.49
BL19	SRR2149896	PRJNA292327	endemic BL	31988648	26564908	83.04
BL15	SRR2149844	PRJNA292327	endemic BL	39765484	34950090	87.89
BL69	SRR2149949	PRJNA292327	endemic BL	41045846	37456160	91.25

Appendix 4 (Chapter 3, section 3.3.1): Clinical characteristics of BL cases, including positivity to EBV, HIV, CMV, KSHV and HTLV-1 viruses.

Sample ID	Age	Sex	Site	EBV status	HIV status	CMV status	KSHV status	HTLV-1 status	Stage	Response to COM	Relapse
BL30	NR	NR	NR	NR	NR	Neg	Neg	Neg	NR	NR	NR
BL84	9	F	Abdomen	Pos	Neg	Neg	Neg	Neg	C	Complete Res	Yes
BL81	7	M	Abdomen	Pos	Neg	Neg	Neg	Neg	C	Complete Res	No
BL80	9	F	LN	Pos	Neg	Neg	Neg	Neg	C	Complete Res	No
BL62	9	M	Jaw	Pos	Neg	Neg	Neg	Neg	D	Complete Res	No
BL60	**7**	**M**	**Jaw**	**Pos**	**Neg**	**Pos**	**Neg**	**Neg**	**C**	**Not Res**	**NR**
BL50	8	F	Pelvic	Pos	Neg	Neg	Neg	Neg	D	Lost to Follow up	NR
BL49	7	M	Jaw	Pos	Neg	Neg	Pos	Neg	C	Lost to Follow up	NR
BL45	3	M	Jaw	Pos	Neg	Neg	Neg	Neg	B	Lost to Follow up	NR
BL43	5	M	Jaw	Pos	Neg	Neg	Pos	Neg	C	Lost to Follow up	NR
BL40	7	M	Jaw	Pos	Neg	Neg	Neg	Neg	C	Lost to Follow up	NR
BL35	7	M	Jaw	Pos	Neg	Neg	Neg	Neg	C	NR	NR
BL27	3	M	Abdomen	Pos	Neg	Neg	Pos	Neg	C	Lost to Follow up	NR
BL23	7	M	Jaw	Pos	Neg	Neg	Neg	Neg	A	Lost to Follow up	NR
BL22	**6**	**M**	**Jaw**	**Pos**	**Neg**	**Pos**	**Neg**	**Neg**	**C**	**Complete Res**	**No**
BL20	10	M	Jaw	Pos	Neg	Neg	Neg	Neg	B	NR	NR
BL19	3	M	Abdomen	Pos	Neg	Neg	Neg	Neg	C	Complete Res	No
BL15	4	M	Neck	NR	NR	Pos	Pos	Pos	C	Lost to Follow up	NR
BL69	4.5	M	Jaw	Pos	Neg	Pos	Neg	Neg	C	Complete Res	Yes

Appendix 5 (Chapter 3, section 3.2.2): RNAseq data generated from in-house BL and DLBCL cell lines. Results from the alignment of processed reads from each sample with the GRCh38 human reference genome cDNA index are also highlighted in this table.

Sample ID	condition	processed reads	aligned reads	% mapped reads
1_GLOR-A	endemic BL	24654695	17929733	72.72
2_GLOR_B	endemic BL	30108460	21951413	72.91
5_BL31-A	endemic BL	31937202	11003289	34.45
6_BL31-B	endemic BL	27333697	15349264	56.16
7_FARAGE-A	DLBCL	27789843	17479248	62.90
8_FARAGE-B	DLBCL	27640976	12137971	43.91
9_SAV-A	endemic BL	27221016	19659432	72.22
10_SAV-B	endemic BL	32245534	22314849	69.20
11_EZEMA-A	endemic BL	24086465	15745497	65.37
12_EZEMA-B	endemic BL	21894206	14371981	65.64
13_SUDHL4-A	DLBCL	30691199	21780609	70.97
14_SUDHL4-B	DLBCL	26730704	18823174	70.42
15_SUDHL5-A	DLBCL	28585297	19561390	68.43
16_SUDHL5-B	DLBCL	29363646	20217535	68.85
17_SUDHL6_A	DLBCL	29720828	16548266	55.68
18_SUDHL6_B	DLBCL	42923524	27822614	64.82

Appendix 6 (Chapter 3, section 3.2.5): Univariate analysis results of the NMR metabolomic data. Table presents the p-value, the false discovery rate (FDR) and the fold change value for each metabolite. P-values from the Shapiro-Wilk test are also highlighted to test for normality.

Name	p-value	q-value (FDR)	Fold Change	BL/DLBCL	Shapiro-Wilk p-value
UDP-GlcNAc	< 0.0001 (W)	<0.001	-1.05	Down	4.66E-08
Creatine	< 0.0001 (W)	<0.001	-1.08	Down	2.28E-06
Taurine	< 0.0001	<0.001	-1.04	Down	0.1054
L-Alanine	< 0.0001 (W)	<0.001	1.03	Up	8.10E-06
Phosphorylcholine	< 0.0001 (W)	<0.001	1.06	Up	1.59E-05
Glycine	< 0.0001 (W)	<0.001	1.02	Up	3.74E-07
Myoinositol	< 0.0001 (W)	<0.001	1.05	Up	1.04E-05
L-Glutamine	0.0005 (W)	0.0014	-1.05	Down	1.17E-07
D-Glutamic acid	0.0148 (W)	0.0347	1.02	Up	2.53E-05
Uridine diphosphate glucose	0.0563	0.1182	-1.01	Down	0.06514
L-Tyrosine	0.1055 (W)	0.2014	-1.01	Down	3.45E-06
L-Asparagine	0.1247 (W)	0.2183	-1.01	Down	1.02E-06
Fumaric acid	0.1708 (W)	0.2759	1.01	Up	1.53E-06
Nicotinuric acid	0.2067 (W)	0.3101	1.02	Up	0.009
Succinic acid	0.2300 (W)	0.322	1	Up	0.0015
L-Leucine	0.6325 (W)	0.7813	-1	Down	0.0001
L-Valine	0.6190 (W)	0.7813	1.03	Up	1.55E-09
Formic acid	0.6912 (W)	0.8064	-1.01	Down	0.0015
L-Isoleucine	0.7405 (W)	0.8184	1.01	Up	1.15E-06
Acetic acid	0.8214 (W)	0.8624	-1.01	Down	3.27E-06
L-Aspartic acid	0.9914 (W)	0.9914	1	Up	0.0013

Appendix 7 (Chapter 3, section 3.3.5): Table presents the 113 common significantly altered genes (q-value < 0.1) as identified from differential expression analyses in both primary tumours and cell lines.

		Primary tumours dataset			Cell lines dataset		
Gene ID	**Entrez ID**	**p-value**	**q-value**	**beta**	**p-value2**	**q-value2**	**beta2**
SCN4A	6329	0.00046	0.00146	0.97	0.000596	0.038575	1.96
E2F2	1870	2.28E-15	6.58E-14	1.68	4.57E-06	0.000726	1.10
CYB561	1534	0.00171	0.00465	-0.94	0.000597	0.038575	-1.97
GRAMD1B	100128242	4.88E-10	5.34E-09	-1.89	4.60E-05	0.005017	-3.78
RNASET2	8635	2.03E-06	1.11E-05	-1.02	2.30E-06	0.000457	-1.99
LZTS1	11178	1.97E-06	1.08E-05	2.34	1.69E-07	5.16E-05	4.70
CASP8	841	0.00338	0.00843	-0.41	0.002464	0.097228	-0.72
RAB27A	5873	9.62E-06	4.54E-05	-0.79	0.000638	0.04073	-1.97
STK10	6793	1.07E-05	5.01E-05	-0.62	3.46E-06	0.000628	-1.44
ACAP1	9744	0.00136	0.00379	-0.40	3.96E-05	0.004678	-0.95
STXBP2	6813	0.00035	0.00113	-0.43	0.001929	0.083595	-0.74
APBB1IP	54518	7.49E-08	5.48E-07	0.89	0.001984	0.084801	0.90
PPP2R5C	5527	0.00031	0.00101	0.47	0.000315	0.02505	0.79
TESC	54997	2.10E-06	1.15E-05	-1.75	0.002365	0.094	-3.76
JAK2	3717	8.44E-09	7.46E-08	-1.57	0.000439	0.031546	-2.61
CEP128	1E+05	3.26E-05	0.00014	0.88	0.000618	0.039616	1.33
RASSF2	9770	0.00192	0.00514	-0.39	0.000676	0.042321	-1.48

		Primary tumours dataset			Cell lines dataset		
Gene ID	**Entrez ID**	**p-value**	**q-value**	**beta**	**p-value2**	**q-value2**	**beta2**
SLCO5A1	81796	1.80E-05	7.97E-05	-2.18	1.80E-10	1.14E-07	-3.82
SQOR	58472	0.00103	0.00297	-1.16	0.002196	0.09008	-2.81
DUSP6	1848	0.001	0.0029	-1.07	0.001448	0.07126	-3.65
RHOF	54509	6.85E-10	7.34E-09	-1.51	0.00023	0.01951	-2.25
CMTM3	123920	0.00173	0.0047	-0.84	0.000713	0.04315	-2.83
SH3BGRL3	83442	0.00029	0.00097	-0.56	0.000786	0.04684	-1.24
MBOAT2	129642	1.37E-07	9.51E-07	1.36	3.66E-07	9.98E-05	2.84
AFF3	3899	0.00035	0.00113	1.02	3.51E-07	9.72E-05	3.92
ST14	6768	0.00017	0.00058	1.57	0.00086	0.04951	4.79
FADS1	3992	0.00094	0.00274	-1.01	2.80E-06	0.00052	-2.87
MTMR12	54545	0.00343	0.00854	0.53	0.00106	0.0582	0.85
TRIM36	55521	1.58E-05	7.09E-05	-1.40	0.0009	0.05143	-2.19
BMP3	651	2.13E-08	1.73E-07	3.47	0.001904	0.0832	5.12
RAD17	5884	4.49E-11	5.97E-10	1.03	8.03E-05	0.00785	3.14
MSI2	124540	0.00033	0.00108	0.63	6.66E-06	0.00101	1.07
TBRG1	84897	3.54E-09	3.34E-08	-0.73	0.001566	0.07422	-0.73
FAM167A	83648	1.65E-06	9.16E-06	2.25	2.33E-05	0.00303	4.38

		Primary tumours dataset			Cell lines dataset		
Gene ID	**Entrez ID**	**p-value**	**q-value**	**beta**	**p-value2**	**q-value2**	**beta2**
HCK	3055	3.59E-06	1.86E-05	-1.17	0.001216	0.063121	-3.90
E2F1	1869	3.42E-12	5.53E-11	0.90	0.002477	0.097416	0.60
PON2	5445	0.00012	0.00043	-0.93	3.80E-05	0.00456	-2.88
PLEKHA8	84725	0.00308	0.00777	0.46	0.000658	0.041628	0.86
AHR	196	0.00281	0.00716	-0.93	4.55E-16	7.71E-13	-4.52
TNFSF8	944	1.70E-14	4.18E-13	2.01	0.000949	0.053417	4.29
ZMIZ1	57178	0.00112	0.00321	-0.56	0.000498	0.034359	-2.30
MFSD10	10227	1.90E-06	1.04E-05	-0.58	1.56E-07	4.85E-05	-1.28
TCIRG1	10312	2.31E-17	9.35E-16	-1.27	0.001315	0.066315	-0.96
PTPN6	5777	4.00E-05	0.00016	-1.15	0.000793	0.046887	-2.00
AICDA	57379	2.86E-10	3.26E-09	2.97	2.65E-06	0.000513	3.51
BACH2	60468	1.27E-12	2.22E-11	2.60	4.40E-10	2.58E-07	4.62
ALDH5A1	7915	9.35E-09	8.18E-08	1.34	0.000273	0.022126	1.61
EPM2A	7957	1.86E-06	1.03E-05	0.72	0.002137	0.088368	0.82
GMDS	2762	0.00349	0.00866	-0.60	3.53E-05	0.004275	-1.25
TFDP2	7029	2.63E-14	6.19E-13	1.87	0.00213	0.088368	1.63
KDM5B	10765	3.53E-06	1.83E-05	-1.26	0.001941	0.08383	-3.62
ID3	3399	7.97E-08	5.79E-07	1.57	1.32E-06	0.000315	4.44
CXCR4	7852	1.70E-06	9.45E-06	0.90	0.001866	0.081802	2.00
BHLHE41	79365	6.36E-10	6.86E-09	-2.37	0.002333	0.093671	-2.07
BATF3	55509	3.17E-06	1.67E-05	-1.62	0.001489	0.072362	-3.36

		Primary tumours dataset			Cell lines dataset		
Gene ID	**Entrez ID**	**p-value**	**q-value**	**beta**	**p-value2**	**q-value2**	**beta2**
SH3RF1	57630	0.00404	0.00984	-1.08	2.32E-16	4.43E-13	-4.83
PLCL2	23228	2.01E-15	5.89E-14	1.52	8.76E-16	1.21E-12	1.95
BATF	10538	2.12E-08	1.72E-07	-1.93	0.00017	0.01521	-4.20
PRXL2C	195827	0.00013	0.00048	-1.05	7.49E-11	5.61E-08	-3.04
THEM4	117145	0.00128	0.00358	-0.71	0.000792	0.04689	-2.49
ZYX	7791	3.16E-05	0.00013	-1.05	0.000544	0.03657	-4.24
ZNF382	84911	2.11E-06	1.15E-05	-1.35	0.000812	0.0474	-1.91
ARHGAP25	9938	1.81E-17	7.41E-16	-1.25	0.001118	0.05965	-1.34
EOMES	8320	3.44E-07	2.20E-06	-1.55	5.22E-11	4.19E-08	-3.41
SMIM14	201895	0.00379	0.0093	1.45	0.001091	0.05925	3.78
WNK2	65268	6.99E-07	4.18E-06	2.64	1.41E-09	7.42E-07	5.04
RIMKLB	57494	1.13E-07	7.98E-07	1.41	9.85E-09	4.17E-06	4.48
TMC8	147138	3.30E-11	4.51E-10	-0.97	0.001115	0.05965	-0.77
VPREB1	7441	3.59E-06	1.86E-05	3.19	4.20E-13	4.27E-10	4.63
TAPT1	202018	0.00018	0.00062	0.63	2.69E-05	0.00342	1.11
TCEA2	6919	0.00017	0.00059	0.64	0.0017	0.07885	0.88
BCL2	596	1.27E-12	2.22E-11	-2.16	5.38E-07	0.00014	-2.95
ISG20	3669	1.30E-06	7.41E-06	1.03	7.27E-05	0.00727	2.13
NADSYN1	55191	1.03E-11	1.53E-10	-0.83	0.000846	0.04892	-0.79
SPTBN2	6712	0.00048	0.00151	1.50	0.002132	0.08837	2.93
FAM241A	132720	9.09E-11	1.14E-09	1.25	4.00E-05	0.00469	1.24

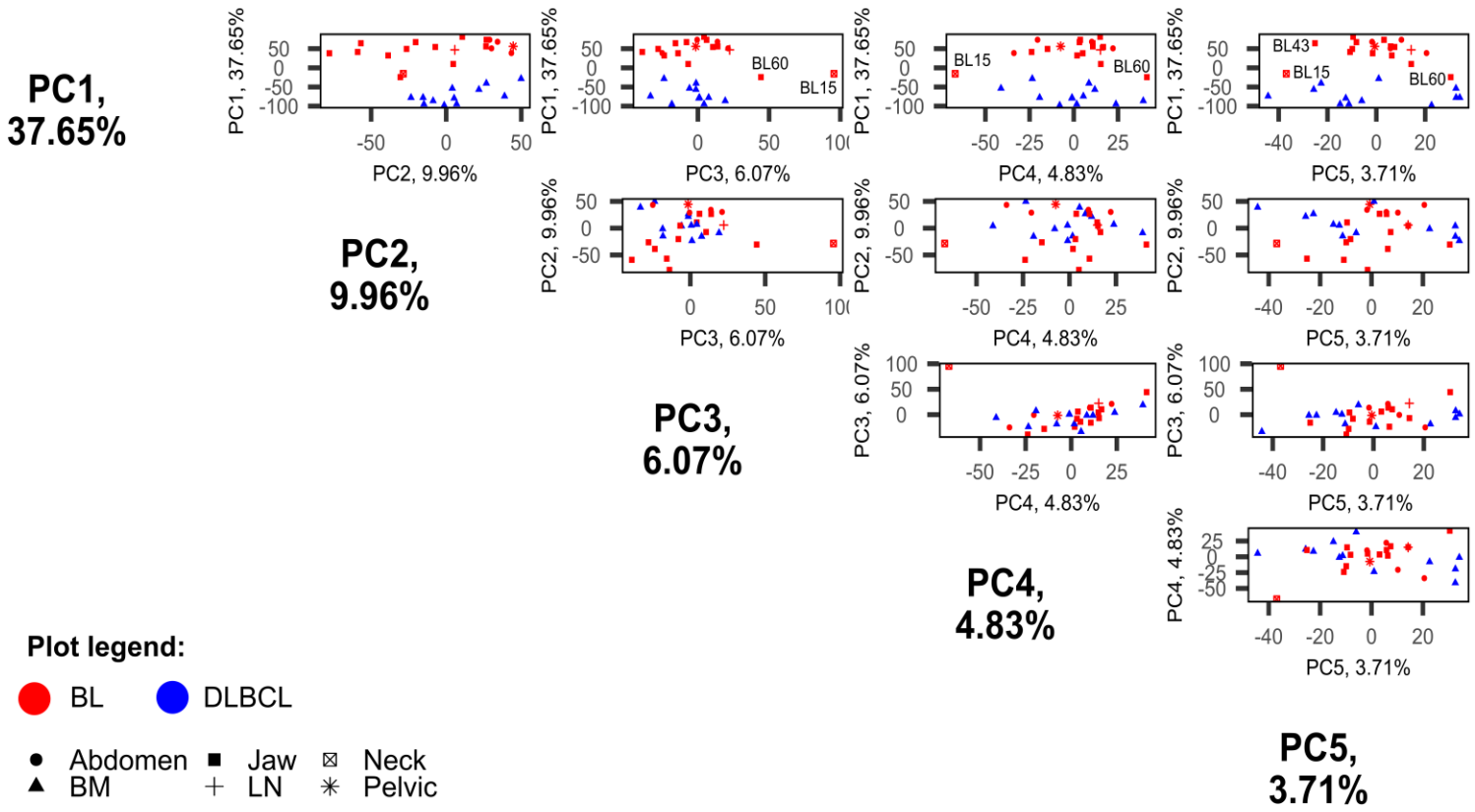
		Primary tumours dataset			Cell lines dataset		
Gene ID	**Entrez ID**	**p-value**	**q-value**	**beta**	**p-value2**	**q-value2**	**beta2**
GNAZ	2781	0.00013	0.00046	1.76	5.29E-05	0.005565	2.54
ACKR4	51554	0.00157	0.0043	1.36	8.65E-11	6.00E-08	4.36
SLC44A2	57153	0.00054	0.00169	0.54	0.001079	0.058768	1.12
SNX9	51429	0.00036	0.00117	-1.10	2.72E-05	0.003431	-3.07
ENAM	10117	4.36E-06	2.23E-05	1.64	0.000396	0.029153	3.20
JCHAIN	3512	8.37E-08	6.05E-07	3.04	5.91E-05	0.006048	5.16
LARGE1	9215	2.81E-06	1.49E-05	1.69	6.25E-05	0.006313	1.95
E2F5	1875	8.10E-07	4.78E-06	1.28	0.00014	0.012881	1.84
ARNTL	406	3.42E-06	1.78E-05	-1.38	8.30E-05	0.00801	-2.33
LPIN1	23175	1.50E-17	6.26E-16	-1.47	0.001509	0.073005	-0.99
FHOD3	80206	1.39E-10	1.68E-09	3.18	3.84E-09	1.67E-06	5.71
FADS2	9415	5.74E-05	0.00023	-1.39	4.67E-05	0.005058	-3.77
HRK	8739	2.99E-07	1.93E-06	2.32	0.00114	0.060395	2.56
MDFIC	29969	1.15E-06	6.62E-06	-1.69	1.24E-18	3.15E-15	-5.05
LMO2	4005	3.14E-11	4.31E-10	-2.57	2.65E-05	0.0034	-4.40
HEY2	23493	3.30E-11	4.51E-10	3.06	4.13E-05	0.004809	3.97
VILL	50853	1.16E-05	5.38E-05	-0.90	1.62E-09	7.98E-07	-3.67
RNF144B	3E+05	6.97E-05	0.00027	0.94	0.000686	0.042357	4.37

		Primary tumours dataset			Cell lines dataset		
Gene ID	**Entrez ID**	**p-value**	**q-value**	**beta**	**p-value2**	**q-value2**	**beta2**
PCCA	5095	2.72E-05	0.00012	-1.13	0.001709	0.07893	-2.03
RMI2	116028	1.55E-05	6.96E-05	1.02	0.001518	0.07306	0.85
ACBD7	414149	0.00011	0.0004	1.06	0.001582	0.0745	1.79
GNG7	2788	6.65E-13	1.22E-11	2.78	1.27E-15	1.61E-12	2.46
KCNA3	3738	0.00295	0.00748	0.62	3.46E-07	9.72E-05	4.94
MAGEF1	64110	3.14E-05	0.00013	0.69	8.92E-07	0.00023	1.15
TAF7	6879	5.08E-09	4.68E-08	0.70	0.002255	0.09105	0.86
UBE2E2	7325	3.12E-07	2.01E-06	0.98	6.19E-06	0.00095	1.30
LHFPL6	10186	2.19E-06	1.19E-05	1.05	0.00256	0.09965	2.38
CRELD2	79174	0.00125	0.00353	-0.41	1.74E-05	0.00237	-1.24
ROR1	4919	1.06E-06	6.13E-06	1.88	4.37E-08	1.59E-05	4.76
ZNF626	199777	0.00012	0.00044	0.87	0.001211	0.06312	2.16
NUGGC	389643	5.27E-05	0.00021	1.67	0.001945	0.08383	1.52
SIRPA	140885	4.56E-05	0.00018	-1.17	8.67E-20	4.41E-16	-6.69
RCSD1	92241	0.00089	0.00262	0.62	0.000352	0.02673	1.82
NAP1L4	4676	2.34E-09	2.29E-08	1.04	0.001909	0.0832	1.46
TUBB3	10381	7.55E-07	4.48E-06	2.76	8.21E-05	0.00798	3.99

Appendix 8 (Chapter 3, section 3.2.6): Pathway-based integration analysis with the 31 mapped KEGG metabolic pathways for the NHL datasets. Impact values represent the degree centrality from topology analysis.

Metabolic pathways	Total	Expected	Hits	Raw p	-log(p-value)	q-value	Impact
Alanine, aspartate and glutamate metabolism	61	0.50	4	0.0013	6.67	0.107	0.233
D-Glutamine and D-glutamate metabolism	10	0.08	2	0.0028	5.89	0.116	0.333
Glyoxylate and dicarboxylate metabolism	56	0.46	3	0.0099	4.62	0.277	0.109
Aminoacyl-tRNA biosynthesis	74	0.61	3	0.0211	3.86	0.442	0.041
Glycerophospholipid metabolism	86	0.70	3	0.0312	3.47	0.452	0.129
Glycerolipid metabolism	35	0.29	2	0.0323	3.43	0.452	0.147
Nitrogen metabolism	10	0.08	1	0.0792	2.54	0.950	0.111
Ascorbate and aldarate metabolism	13	0.11	1	0.1017	2.29	0.967	0.083
Glycine, serine and threonine metabolism	68	0.56	2	0.1054	2.25	0.967	0.284
Taurine and hypotaurine metabolism	16	0.13	1	0.1237	2.09	0.967	0.267
Amino sugar and nucleotide sugar metabolism	79	0.65	2	0.1352	2.00	0.967	0.051
Sulfur metabolism	18	0.15	1	0.1382	1.98	0.967	0.118
alpha-Linolenic acid metabolism	22	0.18	1	0.1663	1.79	1.000	0.095
Primary bile acid biosynthesis	92	0.75	2	0.1726	1.76	1.000	0.044
Arginine biosynthesis	27	0.22	1	0.2002	1.61	1.000	0.077
Butanoate metabolism	29	0.24	1	0.2134	1.54	1.000	0.071
Mannose type O-glycan biosynthesis	30	0.25	1	0.2200	1.51	1.000	0.138
Selenocompound metabolism	35	0.29	1	0.2518	1.38	1.000	0.029
Fructose and mannose metabolism	40	0.33	1	0.2824	1.26	1.000	0.051
Nicotinate and nicotinamide metabolism	42	0.34	1	0.2944	1.22	1.000	0.049
Propanoate metabolism	48	0.39	1	0.3290	1.11	1.000	0.043
Galactose metabolism	51	0.42	1	0.3457	1.06	1.000	0.040
Porphyrin and chlorophyll metabolism	53	0.43	1	0.3566	1.03	1.000	0.019

Metabolic pathways	Total	Expected	Hits	Raw p	-log(p-value)	q-value	Impact
Glutathione metabolism	56	0.46	1	0.3726	0.99	1.000	0.091
Inositol phosphate metabolism	69	0.57	1	0.4379	0.83	1.000	0.074
Phosphatidylinositol signaling system	74	0.61	1	0.4612	0.77	1.000	0.055
Fatty acid elongation	75	0.61	1	0.4658	0.76	1.000	0.014
Arginine and proline metabolism	78	0.64	1	0.4792	0.74	1.000	0.026
Valine, leucine and isoleucine degradation	88	0.72	1	0.5217	0.65	1.000	0.023
Pyrimidine metabolism	99	0.81	1	0.5647	0.57	1.000	0.010
Purine metabolism	166	1.36	1	0.7570	0.28	1.000	0.006



Appendix 9 (Chapter 3, section 3.3.1): Scatter plots of the first 5 principal components from PCA with the NHL primary tumours RNAseq data. The red colour circles representing the BL cases and the blue colour circles the DLBCL cases. The marker shapes represent the origin of isolated malignant B-cells: abdomen (circle), bone marrow (BM, triangle), jaw (square), lymph nodes (LN, cross), neck (square X) and pelvic (star).

Appendix 10 (Chapter 3, section 3.3.5): Top 50 (out of 180) KEGG metabolic and regulatory pathways with the NHL data integration analysis. Impact values represent the degree centrality from topology analysis.

Pathways	Total	Expected	Hits	Raw p	-log(pval.)	q value	Impact
Alanine, aspartate and glutamate metabolism	64	0.36	4	0.000427	7.76	0.107	0.311
Kaposi sarcoma-associated herpesvirus infection	191	1.06	6	0.000644	7.35	0.107	0.053
PD-L1 expression and PD-1 checkpoint pathway in cancer	93	0.52	4	0.001741	6.35	0.171	0.138
Hepatitis B	163	0.90	5	0.002065	6.18	0.171	0.058
ABC transporters	183	1.02	5	0.003408	5.68	0.226	0.000
D-Glutamine and D-glutamate metabolism	18	0.10	2	0.004383	5.43	0.233	0.188
Cell cycle	124	0.69	4	0.004922	5.31	0.233	0.091
Human cytomegalovirus infection	231	1.28	5	0.009025	4.71	0.341	0.079
Pathways in cancer	562	3.12	8	0.012084	4.42	0.341	0.074
Mineral absorption	87	0.48	3	0.012415	4.39	0.341	0.000
Apoptosis - multiple species	32	0.18	2	0.013517	4.30	0.341	0.152
Taurine and hypotaurine metabolism	33	0.18	2	0.014339	4.24	0.341	0.138
Circadian rhythm	33	0.18	2	0.014339	4.24	0.341	0.550
Glyoxylate and dicarboxylate metabolism	92	0.51	3	0.014425	4.24	0.341	0.080
Small cell lung cancer	95	0.53	3	0.015714	4.15	0.347	0.039
Central carbon metabolism in cancer	106	0.59	3	0.020984	3.86	0.351	0.000
Bladder cancer	41	0.23	2	0.021646	3.83	0.351	0.031
Prostate cancer	108	0.60	3	0.022034	3.82	0.351	0.032
Endocrine resistance	108	0.60	3	0.022034	3.82	0.351	0.057
Chemokine signaling pathway	194	1.08	4	0.022539	3.79	0.351	0.313
Sulfur metabolism	43	0.24	2	0.023669	3.74	0.351	0.070
Tuberculosis	197	1.09	4	0.023686	3.74	0.351	0.086
Toxoplasmosis	115	0.64	3	0.025933	3.65	0.351	0.098
Epstein-Barr virus infection	204	1.13	4	0.026499	3.63	0.351	0.048
Pathogenic Escherichia coli infection	204	1.13	4	0.026499	3.63	0.351	0.064
Aminoacyl-tRNA biosynthesis	118	0.66	3	0.027711	3.59	0.353	0.031
Cholinergic synapse	124	0.69	3	0.031457	3.46	0.371	0.180
Human immunodeficiency virus 1 infection	217	1.20	4	0.032239	3.43	0.371	0.104
Intestinal immune network for IgA production	51	0.28	2	0.03249	3.43	0.371	0.019
Human papillomavirus infection	333	1.85	5	0.037141	3.29	0.410	0.063

Pathways	Total	Expected	Hits	Raw p	-log(pval.)	q value	Impact
Apoptosis	140	0.78	3	0.042677	3.15	0.451	0.186
Protein digestion and absorption	142	0.79	3	0.044204	3.12	0.451	0.000
Dopaminergic synapse	143	0.79	3	0.044978	3.10	0.451	0.139
PI3K-Akt signaling pathway	358	1.99	5	0.048169	3.03	0.452	0.078
Primary bile acid biosynthesis	64	0.36	2	0.049079	3.01	0.452	0.042
Glycerophospholipid metabolism	149	0.83	3	0.04976	3.00	0.452	0.295
Breast cancer	150	0.83	3	0.050581	2.98	0.452	0.058
Gastric cancer	153	0.85	3	0.053082	2.94	0.462	0.019
Hepatitis C	157	0.87	3	0.056509	2.87	0.471	0.052
Non-small cell lung cancer	72	0.40	2	0.060509	2.81	0.471	0.018
p53 signaling pathway	72	0.40	2	0.060509	2.81	0.471	0.061
JAK-STAT signaling pathway	162	0.90	3	0.060938	2.80	0.471	0.306
Melanoma	73	0.41	2	0.061997	2.78	0.471	0.030
Cellular senescence	165	0.92	3	0.063672	2.75	0.471	0.026
Cushing syndrome	168	0.93	3	0.066463	2.71	0.471	0.026
Platinum drug resistance	76	0.42	2	0.066534	2.71	0.471	0.200
Chronic myeloid leukemia	77	0.43	2	0.068071	2.69	0.471	0.018
Pancreatic cancer	78	0.43	2	0.069619	2.66	0.471	0.018
Necroptosis	172	0.95	3	0.07027	2.66	0.471	0.085
Glioma	79	0.44	2	0.071179	2.64	0.471	0.027

Appendix 11 (Chapter 4, section 4.2.1): Names and origin of the CCLE cell lines.

NAME	ORIGIN	NAME	ORIGIN	NAME	ORIGIN	NAME	ORIGIN	NAME	ORIGIN	NAME	ORIGIN
22RV1	Others	HCC1187	Breast	KPNSI9S	Others	NCIH1792	Lung	P12ICHIKAWA	Leukemia	SNU503	Large intestine
697	Leukemia	HCC1195	Lung	KPNYN	Others	NCIH1838	Lung	P31FUJ	Leukemia	SNU520	Stomach
5637	Urinary tract	HCC1359	Lung	KU1919	Urinary tract	NCIH1869	Lung	PANC0203	Pancreas	SNU601	Stomach
2313287	Stomach	HCC1395	Breast	KU812	Leukemia	NCIH1915	Lung	PANC0504	Pancreas	SNU61	Large intestine
769P	Others	HCC1419	Breast	KURAMOCHI	Ovary	NCIH1930	Lung	PC14	Lung	SNU620	Stomach
786O	Others	HCC1428	Breast	KYSE140	Esophagus	NCIH1944	Lung	PEER	Leukemia	SNU626	CNS
A2780	Ovary	HCC1438	Lung	KYSE180	Esophagus	NCIH196	Lung	PF382	Leukemia	SNU668	Stomach
A3KAW	Lymphoma	HCC15	Lung	KYSE270	Esophagus	NCIH1963	Lung	PFEIFFER	Lymphoma	SNU685	Endometrium
A4FUK	Lymphoma	HCC1500	Breast	KYSE410	Esophagus	NCIH1975	Lung	PK1	Pancreas	SNU719	Stomach
ACCMESO1	Others	HCC1569	Breast	KYSE510	Esophagus	NCIH2030	Lung	PK45H	Pancreas	SNU738	CNS
ALLSIL	Leukemia	HCC1588	Lung	KYSE520	Esophagus	NCIH2052	Others	PK59	Pancreas	SNU761	Liver
ASPC1	Pancreas	HCC1599	Breast	KYSE70	Esophagus	NCIH2073	Lung	PSN1	Pancreas	SNU8	Ovary
AU565	Breast	HCC1806	Breast	L1236	Lymphoma	NCIH209	Lung	QGP1	Pancreas	SNU81	Large intestine
BCPAP	Others	HCC1833	Lung	L33	Pancreas	NCIH211	Lung	RAJI	Lymphoma	SNU840	Ovary
BDCM	Leukemia	HCC1937	Breast	L428	Lymphoma	NCIH2110	Lung	REC1	Lymphoma	SNU869	Others
BL41	Lymphoma	HCC1954	Breast	LAMA84	Leukemia	NCIH2122	Lung	REH	Leukemia	SNU878	Liver
BL70	Lymphoma	HCC202	Breast	LC1F	Lung	NCIH2170	Lung	RERFGC1B	Stomach	SNU886	Liver
BT549	Breast	HCC2108	Lung	LCLC103H	Lung	NCIH2172	Lung	RERFLCAD1	Lung	SNU899	UAT

NAME	ORIGIN	NAME	ORIGIN	NAME	ORIGIN	NAME	ORIGIN	NAME	ORIGIN	NAME	ORIGIN
BXPC3	Pancreas	HCC2218	Breast	LI7	Liver	NCIH2228	Lung	RERFLCAD2	Lung	SNUC1	Large intestine
CA46	Lymphoma	HCC2279	Lung	LK2	Lung	NCIH226	Lung	RERFLCKJ	Lung	SNUC4	Large intestine
CADOES1	Bone	HCC2935	Lung	LNCAPCLONEFGC	Others	NCIH2286	Lung	RERFLCSQ1	Lung	SNUC5	Large intestine
CCFSTTG1	CNS	HCC366	Lung	LOUCY	Leukemia	NCIH2291	Lung	RH30	Others	SQ1	Lung
CHAGOK1	Lung	HCC38	Breast	LOXIMVI	Skin	NCIH23	Lung	RH41	Others	ST486	Lymphoma
CMLT1	Leukemia	HCC4006	Lung	LS1034	Large intestine	NCIH2347	Lung	RI1	Lymphoma	SU8686	Pancreas
COLO201	Large intestine	HCC44	Lung	LS411N	Large intestine	NCIH2405	Lung	RPMI8226	multiple myeloma	SUDHL1	Leukemia
COLO320	Large intestine	HCC56	Large intestine	LS513	Large intestine	NCIH2444	Lung	RPMI8402	Leukemia	SUDHL5	Lymphoma
COLO668	Lung	HCC70	Breast	LU65	Lung	NCIH2452	Others	RS411	Leukemia	SUPT1	Leukemia
COLO678	Large intestine	HCC78	Lung	LU99	Lung	NCIH28	Others	RT112	Urinary tract	SUPT11	Leukemia
COLO679	Skin	HCC827	Lung	LUDLU1	Lung	NCIH292	Lung	RVH421	Skin	SW1088	CNS
COLO680N	Esophagus	HCC95	Lung	M07E	Leukemia	NCIH322	Lung	SET2	Leukemia	SW1353	Bone
COLO684	Endometrium	HCT15	Large intestine	MDAMB157	Breast	NCIH3255	Lung	SF295	CNS	SW1463	Large intestine
COLO741	Skin	HDLM2	Lymphoma	MDAMB175VII	Breast	NCIH358	Lung	SH10TC	Stomach	SW1783	CNS
COLO783	Skin	HDMYZ	Lymphoma	MDAMB231	Breast	NCIH441	Lung	SHP77	Lung	SW1990	Pancreas
COLO792	Skin	HEL	Leukemia	MDAMB435S	Skin	NCIH446	Lung	SIMA	Others	SW837	Large intestine
COLO800	Skin	HEL9217	Leukemia	MDAMB453	Breast	NCIH460	Lung	SJSA1	Bone	T47D	Breast
COLO829	Skin	HH	Leukemia	ME1	Leukemia	NCIH508	Large intestine	SKMEL30	Skin	TALL1	Leukemia
CORL105	Lung	HMC18	Breast	MEG01	Leukemia	NCIH520	Lung	SKMEL5	Skin	TCCPAN2	Pancreas
CORL23	Lung	HPBALL	Leukemia	MELHO	Skin	NCIH522	Lung	SNU1	Stomach	TE1	Esophagus

NAME	ORIGIN	NAME	ORIGIN	NAME	ORIGIN	NAME	ORIGIN	NAME	ORIGIN	NAME	ORIGIN
CORL24	Lung	HT	Lymphoma	MELJUSO	Skin	NCIH524	Lung	SNU1033	Large intestine	TE10	Esophagus
CORL279	Lung	HUCCT1	Others	MHHES1	Bone	NCIH526	Lung	SNU1040	Large intestine	TE11	Esophagus
CORL311	Lung	HUG1N	Stomach	MHHNB11	Others	NCIH596	Lung	SNU1041	UAT	TE14	Esophagus
CORL47	Lung	HUH28	Others	MKN1	Stomach	NCIH647	Lung	SNU1066	UAT	TE15	Esophagus
CORL88	Lung	IGROV1	Ovary	MKN45	Stomach	NCIH661	Lung	SNU1076	UAT	TE4	Esophagus
CORL95	Lung	IPC298	Skin	MKN7	Stomach	NCIH69	Lung	SNU1077	Endometrium	TE5	Esophagus
COV434	Ovary	JMSU1	Urinary tract	MKN74	Stomach	NCIH716	Large intestine	SNU1079	Others	TE6	Esophagus
CW2	Large intestine	JURKAT	Leukemia	MM1S	multiple myeloma	NCIH727	Others	SNU1105	CNS	TE8	Esophagus
DANG	Pancreas	JURLMK1	Leukemia	MOLT3	Leukemia	NCIH747	Large intestine	SNU119	Ovary	TE9	Esophagus
DAUDI	Lymphoma	JVM2	Lymphoma	MONOMAC1	Leukemia	NCIH82	Lung	SNU1196	Others	TF1	Leukemia
DB	Lymphoma	JVM3	Lymphoma	MONOMAC6	Leukemia	NCIH838	Lung	SNU1214	UAT	THP1	Leukemia
DBTRG05MG	CNS	K029AX	Skin	MORCPR	Lung	NCIH854	Lung	SNU1272	Others	TOLEDO	Lymphoma
DEL	Leukemia	K562	Leukemia	MSTO211H	Others	NCIH889	Lung	SNU16	Stomach	TUHR10TKB	Others
DKMG	CNS	KALS1	CNS	NALM6	Leukemia	NCIH929	multiple myeloma	SNU175	Large intestine	TUHR14TKB	Others
DMS79	Lung	KASUMI2	Leukemia	NAMALWA	Lymphoma	NCIN87	Stomach	SNU182	Liver	U937	Lymphoma
DND41	Leukemia	KCL22	Leukemia	NB1	Others	NCO2	Leukemia	SNU201	CNS	UACC257	Skin
DOHH2	Lymphoma	KE37	Leukemia	NB4	Leukemia	NOMO1	Leukemia	SNU213	Pancreas	UACC62	Skin
DU4475	Breast	KE39	Stomach	NCIH1048	Lung	NUGC3	Stomach	SNU216	Stomach	WM1799	Skin
DV90	Lung	KE97	multiple myeloma	NCIH1184	Lung	NUGC4	Stomach	SNU245	Others	WM793	Skin
EB1	Lymphoma	KELLY	Others	NCIH1299	Lung	OC314	Ovary	SNU283	Large intestine	WM88	Skin

NAME	ORIGIN	NAME	ORIGIN	NAME	ORIGIN	NAME	ORIGIN	NAME	ORIGIN	NAME	ORIGIN
ECC10	Stomach	KHM1B	multiple myeloma	NCIH1373	Lung	OE19	Esophagus	SNU308	Others	WM983B	Skin
ECC12	Stomach	KIJK	Leukemia	NCIH1385	Lung	OE21	Esophagus	SNU324	Pancreas	WSUDLCL2	Lymphoma
EFE184	Endometrium	KMH2	Lymphoma	NCIH1395	Lung	OE33	Esophagus	SNU349	Others	YAPC	Pancreas
EFM19	Breast	KMM1	multiple myeloma	NCIH1435	Lung	ONCODG1	Ovary	SNU387	Liver	YD10B	UAT
EHEB	Lymphoma	KMS20	multiple myeloma	NCIH1437	Lung	ONS76	CNS	SNU398	Liver	YD15	Others
EM2	Leukemia	KMS21BM	multiple myeloma	NCIH146	Lung	OPM2	multiple myeloma	SNU407	Large intestine	YD38	UAT
EOL1	Leukemia	KMS26	multiple myeloma	NCIH1563	Lung	OSRC2	Others	SNU410	Pancreas	YD8	UAT
GA10	Lymphoma	KMS27	multiple myeloma	NCIH1568	Lung	OVCAR4	Ovary	SNU423	Liver	ZR751	Breast
GDM1	Leukemia	KMS28BM	multiple myeloma	NCIH1623	Lung	OVCAR8	Ovary	SNU449	Liver	ZR7530	Breast
GSS	Stomach	KMS34	multiple myeloma	NCIH1650	Lung	OVISE	Ovary	SNU46	UAT		
GSU	Stomach	KOPN8	Leukemia	NCIH1703	Lung	OVKATE	Ovary	SNU466	CNS		
HARA	Lung	KP2	Pancreas	NCIH1734	Lung	OVMANA	Ovary	SNU475	Liver		
HCC1143	Breast	KP3	Pancreas	NCIH1755	Lung	OVSAHO	Ovary	SNU478	Others		
HCC1171	Lung	KPNRTBM1	Others	NCIH1781	Lung	OVTOKO	Ovary	SNU489	CNS		

Appendix 12 (Chapter 4, section 4.2.3.2): Feature selection in CCLE RNAseq dataset with the Elastic net regularization method.

#	Gene name	Elastic net score	#	Gene name	Elastic net score	#	Gene name	Elastic net score
1	OR4C7P	-41.42	31	IL21R	-0.05	61	CORO1B	0.01
2	OR8X1P	-1.46	32	SAMSN1	-0.05	62	SH3D19	0.01
3	AL161932.3	-0.65	33	ST8SIA4	-0.04	63	NCKAP1	0.01
4	AC022601.1	-0.48	34	CSTF3-DT	-0.04	64	MEF2C-AS1	-0.01
5	ABCB10P3	-0.13	35	NCKAP1L	-0.04	65	FERMT3	-0.01
6	IKZF1	-0.12	36	TRGV4	-0.04	66	DDAH1	0.01
7	AC004687.1	-0.11	37	ATP1B1	0.04	67	GNA11	0.01
8	WAS	-0.11	38	RCSD1	-0.04	68	LAIR1	-0.01
9	PLEKHG3	0.09	39	MAP4K1	-0.04	69	SEPTIN10	0.01
10	PIK3CG	-0.09	40	ARHGAP30	-0.04	70	LCP2	-0.01
11	LINC00528	-0.09	41	CHMP3	0.03	71	CD48	-0.01
12	CARD8-AS1	-0.09	42	SPN	-0.03	72	ASAP2	0.01
13	SH3BP4	0.08	43	CXorf21	-0.03	73	SELPLG	-0.01
14	KLHL6	-0.08	44	RHOH	-0.03	74	CKAP4	0.01
15	TSPAN6	0.08	45	JAK3	-0.03	75	PDE1B	-0.01
16	TJP1	0.08	46	PTPRF	0.03	76	ANKLE1	-0.01
17	FKBP9	0.07	47	CTNND1	0.03	77	LAMB2	0.01
18	PTPN22	-0.07	48	WASL	0.03	78	FAM78A	-0.01
19	RASAL3	-0.07	49	ARPIN-AP3S2	0.03	79	DAGLA	0
20	MANBAL	0.07	50	AMOTL2	0.03	80	AC008957.1	0
21	COL6A4P2	-0.06	51	AK1	0.02	81	ATP9A	0
22	RRN3P2	-0.06	52	DDR1	0.02			
23	PLEKHA1	0.06	53	ARPIN	0.02			
24	TSPOAP1-AS1	-0.06	54	PTPRC	-0.02			
25	GMFG	-0.06	55	CTBP2	0.02			
26	PARVA	0.06	56	CD276	0.02			
27	AGAP2-AS1	0.06	57	CSF2RB	-0.02			
28	PIM2	-0.06	58	ARHGAP15	-0.02			
29	CD53	-0.05	59	IL2RG	-0.01			
30	PTPN7	-0.05	60	DOCK2	-0.01			

Appendix 13 (Chapter 4, section 4.2.3.2): Feature selection in CCLE LC-MS dataset with rho values calculated with Spearman correlation.

#	Metabolite name	rho	p.value	FDR	#	Metabolite name	rho	p.value	FDR	#	Metabolite name	rho	p.value	FDR
1	xanthine	0.56	5.54E-36	1.25E-33	64	3-phosphoglycerate	0.38	5.33E-16	1.88E-15	127	heptanoylcarnitine	-0.23	1.88E-06	3.33E-06
2	creatinine	0.54	1.37E-33	1.03E-31	65	C54:7 TAG	-0.38	8.19E-16	2.83E-15	128	lactate	0.23	2.15E-06	3.78E-06
3	C20:4 CE	0.54	1.37E-33	1.03E-31	66	cytidine	-0.37	1.16E-15	3.95E-15	129	glucuronate	0.22	4.06E-06	7.08E-06
4	C20:5 CE	0.53	1.84E-32	1.04E-30	67	C40:6 PC	-0.37	1.18E-15	3.96E-15	130	C18:1 SM	-0.22	4.18E-06	7.18E-06
5	taurodeoxycholate	0.53	3.73E-32	1.68E-30	68	C56:4 TAG	-0.37	1.57E-15	5.20E-15	131	alanine	0.22	4.18E-06	7.18E-06
6	phosphocreatine	0.52	6.74E-31	2.53E-29	69	F1P/F6P/G1P/G6P	0.37	2.37E-15	7.73E-15	132	C46:1 TAG	-0.22	5.48E-06	9.34E-06
7	thiamine	0.52	9.47E-31	3.04E-29	70	C54:5 TAG	-0.37	3.63E-15	1.17E-14	133	C50:0 TAG	-0.22	6.25E-06	1.06E-05
8	carnosine	0.51	3.39E-30	9.55E-29	71	hippurate	0.36	6.81E-15	2.16E-14	134	niacinamide	0.22	6.43E-06	1.08E-05
9	C18:2 CE	0.51	1.64E-29	4.10E-28	72	C38:6 PC	-0.36	1.08E-14	3.38E-14	135	malonylcarnitine	0.21	1.08E-05	1.79E-05
10	oxalate	-0.51	1.87E-29	4.20E-28	73	urate	0.36	1.12E-14	3.46E-14	136	C16:0 SM	-0.21	1.45E-05	2.40E-05
11	C16:0 CE	0.51	2.34E-29	4.78E-28	74	C58:7 TAG	-0.36	1.74E-14	5.29E-14	137	C36:2 PC	-0.21	1.66E-05	2.73E-05
12	kynurenic acid	0.5	5.49E-29	1.03E-27	75	carnitine	0.36	1.77E-14	5.31E-14	138	butyrobetaine	0.21	1.82E-05	2.96E-05
13	leucine	0.5	1.70E-28	2.93E-27	76	acetylglycine	-0.35	7.47E-14	2.21E-13	139	C38:2 PC	-0.2	2.02E-05	3.27E-05
14	isoleucine	0.5	2.10E-28	3.37E-27	77	C22:1 SM	-0.35	8.29E-14	2.42E-13	140	C18:0 SM	-0.2	3.22E-05	5.17E-05
15	methionine	0.5	3.91E-28	5.86E-27	78	C54:6 TAG	-0.35	9.06E-14	2.61E-13	141	C36:2 DAG	-0.2	3.68E-05	5.87E-05
16	C16:1 CE	0.49	1.53E-27	2.15E-26	79	betaine	0.35	2.12E-13	6.05E-13	142	C52:2 TAG	-0.2	4.02E-05	6.36E-05
17	C18:1 CE	0.49	1.65E-27	2.18E-26	80	C32:0 PC	-0.34	2.51E-13	7.05E-13	143	C24:1 SM	-0.2	4.31E-05	6.78E-05
18	uracil	0.49	2.28E-27	2.85E-26	81	C18:0 CE	0.34	6.99E-13	1.94E-12	144	C20:4 LPC	0.2	4.57E-05	7.15E-05
19	C18:3 CE	0.49	7.97E-27	9.44E-26	82	cotinine	-0.34	9.58E-13	2.63E-12	145	alpha-ketoglutarate	0.2	4.61E-05	7.16E-05

#	Metabolite name	rho	p.value	FDR	#	Metabolite name	rho	p.value	FDR	#	Metabolite name	rho	p.value	FDR
20	pyroglutamic acid	0.49	1.17E-26	1.32E-25	83	C34:1 PC	-0.34	1.17E-12	3.17E-12	146	pipecolic acid	-0.19	6.51E-05	0.0001
21	tyrosine	0.49	1.25E-26	1.34E-25	84	C18:1 LPC	0.33	1.51E-12	4.05E-12	147	guanosine	-0.19	6.79E-05	0.0001
22	allantoin	0.48	5.11E-26	5.22E-25	85	DHAP/glyceraldehyde 3P	0.33	3.88E-12	1.03E-11	148	C18:2 LPC	-0.19	7.21E-05	0.00011
23	hexoses (HILIC neg)	0.48	8.71E-26	8.52E-25	86	succinate/methylmalonate	0.33	4.40E-12	1.15E-11	149	isocitrate	0.19	8.21E-05	0.00012
24	lysine	0.48	1.26E-25	1.18E-24	87	C36:4 PC-B	-0.32	6.54E-12	1.69E-11	150	glutathione reduced	0.19	9.63E-05	0.00014
25	phenylalanine	0.47	4.18E-25	3.76E-24	88	3-methyladipate/pimelate	-0.32	1.16E-11	2.97E-11	151	serine	0.19	0.0001	0.00016
26	methionine sulfoxide	0.47	5.65E-25	4.89E-24	89	citrulline	0.32	1.77E-11	4.48E-11	152	asparagine	0.19	0.00011	0.00016
27	C20:3 CE	0.46	2.81E-24	2.34E-23	90	C54:4 TAG	-0.32	2.48E-11	6.19E-11	153	erythrose-4-phosphate	0.18	0.00021	0.0003
28	UMP	-0.46	8.00E-24	6.43E-23	91	aconitate	0.31	3.33E-11	8.23E-11	154	C50:1 TAG	-0.18	0.00024	0.00035
29	thyroxine	0.45	1.76E-22	1.37E-21	92	GMP	-0.31	4.50E-11	1.10E-10	155	adenosine	0.18	0.00027	0.0004
30	dCMP	-0.45	2.19E-22	1.65E-21	93	C36:1 DAG	-0.31	7.36E-11	1.78E-10	156	choline	-0.17	0.00031	0.00045
31	trimethylamine-N-oxide	-0.45	2.32E-22	1.68E-21	94	C18:2 SM	-0.31	7.62E-11	1.82E-10	157	2-deoxyadenosine	0.17	0.00034	0.00049
32	NMMA	0.45	3.03E-22	2.13E-21	95	5-HIAA	-0.31	9.82E-11	2.33E-10	158	acetylcarnitine	0.17	0.00041	0.00058
33	ribose-5-P/ribulose5-P	0.44	5.16E-22	3.52E-21	96	5-adenosylhomocysteine	-0.3	1.64E-10	3.83E-10	159	aspartate	0.17	0.00042	0.00059
34	dimethylglycine	0.44	5.79E-22	3.83E-21	97	C46:0 TAG	-0.3	4.38E-10	1.02E-09	160	C32:2 PC	0.17	0.0005	0.00071
35	CMP	-0.44	9.34E-22	6.00E-21	98	C54:1 TAG	-0.29	7.01E-10	1.61E-09	161	C48:1 TAG	-0.17	0.0006	0.00084
36	AMP	-0.44	3.34E-21	2.09E-20	99	C22:6 LPC	0.29	8.33E-10	1.89E-09	162	C22:0 SM	-0.16	0.00078	0.00109
37	histidine	0.43	8.97E-21	5.45E-20	100	C56:2 TAG	-0.29	9.69E-10	2.18E-09	163	adipate	-0.16	0.00089	0.00123
38	SDMA/ADMA	0.43	9.45E-21	5.60E-20	101	hypoxanthine	-0.29	1.08E-09	2.41E-09	164	C48:0 TAG	-0.16	0.001	0.00137
39	C58:6 TAG	-0.43	1.05E-20	6.06E-20	102	ornithine	0.29	1.26E-09	2.78E-09	165	PEP	0.16	0.00106	0.00144
40	arachidonyl_carnitine	-0.43	2.24E-20	1.26E-19	103	C18:1 LPE	-0.28	2.34E-09	5.12E-09	166	C14:0 CE	0.16	0.00111	0.00151

41	sorbitol	0.43	2.65E-20	1.45E-19	104	glutamate	-0.28	3.28E-09	7.10E-09	167	C18:0 LPE	-0.15	0.00146	0.00197
42	C22:6 CE	0.42	6.32E-20	3.39E-19	105	cAMP	-0.28	5.71E-09	1.22E-08	168	C16:0 LPC	0.15	0.00177	0.00237
43	glycodeoxycholate	0.42	1.33E-19	6.98E-19	106	glutathione oxidized	0.28	5.92E-09	1.26E-08	169	C46:2 TAG	-0.14	0.00269	0.00359
44	anserine	0.42	2.13E-19	1.09E-18	107	C52:4 TAG	-0.28	6.49E-09	1.36E-08	170	C54:3 TAG	-0.14	0.0029	0.00384
45	C58:8 TAG	-0.41	3.81E-19	1.91E-18	108	C34:2 DAG	-0.27	1.06E-08	2.21E-08	171	taurocholate	0.14	0.0037	0.00487
46	alpha-glycerophosphate	-0.41	5.20E-19	2.54E-18	109	anthranilic acid	0.27	1.16E-08	2.39E-08	172	2-hydroxyglutarate	-0.14	0.00449	0.00588
47	C56:5 TAG	-0.41	5.97E-19	2.86E-18	110	C36:3 PC	-0.27	1.44E-08	2.95E-08	173	glycine	-0.14	0.00475	0.00618
48	arginine	0.41	9.52E-19	4.46E-18	111	adenine	-0.27	1.48E-08	3.01E-08	174	C34:4 PC	-0.14	0.0049	0.00633
49	alpha-hydroxybutyrate	-0.41	9.90E-19	4.55E-18	112	NADP	0.27	1.73E-08	3.48E-08	175	propionylcarnitine	0.13	0.00672	0.00864
50	lactose	0.41	1.23E-18	5.53E-18	113	inosine	-0.26	3.03E-08	6.04E-08	176	C34:1 DAG	-0.12	0.01018	0.01301
51	hexoses (HILIC pos)	0.41	1.69E-18	7.46E-18	114	C54:2 TAG	-0.26	4.23E-08	8.36E-08	177	homocysteine	0.12	0.0142	0.01805
52	xanthosine	0.4	2.79E-18	1.21E-17	115	C14:0 LPC	-0.25	1.02E-07	2.00E-07	178	thymine	-0.12	0.01443	0.01824
53	C56:6 TAG	-0.4	4.03E-18	1.71E-17	116	C56:3 TAG	-0.25	1.55E-07	3.01E-07	179	cystathionine	-0.12	0.01474	0.01853
54	C56:8 TAG	-0.4	9.64E-18	4.02E-17	117	C52:1 TAG	-0.25	1.75E-07	3.36E-07	180	sucrose	0.12	0.01496	0.0187
55	tryptophan	0.4	1.92E-17	7.86E-17	118	C16:0 LPE	-0.24	3.26E-07	6.21E-07	181	palmitoylcarnitine	-0.12	0.01559	0.01938
56	valine	0.39	2.45E-17	9.85E-17	119	citrate	0.24	3.31E-07	6.21E-07	182	C24:0 SM	-0.12	0.01576	0.01948
57	C56:7 TAG	-0.39	5.61E-17	2.21E-16	120	C16:1 SM	0.24	3.31E-07	6.21E-07	183	taurine	-0.11	0.02528	0.03108
58	creatine	0.39	7.67E-17	2.97E-16	121	C20:4 LPE	-0.24	3.53E-07	6.57E-07	184	hexanoylcarnitine	0.11	0.02806	0.03413
59	1-methylnicotinamide	0.39	9.64E-17	3.68E-16	122	C36:1 PC	-0.24	4.76E-07	8.78E-07	185	butyrylcarnitine	-0.11	0.02806	0.03413
60	C52:5 TAG	-0.39	1.03E-16	3.87E-16	123	4-pyridoxate	-0.24	5.08E-07	9.29E-07	186	beta-alanine	0.11	0.02955	0.03574
61	C36:4 PC-A	-0.39	1.26E-16	4.63E-16	124	C22:6 LPE	-0.24	5.13E-07	9.31E-07	187	malondialdehyde	-0.1	0.03601	0.04333
62	C38:4 PC	-0.38	3.87E-16	1.41E-15	125	glutamine	0.24	8.14E-07	1.46E-06	188	lauroylcarnitine	-0.1	0.03741	0.04477
63	C38:5 PC	-0.38	4.80E-16	1.71E-15	126	threonine	0.23	1.08E-06	1.92E-06					

Appendix 14 (Chapter 4, section 4.3.2): Lists of genes and metabolites involved in positive and negative correlations as identified with the sPLS-DA modelling.

	Negative correlations				Positive correlations			
No.	**Genes**	**Gene expr in HC**	**Metabolites**	**Met expr in HC**	**Genes**	**Gene expr in HC**	**Metabolites**	**Met expr in HC**
1	OR4C7	up	xanthine	down	TJP1	down	xanthine	down
2	AL161932.3	up	creatinine	down	FKBP9	down	creatinine	down
3	AC022601.1	up	C20:4CE	down	SH3BP4	down	betain	down
4	ABCB10P3	up	C16:0CE	down	TSPAN6	down		
5	MEF2C-AS1	up	betain	down	PTPRF	down		
6	CXorf21	up	C46:2TAG	up	PARVA	down		
7	SELPLG	up	C36:1PC	up	GNA11	down		
8	SAMSN1	up	C18:0CE	up	AMOTL2	down		
9	RRN3P2	up			LAMB2	down		
10	CARD8-AS1	up			ATP1B1	down		
11	AC008957.1	up			CTNND1	down		
12	LCP2	up			CHMP3	down		
13	FERMT3	up			AK1	down		
14	PTPN22	up			DDAH1	down		
15	TRGV4	up						
16	LAIR1	up						
17	IL2RG	up						
18	DOCK2	up						
19	RHOH	up						
20	GMFG	up						
21	CD53	up						
22	AC004687.1	up						
23	WAS	up						
24	IKZF1	up						
25	AMOTL2	down						

SCRIPTS

Table with the available scripts and repositories for each computational method presented in this thesis

Method	Interface	Approach	Chapter	GitHub repository
Sleuth in CLL	R	DEA	2	https://github.com/GrigoriosPapatzikas/Thesis_CLL.git
SetRank	R	GSEA	2	https://github.com/GrigoriosPapatzikas/Thesis_CLL.git
rMTA	MATLAB	GSMM	2	https://github.com/GrigoriosPapatzikas/Thesis_CLL.git
gMCSs	MATLAB	GSMM	2	https://github.com/GrigoriosPapatzikas/Thesis_CLL.git
Sleuth in NHL	R	DEA	3	https://github.com/GrigoriosPapatzikas/Thesis_NHL.git
PCA in NHL	R	ML	3	https://github.com/GrigoriosPapatzikas/Thesis_NHL.git
fgsea	R	GSEA	3	https://github.com/GrigoriosPapatzikas/Thesis_NHL.git
PCA in CCLE	Python	ML	4	https://github.com/GrigoriosPapatzikas/Thesis_CCLE.git
tSNE	Python	ML	4	https://github.com/GrigoriosPapatzikas/Thesis_CCLE.git
UMAP	Python	ML	4	https://github.com/GrigoriosPapatzikas/Thesis_CCLE.git
sPLS-DA	R	ML	4	https://github.com/GrigoriosPapatzikas/Thesis_CCLE.git