Imperial College London
Faculty of Medicine
School of Public Health
Department of Epidemiology and Biostatistics

Univariate and Multivariate Statistical Approaches for the Analyses of Omics Data: Sample Classification and Two-block Integration

Javiera Garrido Manríquez

A thesis presented for the degree of Doctor of Philosophy

July 2020

Declaration of Originality

I hereby declare that this thesis contains my own original work, unless explicitly acknowledged. Any contributions made by co-authors are detailed in the text, and any information derived from published or unpublished sources is appropriately referenced as such.

Copyright Declaration

Abstract

The wealth of information generated by high-throughput omics technologies in the context of large-scale epidemiological studies has made a significant contribution to the identification of factors influencing the onset and progression of common diseases. Advanced computational and statistical modelling techniques are required to manipulate and extract meaningful biological information from these omics data as several layers of complexity are associated with them. Recent research efforts have concentrated in the development of novel statistical and bioinformatic tools; however, studies thoroughly investigating the applicability and suitability of these novel methods in real data have often fallen behind.

This thesis focuses in the analyses of proteomics and transcriptomics data from the EnviroGenoMarker project with the purpose of addressing two main research objectives: i) to critically appraise established and recently developed statistical approaches in their ability to appropriately accommodate the inherently complex nature of real-world omics data and ii) to improve the current understanding of a prevalent condition by identifying biological markers predictive of disease as well as possible biological mechanisms leading to its onset. The specific disease endpoint of interest corresponds to B-cell Lymphoma, a common haematological malignancy for which many challenges related to its aetiology remain unanswered.

The seven chapters comprising this thesis are structured in the following manner: the first two correspond to introductory chapters where I describe the main omics technologies and statistical methods employed for their analyses. The third chapter provides a description of the epidemiological project giving rise to the study population and the disease outcome of interest. These are followed by three results chapters that address the research aims described above by applying univariate and multivariate statistical approaches for sample classification and data integration purposes. A summary of findings, concluding general remarks and discussion of open problems offering potential avenues for future research are presented in the final chapter.

This page intentionally left blank.

Acknowledgments

Gracias y lo demás sería lo de menos.

A la Verónica, la Margarita, la Sra. Irma mis hadas madrinas sin cuyos artilugios el machitún no creo que hubiera sido posible.

A los Ivanes, el Simón, la Josefa doctores todos por derecho propio.

Kawthar, Areti, Jelena, Faridah, Saredo los recuerdos se anulan unos a otros pero el rumor del río permanece.

Al Juanin, el flatmate más generoso de todos.

Creo tener bien en claro lo que este título significa.

Y para terminar por donde debí comenzar, gracias a CONICYT por una beca tan contundente como inmerecida.

Chao, buenas noches, etcétera, etcétera.

This page intentionally left blank.

"To get up in the morning, in the fullness of youth, and open a book –now that's what I call vicious!"

F. Nietzsche

This page intentionally left blank.

Contents

This page intentionally left blank.

List of Figures

This page intentionally left blank.

List of Tables

Contribution of Other Researchers

The following analyses described and employed in this thesis were conducted by other researchers:

Imputation of missing values in the proteomics dataset (chapter 3).

Pre-processing analytical steps in the transcriptomics dataset: within- and between-array normalization and imputation of missing values of probe intensity (chapter 3).

Pre-processing analytical steps in the DNA methylation dataset (chapter 3).

Imputation of missing values for covariate of interest (chapter 3).

Estimation of the proportions of the white blood cell subpopulations using the Houseman reference-based deconvolution algorithm (chapter 3).

Allocation of individual gene expression signals into biological pathways using the online bioinformatics tool Database for Annotation, Visualization and Integrated Discovery (DAVID) (chapter 5).

This page intentionally left blank.

Abbreviations

2CV	Double Cross-Validation
5MeC	5-methylcytosines
ABC-DLBCL	Activated B cell DLBCL
AUC	Area Under the Curve
AUROC	Area Under the Receiver Operating Characteristic
BCL	B-cell Lymphoma
BCR	B Cell Receptor
BER	Balanced Error Rate
BMI	Body Mass Index
bp	base pair
CCA	Canonical Correlation Analysis
CIM	Clustered Image Map
CLL	Chronic Lymphocytic Leukaemia
CLR	Conditional Logistic Regression
CMV	Cross-Model Validation
CNV	Copy Number Variation
ComBat	Combatting Batch Effects
CV	Cross-Validation
DC	Dendritic Cell
DLBCL	Diffuse Large B-cell Lymphoma
DNAm	DNA methylation
DQ^2	Discriminant Q^2
DR	Decision Rule
DRT	Dimension Reduction Technique
dsDNA	double stranded DNA
EBV	Epstein-Barr Virus
EDC	Euclidean Distance to Centroids

EDTA	Ethylene Diamine Tetraacetic Acid
EGM	EnviroGenoMarkers
ELISA	Enzyme-Linked Immunosorbent Assays
EPIC	European Prospective Investigation into Cancer and Nutrition
EPIC-Italy	European Prospective Investigation into Cancer and Nutrition - Italy
ER	Error Rate
FDR	False Discovery Rate
FL	Follicular Lymphoma
FN	False Negative
FP	False Positive
FWER	Family Wise Error Rate
GAM	Generalised Additive Model
GC	Germinal Centre
GCB-DLBCL	Germinal Centre B cell DLBCL
GLM	Generalised Linear Model
GLMM	Generalised Linear Mixed Model
GWAS	Genome-Wide Association Study/Studies
HL	Hodgkin's Lymphoma
ICA	Independent Component Analysis
ICD-O3	International Classification of Diseases for Oncology
Ig	Immunoglobulin
ISVA	Independent Surrogate Variable Analysis
LC-MS	Liquid Chromatography-Mass Spectrometry
LDA	Linear Discriminant Analysis
LM	Lagrange Multipliers
LMM	Linear Mixed Model
LOD	Limit of Detection

LOESS	LOcally WEighted Scatterplot Smoothing
LOOCV	Leave-one-out CV
LRT	Likelihood Ratio Test
MAF	Minor Allele Frequency
MALT	Mucosa-Associated Lymphoid Tissue
MBA	Multiplexed Bead Assay
MBL	Monoclonal B Lymphocytosis
MGUS	Monoclonal Gammopathy of Uncertain Significance
miRNA	micro RNA
ML	Maximum Likelihood
MLE	Maximum Likelihood Estimation
MM	Multiple Myeloma
mRNA	messenger RNA
MS	Mass Spectrometry
MSEP	Mean Squared Error of Prediction
MZL	Marginal Zone B-cell Lymphoma
ncRNA	non-coding RNA
NHL	Non-Hodgkin's Lymphoma
NK	Natural Killer
NMC	Number of Misclassifications
NMR	Nuclear Magnetic Resonance
NSHDS	Northern Sweden Health and Disease Study
OLS	Ordinary Least Squares
PBMC	Peripheral Blood Mononuclear Cells
PC	Principal Component
PCA	Principal Component Analysis
pFDR	positive False Discovery Rate
PLS	Partial Least Squares
PLS-DA	Partial Least Squares Discriminant Analysis

PLS-W2A	PLS in mode A
PLS-R	Partial Least Squares Regression
PLS-W2B	PLS in mode B
PRESS	Prediction Error Sum of Squares
QDA	Quadratic Discriminant Analysis
rCCA	regularized CCA
Rd	Redundancy
REAL	Revised European-American Lymphoma
REML	Restricted Maximum Likelihood
RMSEP	Root Mean Squared Error of Prediction
ROC	Receiver Operating Characteristic
rPLS	regularized Partial Least Squares
rPLS-DA	regularized Partial Least Squares Discriminant Analysis
RSS	Residual Sum of the Squares
SCotLASS	Simplified Component Technique-LASSO
SNP	Single Nucleotide Polymorphism
sPCA	sparse Principal Component Analysis
SPCA	Sparse Principal Component Analysis
sPCA-rSVD	sparse PCA via regularized SVD
SVA	Surrogate Variable Analysis
SVD	Singular Value Decomposition
TN	True Negative
TNR	True Negative Rate
TP	True Positive
TPR	True Positive Rate
TtD	Time to Diagnosis
ULR	Unconditional Logistic Regression
VIP	Variable Importance in Projection
WBC	White Blood Cell

Notation

Unless otherwise stated, I employ the common notation where column vectors are denoted by bold lowercase characters and row vectors shown as transposed (e.g. $\boldsymbol{v}$ and $\boldsymbol{v}^T$, respectively). Upper case letters are used to indicate random variables (e.g. V). Bold upper-case characters denote matrices (e.g. $\mathbf{X}$). The superscript T is used to indicate matrix or vector transpose.

More specifically, throughout the chapters of this thesis the matrices $\mathbf{X}$ and $\mathbf{Y}$ typically denote predictor and outcome omics data matrices, respectively. Their corresponding dimensions are $n \times p$ and $n \times q$ where p indicates the number of predictor variables and q the number of response variables and n the number of observations.

This page intentionally left blank.

1

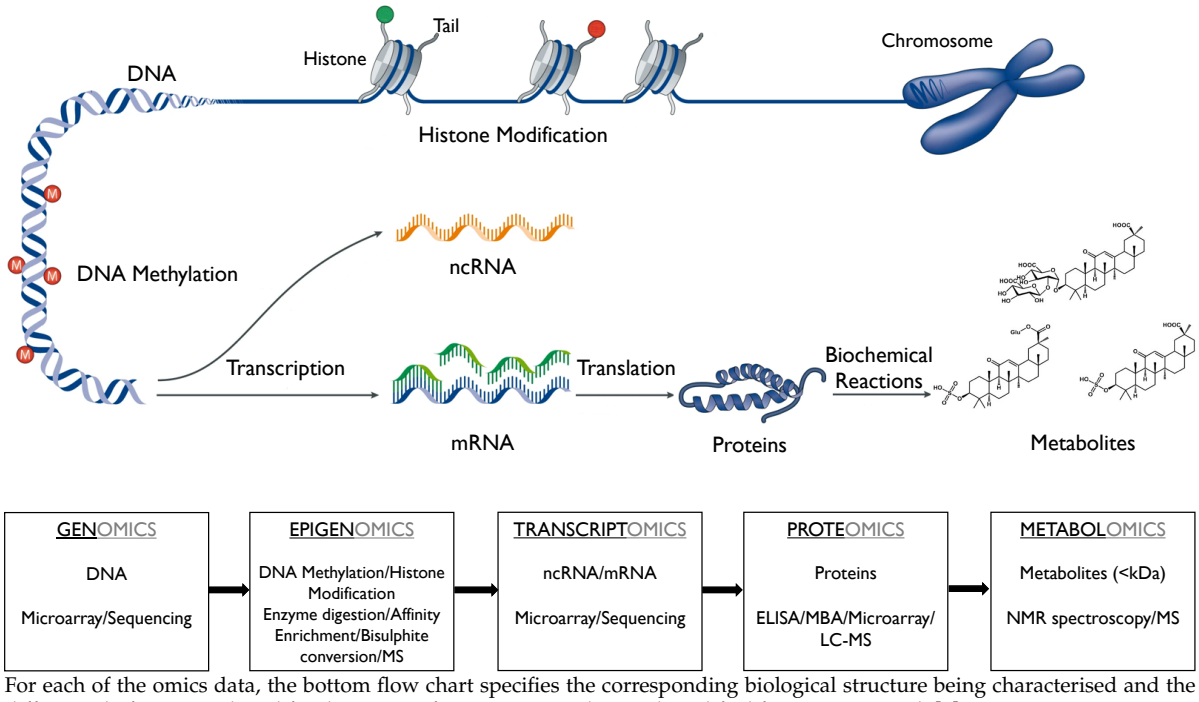
Omics Technologies and the Concept of the Exposome

As the main focus of this this thesis is centred around the application of established and novel statistical methods to data arisen from omics technologies, in this introductory chapter I provide a description of the mainstream platforms, namely genomics, epigenomics, transcriptomics, proteomics and metabolomics, as well as the different tools employed for their quantification. In addition, I elaborate on the exposome concept which is the overarching theoretical framework motivating the research objective related to the discovery of relevant biological patterns for the disease endpoint under study.

1.1 Omics Technologies

Broadly speaking, the term omics can be defined as the quantitative measurement of global sets of molecules in bio-samples using high-throughput techniques coupled with the use of advanced biostatistics and bioinformatics tools to characterise them. In light of recent advancements in this area of research, the term has been expanded to include a wide range of molecule types; however, here I concentrate in the more traditional technologies. An schematic summary of these omics data describing the supporting biological structure they aim to describe and the different platforms em-

Figure 1.1: Graphical and schematic overview of the main omics platforms.



For each of the omics data, the bottom flow chart specifies the corresponding biological structure being characterised and the different platforms employed for their quantification.Image taken and modified from Joosten et al. [1].
ncRNA: non-coding RNA, mRNA: messenger RNA, MS: Mass Spectrometry, ELISA: Enzyme-Linked Immunosorbent Assays, MBA: Multiplexed Bead Assay, LC-MS: Liquid Chromatography-Mass Spectrometry, NMR: Nuclear Magnetic Resonance.

ployed for their quantification is provided in Figure 1.1

1.1.1 Genomics

The genome can be broadly defined as the complete set of the genetic material of an organism and in this section I provide a brief summary on the human genome and the genetic variants that give rise to phenotypic variation. The genome in humans is constituted by 23 chromosome pairs of about three billion base pairs (bps) of DNA and is stored in the nucleus of the cell. Somatic cells have one copy of chromosomes 1 to 22 from each parent, in addition to an X chromosome from the maternal line and either an X or Y chromosome from the paternal line. The human genome also includes the mitochondrial DNA, a small circular molecule which is present in each mitochondrion and that is inherited only from the mother. A *gene* is a region of the DNA that encodes for a function and its specific location in a chromosome is known as *locus*.

The DNA sequence within a gene can be categorized as *exons* or *introns*, depending whether the section of the gene codes for proteins or not. It is estimated that the total number of genes is up to 50,000 [2],[3] while the number of protein-coding genes is about 19,000-20,000 [4]; the latter representing around 1.5% of the whole human genome. Much of the remaining genetic material has no known biological function and it is expected to be involved in transcriptional and translational regulatory functions [5].

There are variations in the human genome across individuals that explain the differences in phenotypic characters. The genomes of two non-related subjects are estimated to differ in around 20 million bps (or 0.6% of the total of three billion bps) [6]. These variations are called *polymorphisms* and the different forms they can take in a given population are known as *alleles*. Genetic variations occur through two main processes, namely genetic recombination and mutation. Recombination is a natural occurring phenomenon through which two homologous chromosomes exchange segments of DNA sequence (meiotic chromosomal crossover) resulting in the production of new genetic material; it is therefore the main source of genetic variation. On the other hand, mutation is the permanent alteration of the nucleotide sequence of the genome, which results from errors during DNA replication or repair processes. The majority of mutations do not produce discernible phenotypic changes and when phenotypic changes do occur, they are based on the accumulation of multiple mutations with small effects [7].

Categorization of genetics variants can be done in terms of the frequency they occur in a specific population. Genetics variants with a Minor Allele Frequency (MAF, frequency of the second most common allele) of 5% or more are known as common genetic variants. The most common among these are *Single Nucleotide Polymorphisms* (SNPs) which are mutations where a single nucleotide has been substituted by another, with a MAF$\geq$1% in at least one population. It is estimated that they occur at a rate of one every 100 to 300 bps, which translates to approximately 10 to 11 million SNPs in the entire human genome, of which seven million are known to have

a MAF>5% [8]. Most SNPs do not alter protein function as they are located in non-coding regions while the ones found within coding or regulatory regions may play a more direct role in phenotypic variation [9]. However, intragenic and synonymous SNPs (that is, SNPs located in exons which result in an allele that encodes for the same amino acid) may still influence the amount and function of gene products, for example by altering RNA splicing and structure, translation rate and protein folding [10].

Other type of genetic variations are *structural variations*, which are large-scale mutations less frequent than SNPs that alter a larger region of the genome (millions to hundreds of millions of bps). An example of structural variations are chromosomal rearrangements which includes deletions, duplications, inversions, substitutions and translocations. Another form of structural variations that has increasingly gained attention in genetic studies are *Copy Number Variations* (CNVs) in which sections of the genome are repeated (either duplicated or deleted) and the number of repeats in the genome varies between individuals; it has been described they play an important role in generating inter-individual variation as well as disease phenotype [11], [12].

Unlike macromutations that are discernible by optical microscopy, the identification of SNPs and CNVs relies on sequencing techniques or microarray technology. The latter is a more affordable tool and therefore more commonly used in large-scale epidemiological studies such as Genome-Wide Association Study/Studies (GWAS). By following an agnostic or hypothesis-free approach, GWAS scan hundreds of thousands of genetic variants across the entire genome of thousands of individuals in order to identify genetic regions that are associated with a trait of interest and thus better improving the understanding of the genetic basis of human diseases.

1.1.2 Epigenomics

The use of genetic material in cells is regulated in multiple ways with epigenetic modifications playing a key role as modulators of transcriptional activity. These biological processes are natural occurring phenomena that can be described as “mitotically

and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence" [13], [14]. The main epigenetics mechanism discovered at present are *DNA methylation* and *chromatin modification*, and I briefly describe them in the following sections.

1.1.2.1 DNA methylation

It involves modification of cytosine bases by covalent addition of a methyl group (CH_3) to carbon 5 of the cytosine ring forming 5-methylcytosines (5MeC) through the action of DNA methyltransferase enzymes. DNA methylation is a vital process for normal development of mammals because of the key role that plays in processes such as normal cell development, control of some tissue-specific gene expression, cellular growth and genomic stability. It is also believed to be a key factor driving carcinogenesis, as it has been shown that tumour suppressor genes are often silenced by aberrant methylations and kept off by epigenetic inheritance in that state [15].

Most of the genome is depleted of CpG sites as they occur with less than one quarter of the expected frequency [16]. However, there are regions where CpG sites cluster together (occur at the expected frequency) which are known as *CpG islands*; they typically remain unmethylated in normal cells, are present in roughly 40% of gene promoters and are associated with active gene expression [17]. CpG island "shores" (regions located approximately 2 kb from CpG islands with comparatively low CpG density) display tissue- and cancer-specific differential methylation patterns and are associated with gene repression [18]. Besides CpG islands and shores, the remainder of genome presents a lower than expected frequency of CpG sites and is typically methylated in normal cells [19].

Three main techniques exist to assess DNA methylation levels on a genome-wide scale based on restriction enzyme digestion, affinity enrichment and bisulphite conversion [20],[21]. In the first case, DNA is digested with methylation-insensitive restriction enzymes, to select genomic regions with moderate to high CpG content (e.g. CpG islands) which are later sequenced to generate a single-base pair resolution DNA

methylation map. Enrichment-based technologies are based on the recognition of 5MeC by either monoclonal antibodies (immunoprecipitation) or methyl-DNA binding protein domain which are later combined with array-based hybridisation or sequencing to determine the proportion of methylated DNA. Lastly, in bisulphite conversion methods DNA is treated with sodium bisulfite to convert cytosine to uracil, which is converted to thymine after an amplification step, whereas 5MeC residues are not converted and remain as cytosines. The proportion of cytosines (5MeC that were not affected by bisulphite conversion) provides an estimate of the DNA methylation levels.

1.1.2.2 Chromatin modification

Chromatin is the ordered structure in which the genetic material of a cell is organised. It is comprised by *nucleosomes* which are formed by an octameric histone protein complex (two of each core histones: H2A, H2B, H3, and H4) with 145-147 bps of DNA wound around it, bound to the outside by a linker histone protein (H1). Nucleosomes are separated by 40-200 bps and they form a characteristic "beads on a string" structure with their coating DNA [22].

It has been discovered that nucleosome position and organization play a key role in the degree of chromatin condensation (heterochromatin) and unfolding (euchromatin) which in turn affects the degree in which DNA is accessed to allow transcription [23]. Post-translational modifications of histones directly regulate the interactions between nucleosomes which activates or represses transcription (depending on the type of chemical modification and its location). Most of the modifications discovered so far have been described to occur on the N-terminal sequences of histones (also known as "tails", a section that project from the nucleosome and are accessible on its surface) or in the central globular domains (core of the nucleosome). For example, covalent addition of an acetyl group to the amino acid lysine of histone N-terminal "tails" reduces chromatin compaction and favours gene transcription while the opposite occurs when the amino acid is deacetylated [24],[25],[26]. In addition to the

"unravelling" effect on the chromatin, covalent modifications of histones also have an indirect effect on gene expression as they are involved in recruiting effector proteins (known as histone readers) by providing ligands for their specific domains. It has been observed that such changes play a role in activating downstream signalling, blocking the access of remodelling complexes, or affecting the recruitment of chromatin modifiers and transcription factors [27].

Characterization of this epigenetic changes in chromatin at a genome-wide scale is hindered by the dynamic location of the histones, in contrast to the fixed and known location of CpG loci. Despite this fact, there are strategies currently available to analyse post-translation modifications of histones which include antibody-based methods (such as western blotting, immunofluorescence analysis and chromatin immunoprecipitation) and Mass Spectrometry (MS)-based proteomics, which allows untargeted high-throughput analysis of histone modifications.

1.1.3 Transcriptomics

The transcriptome aims at describing the full information of all RNA transcribed from the genome in a specific tissue or cell, at a particular developmental stage and under a certain set of conditions, capturing a snapshot in time of the total transcripts present at that moment [28], [29]. The study of the transcriptome is essential not only for the understanding of the human genome at the transcription level but also provides a comprehension of gene structure and function, regulation of gene expression, genome plasticity and consequently development and disease [30]. The catalogue of RNA molecules is incredibly diverse: in addition to the protein-coding *messenger RNA* (mRNA), a myriad of *non-coding RNA* (ncRNA) have been described which play multiple structural and regulatory roles in the molecular biology of the cell [31].

As reported by transcriptomics studies, more than 93% of the human genome is transcribed into RNA where only 2% corresponds to protein-coding mRNA and the remaining percentage consists of ncRNA, mostly represented by ribosomal and transfer RNAs (rRNAs and tRNAs, respectively) [32]. ncRNAs are classified into housekeep-

ing and regulatory categories based on their functions; the former includes the ones with structural and catalytic roles such as rRNAs and tRNA for their function in protein translation and the latter includes micro RNA (miRNA) for their role modulating mRNA activity [30]. The diversity of RNA molecules is further complicated by alternative splicing events affecting both mRNA and ncRNA generating a wide range of transcripts isoforms, a process considered to play a decisive role in increasing cellular and functional diversity in the transcriptomes of higher eukaryotes [33].

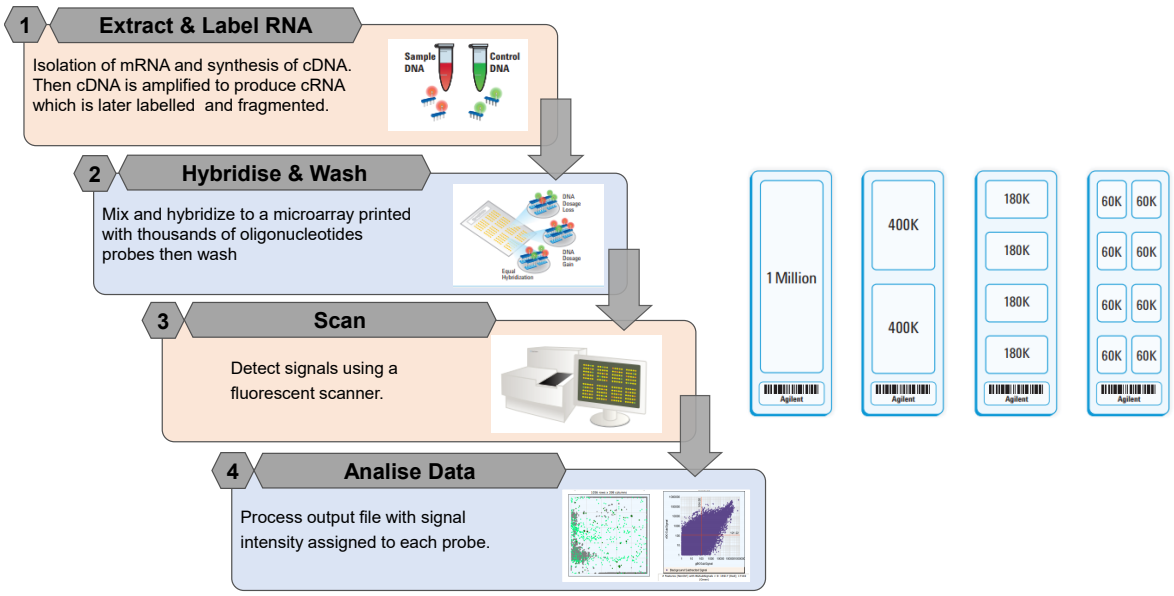
The first attempts to study the whole transcriptome began in the early 1990s on human brain tissue [34], [35] and from that point onwards transcriptomics research has been characterised by the development of new techniques which have redefined what is possible every decade and rendered previous technologies obsolete [36]. As for genotyping, there are two key contemporary techniques in the field allowing interrogation of transcripts at a genome-wide scale: *microarrays* and *RNA sequencing* (RNA-Seq). Microarray is a targeted technique that enables expression analysis of a priori set of genes based on hybridization of fluorescently labelled targets (complementary DNA (cDNA) chains derived from RNA molecules using reverse transcriptase) to probes that are attached to a solid surface. On the other hand, RNA-seq is an untargeted method that uses deep-sequencing technologies in combination with computational algorithms to allow for the reconstruction of the original full-length sequence of the RNA molecules present in the biological sample [37]. Unlike microarrays, where the fluorescence intensity at each probe location on the array indicates the transcript abundance, in RNA-seq abundance is directly derived from the number of counts from each transcript. RNA-seq technologies has recently superseded microarray techniques as the method of choice for transcriptome studies due to the many advantages the single base resolution allows: not limited to detecting transcripts that correspond to existing genomic sequence, it can reveal the precise location of transcription boundaries (i.e. identification of exon and intron boundaries) and it can detect SNPs and other gene variants within transcripts [38], [37]. In addition, it presents technical advantages such as low background signal, large dynamic

range of expression levels as there is not an upper limit for quantification, higher sensitivity and reproducibility and less requirement of RNA sample (nano vs microgram quantity) [36], [37], [37]. However, hybridization-based microarrays have been used for more than 15 years and are still extensively employed in the context of large-scale epidemiological studies as they provide a comparatively inexpensive way to detect and quantify transcripts at a genome-wide level and since they were the method of choice for the quantitative and qualitative characterization of the transcriptome in the EGM project, I provide a more elaborate description of this technology.

The structure of a microarray consists of oligonucleotide sequences (commonly known as probes) which are several dozen nucleotides long attached to the surface of a glass slide. Since probes are formed by attaching one nucleotide at a time, it is possible to construct a microarray with hundreds of thousands of different oligonucleotide sequences which are complementary to characteristic fragments of known RNA sequences [39]. During the experiment, samples containing RNA transcripts are spread on the surface of a microarray and its components hybridize specifically to their complementary probes, located in repeated copies across the glass slide. Although there is not a linear association, the fluorescence intensity reflects the amount of material hybridized to a given probe and therefore the abundance of a given RNA transcript in the sample [40].

Several companies commercialise microarray platforms (Affymetrix, Agilent, Illumina) and the models they provide differ in aspects such as length of probe sequences, number of probes per gene, section of the transcript to which the probe sequence has affinity (i.e. hybridization strength), assessment of nonspecific hybridization, number of genes to be assayed among other traits. These differences across platforms contribute to relatively low accuracy and reproducibility of microarrays and for this reason they are only used to identify potential differentially expressed genes across samples in the studied experimental conditions. Precise assessment of these genes needs to be further analysed using techniques such as quantitative reverse transcription PCR (qRT-PCR) considered the gold standard method for measur-

Figure 1.2: Main steps conducted in an RNA microarray experiment (left panel) and illustration of microarray chips or glass slides (right panel).



The microarray designs differ in the number of arrays per slide (1,2,4 or 8) and number of probes per array (1 million, 400k, 180k or 60k). Images taken and modified from www.agilent.com.

ing transcript levels [39]. Despite differences across platforms, the actual procedures conducted in a microarray experiment are very similar and consist of the following steps: RNA isolation, cDNA synthesis, amplification and labelling, complementary RNA (cRNA) fragmentation and hybridization, washing and scanning [41], [42]. A summary of the main steps involved in an RNA microarray experiment as well as different types of glass slides commonly employed in RNA microarray studies are illustrated in Figure 1.2.

In a first step, the quality and quantity of the sample is analysed after RNA is isolated from the cells. In a high-quality sample rRNA constitutes over 80% of the entire RNA and its concentration is a good indicator of the overall RNA quality, both before and after the experiment. However, the target of interest in most cases is mRNA which, unlike rRNA, is characterized by the presence of poly-A tails. The cDNA synthesis from mRNA is performed using oligo-dT (a primer with a short sequence of deoxy-thymine nucleotides that binds to the poly-A tails) or random primers. The re-

verse transcription process creates double stranded DNA (dsDNA) sequences which are later repeatedly replicated in a process of in vitro transcription to obtain a large quantity of cRNA (at least 100 times the original amount) containing labelled nucleotides. A different oligo-dT primer is used that serves as a promoter for the polymerase that creates cRNA labelled with either biotin or cyanine. cRNA molecules are fragmented to sequences of 50 to 100 nucleotides which are subsequently added to the microarray slide to initiate the hybridization process. During approximately 17 hours, microarrays are incubated in a hybridization oven set to 65°C in which cRNA binds to the specific probes attached to the glass surface. A washing step follows in order to remove cRNA non-specifically bound to the microarray surface during the hybridisation procedure. Finally, the microarray cartridge is placed in scans where the fluorescence of either phycoerythrin (complex bound to biotinylated C and U nucleotides) or cyanine is excited using a laser.

After all laboratory procedures have concluded, the final output of the microarray experiment is a digital file detailing the signal intensity of each probe to which data pre-process analyses are performed. These pre-processing steps include subtraction of background signal to reduce the effect of cross-hybridization, normalization to reduce the differences that originate from variations in experimental conditions and summarization in which a single expression estimate is calculated for each probe-set based on the intensity of the individual probe signals belonging to that set.

1.1.4 Proteomics

Proteomics is the study of the entire protein complement of a cell, tissue or organism under a specific set of conditions. The measurement and characterisation of protein levels, post-translational modifications and protein interactions provides a more direct measure of functional changes in comparison to the previous omics described and it is a well-established tool to assess inflammation, tissue damage, oxidative stress and signalling in epidemiological research.

Targeted proteomics techniques include *enzyme-linked immunosorbent assays* (ELISA)

usually used for the study of a single protein in biomedical research and clinical settings and *Multiplexed Bead Assay* (MBA) [43], [44]. Both are suitable for high throughput analyses as they allow to measure more than 50 proteins employing a small amount of biological material, making them useful in large-scale research laboratories. The two methods work under similar principles whereby an immobilized antibody is used to capture a soluble ligand (antigen) with subsequent detection of the captured ligand by a second detection antibody. A reporter molecule is then added to the mixture to obtain antigen quantification by measurement of its fluorescent signal. Unlike ELISA which relies on the use of flat surfaces, MBA is based on fluorescent-coded magnetic beads which are coated with different capture antibodies allowing the simultaneous detection of multiple proteins upon addition of the biological sample. Targeted proteomics can also be performed using protein microarrays [45],[46] whose main advantage lays on being able to characterise other variables in addition to quantity such as binding affinity, specificity, post-translational modifications and protein interactions. However, their use is not as widespread as DNA and RNA microarrays because of the difficulties associated with their elaboration, the unstable nature of proteins under different microenvironments and a reduced sensitivity due to cross-reactivity.

Untargeted proteomic analysis is typically performed using Liquid Chromatography-Mass Spectrometry (LC-MS) following a bottom-up method where proteins are first digested into peptides; they are later injected into the mass spectrometer and a putative list of proteins is constructed by comparing the spectra output against peptide databases [47]. Top-down proteomics has emerged as an alternative to the previous approach that forgoes the digestion step and allows the study of the different molecular forms in which the protein product of a single gene can be found, encompassing all forms of genetic variation, splice variants and post-translational modifications [48].

1.1.5 Metabolomics

The characterisation of the full suite of metabolites in a biological sample (cell, tissues, or fluids) in a physiological or developmental state is known as the metabolome [49]. Metabolites are low molecular weight chemicals (less than 2 kDa in size) that play critical roles in the correct functioning of the cell, acting as intermediate and final products of all cellular activities. The word metabonomics, although sometimes used interchangeable with the word metabolomics, places a particular emphasis on dynamics and sensitivity, and was coined to refer to the characterisation of the metabolome in response to a pathophysiological stimuli or genetic modifications (i.e. a particular disease or changes in diet or environment) [50].

It was recently estimated that the collective spectrum of chemicals in the human metabolome may include one million or more compounds, which present a wide-range of physicochemical properties [51]. Two complementary platform are currently used for the characterization of metabolic profiles, namely Nuclear Magnetic Resonance (NMR) spectroscopy and Mass Spectrometry (MS) [52]. NMR is used to detect compounds that contain atomic nuclei with an intrinsic angular momentum (atoms with an odd number of the sum of protons and neutrons); these atomic nuclei spin when an external magnetic field is applied. The most common atomic nuclei used in NMR are ^{1}H, ^{13}C and ^{15}N because they are present in nearly all naturally occurring compounds. MS is an analytical technique based on the principle of ions separation according to their mass-to-charge ratio. It has a higher sensitivity than NMR and it is usually the method of choice for the quantification of specific compounds.

These two high-resolution techniques allow the targeted and untargeted characterisation of metabolic profiles in a biological sample. Untargeted measurement can identify up to 20,000 chemical signals, covering endogenous metabolites, dietary chemicals, microbiome-derived metabolites, environmental chemicals, commercial products, and drugs [53]. These changes in the metabolic level reflect variations at any level of the biological system (from genotype to environmental exposures); thus, the

study of the metabolome becomes the most integrative profile for the representation of a biological state.

1.2 The Concept of the Exposome

The term *exposome* was proposed in an attempt to draw attention away from the study of the genetic basis of human diseases and redirect the efforts to the urgent need for a more complete environmental exposure assessment in epidemiological studies. It was first defined as "the life-course environmental exposures (including lifestyle factors) that affect an individual from the prenatal period onwards" [54]. The term was later broadened from this apparent focus on exposure assessment to a definition that encompasses not only the study of chemical exposures but also dietary and behavioural changes and other exogenous and endogenous agents, and the physiological alterations that are induced as a result of these environmental exposures. This new definition pays additional focus to the summation and integration of external forces and the way that they act upon the genome throughout the course of our lifetime.

Consequently, the exposome has been formally defined as: "The cumulative measure of environmental influences and associated biological responses throughout the lifespan, including exposures from the environment, diet, behaviour, and endogenous processes" [55]. Although the complex set of exposures to which an individual is subject is still at the heart of the definition, the lingering internal damage they produce is just as important as the chemicals themselves. These biological responses can be reflected as metabolic changes, protein modifications, alteration of expression of genes, epigenetic alterations, DNA mutations and adducts, and perturbations of the microbiome; they provide evidence of an actual internal effect which can be investigated through the use of high-throughput omics platforms [56], [57]. In fact, the term *exposomics* has been coined to refer to the characterisation of this internal biological effect to endogenous and exogenous exposures based on the use of omics technologies

Figure 1.3: Graphical representation of the exposome.

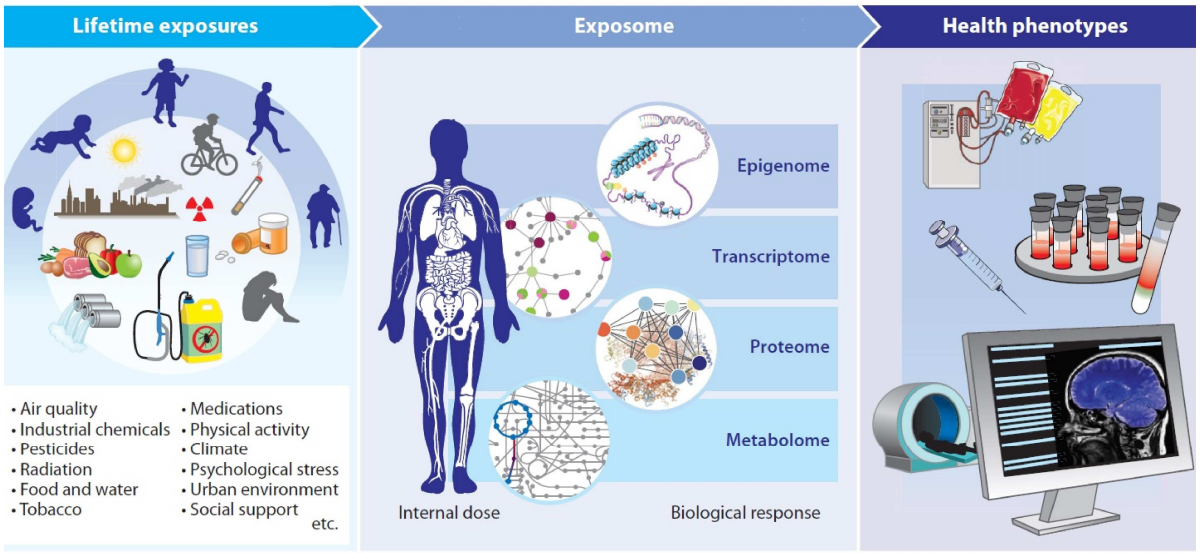


Image taken and modified from Niedzwiecki et al.[53].

[58], [59], [60].

Importantly, the concept of the exposome offers a theoretical leap in studying the role of the environment in human disease. Thus far, omics technologies have been mainly applied to the understanding of disease mechanisms and diagnosis. The study of the exposome invites to the application of omics technologies to investigate the relationship between human exposure and diseases that aims at the identification of alteration in omics profiles reflecting i) a direct response to exposures and ii) the characterisation of the downstream biological effects that are produced as a consequence to that initial response. In addition, it can provide information not only on acute biological responses that occur at a biologically relevant dose but also on the possible effect of long-term alterations in physiology from environmental stressors occurring years or decades before the clinical onset of a disease. Figure 1.3 exemplifies this interplay between lifetime external and internal exposures, omics technologies characterising internal dose and biological response and phenotypic alterations exerted as a result.

Moving away from the theorical concept to more concrete applications, the clearest example provided at present by the scientific literature has been the research link-

ing smoking exposure and tobacco-related cancers. Evidence from epidemiological studies strongly supports the finding that the effect of tobacco smoking on the risk of developing lung cancer is partially reversible, which has led to suggest that the mechanisms of carcinogenesis are likely to include epigenetic events (as genetic variations are fixed after cell replication) [61]. An exposome approach to the study of this exposure-disease association has been applied to test the plausibility of this hypothesis. On the one hand, epigenome-wide methylation studies have found that the methylation of certain CpG sites are strongly associated with long term exposure to smoking. On the other hand, mathematical models derived from observational study data and experimental work on tobacco-related cancers suggest that epigenetic alterations lead to the mechanisms driving cancer onset [62], [63], [64]. Although additional evidence is needed in order to confirm these findings, this example illustrates how alteration in omics profiles (changes in DNA methylation) can detect long term exogenous exposure (tobacco smoking) and how those altered biological responses induce the onset of the disease endpoint (tobacco-related cancers). A critical component of the exposome is therefore to recognise, understand and interpret the interplay between the exposure, the internal biological responses to that exposure and the final disease endpoints that manifest following those responses and omics platforms are instrumental to address that challenge. As exemplified in the smoking-cancer association above, the information gathered through omics technologies can indicate not only the connexion between an exposure and a disease but also provide insights into the mechanisms by which an exposure exerts its effects. Such mechanistic insights may contribute to the weight of evidence in assigning causality to a specific exposure-disease association and open avenues to prevention through modulation of the identified biological pathways [65].

2

Statistical Methods to Analyse Omics Data

Most omics data share a high-dimensional nature in which the number of variables is large and can exceed the number of observations. This problem is known as the "small n, large p situation" and parametric inference using classical statistical methods is either suboptimal or invalid [66]. In addition, variables in this high-dimensional space are often strongly correlated and these correlation patterns reflect complex interaction that characterise biological regulatory processes. For example, in the case of transcriptomics, microarray data present a correlation structure that is highly non-local and reflects expression patterns among sets of genes with similar or complementary functions. On the other hand, proteomics data present a correlation pattern that partially reflects that of the transcriptome (not all mRNA is translated into proteins) but that also depends on the complementary functions of the proteins being expressed. The presence of this *multicollinearity* between variables also imposes restrictions to classical methods of statistical inference. Lastly, omics platforms are characterised by the presence of relatively few features that are of biological relevance which are hidden either by uninformative variables or by technical-induced artefacts introduced during laboratory procedures. While the effect of *technical noise* can be attenuated during the pre-processing steps, statistical methods that are able to address this low signal-to-noise ratio are needed.

A wide range of statistical approaches have been proposed to tackle the three main challenges associated with omics data described above and can be broadly divided into *univariate and multivariate approaches*. Univariate methods separately assess the

association between each variable in the predictor matrix and the outcome of interest and the results are combined and analysed using multiple testing-correction strategies. In contrast, multivariate approach techniques consider all measurements simultaneously and the statistical procedure seeks to identify relevant variables that jointly explain the variation in the data. Multivariate methods have been traditionally classified into variable selection techniques which apply a penalization to the regression coefficients thus reducing the effect of irrelevant measurements and Dimension Reduction Techniques (DRTs) which introduce artificial variables (also known as latent variables or component scores) that summarise most of the information contained in the original data matrices facilitating the identification of the main features driving the variation. The construction of these latent variables can be accomplished under a supervised or unsupervised context, that is with or without respect the outcome of interest, respectively.

Recent developments in this area of research have combined the two approaches, variable selection and dimensionality reduction, in order to give rise to novel statistical techniques that allow the construction of parsimonious models that improve interpretability and ease the extraction of biological relevant information in this high-dimensional setting. To be more precise, penalization constraints are added in the creation of the artificial variables which enhances the detection of the main features driving the variation, a much-needed aim especially in situations where p far exceeds n. Such methodological combination has also been instrumental to make possible the incorporation of prior knowledge related to relevant group structures in blocks of omics data (i.e. subsets of correlated variables). Biological pathways within gene expression signals or inflammatory markers with similar biological functions in proteomics data are common examples of group structures within omics data. Thus, relevant information regarding functional groups while typically ignored in traditional univariate and multivariate approaches can now be incorporated into the construction of the model, further enhancing statistical performance and biological interpretation.

In the following sections of this chapter, I provide an overview of the main statistical

techniques under both univariate and multivariate frameworks. There are advantages and disadvantages to each of these statistical approaches which I discuss in the corresponding sections of the chapter.

2.1 Univariate approaches

2.1.1 Linear and Generalized Linear Models

Let us consider the two data matrices $\mathbf{X}$ and $\mathbf{Y}$ containing n observations of p predictor and q response variables, respectively. Univariate approaches typically accommodate a situation where $q = 1$ and the principle behind them is to independently assess the association between each omics measurement p and the outcome of interest q. The response is then considered to be a one-column matrix or a n-dimensional vector. The same statistical model is repeatedly applied p times, one model per predictor variable, in order to obtain values for a test statistic and the corresponding p-values. For a predictor variable j and for individual i, such statistical model can be formulated as follows:

$$\boldsymbol{y}_i = \beta_0 + \boldsymbol{\beta x}_{ij} + \boldsymbol{\varepsilon}_i,\ \boldsymbol{\varepsilon} \sim \mathcal{N}\left(0, \sigma^2\right) \tag{2.1}$$

where $\boldsymbol{y}_i$ is the measured outcome for individual i, β_0 is the intercept of the model, $\boldsymbol{\beta}$ is the regression coefficient, $\boldsymbol{x}_{ij}$ is the observed value for predictor j and individual i and $\boldsymbol{\varepsilon}_i$ is the residual random error. The previous equation can be equivalently expressed as follows:

$$\mathbb{E}\left(\boldsymbol{y}_i\right) = \beta_0 + \boldsymbol{\beta x}_{ij} \tag{2.2}$$

The type of statistical model depends on the nature of the response variable. A continuous $\boldsymbol{y}$ can be accommodated under a classical *linear regression model* while outcomes of a different nature can be dealt under a more general regression framework called

Generalised Linear Models (GLMs) [67], which link different types of responses to linear combinations of explanatory variables. There are three key components to any GLM:

1.- the probability distribution of the response variable: $f\left(\boldsymbol{y}_i|\ parameters\right)$,

2.- the linear combination of predictor variables: $\boldsymbol{\beta}_0 + \sum_{1=j}^{p} \boldsymbol{\beta}_j \boldsymbol{x}_{ij}$ and

3.- the link function n or $g(\mu)$ specifying how the expected value of the response relates to the linear predictor of explanatory variables: $g\left(\mathbb{E}\left(\boldsymbol{y}_i\right)\ |\boldsymbol{x}_{ij}\right)$.

The probability distribution from which the response variable is assumed to be generated belongs to the exponential family, a large class of probability distributions that includes the normal, binomial, Poisson and gamma distributions, among others. A summary of the most common type of response outcomes alongside their corresponding exponential-family distributions and link functions that are accommodated under the GLM framework is provided below.

Response Variable	**Probability Distribution (parameters)**	**Link function**
Continuous	Gaussian (u)	Identity (u)
Dichotomous	Bernoulli (p)	Logit ($\log\left(\frac{u}{1-u}\right)$)
Counts	Binomial (p)	
Multinomial	Multinomial (p)	
Counts	Poisson (λ)	Log ($\log(u)$)

Regression models that are of routine use in the analysis of omics data such as the logistic and Poisson regression models, proportional hazard model for time-dependent outcomes and beta regression for rate outcomes (used for epigenetics data, for example) are specific types of GLMs. The classical linear model is also a particular type of GLM in which the response is assumed to come from a normal distribution $\boldsymbol{y} \sim \mathcal{N}\left(\mu, \sigma^2\right)$ and the link function is the identity (no transformation on $g\left(\mathbb{E}\left(\boldsymbol{y}_i\right)\ |\boldsymbol{x}_{ij}\right)$ required).

2.1.2 Generalised Additive Models

Further extensions to linear and GLMs are the *Generalised Additive Models* (GAMs) [68], [69] in which the linearity assumptions are relaxed and the relationship between predictor(s) and response is decomposed as a mixture of smooth non-linear functions specific to each feature p. These models can be specified as follows:

$$\boldsymbol{y}_i = \boldsymbol{\beta}_0 + \sum_{j=1}^{p} f_j\left(\boldsymbol{x}_{ij}\right) + \boldsymbol{\varepsilon}_i \tag{2.3}$$

$$= \boldsymbol{\beta}_0 \; + \; f_1(\boldsymbol{x}_{i1}) \; + \; f_2(\boldsymbol{x}_{i2}) \; + \; ... \; + \; f_p(\boldsymbol{x}_{ip}) \; + \boldsymbol{\varepsilon}_i \tag{2.4}$$

They are called additive because a separate function is calculated for each $\boldsymbol{x}_j$ and their respective contributions are subsequently added. The choice of the parameter that optimally defines the flexibility of the fitting curve (for example, the effective number of degrees of freedom in a smoothing spline or the span of neighbouring points used in a local regression) usually relies on a Cross-Validation (CV) procedure.

2.1.3 Linear and Generalized Linear Mixed Models

In the GLM framework observations are assumed to be independent and to have equal variances. However, there are many applications where this independence assumption is not appropriate, for example in the case of grouped, multilevel and longitudinal data. *Linear Mixed Models* (LMMs) are extensions of classic linear models that include random effect terms in order to explicitly account for the similarity and dependency of observations within groups. The inclusion of a random effect term allows a variation in the mean level of the response variable (random intercept) and/or a variation in the association between the response variable and the predictors (random slopes) across groups. The term *"mixed"* refers to the combination of fixed and random effects in the model. When both random intercepts and random slopes are

incorporated, the model for observation j in group i can be specified as follows:

$$\boldsymbol{y}_{ij} = (\boldsymbol{\beta}_0 \; + \; \boldsymbol{u}_i) + (\boldsymbol{\beta}_1 \; + \; \boldsymbol{w}_i)\, \boldsymbol{x}_{ij} + \boldsymbol{\varepsilon}_{ij}, \tag{2.5}$$

where $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ are the fixed-effect coefficients as included in a linear model and are identical for all groups while $\boldsymbol{u}_i$ and $\boldsymbol{w}_i$ are the random-effect coefficients for group i and are assumed to be multivariate normally distributed with mean 0 and variance-covariance matrix $\left(\begin{smallmatrix} \sigma_u^2 & \sigma_{uw}^2 \\ \sigma_{uw}^2 & \sigma_w^2 \end{smallmatrix}\right)$. In contrast to $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ which are parameters to be estimated, $\boldsymbol{u}_i$ and $\boldsymbol{w}_i$ are random variables and the information to fit the multivariate normal distribution comes from all observations in the data. Therefore, the random effect terms vary by group and can be viewed as deviations around the value of $\boldsymbol{\beta}_s$. For the sake of simplicity, random effects are usually assumed to be independent, which effectively reduces the variance-covariance matrix to $\left(\begin{smallmatrix} \sigma_u^2 & 0 \\ 0 & \sigma_w^2 \end{smallmatrix}\right)$. Lastly, the term $\boldsymbol{\varepsilon}_{ij}$ is the random error for observation j in group i, assumed to be multivariate normally distributed. The specific structure of the variance-covariance matrix in this case depends upon context. For example, when observations are sampled independently within groups, the covariance between groups is assumed to be 0 and the only parameter to be estimated is the common error variance σ_e^2. Similarly, if there is dependency within groups, like in the case of longitudinal data, a specific covariance structure can be used in the model to capture the autocorrelation among the errors.

Generalised Linear Mixed Models (GLMMs) are a natural extension of mixed models to accommodate a link function as in Equation 2.2. Since in this case the corresponding marginal model does not have a closed form, their interpretation is not as straightforward as in LMMs: the response variable is modelled as a function of the predictor variables conditional on the attributes of each individual group. In other words, the group-specific effect of a predictor cannot be interpreted as the population average effect.

2.1.4 Model Fitting and Model Choice

2.1.4.1 Parameters Inference: Maximum Likelihood Estimation

For the models specified above, the form of the equations is fixed while parameters are unknown and must be estimated from the available data. In classical inference paradigm (as opposed to Bayesian framework), the most common method used is *Maximum Likelihood Estimation* (MLE), which produces estimators with desirable statistical properties. It is also a general principle that allows to estimate the uncertainty of each parameter (construct confidence intervals around the estimator). The basic intuition behind MLE is to find the parameters that make the observed data most probable, or in other words, to find the parameters such that plugging these estimates into the model yields fitted values as close as possible to the actual observed values of the response variable.

From a mathematical perspective, MLE seeks to maximize the likelihood function, which roughly speaking, can be described as the probability density function or probability mass function of the data seen as a function of the parameters. The log-likelihood is usually considered instead of the likelihood function because it is numerically more stable and because it is the more fundamental quantity of interest. More formally, the likelihood function is defined as:

$$L(\boldsymbol{\theta}|\boldsymbol{x}_1, ..., \boldsymbol{x}_n) = \prod_{i=1}^{n} f(\boldsymbol{x}_i|\boldsymbol{\theta}), \tag{2.6}$$

where $\boldsymbol{\theta}$ is a vector of model parameters, such as β above. The log-likelihood, which is by construction additive for independent observations, is defined as:

$$log\ L(\boldsymbol{\theta}|\boldsymbol{x}_1, ..., \boldsymbol{x}_n) = \sum_{i=1}^{n} log\ f(\boldsymbol{x}_i|\boldsymbol{\theta}) \tag{2.7}$$

The method of MLE finds the value of the parameter $\boldsymbol{\theta}$ that maximizes the log-likelihood function:

$$\hat{\boldsymbol{\theta}}_{MLE} = argmax\ l_n(\boldsymbol{\theta}) \tag{2.8}$$

The longer expression of the log-likelihood is replaced by $l_n(\boldsymbol{\theta})$ to simplify notation. The subscript n is a reminder that the samples are assumed to be independent and identical distributed (i.i.d.). Further mathematical details about MLE are beyond the scope of this thesis.

In the classical linear model, parameters are typically inferred using the method of Ordinary Least Squares (OLS) which minimizes the Residual Sum of the Squares (RSS) $\varepsilon^2 = \sum_{i=1}^{n} (\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i)^2$. As it turns out, the log-likelihood is also a function of the residuals and maximizing the log-likelihood for the parameter μ is identical to minimizing ε^2. Hence, both methods provide identical estimates and OLS is simply a special case of MLE assuming a normal distribution. For GLM, there is no explicit expression for the maximum likelihood estimator and the optimisation of the log-likelihood is done numerically. For LMM and GLMM, the additional variance parameters (σ_u^2, σ_w^2 and σ_{uw}^2, see above) are estimated using Restricted Maximum Likelihood (REML) which is an extension of ordinary MLE that imposes positivity constraints on the variance estimates (MLE can lead to negative estimates, especially when the variation between groups is small) and is the preferred method in models including random effect terms.

2.1.4.2 Hypothesis Testing

After the optimal values of the parameters have been inferred (model fitting), the step that follows is to decide which and how many predictors to include in the term of linear predictors (model choice). That research question can be formulated in statistical terms as assessing whether setting some parameters of interest to the corresponding value under the null hypothesis of no association, leads to a substantial harm of the model fit. That process is called *hypothesis testing* and involves testing the null hypothesis that there is no relationship between the predictors and the response variable ($H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$) versus the alternative hypothesis that there is a relationship ($H_a : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$), where $\boldsymbol{\theta}_0$ is a fixed value typically set to zero. When testing whether a single regression coefficient $\hat{\beta}_i$ is significantly different from zero, the sim-

plest manner to determine if the estimate is significantly far from zero is to compute the t-statistics which is given by:

$$t = \frac{\hat{\boldsymbol{\beta}}_i \ - \ 0}{SE(\hat{\boldsymbol{\beta}}_i)} \tag{2.9}$$

which measures the number of standard deviations that $\hat{\boldsymbol{\theta}}$ is away from the null value. The sampling distribution of the test statistic under the null hypothesis corresponds to a t-distribution with $n - 2$ degrees of freedom. The obtained value of the test statistics is compared against this sampling distribution to calculate a p-value and the null hypothesis is rejected if the p-value is smaller than a significance level, usually set to 5% or 1%.

2.1.4.3 The Wald, Likelihood Ratio and Lagrange Multipliers Tests

In addition to the t-test mentioned above, there are three classical approaches for testing hypotheses about parameters in a likelihood framework: the Wald test, the Likelihood Ratio Test (LRT) and the Lagrange Multipliers (LM) (or score) tests. The *Wald test* statistics takes the following form:

$$W = (\hat{\boldsymbol{\theta}} \ - \boldsymbol{\theta}_0)^T [Var(\boldsymbol{\theta})]^{-1} (\hat{\boldsymbol{\theta}} \ - \boldsymbol{\theta}_0) \tag{2.10}$$

which under the null hypothesis follows a chi-squared distribution with k degrees of freedom X_k^2 ,where k is the number of parameters constrained to take the null value. The Maximum Likelihood (ML) estimate $\hat{\boldsymbol{\theta}}$ and the value of the model parameter under the null hypothesis $\boldsymbol{\theta}_0$ are also referred to as unrestricted and restricted estimates, respectively. Intuitively, the test measures how far the estimated parameters $\hat{\boldsymbol{\theta}}$ are from the null value $\boldsymbol{\theta}_0$ in relation to the variance: the larger the difference between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$ or the smaller the variance in the distribution of the likelihood, the more likely we are to reject the null hypothesis and to prefer the more complex model instead. Such intuition becomes clearer when testing only one parameter (univariate

case) as the formula collapses to:

$$W = \frac{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^2}{Var(\hat{\boldsymbol{\theta}})} \tag{2.11}$$

which is compared against a X_1^2 distribution.

The *LRT* as indicated by its name, is a ratio comparing the maximum of the likelihood function between two different models, the models fitted with $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$:

$$\lambda = \frac{L(\boldsymbol{\theta}_0|\boldsymbol{y})}{L(\hat{\boldsymbol{\theta}}|\boldsymbol{y})} \tag{2.12}$$

For nested models (two of models of different lengths in which the smaller is a subset of the longer one) the value of λ ranges between 0 and 1, reflecting how many times more likely the data are under one model than the other. A value of λ close to 0 indicates that the smaller model is not acceptable compared to the larger model because it makes the data relatively improbable. Conversely, a value of λ close to 1 indicates that the larger model is not better than the smaller one. The LRT can also be specified as the difference in the log-likelihoods as follows:

$$-2\, ln\, (\lambda) = -2\, l(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) \; + \; 2\, l(\boldsymbol{\theta}_0|\boldsymbol{y}) \tag{2.13}$$

which is equivalent to calculate the difference in the deviance for the two models. The deviance is a measure of "distance" (similar to the RSS in the linear model) that compares the log-likelihood of the fitted model with the log-likelihood of the saturated model (i.e. the model where there are as many parameters as data points). LRT calculates the difference in deviance ignoring the term for the saturated model. As with the Wald test, the test statistic is distributed as a chi-squared random variable, with degrees of freedom equal to the difference in the number of parameters between the two models.

In contrast with the Wald test, which is based on unrestricted estimates, and the LRT, which requires both restricted and unrestricted estimates, the *LM* or *score test* requires

the fitting of the restricted model only. Its computation is based on the score or gradient of the likelihood function evaluated at the observed values of the parameters in the restricted model $(S(\boldsymbol{\theta}_0))$. The form of the score test in the univariate case is expressed as:

$$LM = \frac{S(\boldsymbol{\theta}_0)^2}{Var(\boldsymbol{\theta}_0)} \tag{2.14}$$

which is compared against a X_1^2 distribution. Note that none of the terms in the expression above involve the unconstrained maximum likelihood estimate of the parameter $\hat{\boldsymbol{\theta}}$. The test can be explained by the fact that the score function is exactly zero when evaluated at $\hat{\boldsymbol{\theta}}$ but not when evaluated at $\boldsymbol{\theta}_0$; the closer in value $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$ are, the closer $S(\boldsymbol{\theta}_0)$ will be to zero. Thus, larger values of $S(\boldsymbol{\theta}_0)$ will lead to large values of test and rejection of the null hypothesis.

As sample size approaches to infinity, the three tests provide equal results, a property known as asymptotically equivalent. However, for samples of a finite size, the three tests could disagree enough to lead to different conclusions. For linear models, it has been observed that the Wald test statistic will always be greater than or equal to the LRT statistic, which will in turn, always be greater than or equal to the test statistic from the score test (Wald test $\geq$ LRT $\geq$ score test) [70].

2.1.5 Multiple Testing and Correction Strategies

Performing hypothesis testing does not prove that a null hypothesis is either true or false, it can only provide indication of strength of evidence against it. Thus, the decision about whether or not to reject a null hypothesis is subject to errors which can be classified into type I and type II. A type I error occurs when one rejects the null hypothesis when it is true (false discovery or false positive finding) and a type II error occurs when one fails to reject the null hypothesis when it is false (false negative finding). The probabilities of committing type I and II errors are often denoted as α and β, respectively. The former is also the significance level or size of the test and the latter is related to the power of the test; that is, the probability of detecting a true

effect if it exists, given by $1 - \beta$. For any null hypothesis being tested, there is a trade-off between the probability of making a type I and type II error.

Usually in scientific research the interest lies in false positive findings and the probability of making a type I one error as the quantity reported in most studies is the p-value, which explicitly quantifies the probability that the researchers have made a type I error in reporting their conclusions. The type I error is even more important in the analysis of omics data under the univariate framework, as such analyses imply performing *multiple testing* (simultaneously testing a finite number of null hypotheses) leading to an increased number of associations falsely declared as statistically significant. To be more precise, assuming all p variables in the predictor matrix are pairwise independent, the expected number of false positive findings is given $p \times \alpha$. This situation is known as the multiple comparisons problem. The Family Wise Error Rate (FWER) and the False Discovery Rate (FDR) are overall error rates that have been proposed to characterise this increased number of false discoveries. Much of the scientific literature on *correction strategies* for multiple comparisons describes controlling one of these two error rates with different levels of stringency.

2.1.5.1 Definitions: Family Wise Error Rate and False Discovery Rate

There are four possible outcomes when testing multiple null hypotheses, which are commonly tabulated as follows:

	Do not reject	**Reject**	**Total**
$\boldsymbol{H_0}$ **is true**	U True Negative	V Type I Error	n_0
$\boldsymbol{H_a}$ **is true**	T Type II Error	S True Positive	$n - n_0$
Total	$n - R$	R	n

The total number of hypothesis being tested is represented by n and the total number of true null hypotheses by n_0, which is an unobserved quantity. V, S, T and U are

unobserved random variables while R is an observed random variable. The quantities of interest are V, the total number of false discoveries and R, the total number of associations declared statistically significant. The FWER is defined as the probability of making at least one type I error:

$$FWER = Pr\left(V \geq 1\right) = 1 - Pr\left(V = 0\right) \tag{2.15}$$

A more refined definition makes the distinction between the FWER in the strong and in the weak sense. The former is defined as above, regardless of the configuration of true and false null hypotheses or partial null hypotheses. The latter assumes that all null hypotheses being tested are true; that is known as under the complete or global null:

$$FWER_C = Pr\left(V \geq 1 \mid H_0^C \; true\right) \tag{2.16}$$

As an example, consider the scenario where we test 2 null hypotheses $n = 2$, H_0^1 and H_0^2. The possible partial null hypotheses are given by 2^n as follows: $H_0^{P_1} = [H_0^1, H_0^2], H_0^{P_2} = [H_0^1], H_0^{P_3} = [H_0^2], H_0^{P_4} = \emptyset$. Note that $H_0^{P_1}$ is also the complete null H_0^C. The weak sense definition of FWER is the probability of making at least one type I error only in configuration H_0^C. The strong sense definition of FWER is the probability of making at least one type I error in subsets $H_0^{P_1}$ to $H_0^{P_4}$. This obviously implies that correction strategies that control the FWER in the strong sense also ensure control of the FWER in the weak sense. The term family refers to the collection of hypotheses $H_0^1, ..., H_0^n$ that are being considered for joint testing, which tests are to be treated jointly as a family depends on the definition.

On the other hand, the FDR [71] can be described in its most intuitive definition as the expected proportion of errors among all associations declared significant:

$$FDR = E\left(\frac{V}{R}\right), \tag{2.17}$$

however, it cannot be applied when $Pr(R = 0) > 0$. The more extended definition

that accommodates situations where $R = 0$ is specified as:

$$FDR = E\left(V/R|R > 0\right) Pr\left(R > 0\right) + E\left(V/R|R = 0\right) Pr\left(R = 0\right) \tag{2.18}$$

$$= E\left(V/R|R > 0\right) Pr\left(R > 0\right) \tag{2.19}$$

The second term in Equation 2.18 equals zero allowing the reduction of the definition. Another probabilistic quantity of interest is the positive False Discovery Rate (pFDR) [72], [73], defined as the conditional FDR given that at least one hypothesis is rejected:

$$pFDR = E\left(\frac{V}{R} \mid R > 0\right) \tag{2.20}$$

Under the complete null hypothesis the FDR is equal to the FWER weak sense since $V = R$ gives $V/R = 1$ if $V \geq 1$, and replacing those terms in Equation 2.19 gives $FDR = 1 \times Pr\left(V \geq 1\right) = Pr\left(V \geq 1\right) = FWER_C$. In the situation where not all hypotheses are true (i.e. $V < R$), the FDR is less than the FWER, a statement that implies that FWER-controlling strategies also ensure control of the FDR.

In practice, the researcher choses one of these two probabilistic quantities to control and defines an arbitrary overall significance level α (usually 5%) which serves as an upper-bound limit. As an example, if 100 null hypotheses are tested, controlling the FDR at 5% is equivalent to accepting that on average, the 100 tests will result in fewer than five false discoveries, while controlling the FWER at the same level ensures less than five of these tests do not result in any false discoveries. The corrections strategies that can be applied to control either the FDR or FWER have been traditionally classified in *single-step procedures* where a single criterion is used to assess the significance of all test statistics or p-values and *step-wise approaches* involving ordering test statistics or p-values and then using a different criterion for each depending on their rank.

2.1.5.2 Control of Family Wise Error Rate

The most straightforward and popular method is the Bonferroni correction that rejects all $H_0^1, ..., H_0^n$ for which $\alpha' \leq \alpha/n$, controlling the FWER at level α in the strong sense. It is a conservative approach that does not rely on any assumption regarding the dependence structure between the tests. If tests are assumed independent, it is possible to obtain a tighter bound on the FWER by applying a different single-step correction known as Šidák, which sets the per-test significance level α' at $1 - (1 - \alpha)^{1/n}$. In practice, even for moderate values of n, the two corrections yield effectively the same results.

The stepwise approaches to control the FWER are the Holm's [74] and the Hochberg's [75] methods. Both procedures control de FWER in the strong sense and allow more power than the Bonferroni method. Hochberg's is more powerful than Holms' but requires independence or some form of positive dependence between tests. Both procedures use the same set of critical values and can be seen as step-down and step-up version of the Bonferroni test, respectively.

Holms' procedure operates as follows: the p-values are sorted in increasing order $p_{(1)} \leq p_{(2)} \ ... \ \leq p_{(n)}$ with their corresponding hypotheses being $H_0^1, ..., H_0^n$. In the first iteration $i = 1$, the lowest p-value $p_{(i)}$ is compared to $\alpha_i = \alpha/(n - i + 1)$ (the critical value α_1 is equivalent to the Bonferroni threshold α/n). If H_0^1 is rejected the algorithms continues, otherwise no significant findings are declared at a FWER α level and the algorithm stops. Subsequent p-values p_i are contrasted to the critical value $\alpha_i = \alpha/(n - i + 1)$. The procedure stops in the iteration where p_i exceeds the its corresponding cut-off value α_i and $H_0^1, ..., H_0^{i-1}$ are rejected. On the other hand, Hochberg's procedure scans backwards starting with the highest p-value p_n and stops as soon as a p-value p_i succeeds in passing its threshold and $H_0^1, ..., H_0^i$ are rejected.

2.1.5.3 Control of the False Discovery Rate

There are two step-up approaches to control the FDR, namely the Benjamini and Hochberg [71] and the Benjamini and Yekutieli [76] adjustments. The first method controls the FDR under the assumption of independent tests, the second is an adaptation of the original procedure to allow for positive dependence. The algorithm for the first method operates as follows: as before, the p-values are sorted in increasing order $p_{(1)} \leq p_{(2)} ... \leq p_{(n)}$ with their corresponding hypotheses being $H_0^1, ..., H_0^n$. Then, each $p_{(i)}$ is compared to the critical value $q\frac{i}{n}$, where q is the chosen threshold for the FDR. Finally, we define $k = max\left\{i : p_i \leq q\frac{i}{n}\right\}$ and reject $H_0^1, ..., H_0^k$ or do not reject any null hypothesis if no such i exists. Alternatively, the adjusted p-values can be obtained by means of $p_{(i)}^{adj} = p_{(i)}n/i$. However, after these adjustments p-values may no longer be strictly increasing, and reordering is needed to ensure monotonicity. In the Benjamini and Yekutieli procedure, the adaptation is simple to replace q with $\widetilde{q} = q/\sum_{i=1}^{n} i^{-1}$.

The computation of the FDR also leads to the estimation of the q-values [72] , a measure of statistical significance analogous to a p-value that is used to control the FDR rather than the FWER (strictly speaking adjusted p-values control FWER and p-values control the false positive rate). In contrast to the adjusted p-values obtained from the step-up procedures explained above, q-values are based on the pFDR and are calculated as a function of the p-value for each test and the empirical distribution of the entire set of p-values, which is used to estimate the number of true null hypothesis n_0. The q-value of a particular feature is the expected proportion of false discoveries when calling that feature significant. As an example, a q-value of 0.017 means that 1.7% of the variables that show q-values at least as small as that are false positives.

2.1.5.4 Resampling-based Approaches

The assumption about independence between tests seldom holds in the case of omics data as there is an underlying correlation structure between predictors that leads to dependence among hypotheses. Resampling is a general term that encompasses

methods such as permutation, bootstrap and parametric simulation-based analyses [77], [78] and they have become popular in multiple testing applications for omics data because of their ability to account for the unknown correlation structure without making assumptions on the distribution of the test statistic. The general principle behind resampling-based approaches is to take repeated samples from the observed data to simulate the distribution of the p-values under the complete null hypothesis; the per-test significance levels are then ascertained by comparing the observed p-value to the empirical distributions under the null.

More specifically, first step of the procedure is to generate a resampled dataset $\mathbf{Y}^*$ in which the values of the response variable are shuffled and randomly reassigned to observations either using bootstrap or permutation (sampling with and without replacement, respectively). Secondly, using the same method that produced the original p-values from the original data $\mathbf{Y}$, the resampled p-values p^* are obtained from $\mathbf{Y}^*$. Thirdly, the observed p-values of each predictor variable p_i are compared to the resampled ones to assess whether $p_i \geq min\, p^*$. This is because the minimum value of the empirical distribution under the complete null hypothesis corresponds to the maximum significance level that can be considered in order to prevent any false discovery across all tests performed. To put it in other words, any observed p-value that is above the minimum value of the empirical distribution corresponds to a false positive discovery. These three steps are repeated a large number of times B and the adjusted per-test significance levels $\widetilde{p}_i$ are given by the proportion of iterations where the observed p-value was higher than the minimum resampled p-value across the B iterations:

$$\widetilde{p}_i = \frac{1}{B} \sum_{b=1}^{B} I(p_i > min\, p^{*(b)}), \tag{2.21}$$

where I is an indicator variable that takes value of 1 if the condition is met and 0 otherwise [77], [78]. The procedure explained above can be written as a function of test statistics rather than p-values in which case the observed test statistics are contrasted to the maximum resampled values and the algorithm is modified accordingly.

2.2 Multivariate Approaches

The univariate techniques discussed in the previous section are commonly applied to the analysis of omics data as they offer a straightforward way of dealing with situations where $n << p$: each predictor is considered one at a time to assess the relationship with the response variable. Such approaches allow to fit computational efficient and flexible models; although it is worth mentioning that the more flexible the model is, the more computationally demanding it becomes. Examples of this are the use of GAMs to model non-linear relationships and LMMs with random effect terms to assess the effect technical-induced variation. However, exploring the marginal effect of individual variables only uncovers simple patterns in the relationship between $\mathbf{X}$ and $\mathbf{Y}$ and dismisses the joint effect that a subset of relevant variables may have to predict the outcome of interest. In addition, when one variable is independently considered in the analysis, one fails to identify correlation patterns among the predictors, leading to redundancy and unnecessary complexity in the results. Finally, univariate methods do not take into account the presence of functional groups within the data, which constitutes relevant information that eases statistical modelling and biological interpretation if incorporated into the analyses [79]. *Multivariate approaches* address these disadvantages and are discussed in the following sections.

2.2.1 Regularization and Variable Selection

These statistical methods identify the best subset of predictors by fitting a regression model containing all p variables and then estimating the coefficients under a constraint that shrinks them towards zero, a process called *regularization*. If the constraint forces some regression coefficients to be exactly zero, *variable selection* is also performed. The two best-known techniques for shrinking the coefficient estimates are ridge regression and the lasso; the latter method also being a variable selection technique. The name lasso is fact an acronym for Least Absolute Selection and Shrinkage Operator. Extensions of those approaches have been introduced in the literature that

improve over the disadvantages of ridge regression and lasso, which constitute valuable alternatives to deal with the challenges imposed by the high-dimensional omic setting. These techniques are the elastic net and the group and the sparse group lasso.

2.2.1.1 Ridge Regression and Lasso

For classical linear models the regularized regression coefficients are found through the following minimization problem:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2, + f_{pen}(\boldsymbol{\beta}, \lambda) \tag{2.22}$$

where the two terms correspond to the loss and penalty functions, respectively. The regularization or tuning parameter $\lambda \geq 0$ controls the strength of the penalty function. The term $\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$ is the RSS and can also be expressed as $\sum_{i=0}^{n}\left(\mathbf{Y}_i - (\mathbf{X}\boldsymbol{\beta}_i)^2\right) = \sum_{i=1}^{n}\left(\boldsymbol{y}_i - \boldsymbol{\beta}_0 - \sum_{j=1}^{p}\boldsymbol{\beta}_j\boldsymbol{x}_{ij}\right)^2$. Naturally, the penalty term is different for ridge regression and lasso. *Ridge regression* [80] uses the L_2 norm and the penalty term is defined as the sum of the squared values of the regression parameters, multiplied by the tuning parameter $f_{pen}(\boldsymbol{\beta}, \lambda) = \lambda\|\boldsymbol{\beta}\|_2^2 = \lambda\sum_{j=1}^{p}\boldsymbol{\beta}_j^2$. *Lasso* [81] uses the L_1 norm and the penalty is the sum of the absolute values of the regression parameters, multiplied by tuning parameter $f_{pen}(\boldsymbol{\beta}, \lambda) = \lambda\|\boldsymbol{\beta}\|_1 = \lambda\sum_{j=1}^{p}\left|\boldsymbol{\beta}_j\right|$. For the broader case of GLM, the loss function is replaced by the corresponding negative log-likelihood function.

The non-negative tuning parameter λ is directly related to the bias and variance trade-off. Small λ gives more weight to the loss function and the resulting model is characterised by presenting high variance but low bias (overfitting) [82]. As λ increases, the shrinkage of the coefficient estimates leads to a reduction in the variance of the predictions, at the expense of an increase in bias. In particular, when $\lambda = 0$, the penalty term has no effect and Equation 2.22 will produce the unbiased ML regression estimators; when $\lambda = \infty$, the loss function has no weight and the coefficient estimates will be equal to zero (intercept only-model with no variance). The aim is therefore to choose a value that estimates coefficients with a substantial decrease in

variance but that only introduces a small amount of bias in order to fit a model with a significant reduction in prediction error. (This is the reasoning that explains why ridge regression and lasso improve predictive accuracy over OLS). Selecting such an appropriate value of λ typically proceeds by CV.

It is worth mentioning that the intercept term is usually left unpenalized as it is simply a measure of the mean value of the response variable when $\boldsymbol{x}_{i1} = \boldsymbol{x}_{i2} = ... = \boldsymbol{x}_{ip} = 0$ [82]. For that reason, it is customary to centre the columns of $\mathbf{X}$ and $\mathbf{Y}$ to have mean zero before performing ridge regression or lasso, as in that case the estimated intercept is $\hat{\boldsymbol{\beta}}_0 = \bar{\boldsymbol{y}}$. In addition, unlike ML coefficient estimates that are scale equivariant, the regularized coefficients do depend on the scaling of the predictor; therefore, the columns of $\mathbf{X}$ are typically scaled to have sample variance one when performing the regression shrinkage.

Both ridge regression and lasso perform well in terms of prediction accuracy when there is a subset of true coefficients that are small or even zero. If all true coefficients are moderately large, both techniques can still outperform linear regression but only over a restricted range of λ values. However, neither of the techniques will universally provide lower prediction error than the other. In a situation where a relatively small number of predictors have substantial coefficients and the remaining variables have coefficients that are very small or that equal zero, lasso might perform better than ridge regression. In contrast, ridge regression is expected to do better when the response is a function of many predictors with coefficients of similar sizes. The main advantage of lasso resides in its capacity to perform variable selection and therefore to ease model interpretability. As the value of λ increases, more coefficients are set to zero (more sparsity is introduced), and among the non-zero coefficients, more shrinkage is employed [82].

2.2.1.2 Elastic Net

The lasso penalty has been associated with some limitations: first, for $p > n$, it will select at most n variables before the model reaches saturation. Second, in the presence

of correlated relevant predictors, lasso fails to perform grouped selection as it tends to select only one variable from the correlated predictors and ignore the others. These limitations are overcome by the convex combination of the L_2 and L_1 norms:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_2\|\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1 \tag{2.23}$$

which is known as the elastic net regularization [81]. The L_1 part of the penalty introduces sparsity in the model while the L_2 part removes the limitation on the number of selected variables, encourages the grouping effect and stabilizes the L_1 regularization path. Lasso and ridge regression are special cases of Equation 2.23 when $\lambda_2 = 0$ and $\lambda_1 = 0$, respectively. By introducing an additional parameter $\alpha = \lambda_2/(\lambda_2 + \lambda_1)$, the two-penalty term in Equation 2.23 can be combined into one:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\left[(1-\alpha)\|\boldsymbol{\beta}\|_2^2 + \alpha\|\boldsymbol{\beta}\|_1\right],\ \alpha \in [0,1] \tag{2.24}$$

When $\alpha = 0$, the elastic net reduces to ridge regression and when $\alpha = 1$ it becomes the lasso. Thus, elastic net compiles numerical stability and sparsity at the cost of an extra parameter to tune. Tuning of α relies on CV using a two-dimension grid of values, one dimension for each of the parameters to tune (α and λ); alternatively, it can be viewed as a higher-level parameter and defined on subjective grounds. The elastic net regularization has also been extended to accommodate other loss functions.

2.2.1.3 Group and Sparse Group Lasso

Other penalized regression strategies have been proposed by introducing further generalizations of the lasso penalty. One of those techniques is the group lasso [83] which offers a better control over the selected variables by allowing predefined groups of covariates to be jointly selected into or out of the model using the L_2 norm as the penalty term. It is important to clarify that in this case the constraint is applied group-wise and involves the sum of the ordinary L_2 norms $\sum\sqrt{\beta^2}$ as opposed to the squared L_2 norms $\sum\beta^2$ (as in ridge regression and elastic net). Let us suppose that the p predic-

tors in the $\mathbf{X}$ matrix are divided into L non-overlapping groups and that $\mathbf{X}_l$ represents the predictors corresponding to the l^{th} group with corresponding coefficient vector $\boldsymbol{\beta}_l$, the group lasso solves the following penalized least squares in a linear regression scenario:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\| \boldsymbol{y} - \sum_{l=1}^{L} \mathbf{X}_l \boldsymbol{\beta}_l \right\|_2^2 + \lambda \sum_{l=1}^{L} \|\boldsymbol{\beta}_l\|_2, \tag{2.25}$$

As before, λ is non-negative and depending on its value either the entire vector $\boldsymbol{\beta}_l$ is zero or all its elements are non-zero. If all the groups are of size one, the optimization problem reduces to the ordinary lasso since $\|\boldsymbol{\beta}_l\|_2 = \sqrt{\boldsymbol{\beta}_l^2} = |\boldsymbol{\beta}_l|$. It is worth mentioning that the in original formulation of the group lasso a weighting factor $\sqrt{p_l}$ was introduced in the penalty, representing the number of predictors p in group l, which in this case has been omitted for notational simplicity (if there is no factor all groups are equally penalized and larger groups are more likely to be selected). Furthermore, the computation of the original algorithm relied on the assumption of orthonormality of the predictors within a group ($\mathbf{X}_l^T \mathbf{X}_l = \mathbf{I}$), however, alternative approaches were subsequently proposed that deemed such assumption unnecessary. The group lasso has been extended to the logistic regression setting [83] as well as a multivariate outcome [84].

Because of the L_2 norm, the group lasso does not impose sparsity within the selected groups. That is, when a group of predictors is included in the model, all the coefficients in that group are non-zero. The sparse group lasso [85], [86], [87] has been proposed as a penalty that yields sparse solutions at both the group and individual feature levels, and its criterion is defined by augmenting the group lasso problem with an additional L_1 penalty as follows:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\| \boldsymbol{y} - \sum_{l=1}^{L} \mathbf{X}_l \boldsymbol{\beta}_l \right\|_2^2 + \lambda_2 \sum_{l=1}^{L} \|\boldsymbol{\beta}_l\|_2 + \lambda_1 \|\boldsymbol{\beta}\|_1 \tag{2.26}$$

Much like the elastic net scenario, the sparse group creates a compromise between the lasso and group lasso penalties: when $\lambda_1 = 0$ the criterion reduces to the group lasso in Equation 2.25. The two-term penalty can also be expressed as function of the

hyper-parameter α as follows: $\lambda\left[(1-\alpha)\|\boldsymbol{\beta}_l\|_2^2 + \alpha\|\boldsymbol{\beta}\|_1\right]$. Setting $\alpha = 0$ produces the group lasso fit whereas $\alpha = 1$ yields the lasso solution. The factor $\sqrt{p_l}$ has again been removed for notational simplicity. The sparse-group lasso approach has been extended to accommodate a multivariate response [88].

Both group and sparse group lasso are convex optimization problems and their optima are specified by zero sub-gradient equations which are solved using a minimization tool called coordinate descent [84]. More specifically, the equations are solved separately for each group of variables $\boldsymbol{\beta}_l = (\boldsymbol{\beta}_{1l}, \boldsymbol{\beta}_{2l}, ...\boldsymbol{\beta}_{pl})^T$ while holding fixed the vector of parameters of the other groups.

2.2.2 Dimensionality Reduction

These approaches tackle the $n << p$ problem with methods that involve projecting the p predictors into a new low-dimensional space where the most relevant characteristics of the original data are preserved. The process is performed by computing artificial variables (also known as latent variables or component scores), each being a linear combination of the original predictors which are subsequently used for visualisation and analytical purposes. As already mentioned in previous sections, in the context of dimensionality reduction, unsupervised and supervised learning are the classification names given to the methods where the construction of the latent variables is conducted without and with respect to the response matrix, respectively. Principal Component Analysis (PCA) and Partial Least Squares (PLS) are probably the best-known methods in each of these two learning frameworks. In the original formulation of the techniques, the resulting components are linear combinations of *all* p predictors variables, a condition that hampers interpretability of the results. To address this disadvantage, sparse versions of these classical multivariate methods have been introduced by using constraints in the computation of the components, a process that results in the inclusion of the most relevant variables (signals) and the exclusion of the irrelevant features (noise). In the remaining sections of this chapter I provide a description on the main methods that fall under the dimensionality

reduction category.

2.2.2.1 Principal Component Analysis

Principal Component Analysis (PCA) [89] aims at summarising a set of correlated features into a smaller number of representative variables called *principal components* (PCs), that collectively explain most of the variability in the original dataset. Suppose that we have n observations with measurements on only two variables that are being visualized in a two-dimensional scatterplot. The first PC is the direction in the feature space that captures most of the variation in the cloud of data points, or equivalently, it is the direction that is the closest to the collection of points. The second PC is defined as the direction that captures the remaining variability under the condition that is uncorrelated (linearly independent, perpendicular or orthogonal) to the first component. In a two-dimensional space when $p = 2$, there is only one possible direction that satisfies such condition. The components are a linear combination of the two variables, and the weights associated to the variables in each component are known as loading vectors. The two components can also be considered as a new coordinates system onto which the n observations are projected; these projected values are known as PC *scores*.

The definition of the loading vectors (finding the direction in the feature space) can be viewed as an optimization problem that seeks to maximize the sample variance or to minimize the squared Euclidean distance of the n observations subject to normalisation and orthogonality constraints. In practice, standard linear algebra techniques, namely eigen decomposition and Singular Value Decomposition (SVD), are used to solve the PCA problem. The former is applied to the variance-covariance matrix of $\mathbf{X}$ (or correlation matrix if $\mathbf{X}$ is scaled); however, its direct computation is not feasible when p is large, therefore SVD is usually preferred. In detail, let $\mathbf{X}$ be a matrix of size $n \times p$ whose columns have been centred and scaled. SVD decomposes $\mathbf{X}$ into the

product of three matrices as follows:

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T \tag{2.27}$$

where $\mathbf{U}$ is an orthogonal matrix ($\mathbf{U}^T\mathbf{U} = \mathbf{I}_n$) of size $n \times p$, $\mathbf{V}$ is also an orthogonal matrix ($\mathbf{V}^T\mathbf{V} = \mathbf{I}_n$) of size $p \times p$, and $\mathbf{D}$ is a non-negative $n \times p$ diagonal matrix. The diagonal entries of $\mathbf{D}$ are known as singular values while the columns of $\mathbf{U}$ and $\mathbf{V}$ are called the left and right singular vectors, respectively. The loading vectors of the PCs are given by the right singular vectors $\boldsymbol{v}_h$ and the PCs scores are defined by the product $\mathbf{X}\boldsymbol{v}_h$. Thus, the first principal component $\boldsymbol{c}_1 = \mathbf{X}_1\boldsymbol{v}_1$ is the projection of the data $\mathbf{X}$ or linear combination of the columns of $\mathbf{X}$ with highest sample variance among all possible choices of unit vectors $\boldsymbol{v}_h$. The second principal component $\boldsymbol{c}_2 = \mathbf{X}_2\boldsymbol{v}_2$ is the data projection or linear combination with highest sample variance among the ones that are orthogonal with $\boldsymbol{c}_1$, and so on. Similarly, the principal components can be expressed in terms of the left singular vector by the product $\mathbf{C}_h = \mathbf{U}_h\mathbf{D}_{hh}$. Finally, the explained variability by the h^{th} PC is retrieved by $\mathbf{D}_{hh}^2/n$ and it is sorted in increasing order, a property known as successive maximization of variance.

There are some important aspects that are worthy of mention when performing PCA [82]. First, since the procedure seeks to find directions of maximum variation, the variables are typically scaled to have standard deviation one before SVD is applied to $\mathbf{X}$; otherwise, the PCs will be a reflection of the features variances and will undesirably depend on the unit and scales in which they were measured (in setting were all the variables are measured in the same units scaling may not be necessary). Second, since the loading vectors specify directions in the p-dimensional space and the scores the variances of the projected data points along those directions, the signs assigned to the corresponding elements of the vectors can be exchanged. That is, there is no change in the PCs if the sign is flipped on both the loading and score vectors. Third, the maximum number of components is given by the $min(n-1,\ p)$ and the user-defined number of dimensions is typically chosen based on the visual inspection of the curve defined by cumulative proportion of variance explained (detect an

inflection point or elbow in a scree plot). Two or three number of components usually provide a useful lower dimension space to visualise observations and distinguish interesting patterns. Finally, once a lower-dimensional projection has been selected, the PCs can be used as predictors in a regression model in place of $\mathbf{X}$, a method called PC regression (PCR). It presents as obvious advantages that a lower number of variables are used and that the PCs are uncorrelated, and therefore there are no problems with collinearity. In this supervised setting the number of components can also be selected by CV.

2.2.2.2 Sparse Principal Component Analysis

As already mentioned, the PCs are linear combinations of all the original variables and all the elements of the corresponding loading vectors are typically non-zero. Such property makes interpretation difficult, and the identification of variables that play a significant role is usually based on subjective grounds. To make the selection of the relevant features more robust, several approaches have been proposed in the literature which produce modified PCs with sparse loadings. Simple thresholding sets the loadings with absolute values smaller than a certain threshold to zero and is the most obvious and straightforward approach to introduce sparsity [90]. Another simple method is to restrict the loading vectors to take values from a small set of allowable integers such as 0, 1 and -1. More sophisticated methods include *Simplified Component Technique-LASSO* (SCotLASS) [91] which defines the loadings by applying the lasso penalty on the maximization problem of PCA, *Sparse Principal Component Analysis* (SPCA) [92] which reformulates PCA as a regression-type problem and achieves sparsity by imposing the lasso penalty on the regression coefficients and *sparse PCA via regularized SVD* (sPCA-rSVD) [93] which combines the low rank approximation property of SVD with sequential regularization of the loading vectors [1]. Here, I briefly describe these three approaches.

[1]Sparse Principal Component Analysis (SPCA) (uppercase) correspond to the specific technique introduced by Zou et al. [92] while sparse Principal Component Analysis (sPCA) (no uppercase) is a generic name referring to any PCA method that retrieves PCs with sparse loadings.

The successive variance maximization problem that defines the loading vectors in PCA can be expressed as follows:

$$\boldsymbol{a}_k^T(\mathbf{X}^T\mathbf{X})\boldsymbol{a}_k, \tag{2.28}$$

$$\text{subject to } \boldsymbol{a}_k^T\boldsymbol{a}_k = 1 \text{ and (for } k \geq 2) \ \boldsymbol{a}_h^T\boldsymbol{a}_k = 0, h < k;$$

which correspond to the normalization (unit norm to obtain a bounded solution) and orthogonality constraints, respectively; the latter is only applied from the second loading vector onwards. SCotLASS adds an extra constraint:

$$\sum_{j=1}^{p} |\boldsymbol{a}_{kj}| \leq t \tag{2.29}$$

where t is again a tuning parameter. The L_1 constraint encourages some of the elements of the loadings to be zero and hence a to be sparse, for sufficiently small t. The major disadvantage associated with this method is that the choice of the appropriate value for t by CV is computationally expensive. SPCA improves over that difficulty as it transforms PCA to an elastic net optimization problem as follows:

$$\left(\hat{\mathbf{A}}, \hat{\mathbf{B}}\right) = \underset{\mathbf{A},\mathbf{B}}{\operatorname{argmin}} \left\|\mathbf{X} - \mathbf{A}\mathbf{B}^T\mathbf{X}\right\|^2 + \lambda_2 \sum_{j=1}^{k} \left\|\boldsymbol{\beta}_j\right\|_2^2 + \sum_{j=1}^{k} \lambda_{1,j} \left\|\boldsymbol{\beta}_j\right\|_1 \tag{2.30}$$

where both $\mathbf{A}$ and $\mathbf{B}$ are $p \times k$ matrices and k is the number of PCs to be extracted. The first term is the loss function and is the critical part of the criterion as it corresponds to the approximation of PCA into a regression-type problem. The ridge penalty is added in order to handle all kinds of $\mathbf{X}$ matrices ($p >> n$ or $n > p$ in the presence of collinearity); it is not used to penalize the regression coefficients as in Equation 2.23 but to ensure the reconstruction of PCs. The lasso term produces the sparse loadings for high enough values of λ_1. The approximated loading vectors $\hat{\mathbf{V}}_j$ are given by the normalized regression coefficients $\hat{\boldsymbol{\beta}}_j / \left\|\hat{\boldsymbol{\beta}}_j\right\|_2$.

Default values are set for the parameter λ_2 which depend on the dimensionality of $\mathbf{X}$. For $n > p$ data λ_2 can be zero but usually a small positive number is chosen to over-

come potential collinearity problems and for $p > n$ a positive λ_2 is required in order to get an exact PCA solution. In the case of λ_1, different values are allowed for penalizing the different PCs and the decision is made based on a compromise between variance explained and sparsity. Although the algorithm delivers a computationally efficient solution, the need for accommodating it to different characteristics of the $\mathbf{X}$ matrix adds an extra complexity that can potentially be disadvantageous.

sPCA-rSVD provides a uniform treatment of both $n > p$ and $p > n$ situations based on the low rank approximation of matrices or deflation property of SVD. Consider that $rank(\mathbf{X}) = r$, for any value $l < r$, the closest rank-l matrix approximation to $\mathbf{X}$ is given by:

$$\mathbf{X}^{(l)} = \sum_{k=1}^{l} \boldsymbol{d}_k \boldsymbol{u}_k \boldsymbol{v}_k^T \tag{2.31}$$

The term *"closest"* means that $\mathbf{X}^{(l)}$ minimizes the squared Frobenius norm between $\mathbf{X}$ and an arbitrary rank-l matrix $\mathbf{X}^*$. In approximate terms, the principle behind the sPCA-rSVD approach is to find the best rank-one approximation of $\mathbf{X}$, which is given by $(\boldsymbol{u}_1, \boldsymbol{d}_1, \boldsymbol{v}_1)$ and to impose regularization penalties on $\boldsymbol{v}$ in order to obtain sparse loading vectors. Subsequent sparse loadings $\boldsymbol{v}_i (i > 1)$ are obtained sequentially via rank-one approximation of residual matrices. For example, $\mathbf{X}' = \mathbf{X} - \boldsymbol{d}_1 \boldsymbol{u}_1 \boldsymbol{v}_1^T$ with a best rank-one approximation retrieved by $(\boldsymbol{u}_2, \boldsymbol{d}_2, \boldsymbol{v}_2)$. More precisely, sPCA-rSVD seeks to optimize the following penalized sum-of-squares criterion:

$$\underset{\|\boldsymbol{u}\|_2, \widetilde{\boldsymbol{v}}}{\operatorname{argmin}} \left\| \mathbf{X} - \boldsymbol{u} \widetilde{\boldsymbol{v}}^T \right\|_F^2 + P_\lambda(\widetilde{\boldsymbol{v}}) \tag{2.32}$$

where $\|.\|_F^2$ is the squared Frobenius norm, $P_\lambda(\widetilde{\boldsymbol{v}})$ is a penalty function, $\lambda \geq 0$ is again a tuning parameter, $\boldsymbol{u}$ is a unit-norm vector of length n and $\widetilde{\boldsymbol{v}}$ is a vector of length p. Since the original loading vector $\boldsymbol{v}$ is typically constrained to be unit-norm, the direct application of a penalty on $\boldsymbol{v}$ is inappropriate. For that reason, the criterion incorporates the re-scaled versions $\boldsymbol{u}$ and $\widetilde{\boldsymbol{v}}$ such that the first has unit length and the second is free of any scale constraint. The tilde symbol is used to emphasize the fact that vectors are not normed. The sparse and normalised loading vectors are then given by

$\boldsymbol{v} = \widetilde{\boldsymbol{v}}/\|\widetilde{\boldsymbol{v}}\|_2$. Different penalty functions are allowed other than the lasso and similarly to SPCA they can be different for the different PCs. An important observation to highlight here is the fact that the optimization criterion in Equation 2.32 is connected to least squares regression: for a fixed $\boldsymbol{u}$, the optimal solution $\boldsymbol{v}$ is the least squares coefficient vector of regressing the columns of $\mathbf{X}$ on $\boldsymbol{u}$. Thus, imposing sparsity on $\boldsymbol{v}$ becomes similar to a variable selection problem.

Finally, it is worth mentioning that for the sPCA methods discussed in this section the two main properties of PCA, namely orthogonal loading vectors and uncorrelated PCs are lost. The first property may not be desirable in this setting because enforcing orthogonality might result in less sparse solutions [94]. However, if the PCs are correlated the total variance explained by each of them contains partial contributions from previous components, therefore an adjustment is needed to calculate the cumulative percentage of variance explained and both SPCA and sPCA-rSVD provide different methods to estimate it.

2.2.2.3 Partial Least Squares

While PCA defines the PCs as linear combinations of the original variables that maximise the variance within the $\mathbf{X}$ matrix, PLS seeks to construct new variables that are also meaningful linear combinations of the original $\mathbf{X}$ and $\mathbf{Y}$ variables but in this case, the criterion to maximise is the *covariance or correlation* between the two matrices. These new variables are known as component scores or latent variables and, in contrast to PCA, they are searched in an iterative manner for both $\mathbf{X}$ and $\mathbf{Y}$.

As in introductory background, PLS is a technique that was first originated in 1966 by Herman Wold [95] to refer to a class of algorithms developed for the analysis of an arbitrary number of blocks of data by means of latent variables and two different modes were proposed for their computation (Mode A and Mode B) [96]. Since then, several PLS variants have been introduced in the literature. The particular case of two blocks of data encompasses the following four variations: PLS-SVD [97], [98], [99], [100], PLS in mode A (PLS-W2A, for Wold's Two-Block Mode A PLS) [101], [102], [103] ,

PLS in mode B (PLS-W2B) [104], [105], and Partial Least Squares Regression (PLS-R) [106], [107], [108]. PLS-W2B is equivalent to Canonical Correlation Analysis (CCA) [109] where the objective is to maximise the *correlation* between scores. In this section I focus on the other three variations which aim at maximising the *covariance*. Both PLS-SVD and PLS-W2A model a symmetric relationship between the two blocks of data in order to explain the shared information; on the other hand, PLS-R models an asymmetric relationship where one block is used as predictor to explain the variability in the other [101], [110]. Because of the resemblance with CCA and to emphasize the contrast with PLS-R, the variants PLS-SVD and PLS-W2A are sometimes referred in the literature as PLS canonical mode [111].

Entering to more technical details, PLS seeks to decompose the centred (and possibly standardized) data matrices $\mathbf{X}(n \times p)$ and $\mathbf{Y}(n \times q)$ into latent variables denoted by $\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_h$ and $\boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_h$ which are n-dimensional vectors associated with $\mathbf{X}$ and $\mathbf{Y}$ respectively, where H is a small number specifying the total number of components. These scores are estimated as linear combinations of the original variables and the weight associated to each original variable is given by the loading vectors $\boldsymbol{u}_h$ and $\boldsymbol{v}_h$ of length p and q for $\mathbf{X}$ and $\mathbf{Y}$, respectively. The first pair of latent components is defined as $\boldsymbol{\xi}_1 = \mathbf{X}\boldsymbol{u}_1$ and $\boldsymbol{\omega}_1 = \mathbf{Y}\boldsymbol{v}_1$ and correspond to the score vectors with maximal covariance. Orthogonality constraints are imposed in the optimization problem and in order to ensure that solutions meet the required orthogonality, the PLS algorithms are solved iteratively where a deflation step is performed to remove the information stored in the previous iteration. Subsequent pairs of latent components are then defined in terms of the deflated matrices as $\boldsymbol{\xi}_h = \mathbf{X}_{h-1}\boldsymbol{u}_h$ and $\boldsymbol{\omega}_h = \mathbf{Y}_{h-1}\boldsymbol{v}_h$ where $\mathbf{X}_{h-1}$ and $\mathbf{Y}_{h-1}$ are the residual matrices. Thus, the vectors $\boldsymbol{u}_h$ and $\boldsymbol{v}_h$ are the weights of the original variables that define the component scores in terms of the *deflated matrices*. However, the score vectors can also be defined in terms of the original matrices as $\boldsymbol{\xi}_h = \mathbf{X}_h\boldsymbol{w}_h$ and $\boldsymbol{\omega}_h = \mathbf{Y}_h\,\boldsymbol{z}_h$ where $\boldsymbol{w}_h$ and $\boldsymbol{z}_h$ are known as vectors of *adjusted weights*.

More formally, the optimization criterion for the three two-block PLS methods men-

tioned above can be specified as follows:

$$(\boldsymbol{u}_h, \boldsymbol{v}_h) = \underset{\|\boldsymbol{u}\|_2 = \|\boldsymbol{v}\|_2 = 1}{\operatorname{argmax}} Cov(\mathbf{X}_{h-1}, \mathbf{Y}_{h-1}) \tag{2.33}$$

subject to $\boldsymbol{u}^T\boldsymbol{u}_j = \boldsymbol{v}^T\boldsymbol{v}_j = 0, 1 \leq j < h$, which is the basic orthogonality constraint common for the three methods. For both PLS-W2A and PLS-R, additional orthogonality constraints are added into the optimization problem. The former searches for successive $\mathbf{X}$ score vectors (resp. $\mathbf{Y}$ score vectors) that are orthogonal to the previous ones:

$$Cov(\boldsymbol{\xi}_h, \boldsymbol{\xi}_j) = Cov(\boldsymbol{\omega}_h, \boldsymbol{\omega}_j) = 0,\ 1 \leq j < h, \tag{2.34}$$

whereas PLS-R searches for successive $\mathbf{X}$ scores $\boldsymbol{\xi}_h$ that are orthogonal to the previous ones (as in Equation 2.34) in addition to $\mathbf{Y}$ scores $\boldsymbol{\omega}_h$. As already mentioned, a matrix deflation step is conducted in the computation of the component scores to ensure that the orthogonality constraints are satisfied. To guarantee that the additional constraints imposed in the optimization problems for PLS-W2A and PLS-R are met, a modification is needed in the deflation steps of those two algorithms; this in turn means that all three PLS methods conduct the matrices deflation in a different manner. That being said, the deflation is needed only for $h > 1$; therefore, the solution for the three PLS approaches is the same for the first iteration $h = 1$. Alternatively, the optimization criterion in Equation 2.33 can be formulated in terms of variance-covariance matrix $\mathbf{M} = \mathbf{X}^T\mathbf{Y}$ or in terms on the adjusted weight vectors $\boldsymbol{w}_h$ and $\boldsymbol{z}_h$ with the corresponding modification in the constraints. Thus, multiple equivalent objective functions to Equation 2.33 can be encountered in the literature.

The solution for PLS-SVD is the most straightforward of the three approaches and it is fully given by the SVD of $\mathbf{M}$. That is, the pair of loading vectors $(\boldsymbol{u}_h, \boldsymbol{v}_h)$ are the h first columns of the matrices $\mathbf{U}$ and $\mathbf{V}$, which are respectively the left and right singular vectors of $\mathbf{M}$. Since the loading vectors define the component scores in terms of the residual matrices, an iterative procedure is needed to retrieve them. The deflated matrices are specified by $\mathbf{X}_h = \mathbf{X}_{h-1} - \boldsymbol{\xi}_h\boldsymbol{u}_h^T$ and $\mathbf{Y}_h = \mathbf{Y}_{h-1} - \boldsymbol{\omega}_h\boldsymbol{v}_h^T$

which in turn define the component scores as $\boldsymbol{\xi}_h = \mathbf{X}_{h-1}\boldsymbol{u}_h$ and $\boldsymbol{\omega}_h = \mathbf{Y}_{h-1}\boldsymbol{v}_h$. A different approach to find the optimal loadings is to employ the deflation property of SVD, whereby the cross-product $\mathbf{M}$ is directly deflated by its rank-one approximation $\mathbf{M}_h = \mathbf{M}_{h-1} - \boldsymbol{d}_h\boldsymbol{u}_h\boldsymbol{v}_h^T$. SVD is then performed on the delated matrix and the pair $(\boldsymbol{u}_h, \boldsymbol{v}_h)$ is given by only the first pair of singular vectors. Thus, in this PLS variant the SVD of $\mathbf{M}$ is either performed only once and the solution is the first h pairs of left and right singular vectors or applied iteratively on the deflated cross-product where only the first pair of vectors is stored.

For PLS-W2A, the initial step is to define the first pair of loading vectors $\boldsymbol{u}_1$ and $\boldsymbol{v}_1$, which are the left and right singular vectors from the SVD of $\mathbf{M}$. Those vectors are then used to calculate the first pair of component scores $\boldsymbol{\xi}_1 = \mathbf{X}_0\boldsymbol{u}_1$ and $\boldsymbol{\omega}_1 = \mathbf{Y}_0\boldsymbol{v}_1$. Note that the subscript zero is explicitly written to denote the original data matrices (as opposed to the deflated ones). A regression step follows in which each column of $\mathbf{X}_0$ (resp. $\mathbf{Y}_0$) is regressed onto $\boldsymbol{\xi}_1$ (resp. $\boldsymbol{\omega}_1$), the regression coefficients of those local linear regressions are defined as $\boldsymbol{c}_1 = \mathbf{X}_0^T\boldsymbol{\xi}_1/\boldsymbol{\xi}_1^T\boldsymbol{\xi}_1$ and $\boldsymbol{e}_1 = \mathbf{Y}_0^T\boldsymbol{\omega}_1/\boldsymbol{\omega}_1^T\boldsymbol{\omega}_1$. The subsequent matrix deflation step uses $\boldsymbol{c}_1$ and $\boldsymbol{e}_1$ to obtain the residual matrices $\mathbf{X}_1 = \mathbf{X}_0 - \boldsymbol{\xi}_1\boldsymbol{c}_1^T$ and $\mathbf{Y}_1 = \mathbf{Y}_0 - \boldsymbol{\omega}_1\boldsymbol{e}_1^T$. Finally, the matrix cross-product is updated using the deflated matrices $\mathbf{M}_1 = \mathbf{X}_1^T\mathbf{Y}_1$ and the procedure is reiterated H times.

For PLS-R (also called PLS1 for $q = 1$ or PLS2 for $q > 1$), the aim is to construct latent variables that model $\mathbf{X}$ and simultaneously predict $\mathbf{Y}$. For that purpose, several algorithms have been proposed, the two most well-known are the Nonlinear estimation by Iterative Partial Least Squares (NIPALS) [95], [112], [113] and the Statistically Inspired Modification of PLS (SIMPLS) [114]. They are characterised by the incorporation of an extra relationship that explicitly relates the $\mathbf{X}$ and $\mathbf{Y}$ scores. Such relationship is given by the vectors $\boldsymbol{\omega}_h$ being regressed onto $\boldsymbol{\xi}_1$ yielding the PLS regression coefficients $\boldsymbol{\beta}_h^{PLS} = \boldsymbol{\xi}_h^T\boldsymbol{\omega}_h/\boldsymbol{\xi}_h^T\boldsymbol{\xi}_h$, which can be used for predictions given new observations. The deflation step on the $\mathbf{Y}$ matrix is performed as function of this vector of coefficients $\mathbf{Y}_h = \mathbf{Y}_{h-1} - \boldsymbol{\beta}_h\boldsymbol{\xi}_h\boldsymbol{\omega}_h^T$, ensuring orthogonality of $\boldsymbol{\beta}_h^{PLS}$. Equivalently, the deflation of the matrix $\mathbf{Y}$ can be expressed as function of the local regression coeffi-

cients obtained from the linear regressions of the columns of $\mathbf{Y}_{h-1}$ onto the $\mathbf{X}$ scores $\boldsymbol{\xi}_h$, $\boldsymbol{d}_h = \mathbf{Y}_{h-1}^T \boldsymbol{\xi}_h / \boldsymbol{\xi}_h^T \boldsymbol{\xi}_h$. The residual matrices are then retrieved $\mathbf{Y}_h = \mathbf{Y}_{h-1} - \boldsymbol{\xi}_h \boldsymbol{d}_h^T$. The main difference between the two PLS-R algorithms is that SIMPLS computes the scores directly as linear combinations of the original variables without conducting deflations on the matrices, which provides a computational advantage over NIPLAS; however, the latter remains the most cited and commonly used of the two.

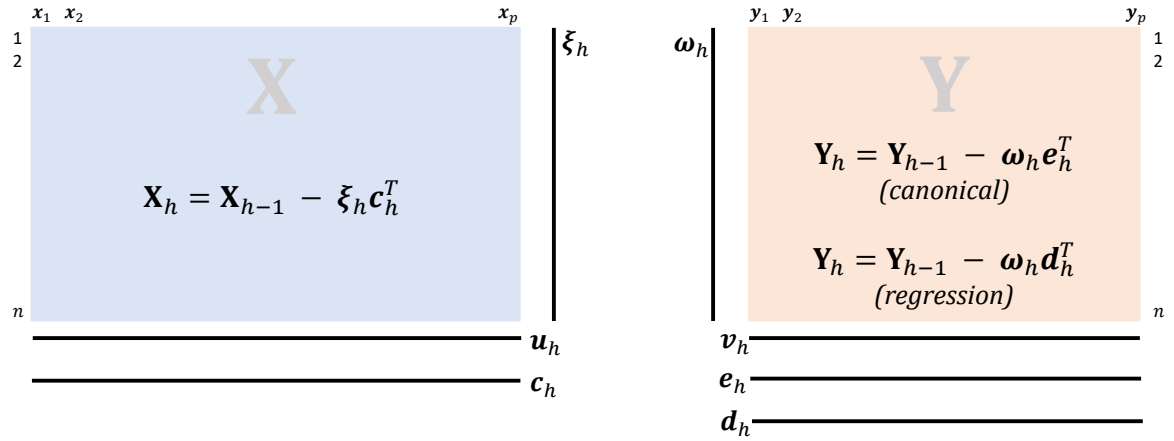
It is worth to summarise the similarities and difference of the PLS variants discussed in this section (see also Figure 2.1). The solutions to the optimization problem in Equation 2.33 are the same for PLS-SVD, PLS-W2A and PLS-R in the first iteration $h = 1$ and it is given by the left and right singular vectors associated with the largest singular value obtained from the SVD of the variance-covariance matrix $\mathbf{M}$. Successive iterations provide different solutions as a consequence of the way the deflation of the data matrices is conducted. PLS-SVD deflates the cross-product directly by subtracting a rank-one estimate from it $\mathbf{M}_h = \mathbf{M}_{h-1} - \boldsymbol{d}_h \boldsymbol{u}_h \boldsymbol{v}_h^T$. PLS-W2A subtract rank-one approximations of the individual data matrices $\mathbf{X}_h$ and $\mathbf{Y}_h$ to obtain $\mathbf{X}_{h+1}$ and $\mathbf{Y}_{h+1}$ and from these residuals a new cross-product $\mathbf{M}$ is computed. PLS-R deflates the $\mathbf{Y}$ matrix in an asymmetric manner using information content from the $\mathbf{X}$ matrix as in $\mathbf{Y}_h = \mathbf{Y}_{h-1} - \beta \boldsymbol{\xi}_h \boldsymbol{\omega}_h^T$ or $\mathbf{Y}_h = \mathbf{Y}_{h-1} - \boldsymbol{\xi}_h \boldsymbol{d}_h^T$ before recalculation of the cross-product. The decomposition of the original data matrices can be written as follows:

$$\mathbf{X} = \boldsymbol{\Xi}\mathbf{C}^T + \mathbf{F}_X \qquad \mathbf{Y} = \boldsymbol{\Omega}\mathbf{E}^T + \mathbf{F}_Y \tag{2.35}$$

$$\mathbf{Y} = \boldsymbol{\Omega}\mathbf{D}^T + \mathbf{F}_Y = \mathbf{X}\hat{\boldsymbol{\beta}}^{PLS} + \mathbf{F}_Y \tag{2.36}$$

where $\boldsymbol{\Xi}$ (resp. $\boldsymbol{\Omega}$) is the matrix of $\mathbf{X}$ (resp. $\mathbf{Y}$) component scores. The matrices $\mathbf{C} = \mathbf{X}^T\boldsymbol{\Xi}, \mathbf{E} = \mathbf{Y}^T\boldsymbol{\Omega}$ and $\mathbf{D} = \mathbf{Y}^T\boldsymbol{\Xi}$ are of dimension $p \times h$ and $q \times h$ and contain the regression coefficients $\boldsymbol{c}_h$, $\boldsymbol{e}_h$ and $\boldsymbol{d}_h$, respectively. $\mathbf{F}_X$ and $\mathbf{F}_Y$ are the $n \times p$ and $n \times q$ residual matrices. Consequently, the first decomposition of $\mathbf{Y}$ corresponds to PLS-W2A and the second to PLS-R. (The decomposition for PLS-SVD requires a more

Figure 2.1: Graphical illustration of the PLS algorithm.



The illustration displays the dimension of the $\mathbf{X}$ and $\mathbf{Y}$ component score vectors ($\boldsymbol{\xi}_h$ and $\boldsymbol{\omega}_h$, respectively), the $\mathbf{X}$ and $\mathbf{Y}$ loading coefficient vectors ($\boldsymbol{u}_h$ and $\boldsymbol{v}_h$, respectively) and the $\mathbf{X}$ and $\mathbf{Y}$ local regression coefficients ($\boldsymbol{c}_h$ and $\boldsymbol{e}_h$ and $\boldsymbol{d}_h$, respectively). The symmetric and assymmetrix delation steps of the $\mathbf{Y}$ matrix are also shown.

elaborate mathematical explanation and is omitted).

As with many techniques discussed in this chapter, the choice of the number of components H is a model selection question that usually proceeds by CV (see details in section 3.6.2.1.1) and the maximum value H can take depends on the PLS variant being performed. For both PLS-SVD and PLS-W2A the algorithms can run for up to $min(p, q)$ iterations but in practice the decision depends on the rank of the matrices [101]. For PLS-SVD, H is given by the rank of M; for iteration beyond that point the pair of score vectors will have zero covariance. On the other hand, in PLS-W2A the restriction is specified by $min(rank(\mathbf{X}), rank(\mathbf{Y}))$ because past that iteration $\mathbf{X}^T\mathbf{Y} = 0$. In PLS-R, in contrast with the other two methods, H is limited by the rank of $\mathbf{X}^T\mathbf{X}$ because of the asymmetric way in which the deflation process is conducted on the $\mathbf{Y}$ matrix. However, if there is a large number of non-relevant noise predictors, the algorithm starts to be deviated from the aim of maximising the variance-covariance between $\mathbf{X}$ and $\mathbf{Y}$ and the covariance between predictors is optimized instead [115].

2.2.2.4 PLS Extensions: Regularized Partial Least Squares

The connexion between the low rank approximation property of SVD and least squares minimisation in linear regression first introduced for sPCA has enabled some authors to extend the technique variants discussed in the previous section to novel *regularized Partial Least Squares* (rPLS) approaches, which include sparse PLS (sPLS) [111], group (gPLS) and sparse group PLS (sgPLS) [116]. Following a similar principle to the variable selection techniques discussed in section 2.2.1, these methods introduce penalty terms to the loading vectors resulting in sparse solutions, a process that in turn improves model interpretability and quality of the estimators.

In contrast to sPCA-SVD, the rPLS approaches aim at penalizing both loadings vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ to perform variable selection in both data sets. The procedure is conducted in an iterative manner for each component at a time, in which the value of one of the loading vectors is fixed while the solution for the other is found. As discussed previously, a penalty term cannot be applied to a unit-norm vector; therefore, the solution is found for an unconstrained vector and the fixed vector is restricted to be unit-norm. After the optimal values have been defined, the solutions are scaled. This procedure thus leads to normed sparse loading vectors. More precisely, the optimization criteria for a fixed unit-norm $\boldsymbol{v}$ (resp. $\boldsymbol{u}$) are specified as follows:

$$\tilde{\boldsymbol{u}_h} \;=\; \underset{\boldsymbol{u}}{\operatorname{argmin}}\{\left\|\mathbf{M}_{h-1} \;-\; \widetilde{\boldsymbol{u}}\boldsymbol{v}^T\right\| \;+\; P_{\lambda 1}(\widetilde{\boldsymbol{u}})\} \tag{2.37}$$

$$\tilde{\boldsymbol{v}_h} \;=\; \underset{\boldsymbol{v}}{\operatorname{argmin}}\{\left\|\mathbf{M}_{h-1} \;-\; \boldsymbol{u}\widetilde{\boldsymbol{v}}^T\right\| \;+\; P_{\lambda 2}(\widetilde{\boldsymbol{v}})\} \tag{2.38}$$

then the normed $\boldsymbol{u}_h$ and $\boldsymbol{v}_h$ are obtained $\tilde{\boldsymbol{u}}_h/\|\tilde{\boldsymbol{u}}_h\|^2$ and $\tilde{\boldsymbol{v}}_h/\|\tilde{\boldsymbol{v}}_h\|^2$. $P_{\lambda 1}(\widetilde{\boldsymbol{u}})$ and $P_{\lambda 2}(\widetilde{\boldsymbol{v}})$ are the penalty functions with tuning parameters λ_1 and λ_2. These are both convex optimization problems and the nature of the penalty terms is modified accordingly. In the case of sPLS, they are specified as follows:

$$P_{\lambda 1}(\widetilde{\boldsymbol{u}}) \;=\; \sum_{i=1}^{p} 2\lambda_1|\widetilde{\boldsymbol{u}}_i| \tag{2.39}$$

$$P_{\lambda 2}(\widetilde{\boldsymbol{v}}) \;=\; \sum_{j=1}^{q} 2\lambda_2 |\widetilde{\boldsymbol{v}}_j| \tag{2.40}$$

This is the L_1 or lasso penalty and for $\lambda_1 \geq 0$ or $\lambda_2 \geq 0$, some elements of the loading vectors $\widetilde{\boldsymbol{u}}$ or $\widetilde{\boldsymbol{v}}$ will be forced to be zero. When both matrices $\mathbf{X}$ and $\mathbf{Y}$ are divided respectively into K and L sub-matrices, the penalty terms for gPLS are specified as follows:

$$P_{\lambda 1}(\widetilde{\boldsymbol{u}}) \;=\; \lambda_1 \sum_{k=1}^{K} \sqrt{p_k} \left\| \widetilde{\boldsymbol{u}}^{(k)} \right\|_2 \tag{2.41}$$

$$P_{\lambda 2}(\widetilde{\boldsymbol{v}}) \;=\; \lambda_2 \sum_{l=1}^{L} \sqrt{q_l} \left\| \widetilde{\boldsymbol{v}}^{(l)} \right\|_2 \tag{2.42}$$

This is the group lasso penalty and depending on the tuning parameters $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$, the entire weight subvector $\widetilde{\boldsymbol{u}}^{(k)}$ or $\widetilde{\boldsymbol{v}}^{(l)}$ will be zero, or non-zero together. While for sgPLS, the penalty terms are:

$$P_{\lambda 1}(\widetilde{\boldsymbol{u}}) \;=\; (1-\alpha_1)\lambda_1 \sum_{k=1}^{K} \sqrt{p_k} \left\| \widetilde{\boldsymbol{u}}^{(k)} \right\|_2 \;+\; \alpha_1 \lambda_1 \|\widetilde{\boldsymbol{u}}\|_1 \tag{2.43}$$

$$P_{\lambda 2}(\widetilde{\boldsymbol{v}}) \;=\; (1-\alpha_2)\lambda_2 \sum_{l=1}^{L} \sqrt{q_l} \left\| \widetilde{\boldsymbol{v}}^{(l)} \right\|_2 \;+\; \alpha_2 \lambda_2 \|\widetilde{\boldsymbol{v}}\|_1 \tag{2.44}$$

where the additional tuning parameters α_1 and α_2 are introduced. This is the sparse group lasso penalty and depending on the combination of α_1 and λ_1 (or α_2 and λ_2) the weight sub-vector $\widetilde{\boldsymbol{u}}^{(k)}$ or $\widetilde{\boldsymbol{v}}^{(l)}$ will be eliminated entirely, or sparsely estimated. As a result of the properties of these penalty functions, sPLS enables individual variable selection, gPLS performs selection at group level and sgPLS enables selection at both group and single feature levels simultaneously.

These regularized versions where originally introduced as extensions of the PLS variants PLS-W2A and PLS-R; however, recent efforts in this area of research have made possible to perform regularization on all four two-block PLS approaches (including PLS-W2B or CCA) using an all-encompassing algorithm [110]. In other words, the deflated matrix $\mathbf{M}_{h-1}$ can be obtained by means of any of the PLS variants previously discussed. This iterative procedure that performs all versions of PLS (here I focus on

PLS-SVD, PLS-W2A and PLS-R) alongside the regularized techniques (sPLS, gPLS, sgPLS), and which optimizes the criteria stated in Equation 2.33,Equation 2.37 and Equation 2.38 can be summarised as follows:

Algorithm 1 Algorithm for regularized versions of PLS-SVD, PLS-W2A and PLS-R

Input: $\mathbf{X}(n \times p), \mathbf{Y}(n \times q), H, \theta_u, \theta_v$.
Output: $\boldsymbol{\Xi}(n \times H), \boldsymbol{\Omega}(n \times H), \mathbf{U}(p \times H), \mathbf{V}(q \times H), \mathbf{C}(p \times H), \mathbf{E}(q \times H), \mathbf{D}(q \times H)$.

$\mathbf{X}_0 \leftarrow \mathbf{X}, \quad \mathbf{Y}_0 \leftarrow \mathbf{Y}$
for $h = 1, \ldots, H$ **do**
  Set $\mathbf{M}_{h-1} \leftarrow \mathbf{X}_{h-1}^T \mathbf{Y}_{h-1}$
  Apply SVD to $\mathbf{M}_{h-1}$ and extract first pair of singular vectors $\boldsymbol{u}_{old} = \boldsymbol{u}_h$ and $\boldsymbol{v}_{old} = \boldsymbol{v}_h$.
  Apply the corresponding soft-thresholding penalty function to weight vectors $\boldsymbol{u}_{old}$ and $\boldsymbol{v}_{old}$ until convergence of $\boldsymbol{u}_{new}$ and $\boldsymbol{v}_{new}$ followed by vector normalization.
  while convergence of $\boldsymbol{u}_{new}$ **do**
    $\widetilde{\boldsymbol{u}}_{new} \leftarrow S_u(\boldsymbol{v}_{old}; \mathbf{M}_{h-1}, \theta_u)$
    $\boldsymbol{u}_{new} \leftarrow \widetilde{\boldsymbol{u}}_{new} / \|\widetilde{\boldsymbol{u}}_{new}\|_2$ ▷ Normalization Step
    $\widetilde{\boldsymbol{v}}_{new} \leftarrow S_v(\boldsymbol{u}_{old}; \mathbf{M}_{h-1}, \theta_v)$
    $\boldsymbol{v}_{new} \leftarrow \widetilde{\boldsymbol{v}}_{new} / \|\widetilde{\boldsymbol{v}}_{new}\|_2$ ▷ Normalization Step
  end while
  Obtain $\mathbf{X}$ and $\mathbf{Y}$ scores.
  $\boldsymbol{\xi}_h \leftarrow \mathbf{X}_{h-1} \boldsymbol{u}_{new}$
  $\boldsymbol{\omega}_h \leftarrow \mathbf{Y}_{h-1} \boldsymbol{v}_{new}$
  Obtain local regression coefficients.
  if PLS-SVD **then**
    $\boldsymbol{c}_h \leftarrow \boldsymbol{u}_{new}$
    $\boldsymbol{e}_h \leftarrow \boldsymbol{v}_{new}$
  end if
  if PLS-W2A **then**
    $\boldsymbol{c}_h \leftarrow \mathbf{X}_{h-1}^T \boldsymbol{\xi}_h / \boldsymbol{\xi}_h^T \boldsymbol{\xi}_h$
    $\boldsymbol{e}_h \leftarrow \mathbf{Y}_{h-1}^T \boldsymbol{\omega}_h / \boldsymbol{\omega}_h^T \boldsymbol{\omega}_h$
  end if
  if PLS-R **then**
    $\boldsymbol{c}_h \leftarrow \mathbf{X}_{h-1}^T \boldsymbol{\xi}_h / \boldsymbol{\xi}_h^T \boldsymbol{\xi}_h$
    $\boldsymbol{d}_h \leftarrow \mathbf{Y}_{h-1}^T \boldsymbol{\xi}_h / \boldsymbol{\xi}_h^T \boldsymbol{\xi}_h$
  end if
  Obtain deflated $\mathbf{X}$ and $\mathbf{Y}$ matrices.
  $\mathbf{X}_h \leftarrow \mathbf{X}_{h-1} - \boldsymbol{\xi}_h \boldsymbol{c}_h^T$
  if PLS-SVD or PLS-W2A **then**
    $\mathbf{Y}_h \leftarrow \mathbf{Y}_{h-1} - \boldsymbol{\omega}_h \boldsymbol{e}_h^T$ ▷ Symmetric Deflation
  else
    $\mathbf{Y}_h \leftarrow \mathbf{Y}_{h-1} - \boldsymbol{\omega}_h \boldsymbol{d}_h^T$ ▷ Asymmetric Deflation
  end if
end for

Note that S_u and S_v are analytical functions introduced to provide the desired sparse

solution for the weight vectors, which depend on the data $\mathbf{M}$, the fixed weight $\boldsymbol{u}$ (or $\boldsymbol{v}$) and additional penalty-specific parameters θ_u (or θ_v). For sPLS $\theta_u = \lambda_1$ and $\theta_v = \lambda_2$, and in the case where there is no sparsity constraint $\lambda_1 = \lambda_2 = 0$, the same results as in classical PLS are obtained. In the other two regularized methods the penalty functions are applied groupwise. Thus, for gPLS these terms are defined as $\theta_u = (p_1...p_K, \lambda_1)$ and $\theta_v = (q_1...q_L, \lambda_2)$, and when $\lambda_1 = \lambda_2 = 0$ classical PLS results are also retrieved. While for sgPLS $\theta_u = (p_1...p_K, \lambda_1, \alpha_1)$ and $\theta_v = (q_1...q_L, \lambda_2, \alpha_2)$, and when $\alpha_1 = \alpha_2 = 0$ the group lasso penalty is obtained and the lasso when $\alpha_1 = \alpha_2 = 1$.

2.2.2.5 PLS Extensions: Partial Least Squares Discriminant Analysis

Another valuable PLS extension proposed in the literature is Partial Least Squares Discriminant Analysis (PLS-DA) [117], [118], [119], which has been introduced to accommodate a situation where the response is a categorical variable, in other words, where $\boldsymbol{y}$ is a vector that takes only one of G possible unordered values $\boldsymbol{y} = 0, ...G-1$ representing the different class categories. It is a technique that can be employed for both exploratory and classification purposes: the former case refers to the identification of variables that are most likely to be responsible for discrimination (for example, by means of graphical display) while the latter refers to the construction of predictive models to determine what class category a sample is most likely to belong to from a set of predictor variables. In this classification setting, PLS-DA shares common attributes with well-known discrimination strategies such as Euclidean Distance to Centroids (EDC) [120], [121] , Linear Discriminant Analysis (LDA) [122] and Quadratic Discriminant Analysis (QDA) [123]. In a situation where there are two groups ($G = 2$) and two predictors ($p = 2$), all four of these methods seek to find a separator or decision function that divides the space into two regions where the two sample groups are located. Of course, when there are more than two variables, the separator will be represented by a hyperplane in a multidimensional space. Furthermore, the four approaches are considered to be supervised techniques because before a decision function is identified and samples assigned to a class, information

regarding the class labels of the samples must be known a priori.

EDC, LDA and QDA employ the same principle to assign a sample to a group and it is based on the distance in the variable space between the sample and the centroids of the group. The group whose distance is smallest is the one that the sample is defined as belonging to and the discriminatory boundary is demarcated to occur when $d_{iA}^2 = d_{iB}^2$, that is, the squared distance between sample i and the centroids of group A is the same to that of group B. (For the sake of brevity, a binary scenario is illustrated). Naturally, the distance measure used for the three different methods is different. EDC employs the Euclidian distance while LDA and QDA use the Mahalanobis distance [124], [125] which uses for its calculation the variance-covariance matrix and different definitions for it are allowed (see section 3.6.2.3.2 for more details). For example, for LDA this is the pooled variance-covariance matrix over all groups while for QDA the variance-covariance matrix is defined for each group and therefore the distance metric differs for each class to be modelled. Because of this difference, QDA yields non-linear separation bounds between the groups and it is preferred over LDA when the variance structure of the groups is different.

In contrast, for PLS-DA the main principle to define a decisions boundary and to assign samples to classes is based on the predicted values from the PLS regression model. To be more precise, in a binary scenario the $y = 0, 1$ vector that specifies class membership (0 for members of group A and 1 for group B) is considered as a one-column matrix and the PLS1 algorithm is conducted as if the outcome variable was continuous. This is a procedure that is referred to in the literature as PLS1-DA. From the PLS regression coefficients, predicted values $\hat{y}$ are retrieved either from the same data used to fit the model (auto prediction) or from a CV procedure. Since PLS is inherently designed for regression purposes and to deal with continuous variables, the resulting predicted values take any values between 0 and 1 instead of an integer. For that reason, a *decision rule* (DR) must be employed in order to translate $\hat{y}$ into meaningful class membership in an accurate manner. The simplest DR is to choose a value halfway between the numerical class labels, following the examples given

in this case, an observation is assigned to class A if the predicted value is below 0.5 and to class B if below. Such simple threshold produces accurate results for a binary classification problem and sample groups of equal size and similar variance; however, it is not an appropriate separator for groups of unequal sizes. Various other DRs have been proposed in the literature to accommodate this scenario and they are discussed in more detail in section 3.6.2.3.2.

When there are more than two groups, it is customary to extend the PLS-DA model so that the vector indicative of class membership $\boldsymbol{y}$ becomes into a dummy block matrix $\mathbf{Y}$. This step is made by using a simple transformation method in which G different indicator variables are created as follows:

$$\boldsymbol{y}_j = \begin{cases} 1 & \text{if } \boldsymbol{y} = j \\ 0 & \text{otherwise} \end{cases} \tag{2.45}$$

In that manner, the response variable is transformed into a qualitative response matrix of 0 and 1 values with G columns representing the outcome categories and with n rows representing the membership of each observation. For example, if there are three groups A, B, and C, the second column of $\mathbf{Y}$ represents class membership of group B and for an observation that belongs to that class the row vector takes values of 0,1 and 0. Following the conversion of $\boldsymbol{y}$ into $\mathbf{Y}$, two different approaches can be employed to conduct PLS-DA. The first method is to perform three PLS1-DA models, one for each column, in what is often called a one-versus-all approach. The second method is to consider the dummy block matrix as a multivariable response and perform the PLS2 algorithm as described in section 2.1.4.3, an approach known as PLS2-DA. From the predicted values, class memberships are ascertained but more elaborate DRs are needed in order to consider the particularities of the data being analysed (i.e. relative group sizes).

A connection has been made between the discriminant methods discussed in this section [126],[119]. For two equal class sizes, PLS-DA with one component provides

equivalent classification performance to EDC and when using two components to LDA. In contrast to PLS-DA, a common cited disadvantage of both LDA and QDA is that the number of variables needs to be less than the number of observations ($n > p$) as it requires the inversion of the within-in group variance-covariance matrix, in which case it will be singular or close to singular. To circumvent this problem, PCA can be used to reduce the dimensions of the data after which LDA or QDA is performed on the PCs. If all non-zero PCs are used in the model, PCA-LDA provides identical results to PLS-DA. However, the main advantage of the latter is its ability to provide good insight into the predictor variables that are behind the discrimination between classes by the examination of the loading vector of the PLS components. The other approaches do not easily relate the classifier to the underlying variables especially when used in conjunction with PCA. It is important to state here that there is another route to discrimination analysis using PLS and it involves a two step-procedure where dimensionality reduction is conducted on the $\boldsymbol{y}$ vector of response values followed by the use of the latent components as predictors in a classical discrimination method [117], [118], [127]. Approaches such as logistic regression, LDA and QDA have been proposed for that purpose.

Finally, since the algorithm for the computation of PLS-DA is the same to PLS-R (either PLS1 or PLS2) sparsity can be imposed as previously described and therefore the corresponding regularized versions (rPLS-DA) have been introduced in the literature: sparse PLS-DA (sPLS-DA) [128], group PLS-DA (gPLS-DA) and sparse-group PLS-DA (sgPLS-DA) [129].

3

The EnviroGenoMarkers Project, Disease Endpoint of Interest and Specifications of Statistical Models

3.1 Study Population

The analysis and results presented in this thesis are based on participants from the *EnviroGenoMarkers* (EGM) project (www.envirogenomarkers.net). EGM is a large-scale EU-funded project aiming at the development of a new generation of biomarkers to study the role of environmental agents in human disease; therefore, it is an exposome project. It corresponds to a case-control study nested within two prospective cohorts: the Italian component of the European Prospective Investigation into Cancer and Nutrition (EPIC-Italy) and the Northern Sweden Health and Disease Study (NSHDS). The main chronic diseases of interest include those where it has been suggested the environment plays an important role in their aetiology, namely breast cancer and Non-Hodgkin's Lymphoma (NHL). The specific disease endpoint of interest in this thesis is B-cell Lymphoma (BCL), a type of NHL.

3.1.1 EPIC-Italy

European Prospective Investigation into Cancer and Nutrition (EPIC) [130] is a multicentre prospective epidemiological cohort study following over half a million healthy participants from middle age onwards. It was established to investigate the relation-

ship between nutritional, lifestyle and environmental factors, and cancer and other chronic diseases. Enrolment of volunteers started in 1992 and a total of 23 centres from ten European countries are currently included in the cohort: Denmark, France, Germany, Greece, Italy, The Netherlands, Norway, Spain, Sweden and the United Kingdom. The Italian sub-cohort, EPIC-Italy, consists of 47,749 volunteers (including 32,157 women) aged 35-70, recruited from five different centres within the country: Varese, Turin, Florence, Naples, Ragusa, between 1993 and 1998 [131]. A standardised self-completed questionnaire was used to collect information on diet and lifestyle, including aspects such as education, socio-economic status, employment, physical activity, reproductive history, disease history, alcohol consumption and tobacco use. Anthropometric measurements were also obtained using standardised methods. In addition, blood samples were extracted in the centres at enrolment and later stored in 0.5ml plastic straws in liquid nitrogen containers at -196°C. More details can be found in [132].

3.1.2 Northern Sweden Health and Disease Study

It includes participants from three different projects: the Västerbotten Intervention Program, the Västerbotten Mammary Screening Program and the Northern Sweden MONICA project [133]. A total of 95,000 healthy individuals aged 40-60 were invited for inclusion in the project between 1985 and 2008. Subjects were asked to complete a self-administered questionnaire to collect demographic, medical and lifestyle information and a separate self-administered food frequency questionnaire at recruitment as well as anthropometric measurements. Blood samples within these cohorts were collected at recruitment in a uniform manner and stored at -80°C within two hours of collection.

3.1.3 Ethical Approval

The EGM study was approved by the committees on research ethics in Florence (EPIC-Italy) and in Umea (NSHDS) in accordance with the Declaration of Helsinki

of the World Medical Association. At recruitment, all participants provided written consent (using centre-specific forms administered in the local language) to provide detailed information on their dietary and lifestyle habits and to have their health status followed for the rest of their lives.

3.1.4 Selection of Cases and Controls

Incident NHL cases were identified through local Cancer Registries (loss to follow-up of enrolled individuals was less than 2%) and occurred between 1 and 17 years after recruitment. For each incident NHL case identified within the two cohorts, one random control was selected among all cohort members alive and free of cancer at the time of diagnosis of the index case matched by cohort, centre, gender, date of blood collection (+/- 6 months), and age at recruitment (+/- 2.5 years). Information from the two studies was integrated into a single database and standardised. NHL cases were classified into subtypes according to the SEER ICD-O-3 morphology codes [133].

The analysis was conducted in two analytical phases: initially 100 case-control pairs were studied (study phase 1), which were supplemented with an additional 181 case-control pairs (147 from NSHDS and 34 from EPIC-Italy) to increase the power of the study (study phase 2). After further subtype characterisation and review, 11 cases were reclassified (Hodgkin's lymphoma (n = 6); T-cell lymphoma (n = 1); and unknown (n = 4)) and excluded from future analysis along with their matched controls. Moreover, two cases without suitable control samples were excluded. A total number of 268 BCL case and 268 healthy controls are retained in the study population. (A descriptive summary of the study population (n = 536) with respect to the main demographic covariates is provided in chapter 4, Table 4.1).

3.2 Disease Endpoint of Interest: B-cell Lymphoma

Lymphomas are solid tumours of the immune system which are traditionally classified as Hodgkin's and non-Hodgkin lymphomas (HL and NHL) accounting for about 10% and 90% of all lymphoma cases, respectively. The type of lymphocyte cell from which they arise is the main characteristic that distinguishes between the two disease entities: while Hodgkin and Reed/Sternberg cells are the hallmark of HL, NHLs arise from either B, T or Natural Killer (NK) lymphocytes with *B-cell Lymphomas* (BCLs) representing the vast majority of NHL cases (85-90%). NHL are also subdivided according to their clinical presentation into indolent (low grade) or aggressive (high grade) which although have a shorter natural history are frequently curable, unlike the indolent tumours which have a long survival but tend to be ultimately fatal. Since B cell malignancies are overwhelming more frequent than T/NK cell neoplasms, BCLs are consequently the disease endpoint of interest of this thesis.

NHL was first described as a separate entity in the late 1940s [134]. Advances in the understanding of biology and genetics as well as the availability of new diagnostic methods and therapies provided a better characterisation of the disease and the ability to identify different subtypes. Several classification systems were proposed that grouped these malignancies according to their histological features including the Rappaport formulation, the Working Formulation, the Kiel classification and the Revised European-American Lymphoma (REAL) classification. In 2001, the WHO produced for the first time a classification that defined all haematological malignancies in terms of immunophenotype, genetic abnormalities and clinical characteristics which was incorporated into the subsequent versions of the International Classification of Diseases for Oncology (ICD-O3) [135], [136]. The WHO classification of haematological malignancies subsequently superseded the previous classification systems and achieved a much-needed international consensus. The 2016 revision [137] is the latest update of the WHO classification and defines a total of 65 different NHL subtypes including 38 BCL subtypes and 27 T- and NK- cell neoplasms.

Thus, NHLs are a heterogenous group of malignancies and include a diverse spectrum of cancers of the immune system. The most common types are Diffuse Large B-cell Lymphoma (DLBCL), Follicular Lymphoma (FL), Chronic Lymphocytic Leukaemia (CLL) and (Mucosa-Associated Lymphoid Tissue (MALT) lymphoma representing 30-40%, 20% and 7% of all B cell malignancies, respectively. While DLBCL has an aggressive presentation causing fulminant symptoms and signs needing prompt assessment and treatment, FL and MALT present an indolent course and can be observed without therapy even for years after first diagnosed; on the other hand, CLL has an indolent presentation in a third of the cases. Often, both FL and indolent CLL progress towards transformation into DLBCL leading to a more aggressive clinical course [138].

3.2.1 Descriptive Epidemiology and Risk Factors

NHL is the most common hematologic malignancy in the world [139]. It is more common in developed countries with an estimated 70,800 new cases in the USA in 2014 and 13,413 new cases in the UK in 2013 [140], [141], [142]. In both countries, this malignancy accounts for around 4% of all cancers and it ranks as the 7th most common cancer among males and the 6th most common cancer among females in the USA while in the UK it occupies the 7th position for both genders [142]. There is a large variation in the geographical distribution of NHL subtypes: a higher incidence of low-grade BCL is seen in high-income regions than in low-income and middle-income regions. In contrast, low- and middle-income regions have a higher incidence of aggressive BCL and T- and NK-cell tumours than high-income regions [143], [144]. For example, extranodal NK–T-cell lymphoma and Burkitt's lymphoma (an aggressive subtype) present a high incidence in east Asia and Africa, respectively and are strongly associated with Epstein-Barr Virus (EBV) infection (i.e. seropositivity prevalence of this virus is high in those regions). On the other hand, a high proportion of FL is observed in North America and Europe.

A long-term increase in the incidence of NHL was observed between the 1950s and

1990s in many high-income countries and is well documented in the literature, however, no further increase has been reported during the last two decades [145]. The reasons for this pattern are unclear, part of the increase has been attributed to AIDS-related tumours following the HIV epidemic. Other factors that have been postulated as potential contributors are changes in classification methods and improvements in diagnostic technologies and cancer registrations [146], [147]. It has been suggested that the introduction of the WHO classification played a role in the upward trend as previously unrecognized lymphoma types were included as NHL subtypes. In addition, refinements in the histomorphological understanding of HL led to a large number of cases being reclassified as NHL [148]. The stabilised incidence observed since early 2000 has been credited to cancer preventative measures and to changes in the prevalence of putative risk factors, such as restrictions in the use of certain chemical carcinogens [149], [150] . Whether this trend pattern reflects overall disease incidence or only changes in specific subtypes remains to be seen as recent evidence shows that the incidence of indolent BCL and T/NK-cell NHL rose considerably whereas that of high-grade BCL remained stable [144], [151].

Factors affecting an individual's risk of developing NHL have been extensively studied. As this is a malignancy of the immune system, the most consistent and strong associated risk factor is congenital or acquired immunosuppression. Individuals with HIV have an increased risk of developing high-grade NHL involving extra nodal sites, some of which are specific to HIV/AIDs patients (plasmablastic DLBCL) and other which also manifest in HIV negative patients [143], [152]. Others at increased risk include organ-transplant recipients, patients who have had high-dose chemotherapy with stem-cell transplantation, and those with inherited immunodeficiency syndromes. Furthermore, several autoimmune disorders have been associated with increased risk of NHL, including rheumatoid arthritis, celiac disease, systemic lupus erythematosus and Sjögren's syndrome; it is unclear whether this increased risk is related to the autoimmune disorder itself or to the immunosuppressive therapies used in its management [153].

Infection also plays a part in development of some lymphomas, either by inhibition of immune function or by other mechanisms, such as induction of chronic inflammatory response. In addition to the previously mentioned associations between NHL subtypes and HIV and EBV infections, other infection agents have been closely connected with the development of NHL. *Helicobacter pylori* causes most gastric MALT lymphomas [154] while Hepatitis C virus has been implicated with splenic Marginal Zone B-cell Lymphoma (MZL) and DLBCL [155], [156]. In these cases, patients with the concomitant infection respond positively (regression of the lymphoma) to the eradication of the virus. *Borrelia burgdorferi* (a tick-borne spirochete that causes Lyme disease) and *Chlamydia psittacosis* (a bacterium transmitted from pet birds to humans causing influenza-like symptoms) are thought to be associated with the development of MZL [157], [158]. *Coxiella burnetii* (a bacterium that infects animals such as goats, sheep, and cattle and that in humans causes Q fever) has been proposed as a risk factor for FL and DLBCL [159].

Other risk factors have been postulated in the literature but either a weak association is shown, evidence implicate them in an inconsistent manner or the percentage of NHL cases they account for is small. Increased risk among individuals with relatives previously diagnosed with NHL or any family history of any hematopoietic malignancy has been reported [160], [161], however familial aggregation is rare. A number of medications including statins, NSAIDs, various antibiotics, phenytoin (a commonly used anticonvulsant drug), antidepressants and benzodiazepine have been implicated by some studies [145], [162], although it is difficult to separate their effects from those of the underlying condition that prompt treatment. Blood transfusion is another medical intervention associated with a higher risk, but evidence remains inconclusive. The effects of some key risk factors such as hair dyes seem to be decreasing as a consequence of the changes in the chemicals used in these products [150]. Finally, lifestyle and anthropometric factors such as cigarette smoking and obesity have been implicated with specific subtypes (FL and DLBCL, respectively) [163], [164].

3.2.2 Pathophysiology and Genetics

To understand the cellular origin of human BCL, the mechanisms that drive normal B cell differentiation and activation should be considered. Normal B cell development starts in the bone marrow where multipotent hematopoietic stem cells give rise to lymphoid precursors that initiate an irreversible differentiation program in order to express a functional and unique *B Cell Receptor* (BCR) through a gene remodelling process known as V(D)J recombination. After a series of selection developments, only the cells with a positively selected functional BCR leave the bone marrow microenvironment while the rest undergoes apoptosis. Mature naïve B cells circulate as small, resting lymphocytes in peripheral blood and secondary lymphoid tissues (spleen, lymph node and mucosa-associated).

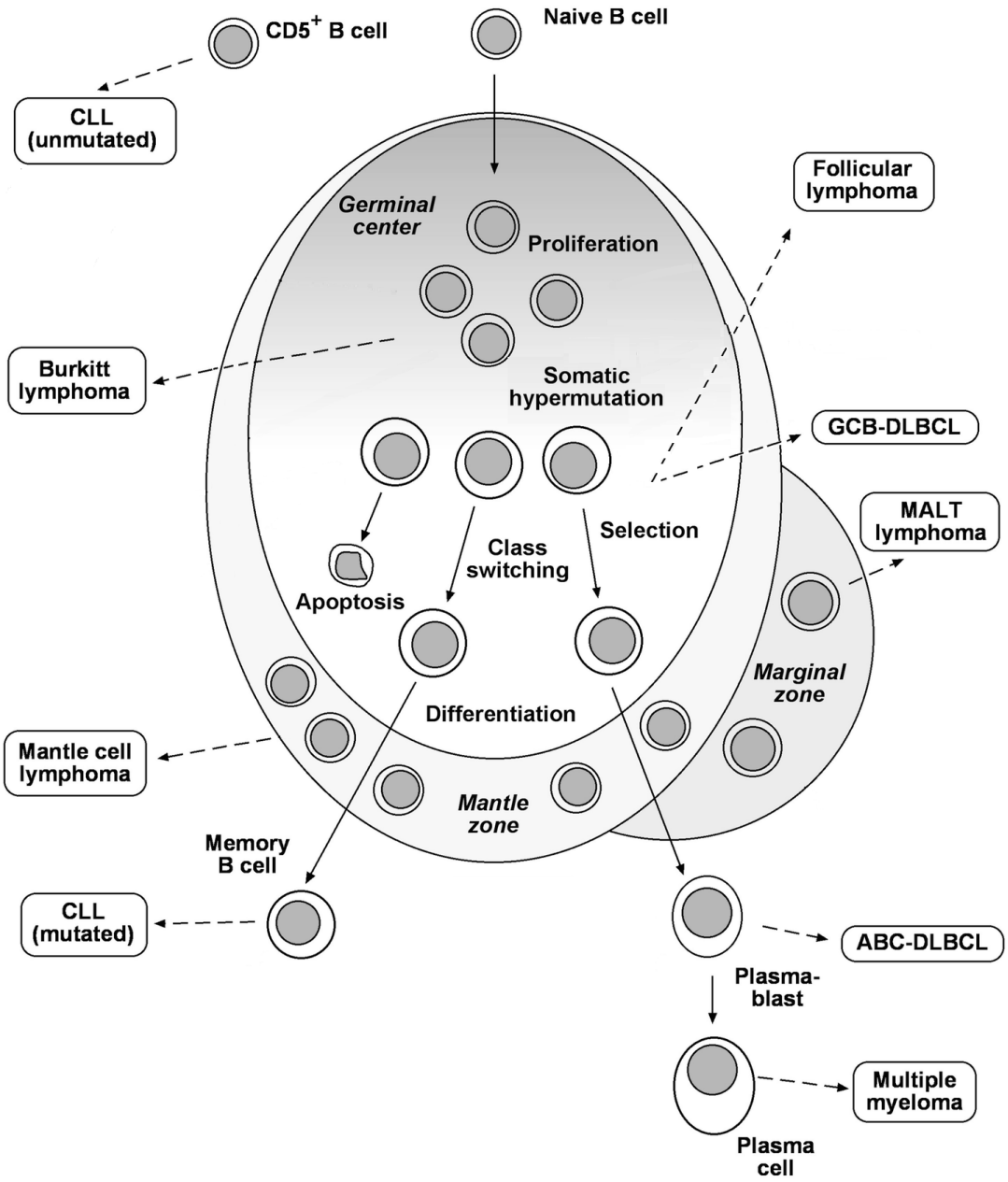
Normal B cell activation occurs in these secondary organs where mature cells encounter antigen and, with the help of T cells, form primary and subsequently secondary lymphoid follicles, the latter being characterised by Germinal Centre (GC)s, which are the histological structures of affinity maturation dedicated to the generation and the selection of B cells that produce high-affinity antibodies. Two different functional compartments can be distinguished in the GC: a dark zone mainly consisting of proliferating GC B cells (centroblasts) and a light zone where GC B cells (centrocytes) are resting. The antibody diversity supported by the GC occurs as a consequence of two additional gene remodelling processes, known as somatic hypermutation (dark zone) and class-switch recombination (light zone). The outcome of the GC reaction is the generation of both memory B cells and plasma cells which leave the GC microenvironment after the differentiation and activation processes have concluded.

Although the gene remodelling processes mentioned above play a central role in normal B cell development, they are mechanisms prone to errors as breaks in the double-stranded DNA (dsDNA) are required [165]. Chromosomal translocations and inactivating mutations in tumour suppressor genes arise when failures in the DNA repair

machinery occur and these genetic aberrations are known to underlie lymphomagenesis [138], [166], [167]. Indeed, unlike most solid tumours which typically have substantial genetic instability, lymphomas generally have a stable genome and several genetic lesions are well-documented to be hallmarks of specific BCL subtypes. The translocations typically implicated involve recombination between Immunoglobulin (Ig) loci on one side and a proto-oncogene locus on the other, with the former being an actively transcribed gene in B cells since the Ig polypeptides are part of the BCR structure. As a result, a proto-oncogene is placed under the influence of active Ig promoters or enhancers deregulating the oncogene expression. Such monotypic translocation patterns are a footprint characteristic in subtypes such as FL, Burkitt and mantle cell lymphomas where the vast majority of tumour cases present specific genetic lesions. For example, over 90% of FLs result from the t(14;18)(q32;q21) translocation which sustains the continuous expression of the BCL2 oncogene. Other subtypes such as DLBCL, MALT lymphoma and Multiple Myeloma (MM) exhibit diverse translocations lesions reflecting a more diverse phenotype of those subtypes.

On the other hand, the GC is known to be the source of many types of lymphoma including DLBCL, FL and Burkitt lymphoma with malignancies arising from GC B cells that are "frozen" at a particular stage of differentiation [138], [166], [167]. For instance, the last subtype mentioned seems to derive from dark zone B cells while in the case of DLBCL two main biological distinct entities have been identified, a subtype resembling light zone GC B cell (Germinal Centre B cell DLBCL (GCB-DLBCL)) and another subtype derived from GC cells arrested during the early stages of post-GC plasma cell differentiation (Activated B cell DLBCL (ABC-DLBCL)). Entities with a post-CG origin include CLL and MM which are presumably derived from memory and plasma cells, respectively. Figure 3.1 provides an illustrative summary of the cellular origin of the most common human BCLs and the stage of the differentiation process from which they arise. The finding that the cellular origin of human BCL resides in the GC can be explained by two critical factors: the vigorous proliferation that B cells undergo during selection and differentiation and the recombination processes

Figure 3.1: B cell maturation in the germinal centre and cellular origin of the most common human BCLs.



BCLs originate from cells that are blocked at different stages of maturation: Burkitt lymphomas resemble dark zone B cells; FL and GCB-DLBCL originate from light zone B cells; and ABC-DLBCL shows characteristics of late GC B cells (plasmablasts) that are committed to plasma cell differentiation. Two CLL subgroups can be distinguished: one with a unmutated IgV genes and one with IgV-mutated lesion arising from different cell differentiation states. Illustration taken and modified from Seifert et al. [167].

of somatic hypermutation and class switching which are mutagenic mechanisms that strongly increase the risk for a B cell to undergo malignant transformation. The last

factor could also partly explain why B cells malignancies are more frequent than T cells [166].

3.3 Omics Measurements in the EGM Project

Fresh blood samples from healthy volunteers were collected using two different anti-coagulants, citrate (EPIC-Italy) or Ethylene Diamine Tetraacetic Acid (EDTA) (NSHDS) and processed for fractionation by centrifugation (15 minutes at 1,500g at room temperature) within two hours after collection for the isolation of buffy coats, erythrocytes and plasma. Fractions were then aliquoted and immediately stored in liquid nitrogen at -196°C (EPIC-Italy) or -80°C (NSHDS) to be later transported on dry ice to the laboratory and stored again for a short period at -80°C before omics analyses. To reduce the impact of technical-induced variation, matched case-control pairs were analysed next to each other in the same plate and batch. As matching was performed by sample date, each case has the same storage time (+/- 6 months) as its matched control. Furthermore, laboratory personnel were blinded in relation to case-control status and quality control samples were included in the analysis with the case-control sets.

From peripheral blood mononuclear cell (PBMC) samples, targeted proteomic, transcriptomic and DNA methylation (DNAm) profiles were acquired; the different platforms used for each of the omics measurements and the number of features they assay are described as follows:

1.- Targeted proteomics:

   – MILLIPLEX®HCYTOMAG-60K and HSCYTMAG-60SK kits.

   – A panel of inflammation-related proteins (n = 32).

2.- Gene expression:

   – Agilent 4x44K human whole genome microarray.

  - A total of 29,662 transcripts.

3.- DNAm:

  - HumanMethylation450 BeadChip (HM450).
  - A total of 485,577 CpG sites.

These platforms correspond to a Multiplexed Bead Assay (MBA), an RNA microarray and a DNA bisulphite conversion method, respectively. (See section 1.1 for more details on the functioning of these platforms). All measurements were conducted according to the protocol described by the manufacturer. All 268 case-control pairs have full proteomic measurements while 232 (86.57%) pairs have full-resolution gene expression data. DNAm profiles were available for a subset of those pair sets: 199 (74.25%) and 176 (75.86%) case-control pairs have complete epigenetics data in addition to the proteomics and transcriptomics data, respectively. Details of the successfully analysed bio-samples of the omics profiles under study by main disease subtypes are summarised in Table 3.1. In the proteins analyses, four analytes (IL-12, IL1-RA, sIL2-RA and Flt3ligand) were excluded from further statistical analyses due to a high rate of non-detects (>75%). The final number of measurements under study corresponds to 28 immune markers: 10 chemokines, 12 cytokines and 6 growth factors. Details of the final analytes measured in the assay are presented in Table 3.2.

Appropriate data pre-processing steps were performed to each of these three omics measurements before analitical research. For proteomics, concentration levels were log-transformed ($\log_2$) to normalise their distributions for all statistical analyses. Measurements out of the range of the calibration curve were imputed based on a Maximum Likelihood Estimation (MLE) method which was informed by the observed correlation structure within the data [168]. Imputation of samples with concentration levels below the Limit of Detection (LOD) (see Table 3.2) was carried out using the empirical LOD across all plates as the upper bound. Imputation of samples with a concentration exceeding the calibration curve was carried out using a value of twice the highest empirical concentration that was not out of range of the calibration curve

Table 3.1: Number of study participants with successfully analysed proteomics and both proteomics and epigenetics samples and with successfully analysed transcriptomics and both transcriptomics and epigenetics samples.

Disease Subtype	*Proteomics n (%)*	*Proteomics and Epigenetics n (%)*	*Transcriptomics n (%)*	*Transcriptomics and Epigenetics n (%)*
CLL	42 (7.84)	24 (6.03)	34 (7.33)	18 (5.11)
DLBCL	44 (8.21)	35 (8.79)	37 (7.97)	30 (8.52)
FL	39 (7.28)	29 (7.29)	37 (7.97)	29 (8.24)
MM	76 (14.18)	62 (15.58)	67 (14.44)	57 (16.19)
Others	67 (12.5)	49 (12.31)	57 (12.28)	42 (11.93)
Controls	268 (50)	199 (50)	232 (50)	176 (50)
Total	536 (100)	398 (100)	464 (100)	352 (100)

As detailed in the following sections, proteomics and transcriptomics measurements are analysed independently in relation to case-control status while epigenetics measurements are employed to indirectly estimate cell-type composition only. Therefore, DNAm profiles are used to assess the possible confounding effect that intra-sample variation introduces in the concentration of inflammatory markers and levels of gene expression signals (see section section 3.4 below).

as the upper bound.

As per transcriptomics, the intensity levels were also log-transformed ($\log_2$) to normalise their distributions before conducting statistical analyses. The technical performance and quality of the microarrays was assessed by visual evaluation of the scan images before and after within- and between-array normalisation which was performed using the LOcally WEighted Scatterplot Smoothing (LOESS) and A-quantile methods, respectively [169]. A description of these approaches is provided in section A.1. The missing values of probe intensity were imputed using the k nearest neighbours approach by which the missing data is replaced by the average value of the k nearest patterns (k = 15, Euclidian metric) [169].

In contrast to RNA microarray experiments where pre-processing steps are well established and documented, there is no clear consensus on how to perform these procedures in the analysis of DNAm using HM450 bead chip arrays [170], [171], [172], [173]. Reasons explaining the difficulties associated with pre-processing steps in DNAm are discussed in section A.1. In the case of the analysis of the DNAm data

Table 3.2: Panel of the final 28 analytes measured in the MBA kit.

Protein (abbreviation)	*LOD (pg/mL)*	*Protein (complete name)*	*Molecular Function*
EGF	1.73	Epidermal growth factor	Growth factor
FGF2	2.21	Fibroblast growth factor 2	Growth factor
GCSF	0.71	Granulocyte colony stimulating factor	Growth factor
VEGF	8.83	Vascular endothelial growth factor	Growth factor
GMSCF	0.08	Granulocyte-macrophage colony stimulating factor	Growth factor
TGFa	0.28	Transforming growth factor alpha	Growth factor
Eotaxin	1.20	—	Chemokine
Fractalkine	1.30	—	Chemokine
GRO	3.80	Growth-related oncogene	Chemokine
MCP1	0.73	Monocyte chemotactic protein 1	Chemokine
MCP3	1.79	Monocyte chemotactic protein 3	Chemokine
MDC	2.25	Macrophage derived chemokine	Chemokine
MIP1a	0.45	Macrophage inflammatory protein 1 alpha	Chemokine
MIP1b	1.75	Macrophage inflammatory protein 1 beta	Chemokine
IP10	2.81	Induced protein 10	Chemokine
IL8	0.07	Interleukin 8	Chemokine
IL1b	0.08	interleukin 1 beta	Cytokine
IL2	0.09	Interleukin 2	Cytokine
IL4	0.11	Interleukin 4	Cytokine
IL5	0.04	Interleukin 5	Cytokine
IL6	0.02	Interleukin 6	Cytokine
IL7	0.08	Interleukin 7	Cytokine
IL10	0.10	Interleukin 10	Cytokine
IL13	0.04	Interleukin 13	Cytokine
INFa	1.54	Interferon alpha 2	Cytokine
INFg	0.14	Interferon gamma	Cytokine
TNFa	0.01	Tumor necrosis factor alpha	Cytokine
sCD40L	—	Soluble CD40 ligand	Cytokine

MBA:Multiplex Bead Assay, LOD: Limit of Detection.

used in the EGM project, the specific steps were carried out as follows: i) exclusion of samples and probes with high rate of non-detected probes and samples (respectively), ii) dye bias correction, iii) normalisation and iv) Combatting Batch Effects (ComBat) correction (see section 3.5 for more details).

Finally, similar to proteins concentration levels the missing values for covariate of interest were also imputed based on a MLE method [168]. These include the following: Body Mass Index (BMI) (n = 8, 1.5%), smoking status (n = 14, 2.6%), education (n = 16, 2.98%), alcohol intake (n = 41, 7.65%) and physical activity (n = 2, 0.37%).

3.4 Cell-type Heterogeneity: White Blood Composition Correction

Omics profiles derived from complex tissues such whole blood or solid tumours represent an average change in concentration or expression over many different cell types. Importantly, the cell-type composition of complex tissues also varies in response to phenotypes such as cancer or age [174]. Cell heterogeneity is thus a confounding factor and correcting for this effect becomes crucial if one wishes to identify potential causal alterations between omics measurements and a specific phenotype of interest. In the case of the EGM project, omics profiles were generated using PBMC samples which is a mixture of different cell types, mainly two types T lymphocytes (CD4 and CD8 cells), B cells, NK cells, monocytes and granulocytes. In addition, the disease endpoint of interest of this study is lymphoma whose target tissue is blood with one clinical manifestation being the pathological alteration of the White Blood Cell (WBC) composition. Therefore, cell-type heterogeneity adjustment is a key step to be considered in the statistical analyses of proteomic and transcriptomics profiles in relation to BCL.

Quantification of the WBC composition could be accomplished with the use of freshly drawn venous blood that is immediately prepared in a specially equipped laboratory [175]. It requires labour-intensive and expensive steps (flow cytometric measurements based on protein membranes to distinguish specific leukocyte subtypes) that become impractical to be used in wide-scale epidemiological studies [175]. In the absent of the actual cell counts for each study sample, the immune cell composition can be estimated using deconvolution approaches that estimate putative numbers

and proportions of cell types making use of either DNAm or gene expression profiles as surrogate variables. Typically, DNAm provides more reliable estimates than RNA-based approaches because of the fundamental role this mechanism plays in lineage cell-type differentiation [176]. Deconvolution algorithms can be classified into reference-based and reference-free depending on whether the statistical approach requires a priori defined reference DNAm profiles of cell types that are present in the tissue of interest (i.e. regions of the genome known to be differentially methylated across cell types).

Houseman et al. [175] introduced the first reference-based deconvolution algorithm which has been demonstrated to perform well on blood (whole and PBMC), a tissue for which the cell-specific composition is well-established and for which reliable DNAm reference profiles have been generated. For this reason, in this thesis the Houseman algorithm was employed in conjunction with the DNAm profile reference database made available by Reinius et al. [177] (purified human leukocytes from six healthy male blood donors) in order to estimate the proportions of the WBC subpopulation of interest: T lymphocytes CD4 and CD8, B cells, NK cells, monocytes and granulocytes. (A brief description of the approach is provided in section A.2). These inferred proportions were then included as additional covariates in the univariate analyses performed using proteomics and transcriptomics in relation to BCL case-control status (see section 3.6.1 below). The WBC correction thus allows the identification of inflammatory markers or gene transcripts whose changes in concentration or expression are not driven by underlying changes in cell-type composition.

3.5 Batch Effects: Correction for Technical-induced Variation

Intra sample heterogeneity is not the only confounding factor that may affect the reliability of omics analysis, systematic variation introduced by technical noise can also dilute the biological effect of interest. These technical artefacts are known as batch ef-

fects and arise by differences in the experimental conditions (e.g. atmospheric ozone levels, temperature, humidity, pH, etc.), reagent quality, laboratory personnel, among others [178]. In an ideal experimental setting, all samples would be processed in one go, however due to the large number of samples and the restrictions imposed by the laboratory equipment, this is seldom the case. For example, in a typical microarray experiment 12 samples are analysed on one chip, eight chips (96 samples) form one plate, and two plates can be placed in the hybridisation oven at the same time. Consequently, batch effects cannot be avoided in studies comprising a large number of subjects and removal of these effects is necessary for a robust analysis. The data pre-processing steps mentioned previously such as subtraction of background noise and within- and between-array normalisation techniques do not adjust for inter-batch variability.

Several approaches have been proposed for batch effect removal, including linear models [179], Linear Mixed Models (LMM)s [180], [181], [182], ComBat [183], Surrogate Variable Analysis (SVA) [184] and Independent Surrogate Variable Analysis (ISVA) [185]. The first three methods assume that the variables responsible for the technical variation are known while the last two approaches model potentially unknown confounding factors by estimating surrogate variables which are derived from a residual matrix (a matrix whose signal due to the primary variable(s) of interest has been removed). The estimated parameters from all these models are used to calculate new concentration or expression levels that are corrected for batch effect or technical variation (in other words, to obtain the batch-adjusted or "de-noised" measurements).

Common variables that have been linked to introduce batch effects are processing groups and dates of the main experimental processes, and their potential impact can be assessed by performing Principal Component Analysis (PCA), hierarchical clustering or by plotting individual features versus these putative technical confounders [178]. These identified variables are then employed as standard covariates in a linear model, random effect terms in a LMM or modelled through ComBat. On the other

hand, approaches such as SVA or ISVA can be employed when the true sources of technical variation are unknown or cannot be adequately identified; they aim at the construction of (independent) surrogate variables responsible for technical induced-variation which can be later included as covariates in future analyses. More details on these three methods is provided in section A.3.

Under the assumption that the main drivers of unwanted variation were known, the method of choice to correct for technical-induced variation in this thesis was LMMs and the model specifications are described in more details in the following section.

3.6 Statistical Methods to Analyse EGM Data

3.6.1 Linear Mixed Model: Model Specifications

In chapter 4, LMMs including a random intercept term only (see section 2.1.3 for more details) were applied to proteomics and transcriptomics datasets in order to i) investigate the relationship between each marker separately and the disease outcome and ii) attenuate the potential effect of technical-induced noise in omics experiments. The general formulation of the model for participant j in group i can be described as follows:

$$\boldsymbol{y}_{ij} = (\boldsymbol{\beta}_0 \ + \ \boldsymbol{u}_{Ai}) \ + \ \boldsymbol{\beta}_1 \boldsymbol{x}_{ij} \ + \ \boldsymbol{\beta}_2 \boldsymbol{FE}_{ij} \ + \ \boldsymbol{\varepsilon}_{ij}, \tag{3.1}$$

where $\boldsymbol{y}_{ij}$ is a continuous variable representing the immune analyte concentrations (proteomics) or gene expression levels of probes (transcriptomics), $\boldsymbol{\beta}_0$ is the intercept of the model, $\boldsymbol{\varepsilon}_{ij}$ is the residual error and $\boldsymbol{x}_{ij}$ is the outcome of interest, a binary variable indicating if individual j is a BCL case or not, with $\boldsymbol{\beta}_1$ being its associated regression coefficient. In the case of transcriptomics data, it is common to express $\boldsymbol{\beta}_1$ as the fold-change (f) using the transformation $f \ = \ 2^{\boldsymbol{\beta}_1}$. $\boldsymbol{FE}_{ij}$ is a vector of fixed effect variables for individual j belonging to group i with the corresponding regression coefficients compiled in the vector $\boldsymbol{\beta}_2$. The following variables are included in the fixed effect term:

- Matching variables for cases and controls: age, gender and cohort (EPIC-Italy or NSHDS).
- The experimental study phase (1 or 2).
- Potential confounders observed in previous analyses of BCL within the EPIC cohort [182], [186], [187], [188]: BMI (continuous variable measured in kg/m2), education (5 classes: none, primary, technical/professional, secondary, university/college), physical activity (4 classes: inactive, moderately inactive, moderately active, active), smoking status at enrolment (3 classes: non-smokers, former smokers, smokers) and alcohol intake at enrolment (continuous variable measured in g/day).

Technical variation was modelled through the random intercept term $\boldsymbol{u}_{Ai}$ which represents the shift associated to A_i, the value of the random effect variable(s) A observed for group i. The following variables were included in the random intercept term:

- The analytical plate on which the sample was processed (proteomics analysis).
- The dates in which the three main laboratory steps of the RNA microarray experiment were conducted: RNA isolation, dye labelling and hybridisation (transcriptomics analysis).

For the first objective, investigate the relationship between each individual marker and BCL status, hypothesis testing was performed using the Likelihood Ratio Test (LRT, see section 2.1.4.3 for more details) comparing the maximum of likelihood of the LMM excluding the outcome of interest to the model including that variable. The test statistics was compared to a chi-squared distribution with one degree of freedom, which is the difference in number of parameters between the two models. For the second objective, reduce the potential effect of technical noise, the estimated parameters from the full LMM specified in Equation 3.1 were used to calculate concentration and expression levels adjusted for technical variation ("de-noised" proteomics and transcriptomics data, respectively) by subtracting the estimated random effect term(s) from the original observed levels.

As a sensitivity analysis, correction for WBC differentials was conducted in the case-control pairs for which DNAm profiles were available (in addition to the proteomics and transcriptomics profiles) by including in the LMM the estimated proportion of the main cell types that conform PBMC samples. The adjustment was performed in two different manners:

- Partial WBC adjustment: each cell proportion (continuous variable from 0 to 1) was included in the fixed effect term along with the covariates described above. In other words, for each protein and for each transcript a total of six models with partial adjustment were fitted. The proportions of the following cells were added in the model: T lymphocytes CD4 and CD8, B cells, NK cells, monocytes and granulocytes.
- Full WBC adjustment: five cell proportions were jointly added as additional confounder variables to the fix effect term along with the covariates described above. Since in this case the proportion of the six cell subpopulations sums to 1, one cell type (here granulocytes) was excluded in order to properly estimate the regression coefficients.

3.6.2 Partial Least Squares: Model Specifications

Partial Least Squares (PLS) approaches are extensively applied to EGM omics data in chapters 5 and 6. In particular, in chapter 5 I employ regularized Partial Least Squares Discriminant Analysis (rPLS-DA) to analyse the proteomics and transcriptomics datasets in order to i) identify proteins and transcripts indicative of future risk of BCL and its main histological subtypes and to ii) investigate the applicability of these novel statistical approaches. Subsequently, in chapter 6 I employ regularized Partial Least Squares (rPLS) to perform two-block integration between the transcriptomics and proteomics datasets in order to i) unravel relevant biological patterns for the disease outcome under study and to ii) thoroughly compare and contrast the applied statistical methods.

The use of these techniques as a statistical methods to analyse high-dimensional data requires the computation of tuning parameters, a process that is usually referred to as model selection or model calibration because every value of the parameter corresponds to a different statistical model with different predictors variables. As outlined in section 2.2.2.3, the number of parameters differs for classical PLS (where no sparsity is imposed in the construction of the components) and for the regularized versions. In the first case, there is only one parameter to tune that corresponds to the number of latent variables, component scores or dimensions H while for the second case there are additional parameters to select that control the regularization level. These extra parameters and the model feature they control can be summarised as follows:

1.- Sparse PLS (sPLS):

  – Number of variables ($\lambda_{1,h}, \lambda_{2,h}$).

2.- Group PLS (gPLS):

  – Number of functional groups ($\lambda_{1,h}, \lambda_{2,h}$).

3.- Sparse-group PLS (sgPLS):

  – Number of functional groups ($\lambda_{1,h}, \lambda_{2,h}$).

  – Number of variables included in the selected functional groups ($\alpha_{1,h}, \alpha_{2,h}$).

In the two-block scenario these model elements must be defined for both the $\mathbf{X}$ and the $\mathbf{Y}$ matrices (hence the subscripts 1 and 2); however, if the $\mathbf{X}$ (or $\mathbf{Y}$) matrix is not divided into K (or L) groups, the number of parameters needed is reduced for gPLS ($\lambda_{1,h}$) and for sgPLS ($\lambda_{1,h}, \alpha_{1,h}$) but remains the same for sPLS. In contrast, for discriminatory purposes (rPLS-DA) such specifications are only required for the predictor matrix $\mathbf{X}$. Both two-block PLS and PLS-DA require these additional parameters to be specified for the H component scores (hence the subscript h). Note that by extension, if λ and α define the number of elements to include in the corresponding models, they also determine the degree of sparsity or parsimony; that is, the number of zero

elements per component.

3.6.2.1 Parameter Tuning

The choice of the optimal values of these three tuning parameters remains an open question as different approaches and metrics of performance are employed in the literature. Commonly, Cross-Validation (CV) is used to choose the optimal value of H, λ and α by means of a sequential strategy in which the value of H is selected first followed by the tuning of λ and α for each latent variable h at a time. In other words, for the regularized versions the optimal degree of sparsity on a given PLS component is dependent on the degree of sparsity selected on the previous components. Depending on the chosen metric of performance, the optimal value of the tuning parameters H, λ and α is defined as the value that minimises, maximises or meets a certain criterion. Several metrics of performance (also called diagnostic statistics or scores) are available for this purpose; they depend on both the PLS variant being performed (discriminant analysis versus two-block) and the PLS mode (regression versus canonical). For the regression case, these can be summarised as follows:

- R^2 or proportion of variance explained in the response $\mathbf{Y}$ (two-block PLS).
- Q^2 statistics (two-block PLS).
- Discriminant Q^2 (DQ^2) (PLS-DA).
- Root Mean Squared Error of Prediction (RMSEP) (two-block PLS).
- Number of Misclassifications (NMC) (PLS-DA).
- Misclassification Error Rate (ER) (PLS-DA).
- Area Under the Curve (AUC) or Area Under the Receiver Operating Characteristic (AUROC) (PLS-DA).

The first three metrics are employed to define the optimal number of components (H) only, either in the classical or the regularized versions. The rest of the statistics have been employed to ascertain the appropriate number of components and/or optimal

degree of sparsity (λ and α). On the other hand, in PLS canonical mode there is a lack of robust statistical criteria to evaluate canonical correlation methods and to determine optimal model specifications; typically, decisions make use of prior biological knowledge and visual inspection and interpretation of the output. In the following sections, I describe in more details the sequential strategy used to optimize the model parameters as well as the metrics of performance used in regression mode (metrics of performance for canonical mode are described in section A.4).

3.6.2.1.1 Sequential Strategy

A thorough parameter tuning strategy is to first define a sufficiently large number of components, for example H = 6 and fit six different PLS models with h = 1 to h = 6 dimensions [189]. CV is then performed on those six models; the one optimizing the value of the chosen diagnostic statistics defines the final number of latent variables. The second step is to determine the optimal degree of sparsity for the chosen number of dimensions which is conducted sequentially for each component at a time. For the particular case of a PLS-DA model, it has been suggested to start the sequential strategy by defining this sufficiently large number of dimensions as G or $G + 2$ components, where G is the total number of classes [190]. Of course, computational constraints must be considered in both scenarios and are of special importance in setting where $n << p$ (PLS-DA) or $n << p + q$ (two-block PLS). As such, there are differences in the sequential strategy depending on the specific technique being applied as well as the way in which sparsity is imposed and they are discussed below.

Imposing sparsity in one block only or PLS-DA : For the sake of simplicity, let address first the two-block PLS scenario where one wishes to impose sparsity on the $\mathbf{X}$ matrix only, which is equivalent to a PLS-DA model. For a sPLS(-DA) model where the aim is to define the optimal number of variables, CV is performed on p different models retaining 1 to p variables, where p is the total number of variables in the predictor matrix. For a gPLS(-DA) model where one seeks to define the optimal number of functional groups, CV is performed on K different models retaining 1 to K

functional groups, where K is the total number of functional groups in the predictor matrix. For a sgPLS(-DA) model, a more complex scenario is faced as the number of different models whose performance is assessed is contingent upon K as well as the desired degree of sparsity inside the selected functional groups (as discussed in section 2.2.2.4, the tuning parameters α can take any values between 0 and 1, $\alpha = 0$ is equivalent to gPLS and $\alpha = 1$ to sPLS). Once the process is concluded for the first component, the procedure is repeated for the second component with an optimal degree of sparsity already defined for the first latent variable. Thus, the whole tuning parameter process finalises with an optimal number of components as well as optimal degree of sparsity for those chosen dimensions.

Expectedly, in a high dimensional setting assessing the performance by means of CV on all p or K models becomes impractical, therefore a grid of values can be used instead. The grid needs to be carefully chosen to achieve a balance between resolution and computational time [191]. On the one hand, one should consider the minimum and maximum values of the selection size that can be handled practically for follow-up analyses. On the other hand, the computational aspect should also be considered as repeated CV is a time-consuming procedure when dealing with high-dimensional data. In the particular case of sgPLS(-DA), a two-dimensional grid is employed also applying a sequential strategy; that is, for any $1 \leq k \leq K$ different values of $\alpha_{1,h}$ are tested [116].

Imposing sparsity on the two blocks of data : When the aim is to introduce sparsity simultaneously on $\mathbf{X}$ and $\mathbf{Y}$, a two-dimensional grid must be used for both sPLS and gPLS while a four-dimensional grid is necessary in the case of sgPLS [116]. Computational practicalities restrict such a search; therefore, a coarse tuning grid can be assessed first to evaluate the likely boundaries of the values that define the model specifications before setting a finer grid [190]. It is worth to mention that prior biological knowledge can also be taken advantaged of to counterbalance these computational limitations and to propose sensible values of the tuning parameters. It

has also been proposed to choose all hyperparameters (sPLS: $\lambda_{1,h}, \lambda_{2,h}$, gPLS: $\lambda_{1,h}, \lambda_{2,h}$ and sgPLS: $\lambda_{1,h}, \lambda_{2,h}, \alpha_{1,h}, \alpha_{2,h}$) exclusively on subjective grounds making use of the available biological information to determine model specifications [84], [116], [128]. This alternative approach avoids the restrictions associated to the sequential strategy previously described, however, it may not be suitable when the prior knowledge is lacking.

Cross-validation : CV [82] is at the core of any parameter tuning process and correspond to a resampling technique mainly used for the purposes of model assessment and model selection. The former refers to the evaluation of the prediction performance of a statistical method, the latter to the selection of the appropriate level of complexity (such as the adequate number of predictor variables, as previously mentioned). Here, it is important to emphasize the distinction between two different types of prediction errors, namely, training and test errors. They both quantify the extent to which the predicted response value differs to the true response value from a given observation, the difference lies in the set of observations used to fit the statistical model. The training error is the average error that results from using a method to predict the response variable on the *same* observations that were used for its training while test error results from the predictions made on *new* observations, that is, measurements that were not used in fitting the statistical model. The subset of samples that contain the observation used to fit the statistical method is known as *training data* and the subset that is solely used for validation purposes is called *test or validation data*. CV is a technique used for estimating the test error using a training data and is one of the multiple methods available to estimate (either directly or indirectly) the test error rate in the absence of an independent designated test set. The type of test error rate estimated depends on the nature of the response, for continuous variables RMSEP is used while for categorical variables the misclassification ER is calculated.

There are two ways of conducting CV: Leave-one-out CV (LOOCV) and k-fold CV. Both approaches involve splitting the set of observations into two parts of different

sizes, where the training set comprises more observations than the test set. In LOOCV only a single observation is used for the validation set $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ and the remaining observations make up the training set. The statistical learning method is fit on the $n-1$ training observations, a prediction $\hat{\boldsymbol{y}}_i$ is made for the excluded observation based on its value $\boldsymbol{x}_i$ and a prediction error is obtained $\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i$. The procedure is repeated n times where each observation is iteratively excluded from the training set producing as a result n values of prediction error. The LOOCV test prediction error is the average across those n values. On the other hand, k-fold CV randomly splits the set of observations into k groups or folds of approximately equal size. The first fold is treated as a test set and the statistical method is fit on the remaining $k-1$ folds. A prediction error is computed for the observations in the held-out-fold and the average is obtained. This procedure is repeated k times; each time, a different fold is left out and treated as a test set. The k-fold CV error estimate is computed by averaging the k mean prediction error. Note that LOOCV can be viewed as a special case of k-fold CV when $k = n$.

Unless the number of samples is small, k-fold CV is usually preferred over LOOCV because it gives more accurate estimates of the test prediction error; the reason for this advantage is rooted in the bias-variance trade-off. Error due to bias is lower in LOOCV because the training dataset is larger than in k-fold CV, the corresponding sizes are $n-1$ and $(k-1)n/k$, respectively. However, for any $k < n$ LOOCV presents higher variance because the prediction error is calculated based on only one observation (for example, for an outlier observation the prediction error will be significantly different from the other $n-1$ estimates). Given these considerations, k-fold CV is typically performed setting k = 5 or k = 10, as it has been empirically shown that these values yield test error estimates that suffer neither from excessively high bias nor from very high variance. In order to obtain more robust estimates, a common strategy is to perform repeated k-fold CV (or M k-fold CV) where in each repetition, the folds are split in a different way and the final estimate of the test error is retrieved by calculating the average over the M replications. Finally, Cross-Model Validation

(CMV) also corresponds to an extension to the CV method that is commonly applied in the literature and is described in more details in section A.5.

3.6.2.2 Metrics of Performance for Regression Mode

The diagnostic statistics that have been proposed for model selection in the context of PLS regression mode are R^2, Q^2 and RMSEP [107], [111], [192], [193]. These three scores have been employed to decide the optimal number of components while only the RMSEP has been additionally implemented to determine the degree of sparsity. The R^2 is a common statistic used to measure model accuracy or goodness of fit in linear regression, that is, the extent to which the model fits the data. It corresponds to the proportion of variance explained in the response $\mathbf{Y}$ that can be explained by the predictor variables or equivalently, the amount of variability that is left unexplained after performing the regression. Its definition is given by:

$$R^2 = \frac{TSS \ - \ RSS}{RSS} \ = \ 1 - \ \frac{RSS}{TSS} \tag{3.2}$$

where $TTS = \sum (y_i \ - \ \bar{y})^2$ is the total sum of squares and $RSS = \sum_{i=1}^{n} (y_i \ - \ \hat{y}_i)^2$ is the residual sum of squares. (I discuss in detail how to obtain the predicted values $\hat{y}_i$ from a PLS-R model in section 3.6.2.3.2 below). As it is a proportion, its value lies between 0 and 1, numbers closer to one indicate better model accuracy. It is somewhat challenging to decide on a "good" value for R^2 as it depends on the particularities of the data. In addition, R^2 is a statistic that measures the training error; therefore, it will always increase as the model complexity increases and choosing this criterion to decide the appropriate H of may lead to overfitting. However, despite those shortcomings R^2 is used as a criterion to decide the optimal H because of its simplicity and computationally efficiency. The statistics is calculated for models with increasing number of components and a decision is made based on whether the addition of an extra dimension substantially increases the proportion of variance explained in the response variable, following a similar principle behind the choice of principal components (PCs) in Principal Component Analysis (PCA).

The cross-validated R^2 or Q^2 statistics (also known as Q^2 criterion) was proposed as an alternative to overcome potential overfitting problems as it measures the overall predictive ability of a model while simultaneously assessing the marginal contribution of the $\mathbf{X}$ scores to the predictive power of a PLS-R model. The Q^2 is calculated for each $\mathbf{Y}$ variable and for models of increasing sizes (number of dimensions) as follows:

Algorithm 2 Computation of Q^2 statistics

Input: $\mathbf{Y}(n \times q), \hat{\mathbf{Y}}(n \times q), H$.
Output: Q^2.

for $h = 1, \ldots, H$ **do**
  for $k = 1, \ldots, q$ **do**
    $RSS_h^k \leftarrow \sum_{i=1}^{n} (y_i - \hat{y}_{hi})^2$
    $PRESS_h^k \leftarrow \sum_{i=1}^{n} (y_i - \hat{y}_{h(-i)})^2$
    Store RSS_h^k and $PRESS_h^k$
  end for
  $PRESS_h \leftarrow \sum_{k=1}^{q} PRESS_h^k$
  $RSS_h \leftarrow \sum_{k=1}^{q} RSS_h^k$
  Evaluate $Q^2 \leftarrow 1 - \frac{PRESS_h}{RSS_{h-1}}$
end for

$PRESS_h^k$ is the Prediction Error Sum of Squares (PRESS) and RSS_h^k is the RSS both defined for the variable k and the PLS dimension h. The term $\hat{y}_{h(-i)}$ is the predicted value for the i^{th} observation estimated from the model fitted to all observations but the i^{th}. As such, its computation is effectively LOOCV but can be equivalently defined for k-fold CV. The criterion to determine if a new $\mathbf{X}$ component is considered significant for the prediction of $\mathbf{Y}$ is:

$$\sqrt{PRESS_h} \leq 0.95\sqrt{RSS_{h-1}} \iff Q_h^2 \geq 0.0975 \tag{3.3}$$

In other words, it has been suggested to define the optimal number of components as the first h that $Q_{h+1}^2 < 0.0975$ [194]. This is because a "good" value for Q^2 is a value that is close to the R^2, which reflects a model with good prediction accuracy independently of the specific data that was used to train it. Note that Q_h^2 refers to

the predictive ability of a PLS model including components 1 to h for any $h < H$, where H is the total number of components being explored in the calibration process. Therefore, in this context it does not refer to the predictive ability of single component (which is computed as $Q_h^2 = 1 - PRESS_h/RSS_h$)

The Q^2 is closely related to the RMSEP, the difference being that the former averages the predictions errors across the variables in $\mathbf{Y}$ and thus gives a more general insight of the model as a whole whereas the latter requires to be computed for each variable k in $\mathbf{Y}$. The RMSEP is defined as follows:

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}} \tag{3.4}$$

To select the optimal H and/or degree of sparsity a validation or calibration curve plot can be used, in which the cross-validated RMSEP is plotted against different values of the tuning parameter [192]. Usually, one takes the first local minimum rather than the absolute minimum in the curve, to avoid over-fitting and to favour sparsity. Note that the Mean Squared Error of Prediction (MSEP) is employed interchangeably with RMSEP.

3.6.2.3 Metrics of Performance for Discrimination

Before describing in more details the different diagnostic statistics used to calibrate a PLS-DA model, it is important to highlight that it is commonly suggested in the literature to define the number of dimensions to $H = min(p, G - 1)$, where p is the total number of predictor variables and G is the total number of classes [128]. This is given by the resemblance between PLS-DA and Linear Discriminant Analysis (LDA) as in the latter case the number of discriminant vectors is chosen following that criterion.

As discussed in section 2.2.2.5, PLS-DA is equivalent to conduct Partial Least Squares Regression (PLS-R) previous transformation of the response vector followed by the application of a Decision Rule (DR) to translate the predicted value into meaningful

class membership. Thus, in theory any of the diagnostic statistics employed in a regression setting could be used for optimization of a PLS-DA model; however, in practice adaptations are needed and different scores are employed.

3.6.2.3.1 Discriminant Q^2 statistics

Consider the two-class problem where the groups are labelled as 0 and 1 and the discrimination border is set to 0.5. When a sample of class 1 receives a class prediction of +1.5, it corresponds to a perfect class prediction, however, it is still considered as an error if computing the Q^2. The Discriminant Q^2 (DQ^2) statistics [195] is an adaptation of the original statistics in which the prediction error is disregarded when the class prediction is beyond the class label. In this manner, correct class predictions do not contribute to the estimate of prediction error. For its calculation, the PRESS is redefined to PRESSD and the computation of DQ^2 is given as before:

$$PRESSD_h = \sum_{-1<\hat{y}_i<+1} (y_i - \hat{y}_{h(-i)}) \tag{3.5}$$

$$DQ_h^2 = 1 - \frac{PRESSD_h}{RSS_{h-1}} \tag{3.6}$$

The criterion to determine the appropriate number of components is the same as per standard Q^2.

3.6.2.3.2 Decision Rules

Recall that in PLS-R and PLS-DA the algorithm returns (among other outputs) a matrix of PLS regression coefficients $\boldsymbol{\beta}^{PLS}(p \times H)$ describing the inner extra relationship between the two matrices in the regression mode. This matrix can be used for prediction purposes given new observations. The predicted values are specified as follows [111], [190]:

$$\hat{\mathbf{Y}}_{new} = \mathbf{X}_{new} * \mathbf{U}\left(\mathbf{C}^T\mathbf{U}\right)^{-1}\mathbf{D} \tag{3.7}$$

$$\mathbf{X}_{new} * \boldsymbol{\beta}^{PLS^T} \tag{3.8}$$

where $\mathbf{U}, \mathbf{C}, \mathbf{D}$ and $\boldsymbol{\beta}^{PLS}$ are derived from the $\mathbf{X}$ and $\mathbf{Y}$ training data sets while $\mathbf{X}_{new}$ is derived from the test set (i.e. a CV procedure is conducted) or from an external validation set. $\mathbf{U}$ is a $p \times H$ matrix containing the loading vectors associated to $\mathbf{X}$, $\mathbf{C}$ is a $p \times H$ matrix containing the local regression coefficients of $\mathbf{X}$ on its H latent components ($\boldsymbol{\xi}_h$) and $\mathbf{D}$ is a $H \times G$ matrix containing the regression coefficients of the columns of $\mathbf{Y}$ on the $\mathbf{X}$ scores ($\boldsymbol{\xi}_h$). (Details were described in section 2.2.2.3 and section 2.2.2.4). Finally, $\hat{\mathbf{Y}}_{new}$ is a matrix of size $n_{new} \times G$ containing the predicted dummy variables.

The predicted values are not bounded to the range of the set of integers $[0, ..., G-1]$; instead, they can take any value in the range of $[-\infty, +\infty]$. In a binary classification problem, a classification threshold of 0.5 can be used when the two classes (i.e. 0 and +1) have similar size and variance [119]; however, other DRs have been proposed for situations where these conditions are not met. Broadly speaking, the diversity of DRs can be divided into two major groups: rules that make use of predicted values $\hat{\mathbf{Y}}_{new}$ or the ones that rely on the predicted $\mathbf{X}$ scores $\boldsymbol{\Xi}_{pred}$. In both groups a particular approach is applied to either $\hat{\mathbf{Y}}_{new}$ or $\boldsymbol{\Xi}_{pred}$ in order to determine the class membership of the test sample. The simplest and most intuitive rule to predict the class of a new observation sample is the naïve method (also called maximum value or maximum distance), which is based on $\hat{\mathbf{Y}}_{new}$. The predicted class is the outcome category whose predicted dummy value is closest to the class labels (PLS1-DA) or the outcome class with the largest predicted dummy value (PLS2-DA) [190]. More sophisticated DRs based on $\hat{\mathbf{Y}}_{new}$ include the identification of either one single fixed point (classification boundary) or two fixed points (boundary line) [196]. In this case, the optimal point(s) are determined arbitrarily or according to specific diagnostic tools such as visual inspection of the predicted value or $\mathbf{Y}$ score plots, Receiver Operating Characteristic (ROC) curve and probability density function. A simple single fixed-point rule in a binary problem is to define a cut-off point halfway between the means of the two groups while the analysis of a ROC curve is better suited for the identification of the ideal threshold in a multi-class problem.

On the other hand, the principle behind the second group of DRs is to convert the predicted $\mathbf{X}$ scores into a particular distance metric which is used to assign samples [190]. The predicted scores are defined as:

$$\mathbf{\Xi}_{pred} \; = \; \mathbf{X}_{new} * \mathbf{U} \left(\mathbf{C}^T \mathbf{U} \right)^{-1} \tag{3.9}$$

where $\mathbf{\Xi}_{pred}$ is a matrix of dimensions $n_{new} \times H$. For the calculation of the distance metric, it is necessary to first define the centroids C_G for all G classes of the observations belonging to the training set, which is conducted based on the $\mathbf{X}$ score matrix $\mathbf{\Xi}_{train}$ ($n_{train} \times H$). Then, the distance between each observation of $\mathbf{\Xi}_{pred}$ and the centroids of each class is calculated using either the Euclidian distance $\sqrt{\sum_{h=1}^{H} \left(\boldsymbol{x}_h \; - \; (C_G)_h \right)^2}$ or the Mahalanobis distance $\sqrt{\sum_{h=1}^{H} \left(\boldsymbol{x}_h \; - \; (C_G)_h \right)^T S^{-1} (\boldsymbol{x}_h \; - \; (C_G)_h)}$, where $\mathbf{S}$ is the variance-covariance matrix of $(\boldsymbol{x}_h - \; (C_G)_h)$ [124], [125]. The predicted class of a new observation is the class for which the distance between its centroid C_G and the predicted scores is minimal.

It has been observed that the naïve method shows good accuracy in a two-class problem where the groups are of equal size but different variance. In a more complex scenario (i.e. two groups of imbalanced sizes) an optimized cut-off point is needed in order to achieve a better performance [196]. In multi-class complex classification problems, it has been described that the $\mathbf{X}$ score approaches (and specifically the Mahalanobis distance metric) provide better predictions than the naïve method [190]. It is important to note here that the score-based DRs consider the prediction in the dimensional space spanned by all H components while the predicted values based DRs consider a single (or two) point(s) estimate(s) using the predicted dummy variables on the last dimension of the model only. In addition, all DRs could achieve zero prediction error if supported by a sufficiently large number of PLS components. The aim is to identify and employ the DR that produces the best performance with the minimum number of dimensions.

3.6.2.3.3 NMC, ERs and AUROC

Once the predicted values of all samples in the dataset have been translated to meaningful class memberships, the assigned class can be compared to the true class membership and classified either as a True Positive (TP), a True Negative (TN), a False Positive (FP) or a False Negative (FN). In that manner, a confusion matrix is computed summarising the prediction ability of the model with the true and predicted classes indicated by the rows and columns, respectively. If there are two classes, the confusion matrix is of dimension 2 x 2 with the number of correct predictions shown on the diagonal elements (class 0 correctly predicted as class 0 and class 1 correctly predicted as class 1; TN and TP, respectively) and the off-diagonal elements are the misclassifications (class 0 predicted as class 1 and vice versa; FP and FN, respectively). More precisely:

		Predicted Classes	
		0	**1**
True Classes	**0**	True Negative (TN)	False Positive (FP)
	1	False Negative (FN)	True Positive (TP)

The number of Number of Misclassifications (NMC) [197] is the sum of the off-diagonal elements: NMC=FP+FN; and it is the most intuitive of all metrics of performance as it simply indicates the number of samples which are wrongly classified by the model. The overall misclassification ER is the sum of the off-diagonal elements divided by the total number of samples

$$ER = (FP + FN)/n \ \in [0, 1] \tag{3.10}$$

while the Balanced Error Rate (BER) [191] is the average of the proportions of wrongly classified samples in each class:

$$BER = \frac{(FP/n_0 + FN/n_1)}{2} \ \in [0, 1] \tag{3.11}$$

where n_0 and n_1 are the number of observations in groups 0 and 1, respectively. Naturally, BER is appropriate in cases of an unbalanced number of samples per class. Apart from these quantities, several other ratios of interest can be derived from the confusion matrix where the two most commonly used are the sensitivity and specificity, especially in assessing the performance of diagnostic tests. The sensitivity (or true positive rate, TPR) of the test is defined as

$$Sensitivity = \frac{TP}{TP + FN} \in [0, 1] \tag{3.12}$$

The specificity (or true negative rate, TNR) of the test is given by

$$Specifity = \frac{TN}{FP + TN} \in [0, 1] \tag{3.13}$$

These definitions show that the sensitivity can be increased by reducing the number of FN and the specificity can be increased by reducing the number of FP. Thus, the sensitivity is a measure of how well the model is able to correctly classify samples of the group 1 (i.e cases) while the specificity measures how well the model can predict samples from the group 0 (i.e controls). Although it would be desirable to maximise both sensitivity and specificity and thus to reduce the total NMC, there exists a trade-off between these two ratios. The ROC curve provides an insight to this trade-off by plotting these two quantities (sensitivity against 1-specificity) for different values of the classification boundary. It provides an spectrum of performance assessments and the corresponding AUC can be interpreted as the probability (values range between 0 to 1) that the model will rank a randomly chosen positive instance higher than a randomly chosen negative one; an AUC of 0.5 (or less) is interpreted as having no practical utility.

3.6.2.4 Performance Assessment of Final Model

In the previous sections, I have discussed how the selection of the optimal tuning parameters (number of dimensions and degree of sparsity) is conducted, a process

that depends on the PLS variant employed to perform the analysis (two-block PLS versus PLS-DA) and the statistical mode (regression versus canonical). As a result of the tuning procedure, the optimal PLS model is specified with the appropriate values of the parameters and the aim then is to assess the performance and interpretability of the calibrated model. In a regression setting, this goal is achieved by assessing the prediction error by means of repeated CV conducted on the complete $\mathbf{X}$ and $\mathbf{Y}$ datasets; any of the diagnostics statistics detailed above can be used for that purpose. In classification problems, it has been observed that CV may lead to overoptimistic result and permutation tests are commonly performed to obtain the distribution of the statistics under the H_0 of no association and thus draw a statistically significance threshold (for a specific statistics to be considered significant, the results obtained from the non-permuted set of samples should fall outside the 95 or 99% confidence bounds of the null distribution obtained from the permuted samples) [197]. Alternatively, when sample size allows CMV can be conducted. On the other hand, when canonical-based methods are applied CV is employed to assess either the correlation or covariance as well as the proportion of variability of the pair of omics matrices explained by the calibrated model.

Furthermore, the relevance of individual features is routinely assessed in order to gain insight into the validity of the final model and to identify the most significant variables explaining the variation between the matrices. Different measures can be used for that purpose, including the standard outputs of the PLS algorithm, namely loading weight vectors and matrix of PLS regression coefficients as well as a statistic introduced specifically for this purpose named Variable Importance in Projection (VIP) [198]. As outlined in section 2.2.2.3, the loading vectors are the weights of the original variables that define the component scores in terms of the deflated matrices, meaning that they determine the contribution of individual variables in each component. Therefore, relevant features are characterised by presenting higher absolute values of their associated loading weights for a given PLS component. However, the interpretation of the loading vectors is impaired by the fact that each feature has H

different weights associated to it. This drawback is avoided when the matrix of PLS regression coefficients is examined, as in this case the relevance of a variable is judged on the basis of the last dimension only. The vector of coefficients is a single measure of association between each predictor and the response and as with the loadings, the variables having small absolute value can be considered as irrelevant. The VIP statistics has the same advantage over the loading weights as the principle behind it is to accumulate the importance of each variable (reflected by the loadings) from each component. For a variable j the VIP score is defined as:

$$VIP_j = \sqrt{p \sum_{h=1}^{H} \left[TSS_h \left(\boldsymbol{u}_{j,h} / \|\boldsymbol{u}_h\|_2 \right)^2 \right] / \sum_{h=1}^{H} (TSS_h)}, \tag{3.14}$$

where TSS_h is the total sum of squares explained by the h^{th} component. Hence, the VIP is a measure of the contribution of each variable according to the variance explained by each dimension where $\boldsymbol{u}_{j,h}/\|\boldsymbol{u}_h\|_2$ represents the importance of the j^{th} variable. Note that $TSS_h = \boldsymbol{v}_h^2 \boldsymbol{\xi}_h^T \boldsymbol{\xi}_h$ [199], therefore, Equation 3.14 can be equivalently expressed by exchanging the terms. The VIP statistic was originally proposed for the classical PLS variant where the identification of the relevant variables in a high-dimensional setting is more challenging, however, it has been extended and applied to the regularized versions of the technique. It is generally accepted that a variable should be selected if $VIP_j > 1$ [199], [200], [201]. In addition, as per loading weights and PLS regression coefficients, a user-defined threshold can be used in order to define a feature as relevant (hard thresholding) and permutation tests can be employed to draw statistically significant conclusions. Note that the examination of the regression coefficients and the computation of the VIP score is only used for the regression mode while assessment of the loading can also be applied on the canonical mode.

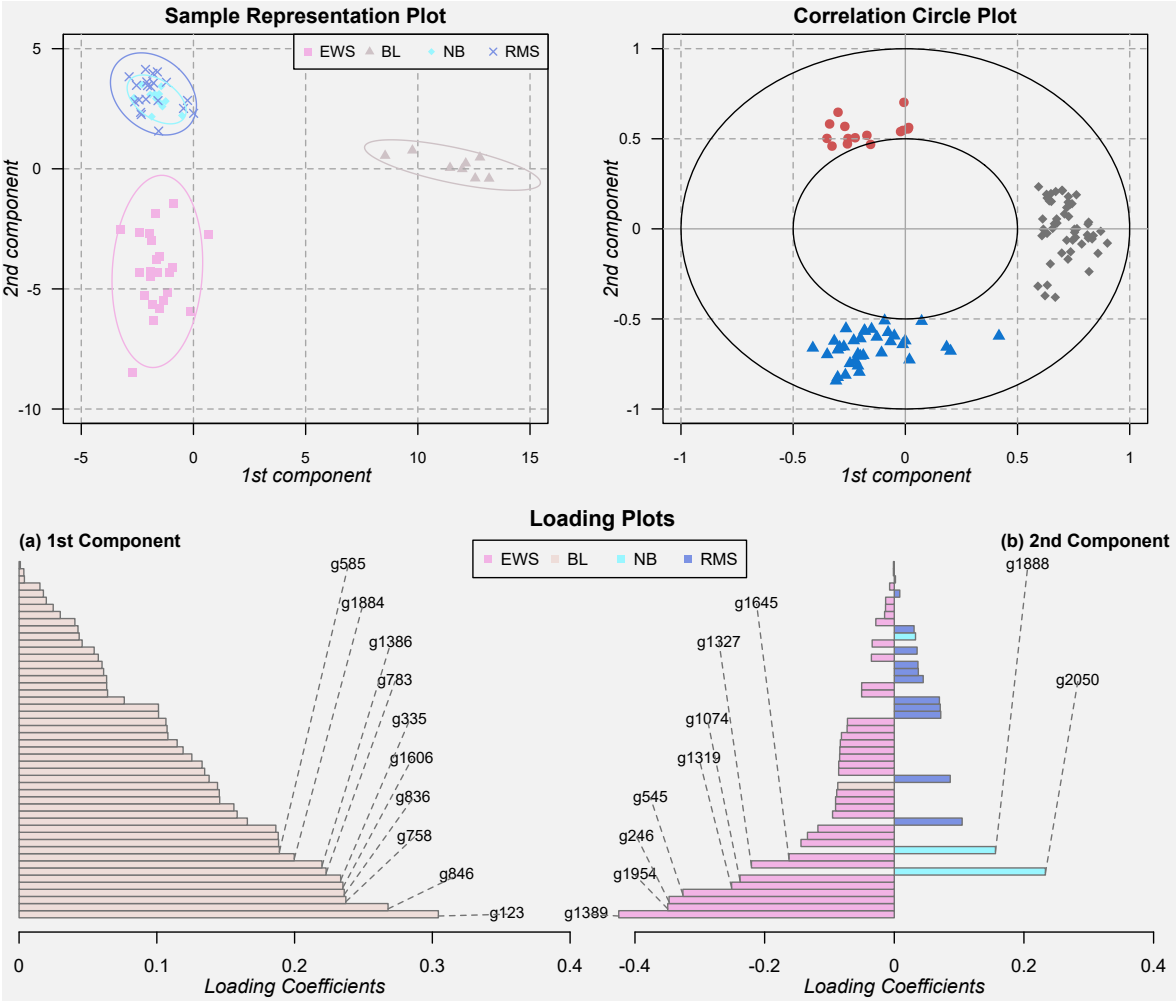
For regularized versions, stability frequency analysis can be performed per PLS component as a way to assess the reproducibility of a molecular signature. Resampling techniques are employed for that purpose: the calibrated model is fit on a subset of the samples (e.g. 80% of the total sample size) and the selected variables are recorded.

Depending on computational constraints, the resampling procedure is repeated a specific number of times yielding the stability frequency of the selected features across models visited. Bootstrap or CV are commonly applied to measure stability frequency [189], [190], in the latter case the output is a by-product of the performance evaluation process in which the features that were selected across the repeated CV runs are recorded.

3.6.2.5 Visualisation: Graphical Outputs

Visualisation outputs are important for the analysis of high-dimensional omics in both two-block and discriminatory contexts as they provide a tool to better unravel the complex associations between biological entities. Various graphical outputs are available, including the feature and sample representation plots which are the simplest and easiest to interpret visualisation tools [191], [195]. The former displays the loading weight of each (selected) variable per dimension, commonly ordered in an increasing order of importance (according to the absolute value of the loading coefficients). Sample representation plots display the component scores and therefore visualise similarities between samples in the reduced dimensional space spanned by the first few latent variables of the model. Figure 3.2 provides illustrative examples of loading coefficient and sample representation plots reproduced on publicly available data. In PLS-DA, the goal of a sample plot is to reveal how well the calibrated model differentiates between the sample classes while in two-block PLS the aim is to show how samples are clustered based on their biological characteristics. In the two-block setting it is also a common practice to represent the samples in a superimposed manner where the x and y axis simultaneously represent the $\mathbf{X}$ and $\mathbf{Y}$ scores [202]. These graphical representations are sometimes referred to as arrow plots as each sample is indicated using an arrow and are a useful tool to assess if the two datasets agree in information content according to the calibrated model being examined. The start of the arrow represents the location of the sample in the $\mathbf{X}$ data set and the tip of the arrow its location in the $\mathbf{Y}$ data set. Hence, short arrows indicate a strong agreement

Figure 3.2: Illustrative examples of common visualisation outputs for the assessment of discriminant and integrative methods.



Visualisation outputs obtained from the analysis of the dataset Small Round Blue Cell Tumours (SRBCT) publicly available from the `mixOmics` R package, which includes the expression levels of 2,308 genes measured on 63 samples. The samples are classified into four classes as follows: Burkitt Lymphoma (BL), Ewing Sarcoma (EWS), neuroblastoma (NB) and rhabdomyosarcoma (RMS). A sPLS-DA model was fitted with two components and 50 signals to retain in each dimension. The sample plot shows a clear separation of cancer types in the two-dimensional space spanned by the model, the correlation circle plot displays three subsets of correlated variables with a similar contribution to define each component and the loading plots of each component easily identifies the highest contributing signals. In this last graphical output, only the first ten contributing variables are named, and colours of the bars indicate the sample group for which the mean expression level is maximum.
sPLS-DA: sparse Partial Least Squares-Discriminant Analysis.

between the data sets while long arrows a disagreement between the data sets.

While the two plots presented above provide an insight into individual variable contribution and identification of sample clusters, they do not allow for the visualisation of pair-wise associations between variables from the two datasets. Three graphical

outputs have been proposed for that purpose, namely relevance networks, Clustered Image Map (CIM) and correlation circle plots [203]. These are complementary outputs that describe co-expression patterns within and between datasets by showing positive, negative, null correlations and/or displaying strong and weak correlation coefficients between variables. They are employed in regularized PLS techniques with a reduced variable selection size as visual interpretation is unfeasible otherwise; alternatively, pre-specified threshold can be employed to remove weaker associations.

More specifically, these three plots correspond to graphical representations of a pairwise similarity matrix which can be viewed as a robust approximation of a Pearson correlation matrix (whose intensive computation on a large dataset makes its calculation prohibitive). The entries of the similarity matrix $\mathbf{M}$ of dimension $p \times q$ are the similarity scores between each variable of $\mathbf{X}$ and $\mathbf{Y}$; their definition is based on the decomposition of the original data matrices in component score and regression coefficient matrices. The similarity scores are thus given by:

$$\mathbf{M}_p^q = \sum_{1=h}^{H} \boldsymbol{u}_h^2 \boldsymbol{c}_p^h \boldsymbol{d}_q^h \approx cor(\mathbf{X}, \mathbf{Y}) \tag{3.15}$$

$$\mathbf{M}_p^q = \sum_{1=h}^{H} \boldsymbol{u}_h^2 \sigma_h^2 \boldsymbol{c}_p^h \boldsymbol{e}_q^h \approx cor(\mathbf{X}, \mathbf{Y}) \tag{3.16}$$

For regression and canonical modes, respectively. The term $\boldsymbol{u}_h^2$ (resp. σ_h^2) is the variance of $\boldsymbol{\xi}_h$ (resp. $\boldsymbol{\omega}_h$), the $\mathbf{X}$-score (resp. $\mathbf{Y}$-score) for the h dimension. The terms $\boldsymbol{c}_p^h$, $\boldsymbol{d}_q^h$ and $\boldsymbol{e}_q^h$ are the local regression coefficients associated to the variable p or q for dimension h. The matrix $\mathbf{M}$ can be factorized as $\mathbf{M} = \mathbf{X}^* \mathbf{Y}^{*T}$ with the $\mathbf{X}^*$ ($p \times h$) and $\mathbf{Y}^*$ ($q \times h$) matrices containing the similarities scores for the original variables of $\mathbf{X}$ and $\mathbf{Y}$, respectively. When only two dimensions are chosen, $\mathbf{M}$ is represented by plotting the rows of $\mathbf{X}^*$ and the rows of $\mathbf{Y}^*$ as vectors in a two-dimensional Cartesian coordinate system, an output that correspond to a correlation circle plot (see below).

Relevance networks : it is a network where nodes represent variables and edges represent associations. A connection is drawn between two variables only if a pre-specified threshold (in absolute value) is exceeded. Positive and negative correlations can be displayed by assigning different colours to the edges; it is a bipartite network and thus the edges are drawn only between features of different datasets.

CIM : it is a heatmap representing the relationship between two groups of variables (selected variables from the $\mathbf{X}$ and $\mathbf{Y}$ matrices) and the colour inside the graph indicates the correlation between the features, as opposed to a standard heatmap that displays the relationship between the samples and one set of variables. The $\mathbf{M}$ matrix is graphically displayed as a two-dimensional image where each entry of the matrix is coloured based on its value and where the rows and columns are reordered according to a hierarchical clustering method.

Correlation circle plots : It highlights subsets of variables from both datasets that are important to define each component while simultaneously displaying the correlation between features within and between datasets. In the plane defined by two chosen dimensions, the coordinates of the features are obtained by calculating (an approximation) of the Pearson correlation between each original variable and their associated component. Two circles of radii 0.5 and 1 are drawn which simplifies interpretation in relation to the scores: variables that play a high contribution are close to the larger circle while variables located close to the origin might not be relevant for the definition of the component. In that latter case, it is advisable to visualise the correlation circles plots in subsequent dimensions to better interpret the information. Cluster of variables can also be identified, and the nature of the correlation can be inferred: the cosine of the angle between two points represent positive, negative or null correlations between features. The correlation is positive if the angle is sharp $\cos(\alpha) > 0$, negative if the angle is obtuse $\cos(\alpha) < 0$, and null if the vectors are perpendicular $\cos(\alpha) \approx 0$. Thus, variables or groups of variables strongly positively correlated are projected closely to each other whereas when the correlation is strongly negative,

the groups of variables are projected at diametrically opposite places on the graph (see Figure 3.2 for an example).

This page intentionally left blank.

4

Analysis of Proteomics and Transcriptomics Data Employing Established Univariate Approaches: Identifying Pre-diagnostic Markers Predictive of B-cell Lymphoma

This chapter is based in part on the publications:

M. Chadeau-Hyam, R. Vermeulen, D.G.A.J. Hebels, R. Castagné, G. Campanella, L. Portengen, R.S. Kelly, I.A. Bergdahl, B. Melin, G. Hallmans, D. Palli, V. Krogh, R. Tumino, C. Sacerdote, S. Panico, T.M.C.M. de Kok, M.T. Smith, J.C.S. Kleinjans, P. Vineis and S.A. Kyrtopoulos. "Prediagnostic transcriptomic markers of Chronic lymphocytic leukemia reveal perturbations 10 years before diagnosis". *Annals of Oncology* **25**.5 (2014), 1065–1072.

R. Vermeulen, F.S. Hosnijeh, B. Bodinier, L. Portengen, B. Liquet, **J. Garrido-Manríquez**, H. Lokhorst, I. A. Bergdahl, S. A. Kyrtopoulos, A. Johansson, P. Georgiadis, B. Melin, D. Palli, V. Krogh, S. Panico, C. Sacerdote, R. Tumino, P. Vineis, R. Castagné, M. Chadeau-Hyam, M. Botsivali, A. Chatziioannou, I. Valavanis, J.C.S. Kleinjans, T.M.C.M. de Kok, H.C. Keun, T. J. Athersuch, R. Kelly, P. Lenner, G. Hallmans, E.G. Stephanou, A. Myridakis, M. Kogevinas, L. Fazzo, M. De Santis, P. Comba, B. Bendinelli, H. Kiviranta, P. Rantakokko, R. Airaksinen, P.

Ruokojarvi, M. Gilthorpe, S. Fleming, T. Fleming, Y. Tu, T. Lundh, K. Chien, W.J. Chen, W. Lee, C.K. Hsiao, P. Kuo, H. Hung and S. Liao. "Pre-diagnostic blood immune markers, incidence and progression of B-cell lymphoma and multiple myeloma: Univariate and functionally informed multivariate analyses". *International Journal of Cancer* **143**.6 (2018), 1335–1347.

All the analyses shown here have been independently replicated by me. A description of the overlapping analysis as well as improvements made by this chapter is provided in section B.1.

4.1 Introduction

In this chapter I apply traditional univariate statistical approaches to analyse EnviroGenoMarkers (EGM) proteomics and transcriptomics profiles in order to investigate the relationship between pre-diagnostic blood levels of inflammatory markers and gene expression signals and future risk of B-cell Lymphoma (BCL) and its main histological subtypes. The possible confounding effect that an heterogenous White Blood Cell (WBC) composition of blood samples may introduce to the identification of predictive disease markers is explored. Furthermore, as omics measurements are susceptible to systemic variability as a consequence of experimental artefacts, I seek to characterise and correct for the technical-induced noise affecting the immune marker concentrations and gene expression levels. Finally, given the prospective nature of the blood samples, I examine the relationship between these omics profiles and disease status as a function of Time to Diagnosis (TtD) which could provide information on biological pathways involved in disease pathogenesis and result in disease biomarkers of prediction.

4.2 Methods

I use the R-statistical package `lme4` to fit the Linear Mixed Model (LMM) described in section 3.6.1. The models were fitted separately on the 28 proteins (inflammatory markers data) and the 29,662 transcripts (gene expression data). Analyses were performed on the full BCL pooled population and subsequently stratified by major histological subtypes: Chronic Lymphocytic Leukaemia (CLL), Diffuse Large B-cell Lymphoma (DLBCL), Follicular Lymphoma (FL), and Multiple Myeloma (MM). Since a larger number of controls increases statistical precision in effect size estimates and tests, disease subtype stratification was conducted including cases of the corresponding subtype and all control subjects (i.e. matched case-controls pairs and the remaining unmatched controls) [1]. The strength of the association between BCL (or the corresponding subtypes) case-control status and each marker level was inferred using a Likelihood Ratio Test (LRT) comparing the model without the disease status variable to the one with it. Multiple testing was accounted for using a stringent Bonferroni correction setting the Family Wise Error Rate (FWER) to 5%. Partial and full adjustment on leukocytes proportion was performed on the participants for which epigenetic information was available: 199 (74.25 %) and 176 (75.86 %) case-control pairs have DNA methylation (DNAm) data in addition to the proteomics and transcriptomics data (268 and 232 case-control pairs), respectively.

The effect of technical-induced noise was assessed by analysing the intercept and variance of random effect terms (microtiter plate number in proteomics and the dates of the major experimental steps in transcriptomics: isolation, labelling and hybridisation). For each of the random effect covariates, I calculate the distribution of the rank of each intercept estimated over all fitted models (28 and 29,662 for proteomics and transcriptomics, respectively). In addition, as an indicator of the extent of nuisance

[1] Improved statistical precision refers to obtaining narrower confidence intervals for a same sample size. Such situation can occur when two or more cases-control pairs have identical values for the matching variables, then combining them into a single group produces an estimator with lower variance. In addition, as a result of increasing statistical precision, power to detect significant associations is also enhanced. [204], [205]

variation, I calculate the frequency that each of the corresponding random effect covariates was estimated to generate negligible noise across all fitted models (i.e. proportion of null variance). The impact of the nuisance variation was further assessed by comparing the distributions of p-values (over the 28 and 29,662 tests) obtained using the LMMs to those obtained under the corresponding linear model (i.e. same predictor variable and fixed effect covariates but without a random effect term).

Proteins that were found to be differentially expressed between cases and controls were further investigated through Unconditional Logistic Regression (ULR) and Conditional Logistic Regression (CLR), in which cases were compared to all controls and to the corresponding matched controls, respectively. Assuming a non-linear relationship between case-control status (response variable) and protein concentration (predictor variable), plasma levels of immune markers were included in the model as categorical predictors and classes were defined according to the quartiles calculated based on the distribution in control subjects [2].Both ULR and CLR models were adjusted for age at enrolment, sex, country, study phase (matching variables), Body Mass Index (BMI), smoking status, education, physical activity and alcohol intake (confounding factors). Tests for trend were calculated comparing models with and without the predictor variable and using the quartile number as a continuous variable. To account for nuisance variation, I remove from the raw concentration levels of each protein the random effect estimates provided by the LMMs. The resulting "de-noised" proteomics data was used in the logistic regression models.

Transcripts identified by the genome-wide screen were further investigated using logistic regression models as well as through gene-enrichment analyses using the openly available Database for Annotation, Visualization and Integrated Discovery (DAVID v6.8, http://david.abcc.ncifcrf.gov/). Analogously to the proteomics analysis, noise variance due to the dates of isolation, labelling and hybridisation

[2]As discussed in section 2.1.1, a logistic regression model assumes that predictors are linearly related to the log odds of the response variable. However, in this case such assumption is expected to be violated and instead of employing a non-linear transformation of the dependent variable (such as fractional polynomials or spline functions), a simpler approach was taken by transforming the continuous predictor into categories [206].

experimental steps was removed before analyses (i.e. "de-noised" transcriptomics data).

4.3 Results

The characteristics of the study population with respect to the main demographic covariates are summarised in Table 4.1. The corresponding distribution of BCL subtypes and genders across phases and countries are shown in Table B.1 to Table B.4. Each study phase includes cases with the main four BCL subtypes as well as subjects from both cohorts and genders.

4.3.1 Proteomics

Table 4.2 shows the median, minimum and maximum concentration levels of all 28 proteins stratified by case-control status, country, phase of study. Table B.5 and Table B.6 display the median, minimum and maximum values across the four BCL subtypes for participants from European Prospective Investigation into Cancer and Nutrition - Italy (EPIC-Italy) and Northern Sweden Health and Disease Study (NSHDS), respectively. As also depicted in Figure B.1 to Figure B.3, the median concentration of most immune markers was higher among controls, study phase 1 and NSHDS subjects compared with cases, phase 2 and EPIC-Italy subjects, respectively. The distribution of concentration levels for all inflammatory markers before and after logarithmic transformation ($\log_2$) is graphically displayed in Figure B.4. The skewed distribution observed in most markers was normalised after transformation which were the values used for the subsequent analyses. In addition, the correlation structure between the 28 proteins under study is represented in Figure 4.1 where the vast majority of pairwise associations displays positive correlation estimates and three groups with a strong clustering pattern can be distinguished.

Table 4.1: Characteristics of the study population with respect to the main demographic variables.

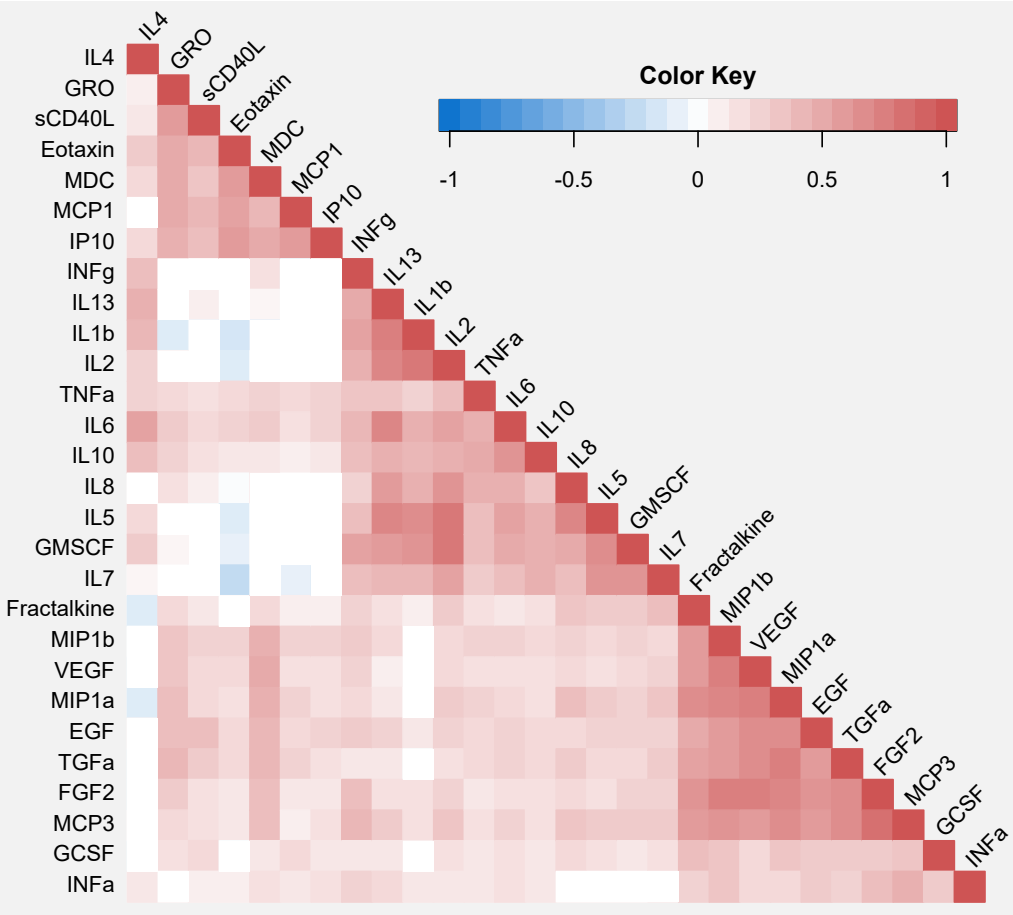
Baseline Variable	*Cases (n=268)*	*Controls (n=268)*	*p-value*
Cohort n (%)			
Epic-Italy	84 (31.34)	84 (31.34)	
NSHDS	184 (68.66)	184 (68.66)	
Phase n (%)			
1	96 (35.82)	96 (35.82)	
2	172 (64.18)	172 (64.18)	
Sex n (%)			
Female	136 (50.75)	136 (50.75)	
Male	132 (49.25)	132 (49.25)	
Age at recruitment (years)	53.1 (7.77)	53.11 (7.75)	
Alcohol intake (g/day)	7.05 (12.51)	8.25 (14.68)	0.31
Body Mass Index (kg/m2)	26.36 (3.83)	26.53 (4.14)	0.62
Smoking Status n (%)			
Current	55 (20.52)	55 (20.52)	0.87
Former	68 (25.37)	63 (23.51)	
Never	145 (54.10)	150 (55.97)	
Education level n (%)			
None	4 (1.49)	1 (0.37)	0.31
Primary	96 (35.82)	104 (38.81)	
Technical/professional	68 (25.37)	56 (20.90)	
Secondary	53 (24.25)	65 (24.25)	
University College	47 (17.54)	42 (15.67)	
Physical Activity n (%)			
Inactive	80 (29.85)	76 (28.36)	0.37
Moderately Inactive	106 (39.55)	95 (35.45)	
Moderately Active	68 (25.37)	74 (27.61)	
Active	14 (5.22)	23 (8.58)	

p-value for difference was calculated using the student's t-test for continuous variables and the X^2 test for categorical variables. Counts and percentages are reported for categorial variables and means and standard deviations for continuous variables.

Table 4.2: Median (minimum - maximum) values of immune markers stratified by case-control status, study cohort and experimental phase.

Protein	*Cases (n=268)*	*Controls (n=268)*	*EPIC (n=168)*	*NSHDS (n=368)*	*Phase 1 (n=192)*	*Phase 2 (n=344)*
EGF	22.31 (0.02-1622.54)	24.97 (0.14-1561.89)	20.31 (0.15-1622.54)	24.23 (0.02-842.01)	40.27 (0.25-1622.54)	19.8 (0.02-550.46)
FGF2	15.44 (0.04-1969.88)	27.12 (0.1-1005.98)	25.7 (0.38-1969.88)	18.84 (0.04-1005.98)	42.09 (0.18-1969.88)	13.79 (0.04-410.95)
GCSF	24.47 (0.56-496)	24.73 (0.77-2649.79)	27.89 (0.56-304.12)	23.04 (0.77-2649.79)	28.23 (1.24-496)	22.72 (0.56-2649.79)
VEGF	118.02 (0.46-6938.42)	204.84 (0.47-5644.06)	143.29 (0.81-4204.53)	162.41 (0.46-6938.42)	409.61 (2.98-6938.42)	82.02 (0.46-2549.75)
GMSCF	2.7 (0.02-285.46)	2.8 (0.01-1566.28)	4.91 (0.04-38)	1.44 (0.01-1566.28)	4.81 (0.03-281.84)	1.41 (0.01-1566.28)
TGFa	1.1 (0-1663.43)	2.25 (0.01-842.46)	1.46 (0-1663.43)	1.69 (0-842.46)	9.07 (0.07-1663.43)	0.83 (0-399.24)
Eotaxin	242.2 (8.4-990.58)	239.31 (11.51-1122.54)	48.02 (8.4-319.46)	333.12 (35.34-1122.54)	156.82 (15.64-1122.54)	255.09 (8.4-1093.35)
Fractalkine	36.62 (0.13-3229.59)	56.8 (0.35-4599.65)	82.54 (0.58-1923.42)	30.46 (0.13-4599.6)	118.39 (0.65-4599.65)	23.14 (0.13-3229.59)
GRO	391.31 (27.15-2188.12)	369.11 (44.38-2914.05)	216.5 (27.15-1860.16)	426.08 (69.12-2914.05)	553.45 (54.3-2914.05)	338.73 (27.15-1319.35)
MCP1	279.4 (1.78-787.6)	297.15 (6.2-1085.85)	213.26 (55.08-634.85)	323.49 (1.78-1085.8)	304.35 (55.08-1085.85)	278.89 (1.78-948.07)
MCP3	6 (0.02-1225.93)	13.42 (0.08-1003.8)	8.18 (0.02-1225.93)	9.43 (0.02-1003.8)	18.04 (0.14-1225.93)	4.32 (0.02-914.3)
MDC	705.34 (42.59-8962.26)	731.64 (50.85-7453.92)	335.82 (50.85-3363.32)	870.55 (42.59-8962.26)	758.34 (91.4-8962.26)	690.65 (42.59-4269.19)
MIP1a	7.28 (0.01-398.74)	12.17 (0.01-276.28)	13.83 (0.09-398.74)	8.32 (0.01-276.28)	43.3 (0.69-398.74)	4.49 (0.01-125.18)
MIP1b	30.84 (0.27-843.45)	36.38 (0.28-1112.56)	33.66 (1.99-572.67)	33.12 (0.27-1112.56)	48.03 (2.95-1112.56)	25.16 (0.27-424.76)
IP10	447.71 (15.86-3374.05)	446.74 (8.88-3766.5)	248.47 (59.19-1294.32)	543.6 (8.88-3766.5)	417.1 (81.68-3766.5)	465.8 (8.88-2738.33)
IL8	4.03 (0.59-304.82)	4.53 (0.67-190.86)	9.74 (0.67-304.82)	4 (0.59-167.32)	9.74 (1.1-304.82)	3.25 (0.59-59.12)
IL1b	0.44 (0-435.52)	0.96 (0-350.58)	1.37 (0-254.81)	0.39 (0-435.52)	0.38 (0-435.52)	0.82 (0.01-350.58)
IL2	2.53 (0.03-627.98)	3.04 (0.01-2224.85)	11.54 (0.02-300.87)	1.99 (0.01-2224.85)	10.45 (0.01-300.87)	1.96 (0.02-2224.85)
IL4	3.19 (0.02-564.88)	6.54 (0.01-1627.27)	0.73 (0.04-66.3)	9.96 (0.01-1627.27)	1.06 (0.02-1627.27)	10.49 (0.01-776.61)
IL5	0.86 (0.01-525.55)	0.86 (0.02-332.06)	2.3 (0.04-525.55)	0.62 (0.01-332.06)	2.3 (0.01-525.55)	0.63 (0.01-49.89)
IL6	3.76 (0.04-480.8)	4 (0.04-1314.71)	2.4 (0.12-45.99)	5.09 (0.04-1314.71)	4.77 (0.04-1314.71)	3.53 (0.09-322.65)
IL7	0.64 (0.01-303.91)	0.75 (0.01-417.49)	1.73 (0.14-18.76)	0.36 (0.01-417.49)	1.6 (0.03-417.49)	0.41 (0.01-140.64)
IL10	12.19 (0.1-1322.5)	10.82 (0.07-2635.7)	11.01 (0.07-297.55)	11.66 (0.1-2635.7)	15.5 (0.07-2635.7)	9.21 (0.32-2333.34)
IL13	2.06 (0.01-929.4)	3.98 (0.01-2474.46)	2.72 (0.02-929.4)	4 (0.01-2474.46)	3.77 (0.01-2474.46)	2.09 (0.02-562.68)
INFa	2.06 (0-2569.21)	4.69 (0-1148.8)	3.93 (0-1034.96)	2.59 (0-2569.21)	1.37 (0-2569.21)	3.58 (0-885.96)
INFg	1.36 (0-2069.35)	2.2 (0.01-1591.65)	1.43 (0-2069.35)	1.9 (0-1591.65)	1.7 (0-2069.35)	1.73 (0.02-1591.65)
TNFa	6.02 (0.3-111.78)	5.11 (0.59-854.31)	4.97 (0.81-38)	5.84 (0.3-854.31)	6.62 (0.81-38)	4.96 (0.3-854.31)
sCD40L	689.32 (12.33-7627.75)	579.19 (4.39-5357.61)	404.41 (4.39-2819.65)	775.1 (12.33-7627.75)	723.67 (4.39-7627.75)	593.06 (12.33-3994.34)

Figure 4.1: Pairwise Spearman correlation coefficients for log-transformed values of the 28 inflammatory markers under study.



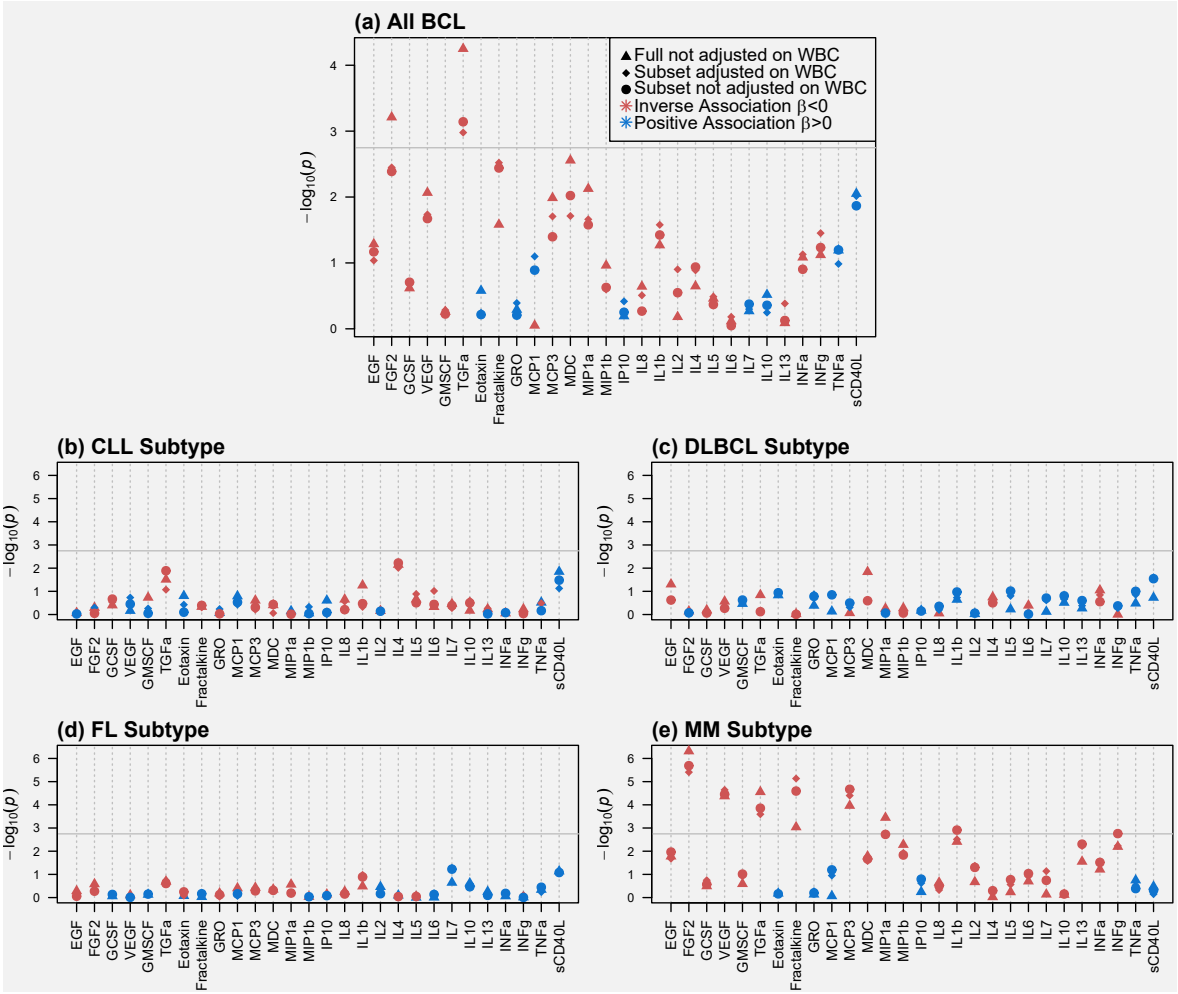
4.3.1.1 Pooled Population

The univariate analysis pooling all BCL cases together revealed general lower level of inflammatory markers among cases compared with controls: 20 out of the 28 showed an inverse association with disease status (Figure 4.2 and Table B.7). Of these, only FGF2 (β=-0.50, p-value=6.15 x 10^{-4}) and TGFa (β=-0.68, p-value=5.62 x 10^{-5}) reached statistical significance at Bonferroni 5%.

4.3.1.2 Subtype Stratified Analysis

Stratified analyses by main histological subtypes did not show any statistically significant associations (FWER<5%) with CLL, DLBCL and FL (Figure 4.2 and Table B.7).

Figure 4.2: Results of the LMM analyses between log-transformed values of proteins and case-control status.



Results are displayed separately for all BCL observations and the four main histological subtypes: CLL (b, n=42 cases and 268 controls), DLBCL (c, n=44 cases and 268 controls), FL (d, n=39 cases and 268 controls) and MM (e, n=76 cases and 268 controls). Strength of association (Y-axis) is measured by -$\log_{10}$ transformed p-values and the grey horizontal line represents the Bonferroni corrected per-test significance level ensuring a FWER control at 5%. Direction of the association is represented in red and blue for the negative and positive regression coefficients, respectively. Results are presented from the WBC unadjusted model (n=536, triangles), from the WBC unadjusted model using observations with epigenetic data available (n=398 case-control pairs, circles) and from the full WBC adjustment model (n=398 case-control pairs, diamonds).
LMM: Linear Mixed Model, WBC: White Blood Cells.

In contrast, six inverse associations for MM subtype were identified: lower concentration levels of FGF2 (β=-1.11, p-value=4.85 x 10^{-7}), VEGF (β=-1.00, p-value=4.23 x 10^{-5}), TGFa (β=-1.08, p-value=2.78 x 10^{-5}), Fractalkine (β=-0.72, p-value=9.14 x 10^{-4}) and MCP3 (β=-0.91, p-value=1.09 x 10^{-4}), MIP1a (β=-0.72, p-value=3.57 x 10^{-4}) were associated with increased risk of MM. The results from this analysis do not support

Table 4.3: Number of significant associations identified in the WBC unadjusted LMM, the six WBC partial adjustment LMMs and the full WBC adjustment LMM categorized by disease type.

Disease Type (n)	*Unadjusted LMM*	*Partial WBC Adjustment*						*Full WBC LMM*
		CD8	**CD4**	**NK cells**	**B cells**	**Monocytes**	**Granulocytes**	
All BCL (536/398/398)	2	1	1	1	1	1	1	0
CLL (310/223/233)	0	0	0	0	0	0	0	0
DLBCL (312/234/234)	0	0	0	0	0	0	0	0
FL (307/228/228)	0	0	0	0	0	0	0	0
MM (344/261/261)	6	8	6	6	5	7	8	6

Results of the analysis including cases and all controls subjects on the immune markers data.
LMM: Linear Mixed Model, WBC: White Blood Cell, NK: Natural Killer.

the presence of a common inflammatory marker across the main four subtypes under study.

4.3.1.3 WBC Correction Analysis

Both full and partial adjustment for leukocyte subpopulations provided consistent results to the unadjusted models. Full WBC adjustment in the analysis where all BCL cases were pooled together revealed TGFa as the only significant association (β=0.65, p-value=1.05 x 10^{-3}) (Figure 4.2, Table 4.3 and Table B.8). In analyses stratified by major histological subtypes, results remained unchanged for CLL, DLBCL and FL as no significant associations emerged while for MM subtype the same six markers were significantly associated with disease status. (Figure 4.2 and Table B.8). Replication of the analysis on the subset of observation with WBC estimates available but without performing leukocyte correction provided consistent results (Table B.9).

The results from the partial WBC adjustment are summarised in Table 4.3. TGFa was found to be the only significant association in the analysis of pooled BCL cases across the six partial WBC correction models and narrow effect size differences were observed (CD8 Lymphocytes: β=-0.65, p-value=7.91 x 10^{-4}; CD4 Lymphocytes: β=-0.66, p-value=6.50 x 10^{-4}; NK: β=-0.66, p-value=7.05 x 10^{-4}; B cells: β=-0.64, p-value=1.11 x 10^{-3}; Monocytes: β=-0.66, p-value=6.25 x 10^{-4}; Granulocytes: β=-0.66, p-value=7.45

x 10^{-4}; full WBC adjustment: β=-0.66, p-value=7.21 x 10^{-4}). In addition, as observed in the unadjusted and in the full WBC adjustment, no significant associations were found for CLL, DLBCL and FL across the six partial adjustment models performed. Finally, MM-specific analyses provided two additional associations after adjustment for CD8 Lymphocytes (IL1b: β=-0.74, p-value=1.18 x 10^{-3} and INFg: β=-0.96, p-value=1.58 x 10^{-3}) and Granulocytes (IL1b: β=-0.76, p-value=8.92 x 10^{-4} and INFg: β=-0.97, p-value=1.27 x 10^{-3}) proportions, one borderline additional association after Monocytes estimates adjustment (INFg: β=-0.95, p-value=1.72 x 10^{-3}) and one less association upon B cells proportion adjustment (MIP1a: β=-0.65, p-value=2.92 x 10^{-3}). Details on the strength of associations and effect sizes of the significant associations found for the BCL pooled population and the main subtypes are presented in Table B.10 and Table B.11, respectively.

4.3.1.4 Predictive Performance Assessment

The results from the multivariable ULR models for MM are presented in Table 4.4 and are consistent with the linear regression analyses. These identified an inverse relationship between risk of MM and plasma levels of FGF2 (OR=0.16, for 4th Q vs. 1st Q, P-trend=0.0001), VEGF (OR=0.16, for 4th Q vs. 1st Q, P-trend=0.0013), TGFa (OR=0.15, for 4th Q vs. 1st Q, P-trend=0.0011), Fractalkine (OR=0.34, for 4th Q vs. 1st Q, P-trend=0.0069), MCP3 (OR=0.33, for 4th Q vs. 1st Q, P-trend=0.0045), MIP1a (OR=0.31, for 4th Q vs. 1st Q, P-trend=0.0001).

4.3.1.5 Time to Diagnosis Analysis

Results of the LMM stratified by median time elapsed between blood samples collection and disease diagnosis (6 years) are shown in Figure 4.3 and Table B.12. Such TtD stratification was conducted for all BCL and the main histological subtypes and revealed similar associations to the analysis pooling all TtD years: inflammatory markers that reached statistical significance were identified for all BCL and MM subtype, the remaining subtype specific analyses did not provide significant findings (results

Table 4.4: Results of the ULR model relating MM subtype case-control status and quantile categories of log-transformed protein concentration levels (n=344).

Protein	*Quantiles Limits*	*Cases/Controls (n)*	*OR*	*Low CI*	*High CI*	*P-trend*
FGF2	Q1=< 2.9	34/67	Ref	Ref	Ref	0.0001
	Q2=2.9 - 4.26	25/67	0.73	0.38	1.39	
	Q3=4.26 - 5.35	11/67	0.32	0.14	0.69	
	Q4=>5.35	6/67	0.16	0.06	0.39	
VEGF	Q1=< 4.44	32/67	Ref	Ref	Ref	0.0013
	Q2=4.44 - 6.15	20/67	0.58	0.29	1.14	
	Q3=6.15 - 7.43	18/67	0.54	0.27	1.07	
	Q4=>7.43	6/67	0.16	0.06	0.41	
TGFa	Q1=<1.27	31/67	Ref	Ref	Ref	0.0011
	Q2=1.27 - 2.65	23/67	0.7	0.35	1.38	
	Q3=2.65 - 3.91	17/67	0.56	0.27	1.11	
	Q4=>3.91	5/67	0.15	0.05	0.38	
Fractalkine	Q1=<4.11	28/67	Ref	Ref	Ref	0.0069
	Q2=4.11 - 5.12	27/67	1	0.52	1.95	
	Q3=5.12 - 6.21	11/67	0.4	0.17	0.89	
	Q4=>6.21	10/67	0.34	0.14	0.74	
MCP3	Q1=<0.97	35/67	Ref	Ref	Ref	0.0045
	Q2=0.97 - 2.48	17/67	0.49	0.24	0.97	
	Q3=2.48 - 3.7	12/67	0.34	0.15	0.72	
	Q4=>3.7	12/67	0.33	0.15	0.68	
MIP1a	Q1=<1.58	27/67	Ref	Ref	Ref	0.0001
	Q2=1.58 - 2.84	32/67	1.27	0.67	2.45	
	Q3=2.84 - 3.64	8/67	0.29	0.11	0.68	
	Q4=>3.64	9/67	0.31	0.12	0.7	

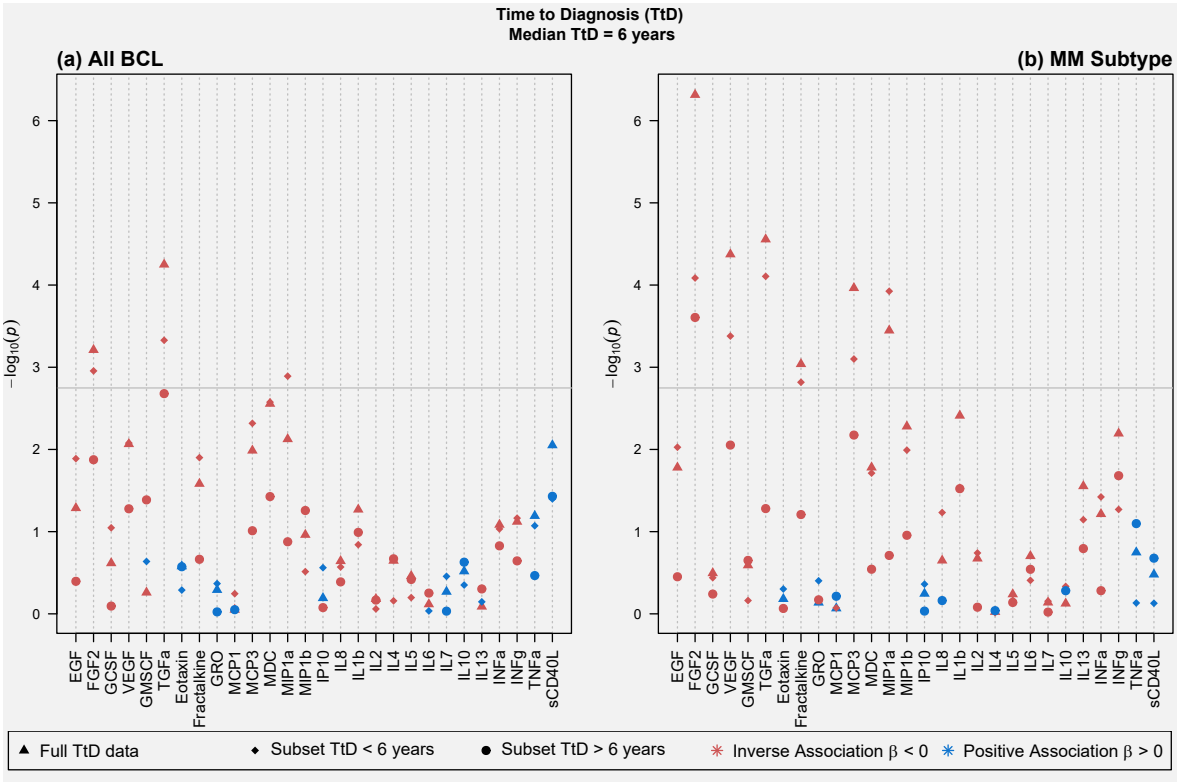
ULR: Unconditional Logistic Regression, OR: Odds Ratio, CI: Confidence Interval, Ref: Reference Category, ULR: Unconditional Logistic Regression.

not shown).

An inverse relationship was identified between BCL case-control status and plasma levels of FGF2 (β=-0.61, p-value=1.11 x 10^{-3}), TGFa (β=-0.77, p-value=4.70 x 10^{-4}) and MIP1a as one borderline additional finding (β=-0.52, p-value=1.28 x 10^{-3}) for the TtD<6 strata; for the TtD>6 strata no significant associations were identified.

Similarly, the MM-specific analysis revealed that the same six markers that reached statistical significance in the pooled TtD analysis were identified as significant associations for the TtD<6 strata (FGF2, β=-1.07, p-value=8.21 x 10^{-5}; VEGF, β=-1.06, p-

Figure 4.3: Results of the LMM analyses between log-transformed values of proteins and case-control status stratified by median TtD.



Results are displayed separately for all BCL observations (a, n=268 pairs) and MM subtype (b, n=76 cases and 268 controls). Strength of association (Y-axis) is measured by -$\log_{10}$ transformed p-values and the grey horizontal line represents the Bonferroni corrected per-test significance level ensuring a FWER control at 5%. Direction of the association is represented in red and blue for the negative and positive regression coefficients, respectively. Results are presented for the pooled analysis (n=536, triangles), for the TtD<6 years strata (n=268, diamond) and for the TtD>6 years strata (n=268, circles). Results for CLL, DLBCL and FL subtypes did not provide significant findings and are not shown.
LMM: Linear Mixed Model, TtD: Time to Diagnosis.

value=4.17 x 10^{-4}; TGFa, β=-1.28, p-value=7.84 x 10^{-5}; Fractalkine, β=-0.83, p-value=1.52 x 10^{-3}; MCP3, β=-0.95, p-value=7.94 x 10^{-4}; MIP1a, β=-0.95, p-value=1.19 x 10^{-4}). FGF2 was the only significant finding for the TtD>6 strata (β=-1.18, p-value=2.48 x 10^{-4}).

4.3.1.6 Sensitivity Analysis

Considering that MM subtype is the most frequent sub-entity in the study population and that the two strongest MM-specific signals correspond to the two associations found in the analysis of all BCL cases and controls (FGF2 and TGFa), I conducted fur-

ther analysis in the pooled BCL population excluding MM observations.These results exposed no significant associations (Figure B.5), a finding that supports the hypothesis that this specific subtype is mainly driving the significant associations found in the BCL analysis.

In addition, since the study participants come from two independent cohorts and from two experimental study phases, I also assessed the robustness of the findings across these two populations (Epic-Italy and NSHDS) and the two analytical phases (1 and 2). Significant findings were found for all BCL and for MM subtype; weaker p-values were observed particularly for EPIC-Italy (n = 21 MM cases vs. n = 55 in the NSHDS cohort) (Table B.13 and Table B.14). There is an overlap of the strongest findings across cohorts and phases (FGF2, VEGF and/or TGFa are consistently found as significant associations) whereas weaker inflammatory markers do not systematically replicate in both cohorts or in both phases (Figure B.6).

Results from the ULR were consistent in both cohorts although strength of the associations was reduced, especially for EPIC-Italy (Fractalkine and MCP3 did not reach statistically significance), possibly owing to the lower number of cases, as previously mentioned (Table B.15). Analysis from the CLR model comparing matched MM case-control pairs provided weaker p-values and revealed FGF2 (OR=0.19, for 4th Q vs. 1st Q, P-trend=0.0093) and TGFa (OR=0.11, for 4th Q vs. 1st Q, P-trend=0.0028), as the only two markers that reached statistical significance, which were among the strongest associations found in the ULR (Table B.16).

4.3.2 Transcriptomics

4.3.2.1 Pooled Population

The LMM fitted to all BCL cases and controls revealed five significant associations at a Bonferroni 5% which are reported in Table 4.5. The corresponding p-values and effect size estimates of those signals obtained from the subtype-specific analyses are also reported in Table 4.5. For CLL, four of those five candidates revealed strong

Table 4.5: Significant associations between gene expression levels and BCL case-control status.

			All BCL (n=464)		*CLL (n=266)*		*DLBCL (n=269)*		*FL (n=269)*		*MM (n=299)*		*All BCL w/o CLL (n=434)*	
	Agilent ID	**Gene Name**	***f***	***p*-value**	***f***	***p*-value**	***f***	***p*-value**	***f***	***p*-value**	***f***	***p*-value**	***f***	***p*-value**
1	A_23_P500400	ABCA6	1.68	9.52E-09	17.40	1.04E-62	0.98	0.823	1.31	0.011	0.99	0.855	1.10	0.075
2	A_23_P26854	ARHGAP44	1.86	3.28E-08	24.49	1.30E-43	1.07	0.633	1.18	0.251	1.17	0.136	1.17	0.043
3	A_23_P210581	KCNG1	0.76	4.78E-07	0.66	0.0002	0.95	0.647	0.82	0.029	0.70	2.16E-06	0.78	4.45E-06
4	A_32_P44394	AIM2	1.26	9.36E-07	2.63	3.78E-24	1.15	0.050	1.13	0.082	1.07	0.236	1.11	0.011
5	A_23_P145889	CDK14	1.28	1.54E-06	3.27	6.16E-28	1.08	0.212	1.12	0.089	0.97	0.578	1.08	0.043

Transcripts were declared significant using a Bonferroni corrected per-test significance level ensuring a FWER control at 5% and are ordered in relation to their corresponding strength of association. The corresponding p-values and effect sizes from the subtype-analyses are also reported. Fold change (f) estimates derived from the regression coefficient (β) obtained from the Linear Mixed Model.

statistical associations and effect sizes while for the other subtypes, the same five signals showed non-significant p-values at a Bonferroni level 5%. Consistently, when the LMM is fitted to the pooled population of BCL observations excluding CLL cases, there are no statistically significant associations and correspondingly the p-values of those five associations increase to non-significant levels. The results from this analysis do not support the presence of a common signal across the main four subtypes under study.

4.3.2.2 Subtype Stratified Analysis

Numerous significant associations were found for CLL subtype (n = 684 at Bonferroni FWER 5%) and the 50 strongest signals are reported in Table 4.6. MM analysis revealed only two borderline significant associations with weak effect sizes (agilent ID=A_23_P164773, gene name=FCER2, fold change=0.75, p-value=3.97 x 10^{-7} and agilent ID=A_32_P8813, gene name=LOC283663, fold change=0.67, p-value=8.19 x 10^{-7}) while the remaining subtypes did not provide any realistic candidate signals. For this reason, subsequent analyses shown in this section are limited to associations found in the CLL subtype.

Table 4.6: First 50 strongest significant associations between gene expression levels and CLL case-control status (n=266).

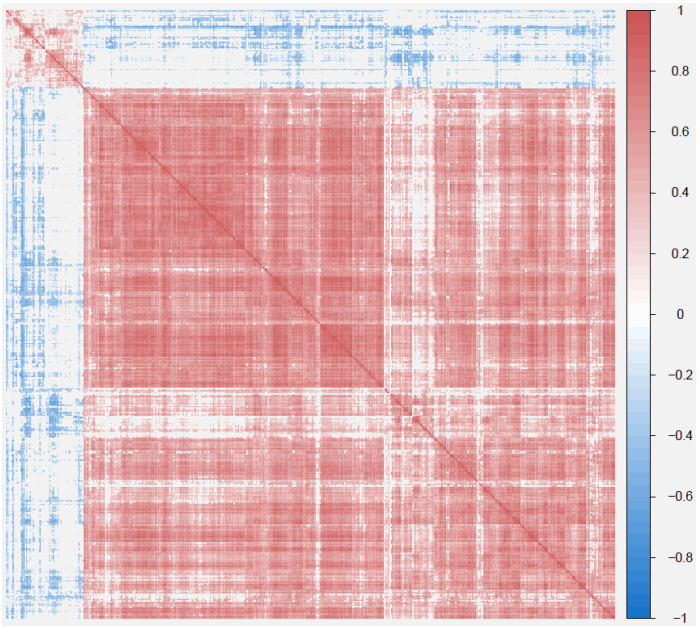
	Agilent ID	*Gene Name*	*f*	*p-value*		*Agilent ID*	*Gene Name*	*f*	*p-value*
1	A_23_P500400	ABCA6	17.39	1.04E-62	26	A_23_P310931	CNR2	2.24	6.71E-26
2	A_32_P53234	—	5.27	1.18E-44	27	A_32_P49854	—	2.03	7.18E-26
3	A_23_P26854	ARHGAP44	24.49	1.30E-43	28	A_24_P319647	FCRL2	3.28	1.76E-25
4	A_23_P27332	TCF4	3.83	7.14E-41	29	A_23_P163697	SYT17	2.62	3.16E-25
5	A_24_P29733	CDK14	3.95	4.43E-39	30	A_23_P31725	BLK	3.01	1.21E-24
6	A_23_P131024	ZBTB32	5.26	5.47E-39	31	A_23_P76402	TCTN1	2.09	1.52E-24
7	A_24_P691826	—	5.83	1.46E-35	32	A_32_P44394	AIM2	2.63	3.78E-24
8	A_23_P130158	WNT3	13.75	5.67E-35	33	A_24_P402588	BCL11A	2.19	4.74E-24
9	A_24_P931428	TCF4	3.79	3.58E-34	34	A_24_P184803	COCH	3.90	6.57E-24
10	A_23_P67529	KCNN4	3.15	3.17E-32	35	A_23_P102113	WNT10A	2.04	1.10E-23
11	A_32_P108156	MIR155HG	3.58	8.20E-31	36	A_23_P164773	FCER2	3.24	2.44E-23
12	A_32_P48054	CNR2	2.76	3.07E-30	37	A_23_P85269	TTN	2.99	2.67E-23
13	A_23_P85250	CD24	3.03	5.25E-29	38	A_32_P2883	—	2.95	6.32E-23
14	A_23_P56553	METTL8	2.71	8.26E-29	39	A_23_P328206	DNMBP	2.62	7.87E-23
15	A_23_P201211	FCRL5	5.19	2.74E-28	40	A_24_P54390	RASGRP3	2.68	1.13E-22
16	A_23_P145889	CDK14	3.26	6.16E-28	41	A_23_P312920	POU2AF1	2.75	1.80E-22
17	A_23_P21758	ADAM28	2.75	1.09E-27	42	A_23_P8961	IL7	2.37	4.83E-22
18	A_24_P376848	FCRL5	4.09	3.18E-27	43	A_23_P39067	SPIB	2.31	5.62E-22
19	A_23_P160751	FCRL2	3.79	3.22E-27	44	A_24_P662636	—	3.64	5.79E-22
20	A_23_P46039	FCRLA	3.32	4.94E-27	45	A_23_P45786	COL9A2	2.22	6.46E-22
21	A_23_P20427	RHOBTB2	2.27	5.35E-27	46	A_32_P72067	—	2.79	9.23E-22
22	A_23_P156907	SOBP	4.12	1.27E-26	47	A_23_P370830	KLHL14	3.87	1.39E-21
23	A_23_P124335	—	2.89	1.63E-26	48	A_23_P4551	SETBP1	2.13	2.09E-21
24	A_23_P17269	CCDC88A	2.07	2.33E-26	49	A_23_P321984	CLECL1	3.08	2.41E-21
25	A_23_P132378	CELSR1	3.61	2.88E-26	50	A_23_P40108	COL9A3	3.62	2.45E-21

Transcripts are ordered in relation to their corresponding strength of association with CLL case-control status. Fold change (f) estimates derived from the regression coefficient (β) obtained from the Linear Mixed Model.

As illustrate in Figure 4.4, high levels of correlation and strong hierarchical clustering are observed among the 684 CLL-specific signals. A positive correlation is observed in the vast majority of the selected probes (79.95%) and the strength of the correlation is higher in positive compared to negative correlated signals.

The relationship between statistical significant associations and effect size estimates is illustrated in Figure 4.5. All CLL-specific signals show gene upregulation in cases (positive fold change), with the 10 strongest associations showing up to 25-fold up-regulation. Gene expression signals were spread evenly across chromosomes, how-

Figure 4.4: Pairwise Spearman correlation coefficients for the 684 candidates identified as differentially expressed in the CLL sub-type analysis.

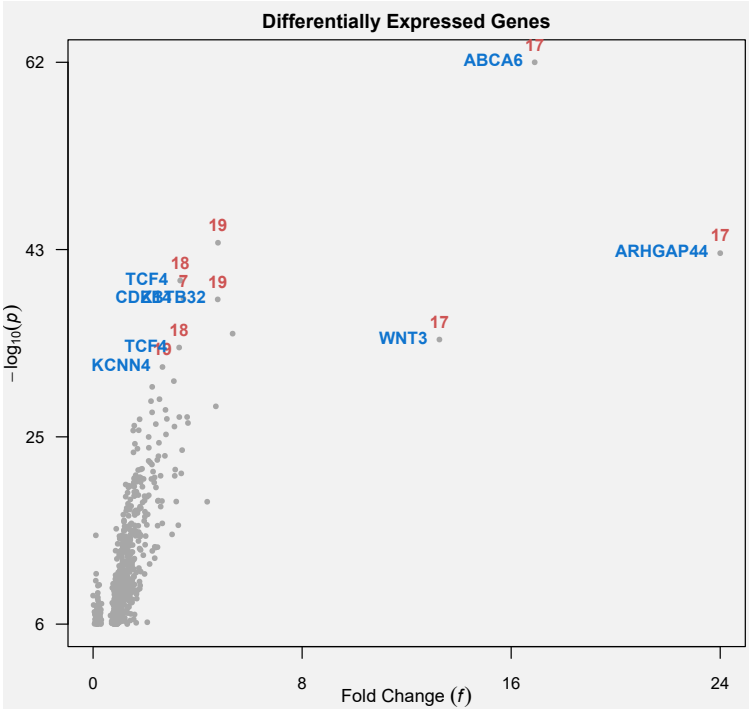


ever two clusters of three strong signals with large effect sizes can be clearly distinguished in chromosome 17 (agilent ID=A_23_P500400, gene name=ABCA6, p-value=1.04 x 10^{-62}, fold change=17.39; agilent ID=A_23_P130158, gene name=WNT3, p-value=5.67 x 10^{-35}, fold change=13.74; agilent ID=A_23_P26854, gene name=ARHGAP44, p-value=1.3 x 10^{-43}, fold change=24.49) and chromosome 19 (agilant ID=A_23_P67529, gene name=KCNN4, p-value=3.17 x 10^{-32}, fold change=3.15; agilent ID=A_23_P131024, gene name=ZBTB32, p-value=5.46 x 10^{-39}, fold change=5.26; agilent ID=A_32_P53234, p-value=1.17 x 10^{-44}, fold change=5.27).

4.3.2.3 WBC Correction Analysis

As depicted in Table 4.7, full and partial adjustment by leukocyte estimates show consistent results compared to the WBC unadjusted models: significant associations were found for the LMM including all BCL cases and controls and for the CLL and MM subtype-specific analyses. The five significant associations found for the pooled population model disappear after full WBC correction while partial WBC adjustment

Figure 4.5: Volcano plot displaying the relationship between the p-values measuring the strength of the association with disease status and their corresponding effect size estimate for each of the 684 differentially expressed genes.



only retains one borderline significant signal for the models adjusted for CD8 Lymphocyte, B cell, Granulocyte proportions (agilent ID=A_23_P210581, gene name=KCNG1, fold change=0.73/0.73/0.73, p-value=1.03 x 10^{-6}/1.49 x 10^{-6}/1.66 x 10^{-6}) and Natural Killer (NK) cell (agilent ID=A_23_P500400, gene name=ABCA6, fold change=1.53, p-value=1.64 x 10^{-6}).

Similarly, the two associations found in the MM subtype analysis do not reach significant levels after full WBC correction while partial correction revealed one borderline significant candidate for the models corrected after CD8 Lymphocytes, B cells and Monocytes estimates (agilant ID=A_23_P149368, gene name=FCRL1, fold change=0.70/0.74/0. p-value=7.51 x 10^{-7}/5.87 x 10^{-7}/9.62 x 10^{-7}), and three significant candidates for the models corrected after Granulocyte proportion (agilant ID=A_23_P149368, gene name=FCRL1, fold change=0.69, p-value=1.15 x 10^{-7}; agilant ID=A_24_P548060, fold change=0.76, p-value=1.24 x 10^{-6}; agilant ID=A_32_P8813, gene name=LOC283663,

Table 4.7: Number of significant associations identified in the WBC unadjusted LMM, the six WBC partial adjustment LMMs and the full WBC adjustment LMM categorized by disease subtype.

Disease Type (n)	*Unadjusted LMM*	*Partial WBC Adjustment*						*Full WBC LMM*
		CD8	**CD4**	**NK cells**	**B cells**	**Monocytes**	**Granulocytes**	
All BCL (464/352/352)	5	1	0	1	1	0	1	0
CLL (266/194/194)	684	415	427	400	18	420	258	18
DLBCL (269/206/206)	0	0	0	0	0	0	0	0
FL (269/205/205)	0	0	0	0	0	0	0	0
MM (299/233/233)	2	1	0	0	1	1	3	0

LMM: Linear Mixed Model, WBC: White Blood Cell, NK: Natural Killer.

fold change=0.63, p-value=2.99 x 10^{-7}). None of the probes identified in the MM-specific partial WBC adjustment matched the associations found in the unadjusted WBC model.

The number of CLL-specific signals decreases from 684 to 18 upon full WBC correction, of which 16 are common with the unadjusted LMM (including the strongest 11 significant candidates). Partial WBC adjustment upon B cell proportion cell count also reveals 18 significant probes, of which 14 are shared with the full WBC adjusted model and 17 with the unadjusted model. A total of 13 gene expression signals are common statistically significant associations across the three models (Table B.17). WBC adjustment upon the remaining cell proportion estimates reveals a number of significant transcripts that ranges between 258 and 427 (Granulocytes and CD4 lymphocytes, respectively) (Table 4.7).

4.3.2.4 Predictive Performance Assessment

In order to account for the possibility that several of the 684 identified CLL-specific signals can jointly contribute to predict the future risk of disease onset, the predictive ability was assessed by performing a stepwise logistic regression procedure. The disease status was considered as the outcome while the expression level of each of the selected candidates was included sequentially as a predictor variable in order to maximise the gain in Area Under the Curve (AUC) for the resulting Receiver Oper-

ating Characteristic (ROC) curve compared to the probe combination retained at the previous iteration. Specifically, the procedure was conducted following these steps:

1.- Include one additional probe in the model from those that are not already in.

2.- Construct the resulting ROC curve.

3.- Derive and store the AUC.

4.- Repeat step 1 to 3 for all the probes that are not already in the model.

5.- Identify the probe that maximises the AUC.

6.- Retain the probe identified in 5 in the model.

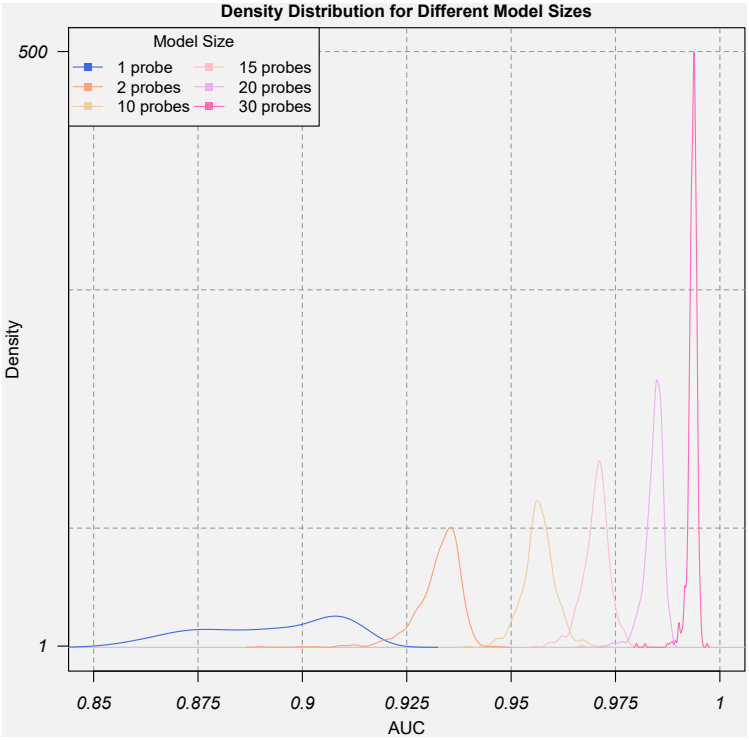
7.- Store the AUC for the model assessed in 6.

Logistic regression models of length 1 to 50 predictor variables were tested. To rule out potential overfitting, a 5-fold Cross-Validation (CV) procedure (repeated 100 times) was performed which allowed to derive density estimates of the distribution of the AUC for the 50 models under study. Results of this iterative procedure are displayed in Figure 4.6. An excellent predictive performance is shown for model of different sizes even when a single transcript is used to predict disease status: the maximum AUC found for a univariate model was based on agilent ID=A_23_P500400, gene name=ABCA6 and was over 90%. Expectedly, as more transcripts are included in the model, the predictive ability improves.

4.3.2.5 Time to Diagnosis Analysis

Stratification upon median TtD analysis reveals significant associations in the two TtD strata for the analysis including all BCL cases and controls as well as for the CLL and MM subtype specific analysis (Table 4.8). A clear overlap between these gene expression signals and the ones identified in the pooled TtD analysis can be observed. For the BCL pooled population, three of the five transcripts identified in the TtD<6 strata were also significant in the analysis including all TtD years; three of the four transcripts identified in the TtD>6 strata were also significant in the analysis includ-

Figure 4.6: Quantitative assessment of the predictive abilities of the CLL-specific signals.



The plot presents the density estimates of the AUC at different steps of the iterative procedure. Larger model sizes have been omitted as they provide consistent results.

Table 4.8: Number of significant associations identified in the WBC unadjusted LMM and the full WBC adjustment LMM categorized by disease subtype for the two TtD strata under study.

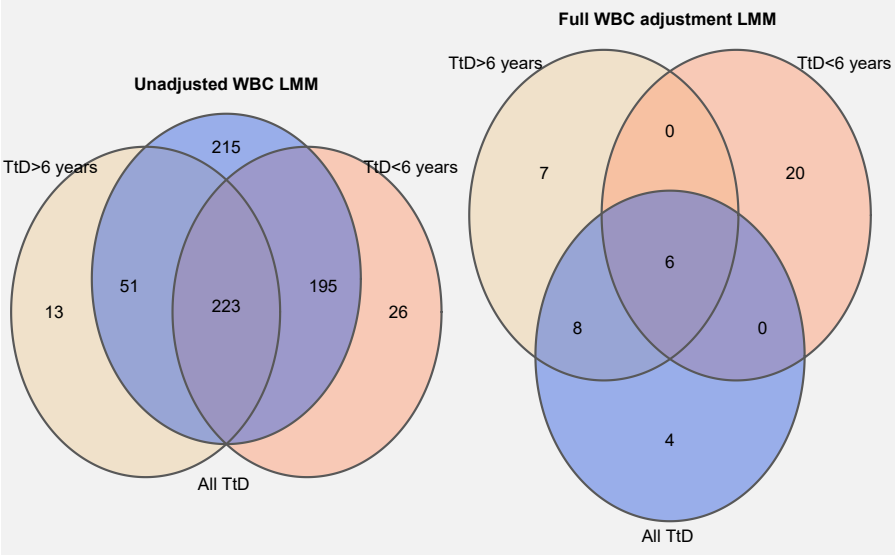
Disease Type (n)	*WBC Unadjusted LMM*		*Full WBC Adjustment*	
	TtD < 6 years	**TtD > 6 years**	**TtD <6 years**	**TtD > 6 years**
All BCL (348/348)	5	4	1	0
CLL (247/251)	444	287	26	21
DLBCL (250/251)	1	0	0	0
FL (249/252)	0	0	0	0
MM (270/261)	1	0	1	0

WBC: White Blood Cell; LMM: Linear Mixed Model, TtD: Time to Diagnosis.

ing all TtD years. Only one gene expression signal is consistently selected across the two TtD strata and the pooled analysis (Table B.18).

Expectedly, a larger overlap is observed in the CLL subtype analysis where the ex-

Figure 4.7: Venn diagrams representing the overlap of candidate signals whose expression was found significantly different in CLL cases and controls for different sub-groups (pooled population, TtD<6 years and TtD>6 years) for the unadjusted WBC LMM (left panel) and the full WBC adjustment LMM (right panel).



TtD: Time to Diagnosis, WBC: White Blood Cell, LMM: Linear Mixed Model.

pression of 223 probes was found to be significantly different between cases and controls for the three sub-groups (pooled population, TtD<6 and TtD>6) (Figure 4.7). Moreover, the effect size and strength of association of the top 10 strongest signals found in the pooled analysis narrowly change in the two TtD strata (Table B.19). Upon WBC full correction, the number of common differentially expressed genes is reduced to six candidates (Figure 4.7). Results from the MM-specific TtD analysis are not shown.

4.3.2.6 Biological Interpretation of the Findings

The transcripts ABCA6, ARHGAP44, TCF4, CDK14, ZBTB32, WNT3 and KCNN4, which are among the strongest identified association, correspond to protein-coding genes [207]. More specifically, the protein encoded by ABCA6 is a membrane-associated transporter involved in the mobilization of various molecules across extra- and intracellular membranes. Although it presents low lineage specificity, its expression is

enhanced in the blood cell types Dendritic Cells (DC) and monocytes. WNT3 belongs to a family of genes which encode for signalling proteins implicated in oncogenesis and other developmental processes such as regulation of cell fate. CDK14 plays a role as a regulator of cell cycle progression and cell proliferation and has been implicated in transcriptional misregulation in cancer; its expression is enriched in blood cells with high specificity for the granulocyte and B cell lineages. TCF4 is an activating transcription factor with a tissue-specificity expression in blood (b cells and DC). Finally, both ZBTB32 (DNA-binding protein with transcriptional repressor effect) and KCNN4 (potassium membrane channel) play a role in the activation of T-lymphocytes.

Furthermore, results from the gene-enrichment analysis showed that, of the 684 significant probes found to be differentially expressed in the CLL population, 574 (83.91%) were mapped to DAVID's database for functional annotation which were grouped into 310 different gene enriched pathways. In order to identify the most relevant to the disease outcome of interest, enriched pathways were further filtered by setting the EASE score$<$0.01, the minimal number of probes per functional group to 5, the fold enrichment value$>$3 and the Bonferroni p-value 10^{-3} (correction according to the minimum number of probes per group). This selection process identified a total of 14 gene enriched biological pathways (208 transcripts) which mainly relate to proliferation and signalling of B cells, Pleckstrin homology domain (intracellular cell signalling) and immune system regulation (Table 4.9).

Table 4.9: Summary of the results from the CLL-specific gene-enrichment analysis.

	Database	*Term*	*Count*	*p-value*	*Fold Enrichment*	*Bonferroni 5%*
1	INTERPRO	IPR011993: Pleckstrin homology-like domain	33	5.07E-09	3.35	4.06E-06
2	INTERPRO	IPR001849: Pleckstrin homology domain	24	8.28E-08	3.84	6.63E-05
3	KEGG_PATHWAY	hsa04662: B cell receptor signalling pathway	12	4.63E-07	7.38	9.27E-05
4	SMART	SM00233:PH	24	6.74E-07	3.36	1.39E-04
5	KEGG_PATHWAY	hsa05340: Primary immunodeficiency	9	8.19E-07	11.24	1.64E-04
6	KEGG_PATHWAY	hsa04672: Intestinal immune network for IgA production	10	1.07E-06	9.03	2.14E-04
7	GOTERM_BP_DIRECT	GO:0030890 positive regulation of B cell proliferation	10	2.44E-07	10.79	4.74E-04
8	UP_KEYWORDS	SH2 domain	13	5.05E-06	5.41	1.70E-03
9	UP_SEQ_FEATURE	domain:PH	21	1.04E-06	3.73	1.73E-03
10	SMART	SM00252:SH2	13	1.98E-05	4.67	4.07E-03
11	KEGG_PATHWAY	hsa04064:NF-kappa B signalling pathway	11	3.17E-05	5.37	6.31E-03
12	UP_SEQ_FEATURE	short sequence motif:ITIM motif 4	5	3.95E-06	36.27	6.53E-03
13	GOTERM_BP_DIRECT	GO:0050853 B cell receptor signalling pathway	10	4.46E-06	7.79	8.65E-03
14	INTERPRO	IPR000980:SH2 domain	13	1.14E-05	4.99	9.13E-03

Biological pathways are ordered in relation to their Bonferroni adjusted *p*-values.

4.3.2.7 Sensitivity Analysis

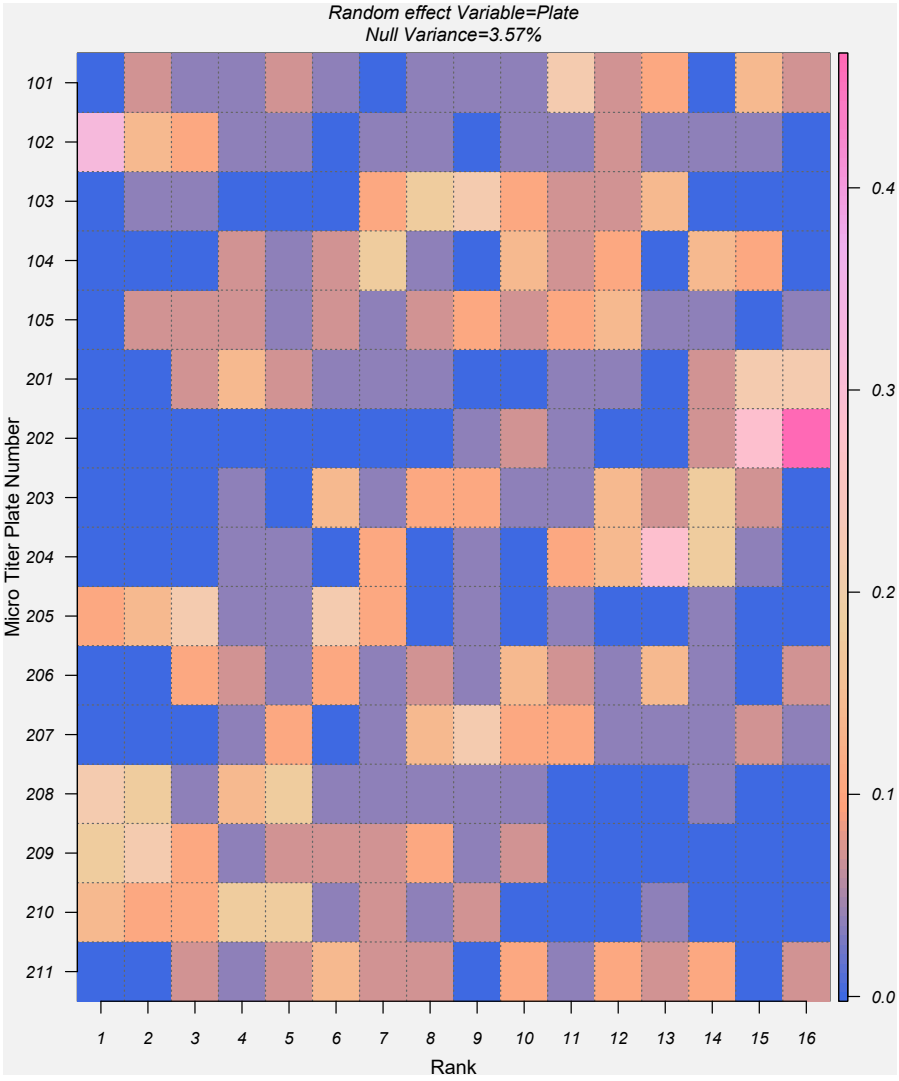
Analogously to the analysis conducted on the proteomics dataset, I assessed the consistency of the findings across the two cohort populations (Epic-Italy and NSHDS) and the two analytical phases (1 and 2). Significant associations were identified in the analysis including all BCL observations (NSHDS, n = 1 and phase 2, n = 2), MM (phase 1, n = 2) and CLL subtypes (Epic-Italy, n = 99; NSHDS, n = 461; phase 1, n = 112 and phase 2, n = 560).

Focusing on the CLL-specific signals, weaker p-values and effect sizes were observed in the transcripts found in the different sub-groups comparing to the ones obtained in the pooled analysis (results not shown). This is particularly true for EPIC-Italy and analytical phase 1, possible to the lower number of observations (11 cases and 76 controls for EPIC-Italy and 9 cases and 107 controls for phase 1). There is an overlap of the strongest CLL-specific gene expression signals across the two study populations (n = 62) and the two experimental phases (n = 92). A total of 46 are significantly differently expressed across the four sub-groups, including the 20 strongest probes found in the pooled analysis. Expectedly, weaker signals do not systematically replicate in the four sub-groups. The findings obtained from this sensitivity analysis provide technical replication of the results as well as validation of the epidemiological study design.

4.3.3 Assessment of Technical-induced Noise

For proteomics and transcriptomics, the analysis of the ranking distribution (over the 28 and 29,662 models, respectively) for each of the levels of the random effect variables clearly shows some plates and dates that were consistently found to generate higher variances and therefore more noise. For example, it seems that samples that were analysed in plate numbers 102, 205, 208, 209 and 210 (five out of the 16 plates) are associated to generate higher noise (Figure 4.8), which is a consistent finding considering the 3.57% percentage of null variance. For transcriptomics, only two RNA

Figure 4.8: Ranking distribution (over the 28 LMM models) of the estimated random intercept for each of the micro titer plate numbers in the proteomics analysis.



isolation dates were estimated to produce marginal noise (2010-12-13 and 2011-02-08) while such experimental dates are more numerous for array hybridisation where 10 analytic days seemed to have been associated with higher variance (Figure 4.9).

Figure 4.9: Ranking distribution (over the 29,662 LMM models) of the estimated random intercept for each of the experimental dates in the transcriptomics analysis.

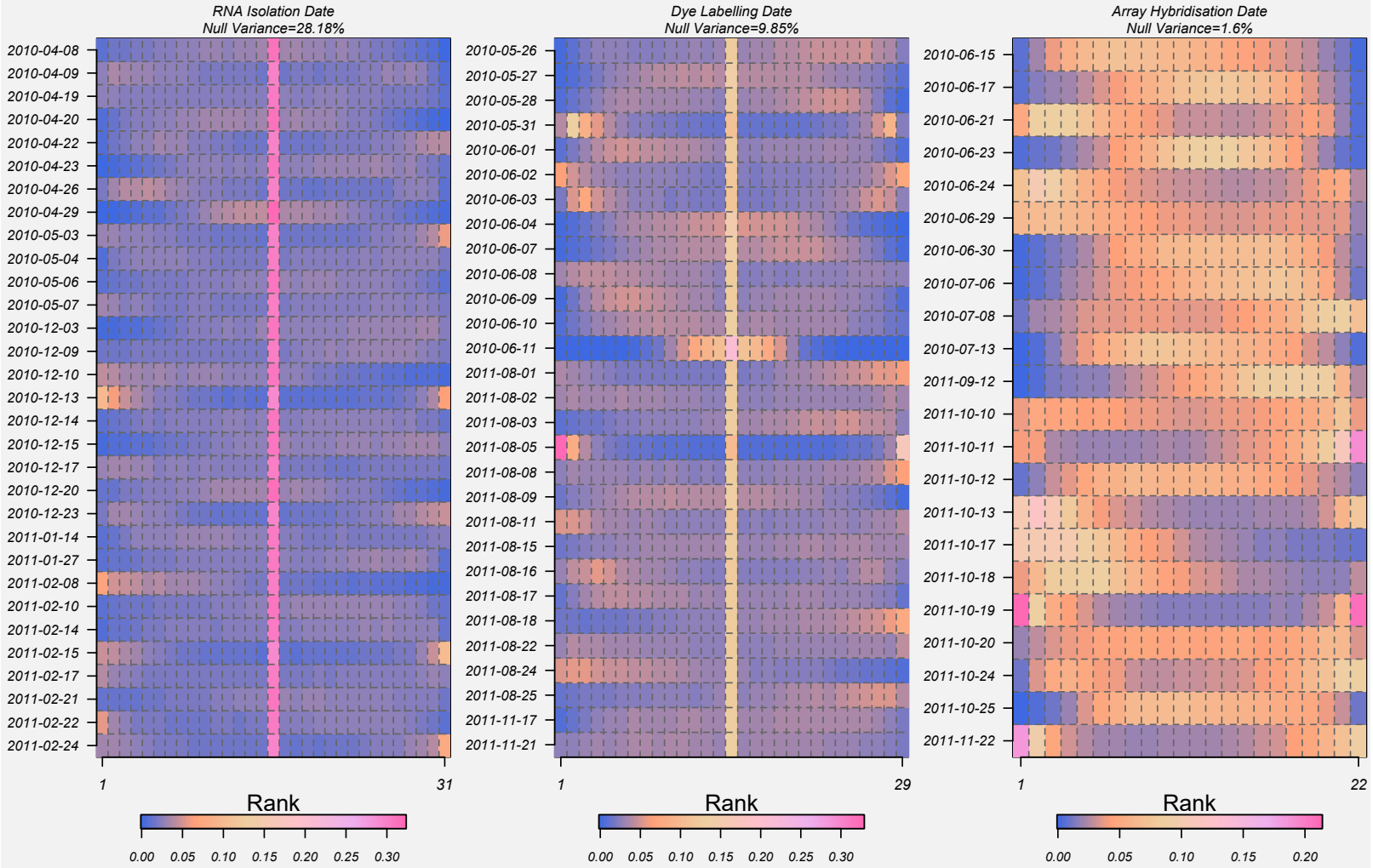
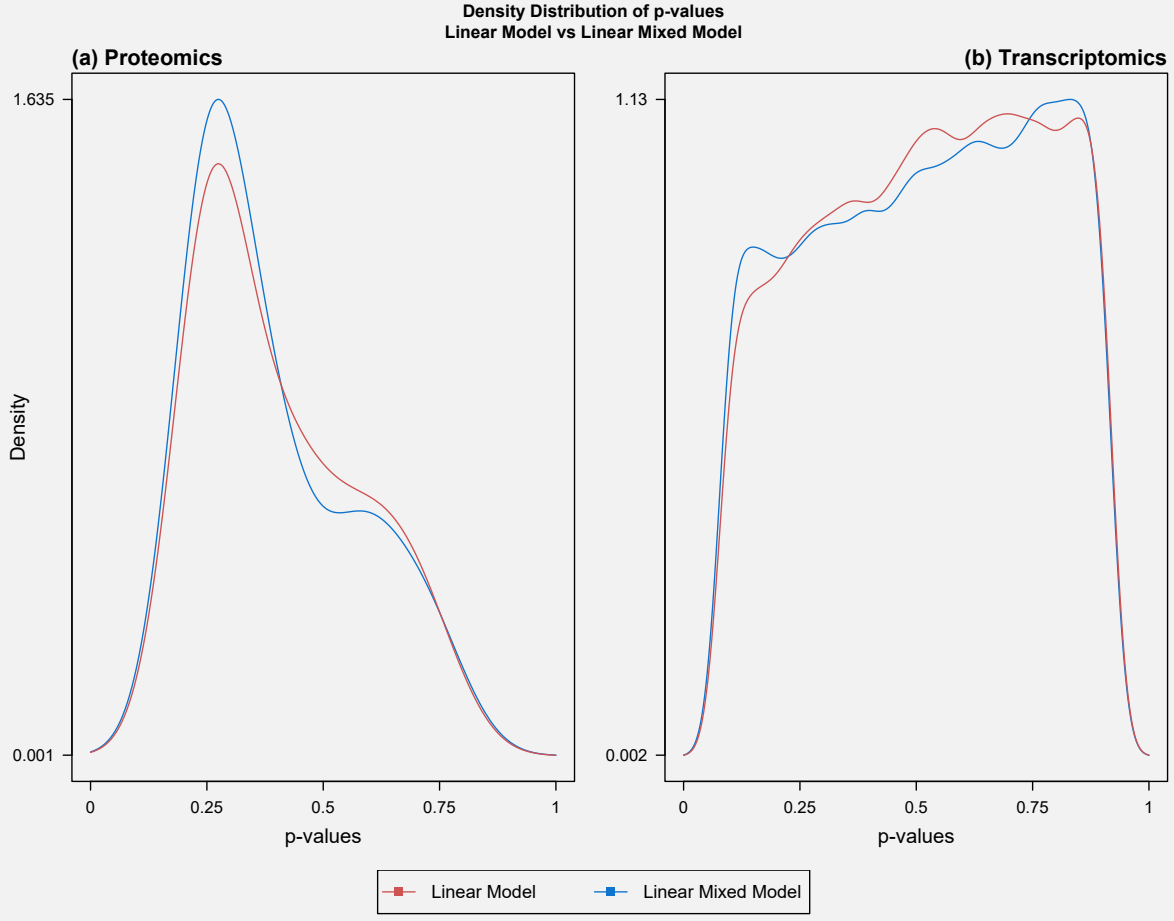


Figure 4.10: Density distribution of p-values for the LMM and the corresponding linear model without the random effect term for the proteomics (panel a) and transcriptomics datasets (panel b).



The comparison of p-value distributions obtained from the LMM to those obtained under the corresponding linear model (setting the random intercept to zero, not accounting for nuisance variation) for the proteomics and transcriptomics data is illustrated in Figure 4.10. Linear models exhibit a more typical null p-value distribution while the inclusion of a random effect term sharpens the distribution for smaller p-values, providing a stronger support for the alternative hypothesis. This finding suggests that random effects successfully estimated technical-induced variation and limited the subsequent dilution effect of nuisance variance, yielding better power to detect significant associations.

4.4 Discussion

In the analyses presented in this chapter, I used blood samples collected years before clinical diagnosis to interrogate the relationship between pre-diagnostic blood levels of immune and transcriptomic markers and future risk of BCL and its main histological subtypes using established univariate statistical approaches. As expected from the biological heterogeneity of BCLs, the results do not support the existence of markers whose change in concentration or expression is common to the pathogenesis of all or multiple histological subtypes of BCL. Instead, and despite the limited number of cases available per disease subtype, the analyses led to the identification of several strong signals associated with prospective MM risk in proteomics and CLL subtype in transcriptomics. In particular, the following immune markers were found consistently inversely associated with MM incidence: FGF2, VEGF, TGFa, Fractalkine, MCP3 and MIP1a while the expression of the following genes was shown to be associated to CLL incidence: ABCA6, ARHGAP44, TCF4, CDK14, ZBTB32, WNT3 and KCNN4 (strongest findings). The strongest associations seem to persist among cases sampled more than six years before clinical diagnosis.

4.4.1 Precursor States

The two BCL subtype for which positive association were found are preceded by asymptomatic pre-malignant precursor states, namely Monoclonal Gammopathy of Uncertain Significance (MGUS) and Monoclonal B Lymphocytosis (MBL) for MM and CLL, respectively. MGUS is characterised by the proliferation of monoclonal plasma cells derived from post-germinal-centre (GC) B cells and is a condition with an incidence of 1% in the population older than 50 years (and up to 10% older than 75 years) [208]. Evidence suggests that most cases of MM are preceded by MGUS whose transformation rate is 1% to 2% per year [209]. On the other hand, MBL is characterised by the presence of circulating monoclonal B cells and no other signs of a lymphoproliferative disorder and is found in approximately 3% of normal people; CLL is always

preceded by MBL [210]. It has also been shown that MBL may progress into CLL with a frequency of 1% to 2% per year but it is not understood whether individual subjects with MBL will or will not progress into CLL or when this event will occur [210]. Thus, the biological alterations found in the analysis conducted in this chapter may be a reflection of these two pre-clinical conditions and these two entities can partially explain why significant associations were not identified for the other subtypes under study as no precursor states are described for DLBCL and FL.

More specifically, gene expression profiles conducted on MBL support the possibility that the CLL-related signals observed in this analysis may arise, at least to some extent, from its pre-clinical condition. This hypothesis is supported by the fact that a number of genes related to WNT signalling (e.g. WNT3, WNT10A, ARHGAP44, TCF4, CDK14, and ZBTB32) have been reported to be activated in MBL [211] which are among the differentially expressed genes with the largest effect sizes observed in this chapter. In addition, of 20 genes reported as being differentially expressed in MBL [211], four (PRKCB, PAG1, TCL1A, ROR1) fall among the CLL-related genes identified in the current analysis. On the other hand, biological pathways reported to be activated in MBL cells (e.g. the MAPKinase and protein kinase A pathways) were not among those indicated by the CLL-related profile observed here. As such, the identified profile seems to be driven only partially by MBL. This is strengthened by the observation that the identified CLL-transcriptomic profile predicts more than 90% of the cases, whereas only $\sim$5%-10% of subjects of the MBL phenotype would be expected to progress to CLL over the six-year average follow-up period of this study. These observations are compatible with the possibility that the CLL differential expression profile detected is due to clones of malignant or premalignant cells, including MBL cells, present at low concentrations in the blood samples several years before clinical onset, which evolve toward CLL via specific transcriptomic signals in an indolent manner at a slow progression rate.

In contrast, the suggestion that the changes in immune marker levels in the MM subtype-specific analyses may partially arise from MGUS seems to be more difficult

to support as the markers revealed in this chapter are inversely related to long-term risk of MM. Although these markers have been reported previously in clinical studies of MM or MGUS [212], [213], [214], [215], [215], the direction of findings reported here is in general opposite to the results found among subjects diagnosed with MM, where higher concentrations seem to be related to generally poorer disease outcome. The reason for this difference in direction of the effect is not known but may direct towards a preclinical deregulation of these important biological systems in individuals developing MM later in life which at the time of clinical manifestation reverse in overexpression. Therefore, external replication of this study is needed either in healthy individuals in an epidemiological setting or in individuals with the precursor condition in clinical studies in order to further clarify these findings. This may be of importance as this disease has a low survival rate (average survival rate is 5 years) [216] and confirmation of the results could lead to the identification of patients at higher risk of progressing to MM and in the long-term could improve individualized surveillance strategies.

4.4.2 Biological Relevance of Findings

For CLL subtype in the transcriptomics analysis, a substantial overexpression (up to 25-fold) was found for the strongest signals in cases compared to controls in addition to a trend towards increased expression while approaching clinical onset demonstrated by the TtD analyses. These two observations suggest that the CLL-related profile reflect to some extent markers of disease progression arising from subpopulations of cells in which disease initiation has occurred long before diagnosis. This is further supported by the fact that some of the strongest associations (e.g. ARHGAP44, ABCA6, and WNT3) are highly overexpressed in CLL malignant cells [217], [218]. This premise is consistent with the clinical progression of the disease as CLL patients present an insidious and indolent course for a third of cases with signs and symptoms taking years to arise.

Several genetics lesions have been reported to play a role in the pathogenesis of

CLL which have been identified using molecular or conventional cytogenetics in case samples and include mainly tumour suppressor gene deletions as a consequence of specific chromosome regions being deleted. The genes involved are ATM and TP53 whose deletion or mutation are found in 30% and 15% of patients (respectively) [167] while the chromosomal abnormalities commonly found are deletion of 13q or trisomy 12, del(13)(q14), del(11)(q22–23), del(17)(p13) and del(6)(q21); these genetic lesions can be acquired during the course of the disease. The findings revealed by the analysis conducted in this chapter do not support strong evidence of chromosome specificity for the signals, except possibly for chromosomes 17, 18, and 19. However, the identified transcripts point toward an important contribution of B-cell signalling, and B-cell activation and proliferation in the aetiology of CLL. For example, the nuclear factor-kappa B (NF-κB) is essential at different stages during mature B cell differentiation in the GC reaction and its activation has been previously reported to be related to the pathogenesis of BCL [138].

For MM subtype in the proteomics analysis, the findings provide evidence for a strong link between FGF2, VEGF, TGFa levels and incidence of this particular subentity; these three inflammatory markers correspond to growth factors. Several clinical studies have reported that the plasma concentrations of FGF2 were elevated in patients with active MM compared to patients with inactive disease, and this correlates with increased bone marrow angiogenesis and lymphangiogenesis [219], [220], [221],[222]. In addition, it has been shown that MM patients who respond to chemotherapy (an immunosuppressed condition) show a significant decrease in serum FGF2 levels, whereas non-responders do not [222]. Similarly, clinical studies have reported that increased serum levels of VEGF are associated with more advanced disease stages and with poor prognosis in BCL and MM cases [219], [220], [222]. It is known that VEGF and its ligands and receptors play a central role in physiological regulation of angiogenesis as well other nonvascular roles including recruitment of inflammatory cells and autocrine and intracrine production of hematopoietic stem cells [223]. Finally, TGFa is an important mitogen that binds to the EGF recep-

tor and has been studied in many other malignancies, but data on MM are limited and no prospective data are available [224],[225], [226]. These observations support a possible role of the growth factors in the pathogenesis of MM. Given their interrelationship and cyclic response, more in-depth monitoring of the FGF2, VEGF, TGFa and its soluble receptors is needed to clarify their possible pre-diagnostic role in this BCL subtype.

4.4.3 WBC Correction Analyses

Both proteomics and transcriptomics analyses correcting for intra sample heterogeneity provided consistent results in comparison to the unadjusted LMMs in the sense that significant associations were found only for MM and CLL subtypes, respectively. For MM, the strongest associations consistently replicate across all models performed. However, the CLL-specific signals show to be more affected by both partial and full WBC adjustment as the number of positive findings was greatly reduced with B cell proportion mainly driven the difference. This suggests that the alteration in B-cell counts as a result of the disease phenotype may have been confounding and driving most of the identified associations with only 18 transcripts surviving the adjustment. As these signals were also the ones showing the strongest associations in the unadjusted model, they may reflect the key genes and relevant biological pathways that are behind the B-cell malignancy process. Moreover, one could argue that this is an expected finding as the pathogenesis of the disease (dysfunctional activation and differentiation of B cells) and its clinical presentation (lymphocytosis) reflect alterations in the normal expression patterns of B cells, in contrast to MM, for example, which affects plasma cells.

4.4.4 Assessment of Technical-induced Noise

The analysis using LMMs suggests that the inclusion of random effect terms effectively restricts the dilution effect of nuisance variation. As discussed in section 3.5, micro titer plate number and experimental dates act as surrogate or proxy variables

for other sources of variation. In the case of proteomics where a higher percentage of null variance is observed, these sources of variation may be partially explained by the use of different blood sample anticoagulants in the two study cohorts (citrate in EPIC-Italy and Ethylene Diamine Tetraacetic Acid (EDTA) in NSHDS) as it is known that different media results in absolute differences in levels of immune markers. On the other hand, in the case of transcriptomics, of the three experimental dates included as random effects, array hybridisation appears to have been associated with higher variance which is, to some extent, an expected output as hybridization is the most time-consuming step of the entire RNA microarray procedure (approximately 17 hours). The dynamics of this process relies on many factors depending on both the reaction conditions and structural properties of the individual RNA molecules which may significantly affect the experimental outcomes. For example, evaporation of some of the water can change the salt concentration in the buffers and significantly affect the efficiency of the process [39].

4.4.5 Study Design Implications

The epidemiological study design employed in this chapter has a number of strengths, including its prospective nature, which limits reverse causation bias that may occur when variation in blood level of circulating levels of biological markers is induced by the disease itself, cancer treatments or lifestyle changes after cancer diagnosis. The availability of two different cohorts allowed for independent confirmation of the observed associations which were consistent for both the proteomics and transcriptomics analyses. In addition, in the specific case of proteomics, a larger panel of immune markers was simultaneously measured which constitutes an advantage in comparison to most previous prospective studies where individual markers are assessed independently.

On the other hand, the study design may be associated with some limitations. Bias may arise from omics profiles measurements of study subjects in two analytical phases despite adjustment in the LMM. However, stratified analyses by cohort and phase

showed overall similar trends for the identified signals despite the associated reduced power. Moreover, omics profiles were measured at a single time point to determine future risk of BCL which may not accurately reflect the long-term immune status and gene expression profile of an individual. Nevertheless, several studies have provided evidence of a reasonable between-to-within person variability ratio suggesting temporal stability for panels of both proteomics and transcriptomics profiles [227], [228], [229]. Finally, it cannot be ruled out that some of the results identified in this chapter such as the inverse association between significant immune markers and disease status or the lack of significant signals for other subtypes, may be explained by the reduced sample size.

4.5 Conclusion

The relationship between EGM omics markers and future risk of BCL was assessed by employing established univariate statistical methods. In particular, LMMs were preferred for their ability to correct for technical-induced variation, a recognized factor to affect the reliability of omics experiments. The results presented in this chapter showed that the proteins FGF2, VEGF and TGFa are inversely associated with MM incidence while the transcripts ABCA6, ARHGAP44, TCF4, CDK14, ZBTB32, WNT3 and KCNN4 are among the strongest signals positively associated with CLL risk. Consequently, LMMs proved to be successful in disclosing relevant associations in the investigated real-world omics data; however, as these statistical approaches model individual markers separately, interesting biological patterns may have been overlooked. In the following chapter I examine this hypothesis by analysing the proteomics and transcriptomics datasets under a dimension-reduction and supervised-learning statistical framework.

This page intentionally left blank.

5

Application of Partial Least Squares Techniques to the Proteomics and Transcriptomics Datasets: Moving from Univariate towards Multivariate Approaches

This chapter is based in part on the publication:

> R. Vermeulen, F.S. Hosnijeh, B. Bodinier, L. Portengen, B. Liquet, **J. Garrido-Manriquez**, H. Lokhorst, I. A. Bergdahl, S. A. Kyrtopoulos, A. Johansson, P. Georgiadis, B. Melin, D. Palli, V. Krogh, S. Panico, C. Sacerdote, R. Tumino, P. Vineis, R. Castagné, M. Chadeau-Hyam, M. Botsivali, A. Chatziioannou, I. Valavanis, J.C.S. Kleinjans, T.M.C.M. de Kok, H.C. Keun, T. J. Athersuch, R. Kelly, P. Lenner, G. Hallmans, E.G. Stephanou, A. Myridakis, M. Kogevinas, L. Fazzo, M. De Santis, P. Comba, B. Bendinelli, H. Kiviranta, P. Rantakokko, R. Airaksinen, P. Ruokojarvi, M. Gilthorpe, S. Fleming, T. Fleming, Y. Tu, T. Lundh, K. Chien, W.J. Chen, W. Lee, C.K. Hsiao, P. Kuo, H. Hung and S. Liao. "Pre-diagnostic blood immune markers, incidence and progression of B-cell lymphoma and multiple myeloma: Univariate and functionally informed multivariate analyses". *International Journal of Cancer* **143**.6 (2018), 1335–1347.

All the analyses shown here have been independently replicated by me. A descrip-

tion of the overlapping analysis as well as improvements made by this chapter is provided in section C.1.

5.1 Introduction

Building upon the findings described previously where established univariate methods were employed for the analyses of omics profiles, in this chapter I use supervised Dimension Reduction Techniques (DRTs) to analyse the proteomics and transcriptomics datasets from the EnviroGenoMarkers (EGM) study. As before, the overarching aim is to identify inflammatory markers and gene expression signals indicative of future risk of B-cell Lymphoma (BCL) and its main histological subtypes; however, here special emphasis is placed on investigating the usability and scalability of the applied statistical approaches. The specific techniques employed are the regularized versions of Partial Least Squares (rPLS) which are used in a discriminatory analysis context in order to identify the markers that best differentiate between the sample groups. Therefore, the process of selecting the most relevant features driving the variation as well as the incorporation of meaningful information related to pre-existing groups of features with a similar biological function are included in the statistical analyses. The strengths and limitations of introducing sparsity at different levels of the feature hierarchy are explored and assessed in terms of interpretability, statistical performance, robustness and biological relevance.

5.2 Methods

The supervised multivariate approaches sparse, group and sparse-group Partial Least Squares-Discriminant Analysis ([s][g][sg]PLS-DA, respectively) were applied separately on the immune markers and gene expression datasets. The predictor matrices employed for the statistical analyses correspond to the "de-noised" immune marker concentration and gene expression level estimates obtained by subtracting the ran-

dom effect term(s) from the original raw levels from the results of the Linear Mixed Model (LMM) described previously (see section 3.5). The outcome matrix employed was a two-column dummy matrix indicating whether each individual observation was a lymphoma case or a control (i.e. binary classification problem, $G = 2$). Statistical analyses were conducted on the full BCL pooled population and stratified by major histological subtypes: Chronic Lymphocytic Leukaemia (CLL), Diffuse Large B-cell Lymphoma (DLBCL), Follicular Lymphoma (FL), and Multiple Myeloma (MM). Disease subtype stratification was performed including cases of the corresponding subtype and its paired control subjects (i.e. there is an equal number of cases and controls in all stratified analyses). In proteomics, the total number of case-control pairs included in the study populations correspond to 536, 84, 88, 78 and 152 for all BCL, CLL, DLBCL, FL and MM, respectively; while for transcriptomics, these numbers correspond to 464, 68, 74, 74 and 134. I use the R-statistical package `mixOmics` to fit the sPLS-DA statistical models and the `sgPLS` package to fit the gPLS-DA and sgPLS-DA models.

For the group and sparse-group analyses, individual variables were grouped by the similarity of their molecular and biological functions. For the proteomics analyses, the 28 immune markers were classified into three different functional modules: growth factors, chemokines and cytokines including 6, 10 and 12 proteins, respectively (see Table 3.2 for details). In the case of the gene expression dataset, transcripts were allocated to biological pathways using the online bioinformatics tool Database for Annotation, Visualization and Integrated Discovery (DAVID v6.8, http://david.abcc.ncifcrf.gov/). The grouping procedure was conducted ensuring that transcripts were solely allocated to a single pathway, thus avoiding the possibility of overlap of probes across pathways. Of the 29,662 transcripts, a total of 14,925 were recognized as being part of a relevant biological pathway accounting for 50.32% of the probes assayed and a total of 849 different groups were identified; 174 of them (0.59%) constitute pathways containing only a single variable. In order to minimise information loss, the remaining 14,737 probes (49.68%) were grouped into a single bi-

ological pathway of non-annotated features. More details on the biological pathways to which individual transcripts were grouped are provided on Table C.1.

5.2.1 Calibration Procedure: Defining the Optimal Value of Parameters

For the three regularized PLS-DA techniques, a model calibration procedure was conducted to optimize the corresponding tuning parameters by means of a sequential strategy where we first seek to determine the optimal number of dimensions followed by the optimization of the parameters defining the degree of sparsity (see section 3.6.2.1.1). Considering that the research question poses a binary classification problem and under the assumption that the inclusion of additional latent variables mainly accounts for variation within the predictor matrix, the total number of components explored was two (H = 2). The R^2 or percentage of explained variation in the outcome matrix and the Discriminant Q^2 (DQ^2) value were the explored diagnostic statistics to determine the optimal number of PLS dimensions.

Once the optimal number of latent variables was defined, the additional model parameters were tuned using M-k-fold Cross Validation (CV, k=5 and M=100) and the overall misclassification Error Rate (ER) averaged across the 100 repetitions was computed for each component as the chosen metric of performance. To avoid the possible introduction of bias, the CV procedure was conducted ensuring that cases-control pairs were allocated to the same fold in each of the iterations. Predictions of new observations were made using the maximum value (or maximum distance) as the Decision Rule (DR) to translate the predicted values into meaningful class memberships; a choice made considering the sample groups were of equal size and to allow for the possibility of unequal variance. One-dimensional grids of values of the tuning parameters were employed for both sPLS-DA and gPLS-DA while a two-dimensional grid was used for sgPLS-DA as the optimal degree of sparsity in the predictor matrix depends on two parameters. The specific additional model parameters required to be calibrated for each of the three statistical methods are detailed as follows:

- sPLS-DA: number of individual variables to retain in the predictor matrix (inflammatory markers in proteomics and gene expression signals in transcriptomics).
- gPLS-DA: number of modules to retain in the predictor matrix (functional groups in proteomics and biological pathways in transcriptomics).
- sgPLS-DA: number of modules to retain in the predictor matrix (functional groups in proteomics and biological pathways in transcriptomics) and degree of sparsity within the groups (mixing parameter α_1).

The reduced size of the proteomics matrix in terms of both individual features and functional groups allows for the complete range of values to be explored during the calibration procedure. Therefore, the specific values examined were:

- sPLS-DA: models retaining 1 to 28 proteins (grid resolution of 28 values).
- gPLS-DA: models retaining 1 to 3 functional groups (grid resolution of 3 values).
- sgPLS-DA: models retaining 1 to 3 functional groups (grid resolution of 3 values) and a within-group degree of sparsity (α_1) ranging from 0.05 to 0.95 (grid resolution of 11 values).

On the other hand, the high-dimensional nature of the gene expression data constrains the possible space of model parameters that can be investigated. Thus, the following values were tested:

- sPLS-DA: models retaining 2 to 1000 transcripts (grid resolution of 40 values).
- gPLS-DA: models retaining 1 to 150 biological pathways (grid resolution of 28 values).
- sgPLS-DA: models retaining 1 to 150 biological pathways (grid resolution of 28 values) and a within-group degree of sparsity (α_1) ranging from 0.05 to 0.95 (grid resolution of 11 values).

These sequences were defined under the main assumptions that the inclusion of ad-

ditional variables or modules produce minimal reductions of the ER, models with an excessive number of features have a reduced capacity to extract information of biological relevance and the retention of more variables produce unnecessary high computational costs. It is worth to point out that the grid employed for the calibration of the number of modules in gPLS-DA is the same as the one employed in sgPLS-DA (for both the proteomics and transcriptomics analyses); likewise, the same grid of values of the mixing parameter α_1 was used in the proteomics and transcriptomics sgPLS-DA analyses.

5.2.2 Assessment of Calibrated Models

Once the corresponding optimal parameters were tuned for the three PLS-DA variants, the classification performance of the models was evaluated by CV under the same criteria applied in the calibration procedure (k=5, M=100 and paired case-control subjects being included in the same fold). The computation of the overall misclassification ER, ER per observation type and the Area Under the Curve (AUC) averaged across CV folds were the metrics of performance of choice to compare the classification abilities of the models for the pooled BCL population and the stratified populations by major histological subtypes.

The relevant graphical outputs that could be displayed in this discriminatory analysis context were assessed. Sample representation plots were employed to examine the classification capacity of optimized models with more than one PLS dimension in a graphical manner. The contribution of individual variables was assessed by the inspection of the loading coefficient plots. Moreover, the stability frequency of the selected features was examined in order to verify the results are consistent within the population and not driven by outlier study participants. The graphical representation of the stability frequency was obtained by repeatedly fitting the regularized models in subsets of the original datasets and recording the frequency the features were selected across the models visited. More specifically, the model fitting procedure was repeated 500 times in 80% of the study population; in each iteration a different pop-

ulation subset was employed. The entire process was applied for different values of the corresponding model parameters. In order words, the stability frequency output was obtained as a by-product of the calibration procedure (see also section 3.6.2.4). In addition, it is important to mention that for models selecting more than one component, the stability frequency of the second component was computed retaining in the first dimension the number of features or modules reported as optimal.

Finally, and when appropriate in the case of the (s)(g)(sg)PLS-DA transcriptomics analyses, the selected gene expression signals were further investigated through gene-enrichment analyses using DAVID.

5.3 Results

5.3.1 Proteomics

5.3.1.1 Calibration Procedure and Assessment of Calibrated Models

The values of R^2 and DQ^2 statistics obtained from a classical PLS-DA model fitted with up to two latent variables are shown in Table C.2. Figure C.1 and Figure C.2 display the calibration curves representing the average overall misclassification ER for all the values of the parameters tested for the sPLS-DA and sgPLS-DA models, respectively. Table C.3 shows the ER obtained from the gPLS-DA model retaining one to three functional groups. Based on these results, a decision was made to define both the optimal number of components and the optimal value of the additional parameters determining model sparsity. Table 5.1 details the values of the optimal parameters employed in the calibrated models, the resulting number of inflammatory markers and functional groups selected per component as well as the total number of unique proteins and modules selected across components. Such information is presented for the three statistical methods being applied and for both the pooled BCL analysis and the four stratified analyses by major disease subtype.

Table 5.1: Parameters used in the calibrated model, number of unique inflammatory markers and functional groups selected per component and total number of unique proteins and modules selected across components for the three regularized approaches and for the five study populations.

	All BCL	CLL	DLBCL	FL	MM
sparse PLS-DA					
N Components	2	2	1	1	2
N Variables 1st component	12	3	4	1	2
N Groups 1st component	3	2	2	1	1
N Variables 2nd component	3	4	—	—	1
N Groups 2nd component	2	3	—	—	1
Total N Variables	13	7	4	1	3
Total N Groups	3	3	2	1	2
group PLS-DA					
N Components	1	1	1	1	1
N Variables 1st component	28	28	10	6	28
N Groups 1st component	3	3	1	1	3
sparse group PLS-DA					
N Components	2	1	1	1	1
N Variables 1st component	26	2	9	4	10
N Groups 1st component	3	1	2	1	3
α_1 1st component	0.3	0.9	0.9	0.7	0.95
N Variables 2nd component	4	—	—	—	—
N Groups 2nd component	1	—	—	—	—
α_1 2nd component	0.9	—	—	—	—
Total N Variables	26	2	9	4	10
Total N Groups	3	1	2	1	3

Cell colours indicate the functional group(s) which was (were) selected in the component or final model: orange for growth factor, blue for chemokine and yellow for cytokine.
PLS-DA: Partial Least Squares-Discriminant Analysis.

5.3.1.1.1 Pooled Population

The multivariate sPLS-DA analysis pooling all BCL cases together retained two PLS dimensions with 12 and three proteins being selected in each component. The first component selected the proteins TGFa, FGF2, VEGF and EGF which belong to the

growth factor category; MDC, MIP1a, MCP3 and Fractalkine which belong to the chemokine group and sCD40L, IL1b, TNFa and INFg which are part of the cytokine module. On the other hand, the three inflammatory markers selected in the second component correspond to two cytokines (sCD40L and TNFa) and one chemokine (Eotaxin). Since the cytokines sCD40L and TNFa overlap between components, a total of 13 unique proteins were selected in the calibrated sPLS-DA model.

The gPLS-DA analysis selected the model with one component and with the three functional groups being retained in it; therefore, all proteins were selected, and no sparsity was effectively imposed in the calibrated model.

Two components were chosen in the sgPLS-DA model with the three functional groups being retained in the first dimension and one in the second, the cytokine group was selected in the latter. As a result of including the optimized value of the mixing parameter α_1, 26 proteins were retained in the first component and four in the second. All features but MCP1 (a chemokine) and IL13 (a cytokine) were kept in the first component while in the second latent variable the cytokines sCD40L, TNFa, IL7 and IL10 were retained, which comprise four of the 12 markers that are part of the mentioned functional group. A complete overlap of features is observed between components; thus, 26 unique proteins are part of the optimized sgPLS-DA model.

5.3.1.1.2 Subtype Stratified Analysis

The sPLS-DA subtype-specific analyses selected two components for the CLL and MM sub-entities and one for the DLBCL and FL. Seven unique proteins were retained in the CLL model, three kept in the first dimension and four in the second, which are part of the three functional groups. sCD40L, IL4 (cytokines) and Eotaxin (chemokine) were selected in the first component while TGFa, GCSF (growth factors), TNFa (cytokine) and IL8 (chemokine) were selected in the second component. The MM-specific sparse model selected three variables in total: two in the first dimension (the growth factors TGFa and FGF2) and one in the second (the chemokine IP10). For the disease entities DLBCL and FL the variables MDC, Eotaxin (growth

factors) and sCD40L, IL13 (cytokines) are part of the first subtype sparse model while EGF (growth factor) is part of the second subtype sparse model.

Comparable to what was found in the analysis pooling all BCL cases, the gPLS-DA calibrated model in all four subtype-specific analyses selected only one component. The three functional groups were retained in the CLL and MM analyses and one in the DLBCL and FL analyses (chemokine and growth factor, respectively).

Similarly, one latent variable was chosen in all four stratified sgPLS-DA models with one functional group being retained in the CLL- and FL-, two in DLBCL- and three in the MM-specific analyses. In the CLL model, the cytokine group was selected in the model and as a result of imposing within-group sparsity two variables were retained which correspond to sCD40L and IL4. The FL model retained four out the six growth factors while the DLBCL model kept proteins from the chemokine (four markers) and cytokine (five markers) groups. Finally, 12 proteins were selected in the MM-specific analysis comprising three growth factors (TGFa, FGF2 and VEGF), four chemokines (MCP3, Fractalkine, MIP1a and MIP1b) and three cytokines (IL1b, IL13 and INFg).

5.3.1.2 Model Performance

The classification abilities of the three optimized models for both the BCL pooled population and subtype stratified study populations are shown in Table 5.2. For the disease subtypes DLBCL and FL, the overall misclassification ER and the ER per observation type is equal or above the threshold value of 0.5, which implies that models perform no better than chance to correctly classify observations. Given those results, subsequent analyses focus on the BCL pooled population and the CLL and MM observation types (n = 536, n = 84 and n = 152 case-control pairs, respectively). In terms of both overall misclassification ER and AUC, the CLL-specific analysis exhibits the best classification performance followed by the pooled and MM-specific analyses. This observation holds for the three regularized models, with the exception of the gPLS-DA model where the overall ER is lowest for the pooled population followed by CLL and MM subtypes. Moreover, it is noted that the ER is consistently lower

Table 5.2: Classification performances of the three calibrated PLS-DA models for the five study populations (proteomics dataset).

	sparse PLS-DA				*group PLS-DA*				*sparse group PLS-DA*			
	Overall ER	**ER Control**	**ER Cases**	**AUC**	**Overall ER**	**ER Control**	**ER Cases**	**AUC**	**Overall ER**	**ER Control**	**ER Cases**	**AUC**
All BCL	0.406	0.419	0.393	0.594	0.414	0.426	0.402	0.587	0.394	0.402	0.387	0.606
CLL	0.368	0.425	0.311	0.646	0.456	0.501	0.411	0.588	0.357	0.380	0.334	0.668
DLBCL	0.530	0.503	0.557	0.574	0.525	0.522	0.529	0.579	0.523	0.498	0.548	0.574
FL	0.542	0.532	0.552	0.559	0.534	0.547	0.521	0.565	0.531	0.537	0.524	0.569
MM	0.440	0.428	0.452	0.576	0.421	0.464	0.377	0.586	0.413	0.413	0.413	0.593

PLS-DA: Partial Least Squares-Discriminant Analysis, ER: Error Rate; AUC: Area Under the Curve.

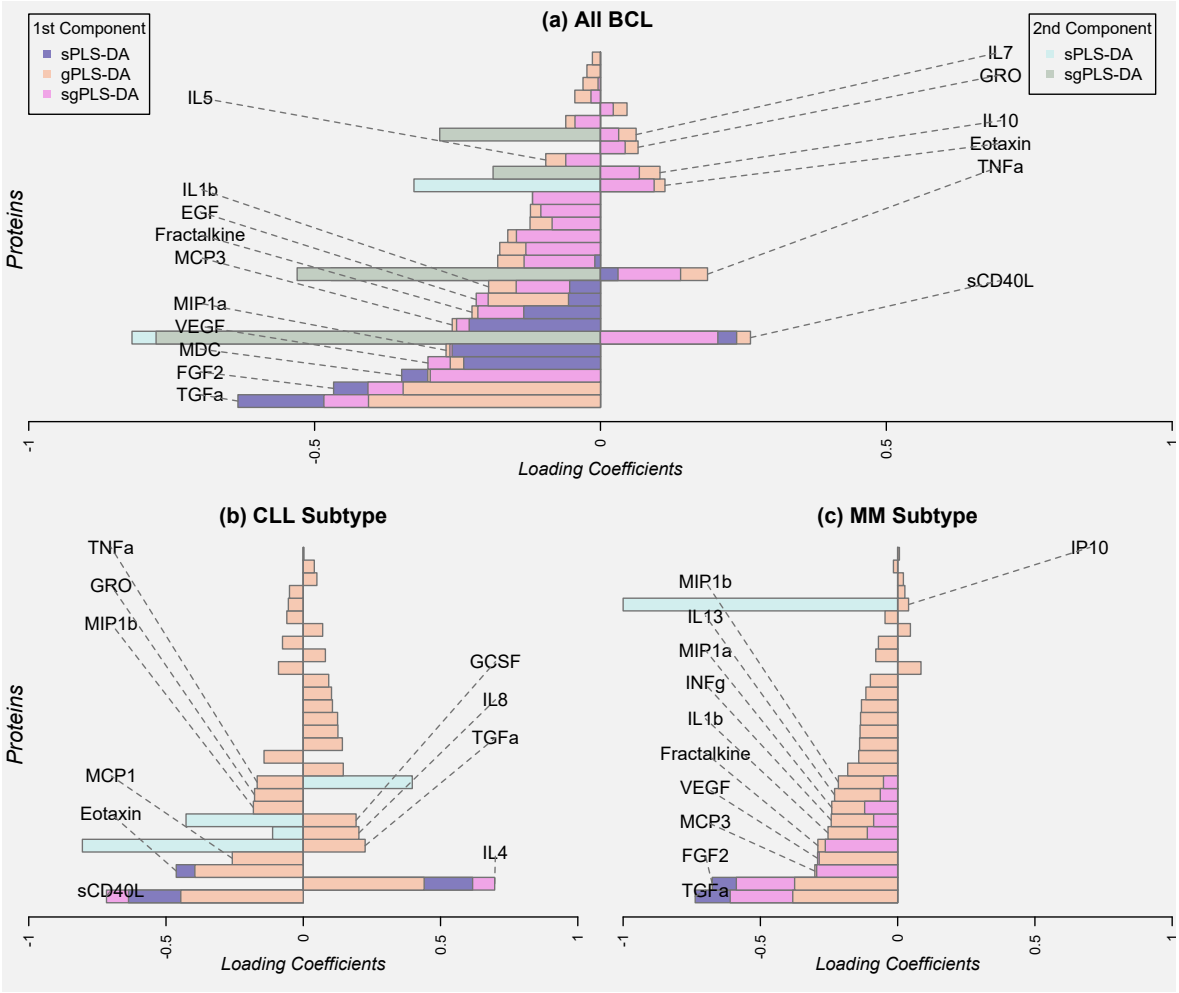
for cases in comparison to controls in the pooled population and the CLL subtype across the three regularized models, with more marginal differences in the sgPLS-DA model. For the MM subtype, there seems to exist a similar classification ability for the two sample types. Lastly, when comparing the classification abilities across statistical methods, it is observed that sgPLS-DA shows the best model performance of the three approaches followed by sPLS-DA and gPLS-DA in the third place, a statement that holds for the three study populations.

5.3.1.3 Assessment of Visualization Tools

5.3.1.3.1 Loading Coefficients Plots

Figure 5.1 exhibits the individual contribution of the selected variables in terms of the loading values for each of the regularized optimized PLS-DA models for the BCL pooled study population and the CLL and MM subtype-specific analyses. The three models consistently agree in identifying the most significant variables as (at least) the first two variables with the highest absolute loading values are the same across statistical methods. There is also an agreement in relation to the sign associated to the corresponding coefficients. These observations seem to hold for the three observation-specific analyses. In addition, it is noticed that i) opposite signs are assigned to coefficients of a given variable when an overlapping is observed between components and

Figure 5.1: Loading coefficients of the selected variables for the three regularized PLS-DA and for the pooled BCL population and the CLL and MM disease subtypes (proteomics dataset).



Loading coefficients of the variables retained in the two components are shown simultaneously.
PLS-DA: Partial Least Squares-Discriminant Analysis.

ii) the more proteins are attained in the calibrated models, the lower their individual contribution are.

The analysis including all BCL case-control pairs shows that the growth factors TGFa, FGF2 and VEGF, the chemokines MDC and MIP1a and the cytokines sCD40L and TNFa are the most important variables for the correct separation of samples. For the subtype-specific analyses, these variables correspond to the cytokines sCD40L and IL4 and the chemokine Eotaxin for CLL and the growth factors TGFa, FGF2 and

VEGF and the chemokines Fractalkine and MCP3 for the MM subtype.

5.3.1.3.2 *Stability Frequency Plots*

The results of the stability frequency analyses from the sPLS-DA and sgPLS-DA models are shown in Figure 5.2 and Figure 5.3, respectively. As discussed in the method section, different values of the model parameters were tested; however, to ease interpretation the specific frequencies reported in here correspond to those retrieved by the optimal values of the model parameters. Overall, the stability results validate those obtained in the assessment of the loading coefficients. In the sPLS-DA pooled BCL analysis, the proteins FGF2, TGFa, VEGF, MCP3, MDC, MIP1a, Fractalkine and sCD40L present the highest stability frequency for a sparse model retaining one component and 12 variables, which correspond to 1.0, 1.0, 0.99, 0.99, 0.99, 0.98, 0.95 and 0.90, respectively. The sparse BCL model retaining two components and three variables displays high frequencies for two proteins: sCD40L and TNFa (0.99 and 0.90, respectively). On the other hand, the subtype-specific stability analyses show an elevated frequency for the proteins sCD40L and IL4 (0.98 and 0.76, respectively) in a model of one component and TGFa (0.79) in a model of two components for the CLL observations, and FGF2 and TGFa (0.96 and 0.63, respectively) in the one-component model and GRO, TNFa and IP10 (0.24, 0.23 and 0.20, respectively) in the two-component model for the MM observations.

The stability frequency analysis of sgPLS-DA facilitates the identification of the most relevant functional groups as well as the most significant proteins within the groups. The analysis on the BCL pooled population reveals the growth factor module as the most important, followed by the chemokine and the cytokine categories. The most stable variables within groups largely agree with those reported in the sPLS-DA analysis: the growth factors FGF2, TGFa and VEGF, the chemokines MCP3, MDC, MIP1a, Fractalkine and the cytokines sCD40L and INFg present the highest stability frequencies. The analyses on the CLL and MM disease subtypes reveal that the most important groups are cytokine and growth factor, respectively; sCD40L and IL4 stand

Figure 5.2: Stability frequency plots from the sPLS-DA models for the pooled BCL population and the CLL and MM disease subtypes (proteomics dataset).

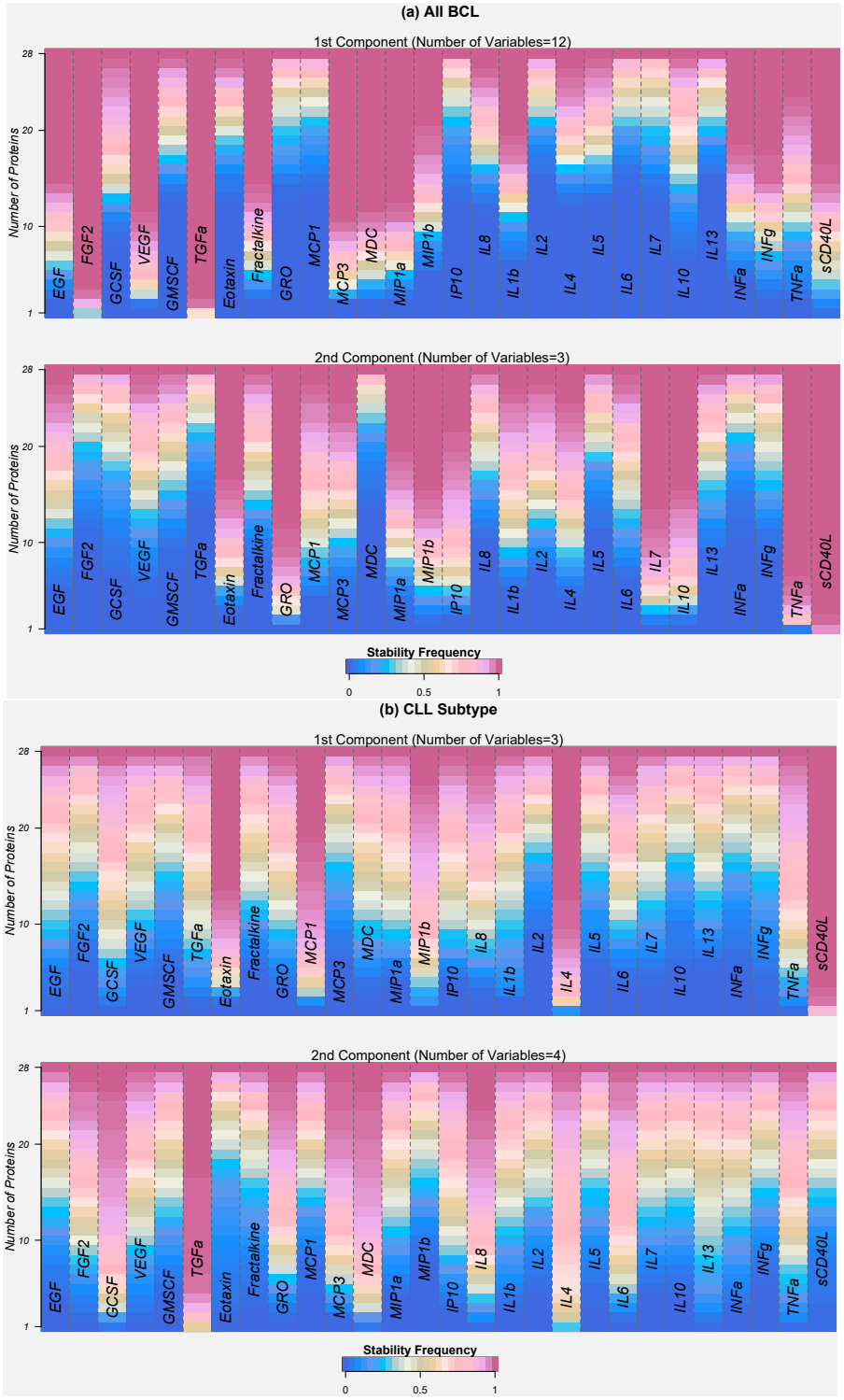
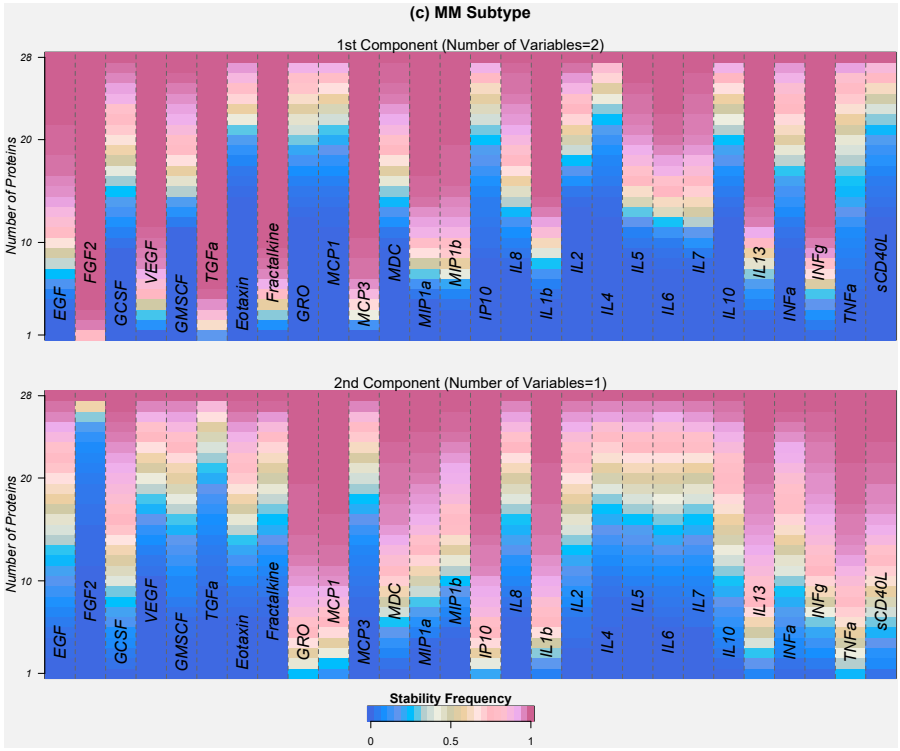


Figure 5.2: Stability frequency plots from the sPLS-DA models for the pooled BCL population and the CLL and MM disease subtypes (proteomics dataset) *(cont.)*.



The stability frequency of sPLS-DA models retaining 1 to 28 variables was assessed. The analysis of the second component was conducted retaining in the first dimension the number of features previously reported as optimal, which is specified in the title of the top panel plot.
sPLS-DA: sparse Partial Least Squares-Discriminant Analysis.

out as the most relevant cytokines and FGF2, TGFa and VEGF as the most relevant growth factors.

5.3.1.3.3 Sample Representation Plot

Figure 5.4 exhibits the sample representation plots for the four models for which two components were considered as optimal, which include the pooled BCL analyses (both sPLS-DA and sgPLS-DA) and the CLL- and MM- specific analyses (only sPLS-DA). The best separation of observation types in the two-dimensional space spanned by the components is appreciated for CLL as case samples cluster in a more constrained area than their matched controls. This finding agrees with those described in the model performance section as the ER in cases is substantially different to the

Figure 5.3: Stability frequency plots from the sgPLS-DA models for the pooled BCL population and the CLL and MM disease subtypes (proteomics dataset).

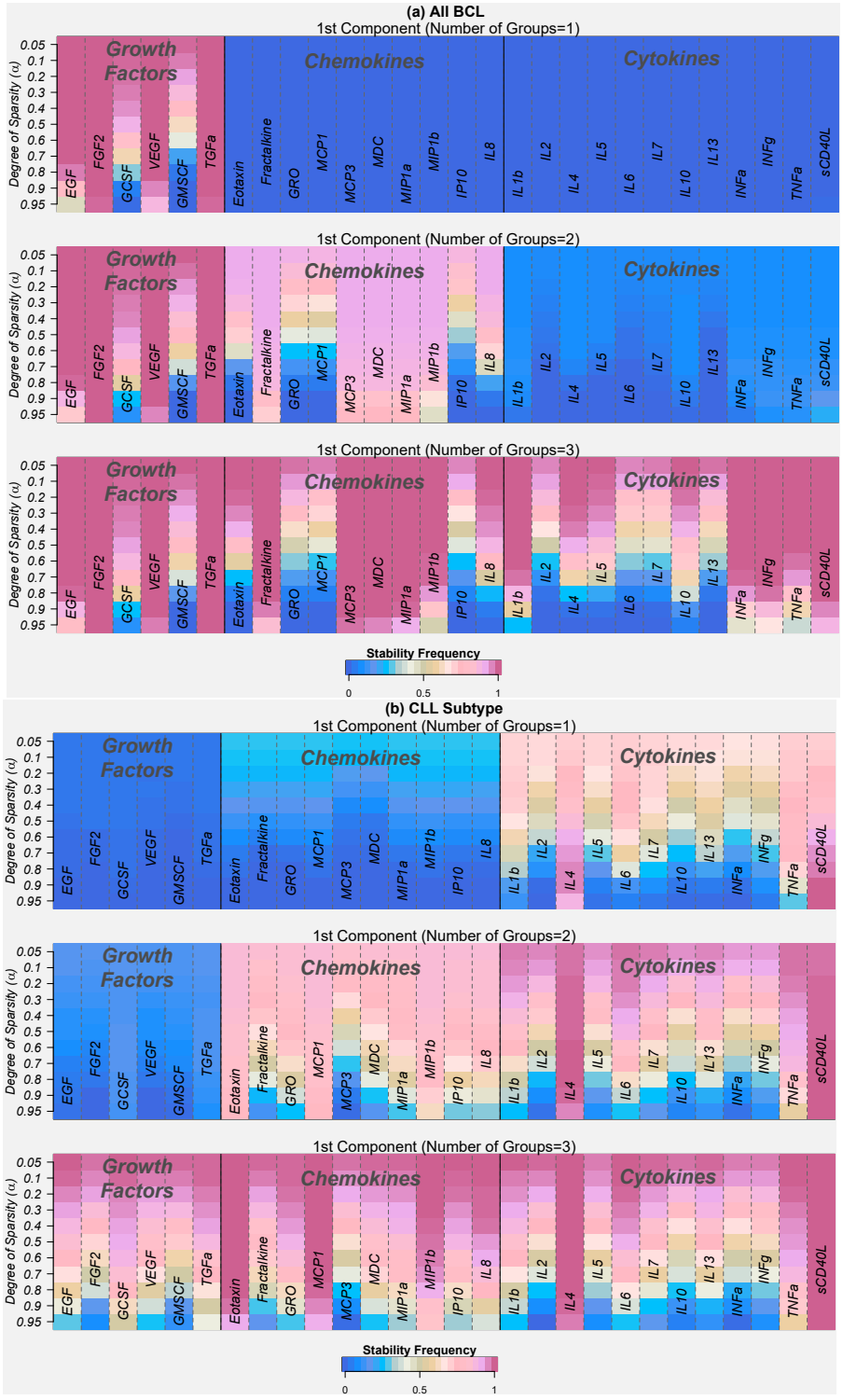
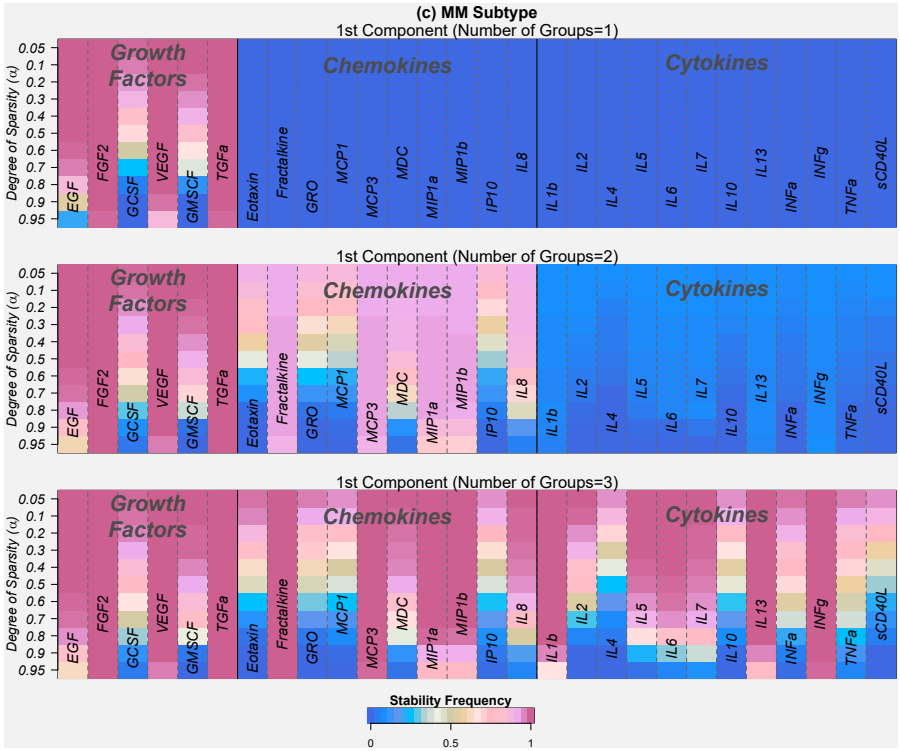


Figure 5.3: Stability frequency plots from the sgPLS-DA models for the pooled BCL population and the CLL and MM disease subtypes (proteomics dataset).*(cont.)*



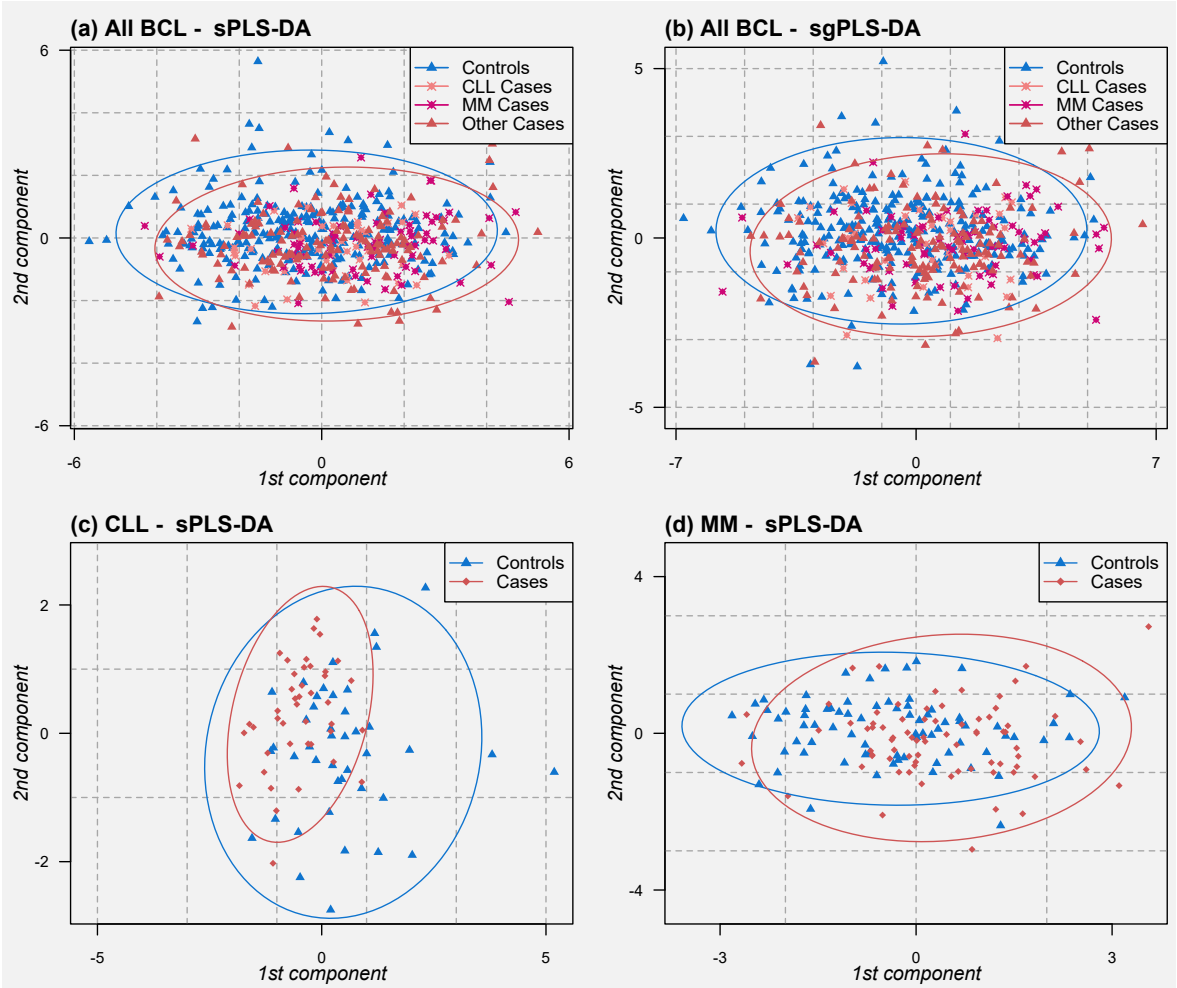
Stability frequency analysis of the sgPLS-DA models simultaneously assessing the two tuning parameters controlling model sparsity: a grid of 11 different values of the mixing parameter α_1 was tested for models retaining 1 to 3 functional groups. For each study population, only results from the first component are displayed.
sgPLS-DA: sparse group Partial Least Squares-Discriminant Analysis.

ER in controls (0.31 vs. 0.43). A less clear separation of samples is observed for MM observation as controls are mainly spanned in the first component while cases cover a greater area in the second dimension. Finally, distinction of cases and controls is more challenging when all BCL cases are pooled together and a marginal distinction is observed between CLL and MM sample types.

5.3.1.4 Sensitivity Analysis

Considering that the results described so far expose positive findings for the disease subtypes CLL and MM, a sensitive analysis was conducted whereby these particular sub-entities (and their matched controls) were independently excluded from the pooled analysis. The classification performance of the optimized models is shown in

Figure 5.4: Sample representation plots displaying the location of the observations on the spaces spanned by the first and second **X** components of the regularized PLS-DA models (proteomics dataset).



Sample representation for the four models for which two components were included in the optimized models. For the pooled BCL study population, a distinction is made between case types. For each sample group, ellipses of the confidence region were drawn employing the R package `ellipse` based on the variance and mean of the matrix of **X** components (the mean defines the location of the ellipse centre). The confidence level controlling the ellipse size was 0.95 and a total of 100 points were used. PLS-DA: Partial Least Squares-Discriminant Analysis.

Table C.4. Model performance decreases as a result of removing CLL observations for all statistical approaches and metric of performance tested, whereas marginal and inconsistent changes are observed when MM samples are excluded. In addition, and following the same methodology described above, all analyses were replicated on the proteomics dataset after correction for WBC differentials (see section 3.4). The classification abilities of the models are summarised in Table C.5. As observed in the

WBC unadjusted analyses, satisfactory classification abilities are observed for CLL and MM and deficient performances for DLBCL and FL; in addition, sgPLS-DA approach tend to outperform both gPLS-DA and sPLS-DA. Lower ER and higher AUC are described for the four subtypes stratified analyses, which is not the case for the BCL pooled population. (For these two sensitivity analyses, the results from the calibration procedure for the regularized models and the details on model parameters of the calibrated models are not shown).

5.3.2 Transcriptomics

5.3.2.1 Calibration Procedure and Assessment of Calibrated Models

The outputs from the model calibration process are presented in Table C.6 and Figure C.3 to Figure C.5, which correspond to the R^2 and DQ^2 statistics and the set of calibration curves, respectively. Table 5.3 shows the values of the model parameters employed in the optimized models, the number of gene expression signals and biological pathways retained per component and the overall number of unique features and modules retained in the complete optimized model for the three PLS-DA approaches and for the analyses including all BCL case-control pairs and stratified by the four main histological subtypes.

5.3.2.1.1 Pooled Population

The sPLS-DA model selected one component and seven transcripts which are part of five different biological pathways, with the group of un-annotated probes being the most frequent (three signals). In contrast, the gPLS-DA model selected a model with two dimensions, 35 pathways were retained in the first and 25 in the second accounting for a total of 96 unique signals and 57 pathways; there is an overlap of seven transcripts and three pathways between components. The majority of those pathways (n = 30, 52.63%) are constituted by only one probe while the remaining groups are comprised of two up to four signals. Lastly, two components were also chosen in the sgPLS-DA model, with nine and 20 pathways retained in each. As a

Table 5.3: Parameters used in the calibrated model, number of unique gene expression signals and biological pathways selected per component and total number of unique transcripts and pathways selected across components for the three regularized approaches and for the five study populations.

	All BCL	CLL	DLBCL	FL	MM
sparse PLS-DA					
N Components	1	1	2	1	2
N Variables 1st component	7	6	90	20	7
N Groups 1st component	5	4	37	8	2
N Variables 2nd component	—	—	25	—	7
N Groups 2nd component	—	—	8	—	3
Total N Variables	7	6	115	20	14
Total N Groups	5	4	42	8	4
group PLS-DA					
N Components	2	1	1	1	1
N Variables 1st component	64	5	22	67	128
N Groups 1st component	35	4	15	35	50
N Variables 2nd component	29	—	—	—	—
N Groups 2nd component	25	—	—	—	—
Total N Variables	96	5	22	67	128
Total N Groups	57	4	15	35	50
sparse group PLS-DA					
N Components	2	2	1	2	2
N Variables 1st component	15	36	15	56	55
N Groups 1st component	9	25	9	30	35
α_1 1st component	0.95	0.95	0.1	0.1	0.95
N Variables 2nd component	34	71	—	2	5
N Groups 2nd component	20	35	—	2	5
α_1 2nd component	0.95	0.05	—	0.95	0.9
Total N Variables	48	107	15	58	60
Total N Groups	28	60	9	32	40

PLS-DA: Partial Least Squares-Discriminant Analysis.

result of imposing within-group sparsity, 28 pathways and 48 individual signals are part of the calibrated model, a one-probe module overlaps between dimensions. Four of the selected modules are one-probe pathways and the rest of the chosen pathways are constituted by two up to 57 individual signals; the relative frequency of selected signals per pathway for these 24 modules ranges from 5.56% to 66.67%. A descriptive summary of the selected gene expression signals in terms of the most frequent biological pathways to which they belong for the three regularized approaches is presented in Table C.7.

5.3.2.1.2 Subtype Stratified Analysis

The sPLS-DA models performed on the study populations stratified by major histological subtypes selected one component for CLL and FL and two dimensions for DLBCL and MM. For the two subtypes retaining one component, six and 20 signals were selected which comprise four and eight pathways, respectively. For DLBCL, 90 and 25 variables were selected in the first and second dimensions, respectively, accounting for a total of 115 unique signals and 42 unique pathways (three modules overlap between components). For MM, seven signals were chosen in each component accounting for a total 14 and four unique variables and modules, respectively.

The four stratified gPLS-DA models selected only one dimension, with CLL yielding the most parsimonious model and MM the most abundant one. A total of four pathways were selected in the former sub-entity which account for five different signals, while the latter subtype selected 50 pathways which translate to 128 different transcripts. The models for DLBCL and FL retained 15 and 35 pathways accounting for 22 and 67 transcripts, respectively.

Finally, and with the exception of DLBCL, two dimensions were selected in the stratified sgPLS-DA models. For CLL, a total of 107 and 60 unique signals and pathways were retained in the optimized model and no overlap is seen between components; 27 of the modules are constituted by only one probe and the remaining 33 pathways are comprised by two up to 57 signals with a within-group relative frequency of se-

lection ranging from 5.56% to 100%. Sparser models are observed for FL and MM sub-entities as 58 and 32 and 60 and 40 gene expression signals and biological pathways are included in the calibrated models, respectively; similar to CLL, there is no overlap of features between dimensions. Nine pathways were selected in the sole dimension for DLBCL which translates to 15 different transcripts.

5.3.2.2 Model Performance

Table 5.4 shows the classification performance of the three optimized models for the analyses including all BCL cases and for the stratified study populations. The results on CLL observation substantially outperforms those conducted on the other populations. In contrast, the abilities of the models to correctly separate DLBCL samples is poor, with an overall misclassification ER and ER per observation type equal or above 0.5 in all statistical approaches. On the other hand, similar metric values are observed for the all BCL, FL and MM populations with overall ER of about 0.4 and AUC below 0.6. In addition, it is noticed a difference in ER per observation type as control samples are consistently classified better than their counterparts across the five study populations and PLS-DA models (exceptions to this are FL subtype gPLS-DA model and DLBCL subtype all PLS-DA approaches); a more accentuated difference in ER per sample type is observed in the CLL and pooled analyses. Finally, and with the exception of FL, a comparison between regularized statistical methods reveals that sPLS-DA yields the best model performance of the three approaches followed by sgPLS-DA and gPLS-DA. Given that CLL steadily shows the best classification performance, subsequent analyses are focused on this particular observation type (n = 68 case-control pairs) and outputs for the rest of the study populations are described only when relevant findings are observed.

Table 5.4: Classification performances of the three calibrated PLS-DA models for the five study populations (transcriptomics dataset).

	sparse PLS-DA				*group PLS-DA*				*sparse group PLS-DA*			
	Overall ER	**ER Control**	**ER Cases**	**AUC**	**Overall ER**	**ER Control**	**ER Cases**	**AUC**	**Overall ER**	**ER Control**	**ER Cases**	**AUC**
All BCL	0.400	0.259	0.541	0.601	0.463	0.403	0.524	0.541	0.430	0.346	0.513	0.572
CLL	0.107	0.001	0.215	0.896	0.214	0.131	0.296	0.787	0.159	0.042	0.276	0.847
DLBCL	0.530	0.568	0.492	0.561	0.556	0.567	0.545	0.563	0.546	0.564	0.529	0.566
FL	0.456	0.448	0.465	0.592	0.453	0.467	0.439	0.586	0.401	0.384	0.418	0.634
MM	0.412	0.388	0.436	0.600	0.524	0.462	0.585	0.545	0.466	0.453	0.480	0.564

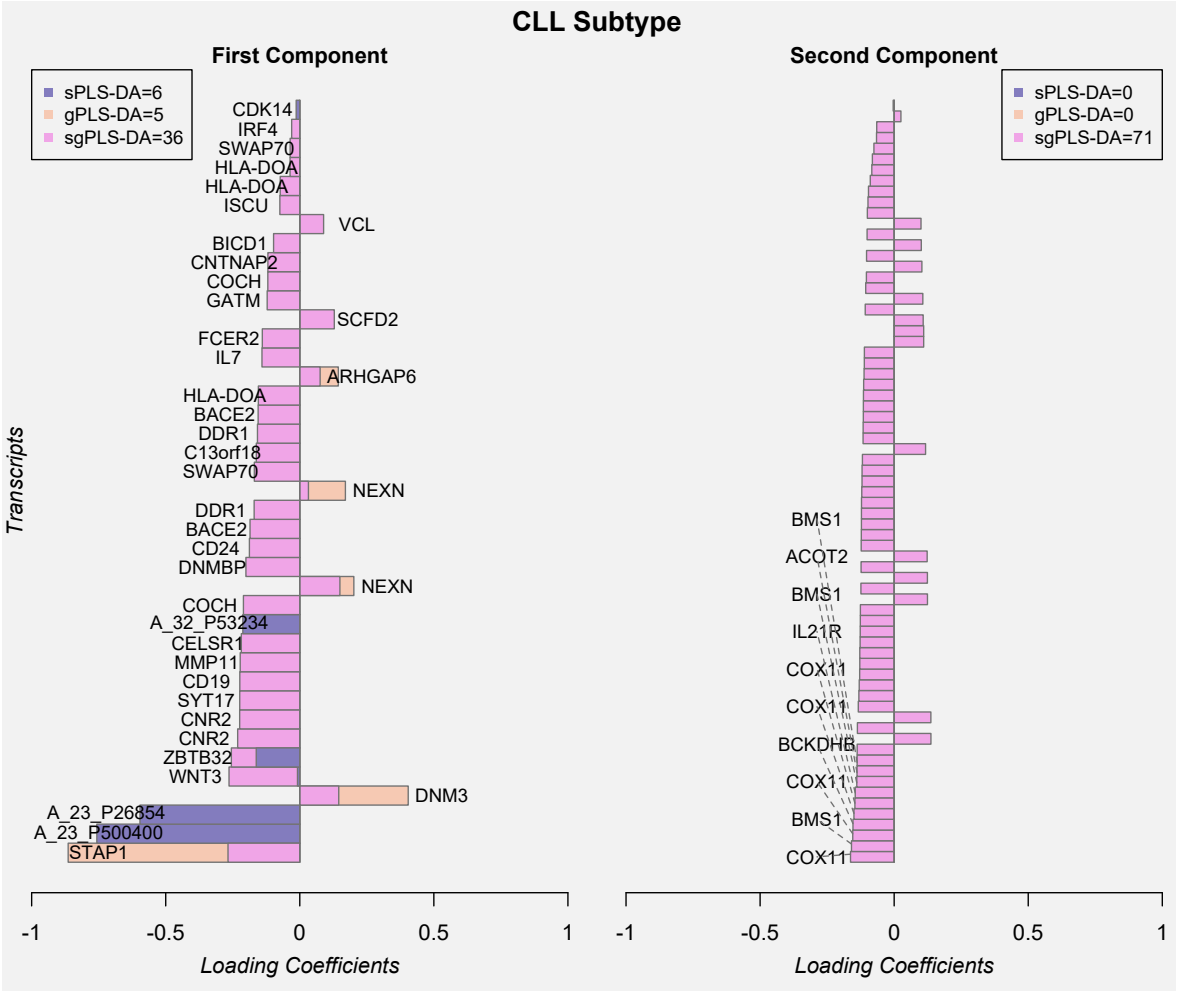
PLS-DA: Partial Least Squares-Discriminant Analysis, ER: Error Rate, AUC: Area Under the Curve.

5.3.2.3 Assessment of Visualisation Tools

5.3.2.3.1 Loading Coefficients Plots

Figure 5.5 displays the loading coefficients associated to the selected variables in the three statistical models for the CLL study population. Most of the loading coefficients are of negative value: only six of the 40 different signals selected in the first component and 14 of the 71 signals selected in the second component have a positive coefficient. Despite the difference in number of selected features across statistical approaches, a greater congruence is seen between the gPLS-DA and sgPLS-DA models since i) a complete overlap of selected features is observed (the five transcripts selected in gPLS-DA) and ii) similar loading values were assigned to the retained transcripts. In contrast, for the sPLS-DA model, two of the six variables (WNT3 and ZBTB32) overlap with the sgPLS-DA model only and the respective loading coefficient values differ considerably. By visual inspection of the loading plot it can noticed that the highest contributing variables in terms of absolute value of the coefficients are STAP1, A_23_P500400 (ABCA6), A_23_P26854 (ARHGAP44), DNM3, WNT3, ZBTB32. On the other hand, none of the variables selected in the second component stand out as a substantially relevant as all the loading coefficients were a assigned a similar value.

Figure 5.5: Loading coefficients of the selected variables for the three regularized PLS-DA for the CLL study population (transcriptomics dataset).



The loading coefficients of the variables retained in the three PLS-DA models are shown simultaneously. The left panel displays the coefficients of the variables retained in the first component and the right panel of the variables retained in the second component. The sgPLS-DA model was the only to select two dimensions.
PLS-DA: Partial Least Squares-Discriminant Analysis, sPLS-DA: sparse Partial Least Squares-Discriminant Analysis, gPLS-DA: group Partial Least Squares-Discriminant Analysis, sgPLS-DA: sparse group Partial Least Squares-Discriminant Analysis.

5.3.2.3.2 Stability Frequency Plots

Figure 5.6 shows the results of the stability frequency analyses from the three regularized models for the CLL study population. The most conspicuous finding is that the sPLS-DA model yields the highest stability frequencies for all the variables selected in the model; in particular, there are two signals (A_23_P26854 [ARHGAP44] and A_23_P500400 [ABCA6]) which present a stability of one for all different values of the

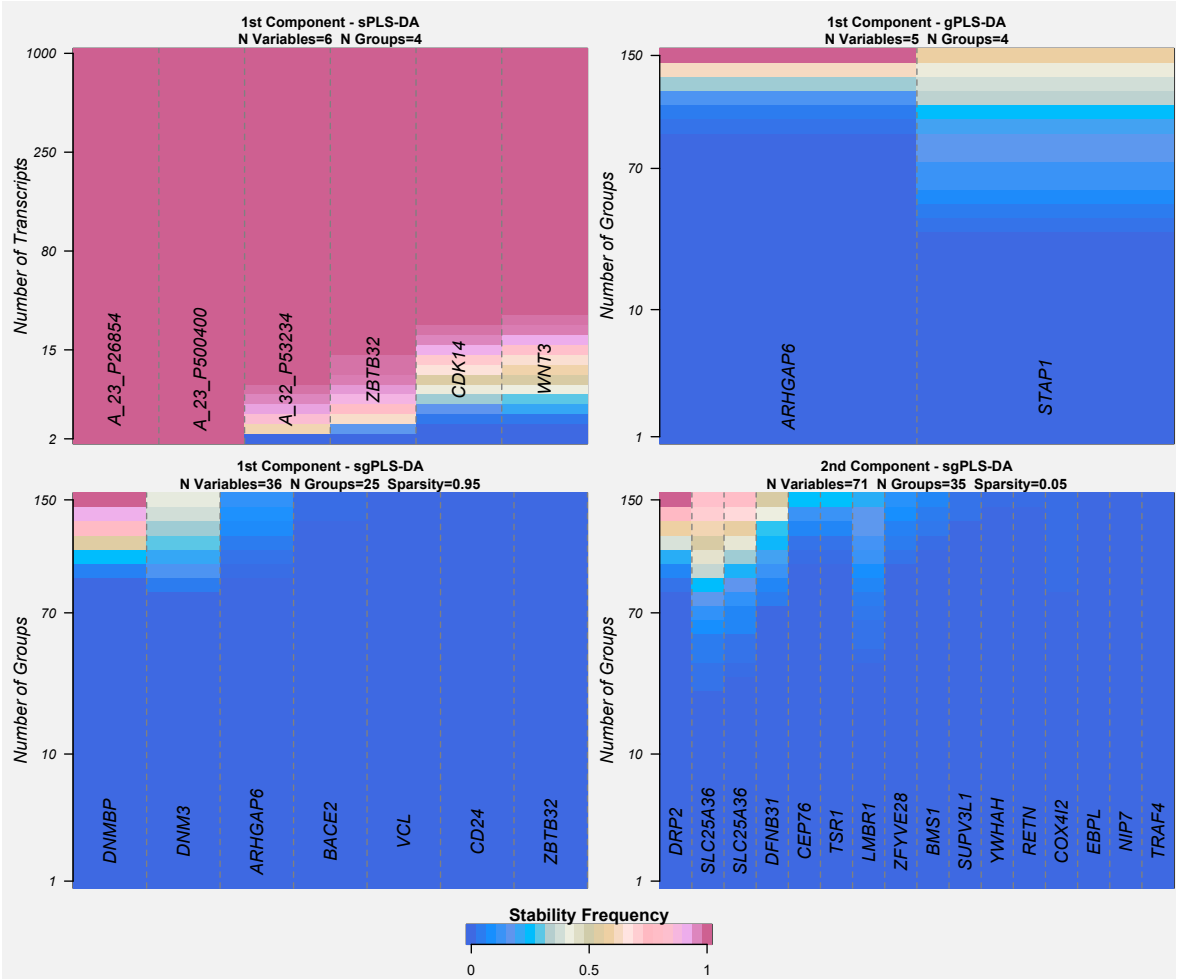
model parameters visited. The six transcripts reach the highest frequency when the sparse model is set to retain 60 variables. In contrast, the output from both gPLS-DA and sgPLS-DA provide low stability frequencies, with some of the selected variables presenting a frequency of zero. For gPLS-DA, only two of the five retained variables present a frequency higher than zero, which correspond to ARHGAP6 and STAP1; the highest achieved frequencies are 0.16 and 0.09, respectively. For sgPLS-DA, seven of the 36 and 16 of the 71 selected transcripts present frequencies higher than 0. In the first component, the highest observed frequencies are for the transcripts DNMBP, DNM3 and ARHGAP6 which correspond to 0.85, 0.76, 0.66, respectively; while in the second component the most stable transcripts are DRP2, SLC25A36, SLC25A36 with frequencies of 0.59, 0.45, 0.34, respectively.

Upon inspection of the stability outputs from the rest of the study populations, similar findings are obtained as relatively high frequencies are observed for the selected variables of sPLS-DA, which is not the case for the gPLS-DA and sgPLS-DA models. The MM stratified analyses yield the highest frequencies in comparison with the other two subtypes, while the pooled BCL analyses largely resemble those described for CLL as transcripts such as A_23_P500400 (ABCA6), A_23_P26854 (ARHGAP44), DNM3 are among the most stable ones (results not shown).

5.3.2.3.3 Sample Representation Plot

The sample representation plots for the regularized PLS-DA models for which two components were selected are displayed in Figure 5.7. Two dimensions were chosen for at least one statistical method in the four subtypes and the pooled population analyses; therefore, the five observation types are represented in the two-dimensional space spanned by the latent variables. Apart from FL and MM (sgPLS-DA), a noticeable finding is that control observations cluster closer together than their matched cases across all representations. This discovery is more conspicuous in the CLL subtype analyses as control samples mainly span a reduced area alongside the second component. As such, a plausible explanation is offered as to why the ER for controls

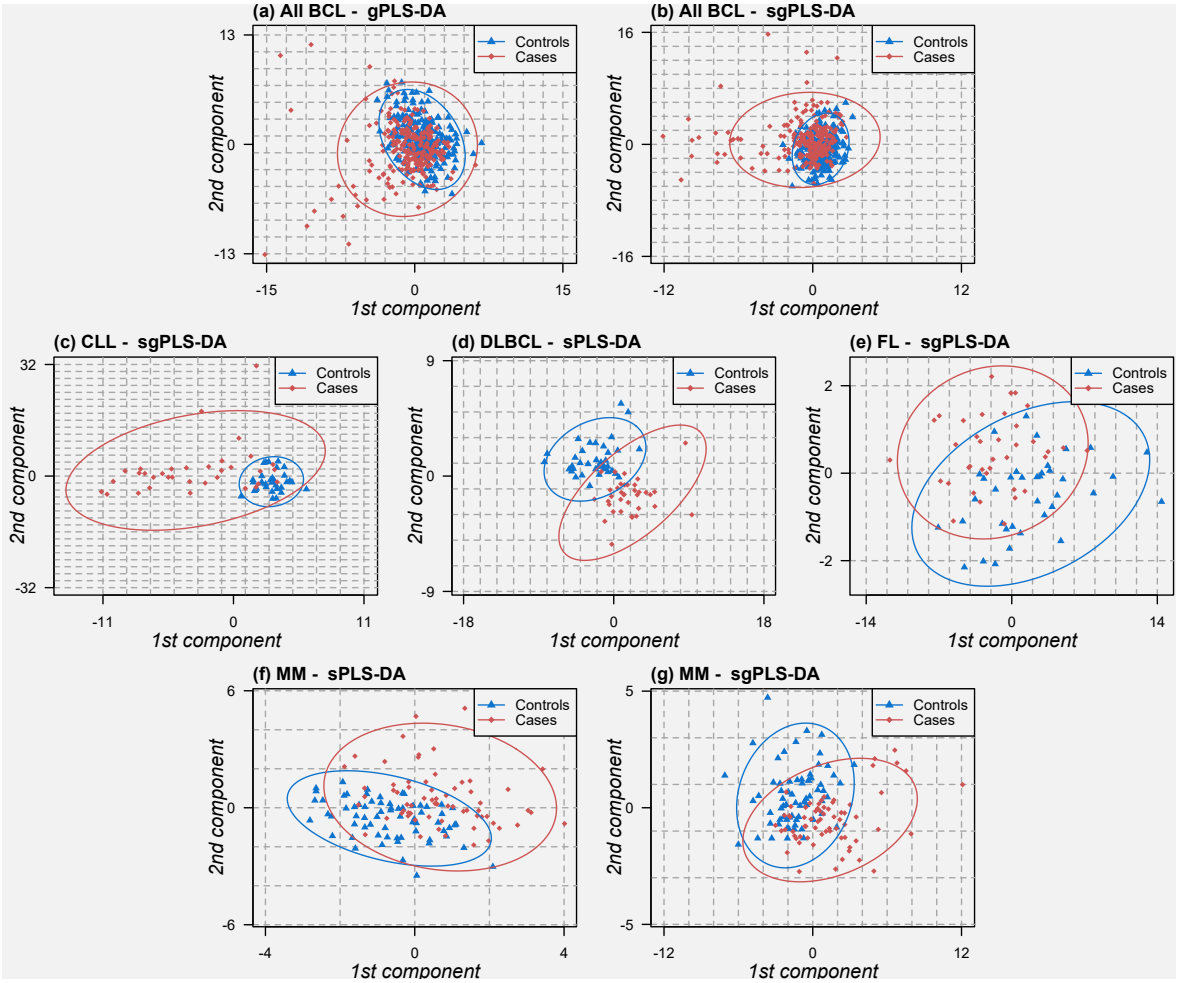
Figure 5.6: Stability frequency plots from the three regularized PLS-DA models for the CLL study population (transcriptomics dataset).



The sPLS-DA stability frequency analysis was conducted with models retaining 2 to 1000 transcripts (grid resolution of 40 values) while the gPLS-DA and sgPLS-DA analyses were conducted with models retaining 1 to 150 biological pathways (grid resolution of 28 values). In the latter case, the plots exhibit results for a fixed value of the mixing parameter α_1 which correspond to the value employed in the optimized sgPLS-DA model (specified in the title). For the gPLS-DA and sgPLS-DA models, only the transcripts showing a stability frequency greater than zero are displayed.
PLS-DA: Partial Least Squares-Discriminant Analysis, sPLS-DA: sparse Partial Least Squares-Discriminant Analysis, gPLS-DA: group Partial Least Squares-Discriminant Analysis, sgPLS-DA: sparse group Partial Least Squares-Discriminant Analysis.

is significantly lower than for case samples (as shown in Table 5.4). The representation of the pooled population seem to be partially driven by what is observed in the CLL samples. Lastly, given the results described in the model performance section, a greater overlap of samples is observed for the rest of the disease subtype study populations, which makes a visual distinction between cases and controls more challenging.

Figure 5.7: Sample representation plots displaying the location of the observations on the spaces spanned by the first and second **X** components of the regularized PLS-DA models (transcriptomics dataset).



Sample representation for the seven models for which two components were included in the optimized models. For each sample group, ellipses of the confidence region were drawn employing the R package `ellipse` based on the variance and mean of the matrix of **X** components (the mean defines the location of the ellipse centre). The confidence level controlling the ellipse size was 0.95 and a total of 100 points were used.
PLS-DA: Partial Least Squares-Discriminant Analysis, sPLS-DA: sparse Partial Least Squares-Discriminant Analysis, gPLS-DA: group Partial Least Squares-Discriminant Analysis, sgPLS-DA: sparse group Partial Least Squares-Discriminant Analysis.

5.3.2.4 Biological Interpretation of the Findings

Since individual regularized PLS-DA models yielded sparse results, functional annotation with DAVID on the CLL-specific analysis was conducted pooling together the selected gene expression signals from the three regularized models. The 111 unique transcripts were mapped to 78 (86.58%) different DAVID IDs which were grouped

into 72 different gene enriched biological pathways. This number was reduced to three after more stringent filtering (EASE score<0.01, setting the minimal number of probes per functional group to 5 and fold enrichment value>3); these enriched pathways relate to mitochondrion and transit peptide.

5.3.2.5 Sensitivity Analysis

As in the proteomics analyses, two types of sensitivity analyses were performed: i) the subtype for which the best classification performance was observed was excluded from the pooled analysis including all BCL case-control pairs and ii) the statistical analyses were conducted on the transcriptomics data after expression levels were adjusted for WBC sub-populations. Table C.8 shows the statistical performance of the calibrated models performed on the pooled study population removing CLL case-control pairs and a significant decline in all assessed metric of performance is observed (except ER in cases). On the other hand, correction for WBC differentials does not improve statistical performances (Table C.9). More specifically, for the CLL subentity the three regularized models produce higher ER and lower AUC estimates, which is particularly true for the gPLS-DA and sgPLS-DA models as the overall ER is close to 0.5 and the ability to correctly classify cases and controls is approximately the same. For DLBCL, the WBC corrected analyses confirm what was observed in the unadjusted models as the statistical methods are unable to correctly separate sample types, while for the remaining of the study populations a marginal statistical worsening is observed (with the exception FL sPLS-DA model). Finally, and similar to the WCB unadjusted models, a comparison between regularized statistical methods shows that sPLS-DA tends to outperform the other two approaches. (As in the proteomics sensitivity analyses, the results from the calibration procedure for the regularized models and the details on model parameters of the calibrated models are omitted).

5.3.3 Comparative Assessment of Univariate and Multivariate Statistical Approaches

The univariate analyses on the proteomics and transcriptomics datasets were conducted on chapter 4. As a brief summary, the most relevant results were observed for MM in proteomics and CLL in transcriptomics as six proteins and 684 transcripts show statistically significant associations with the respective disease outcome of interest. Thus, the most evident difference between the results obtained from the univariate and multivariate approaches is in relation to the CLL subtype and inflammatory markers: while the LMM analyses failed to identify proteins predictive of CLL, the regularized PLS-DA models were able to find a subset of proteins that separate CLL cases from their paired controls with a satisfactory classification performance; in fact multivariate approaches show a better statistical performance for CLL than for MM observations. Another noticeable difference between univariate and multivariate methods refers to the parsimony of the gene expression findings for CLL: the (s)(g)(sg)PLS-DA models selected 6,5 and 107 transcripts (respectively) as opposed to the 684 significant associations described in the LMM.

On the other hand, overlapping findings are observed for the two main disease subtypes across univariate and multivariate approaches. Figure 5.8 corresponds to the Venn diagrams illustrating the proteins and functional groups shared across approaches for MM and the transcripts and biological pathways shared across approaches for CLL. For MM and proteomics, the growth factors FGF2 and TGFa are common to all four statistical approaches, while four markers are shared across all models but sPLS-DA which are the growth factor VEGF and the chemokines Fractalkine, MCP3 and MIP1a. Consequently, the growth factor and the chemokine groups are common to the four statistical methods while the cytokine group is additionally selected by the gPLS-DA and sgPLS-DA models. Furthermore, for CLL and transcriptomics, as it was already discussed in previous sections, an intersection of individual signals is observed between sgPLS-DA and gPLS-DA with five common features and sgPLS-

Table 5.5: Biological pathways that are common across the univariate approach and the three regularized models for the CLL study population.

gPLS-DA and sgPLS-DA (n=4)			
Biological pathway	**Total**	**Selected sgPLS-DA (%)**	
Regulation of organelle organization	2	2 (100)	
Membrane fusion	1	1 (100)	
Negative regulation of cell-matrix adhesion	1	1 (100)	
Synaptogenesis	1	1 (100)	
sPLS-DA and sgPLS-DA (n=2)			
Biological pathway	**Total**	**Selected sPLS-DA (%)**	**Selected sgPLSDA (%)**
Cell activation	57	1 (1.754)	4 (7.018)
Mesoderm formation	18	1 (5.556)	1 (5.556)
sPLS-DA and LMM (n=4)			
Biological pathway	**Total**	**Selected sPLS-DA (%)**	**Selected LMM (%)**
Non-annotated probes	14737	3 (0.020)	307 (2.083)
Phosphorus metabolic process	426	1 (0.235)	19 (4.460)
Cell activation	57	1 (1.754)	10 (17.544)
Mesoderm formation	18	1 (5.556)	1 (5.556)

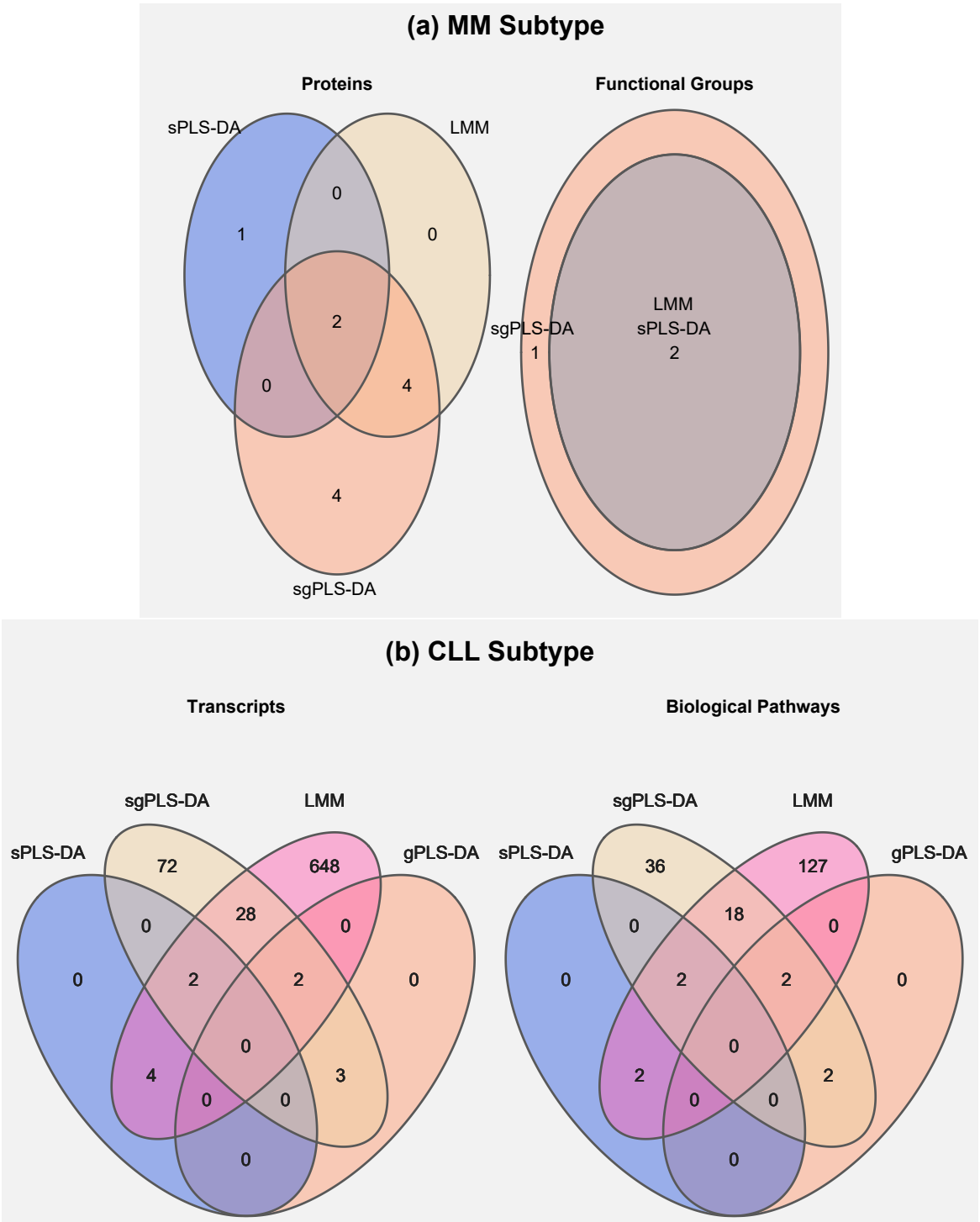
DA and sPLS-DA with only two; similar numbers are observed in terms of biological pathways with four and two common modules, respectively. When incorporating the results from the univariate approaches, a greater overlap is seen between the results from the LMM and sgPLS-DA with 30 individual transcripts and 20 biological pathways shared between the two methods. Marginal intersections are noticed between the results from the LMM and sPLS-DA and gPLS-DA: six and two transcripts and four and two pathways, respectively. A descriptive summary of the gene expression signals in terms of the selected biological pathways that are common across the four CLL statistical approaches is presented in Table 5.5.

Table 5.5: Biological pathways that are common across the univariate approach and the three regularized PLS-DA models for the CLL study population *(cont.)*.

gPLS-DA and LMM (n=2)			
Biological pathway	**Total**	**Selected LMM (%)**	
Membrane fusion	1	1 (100)	
Synaptogenesis	1	1 (100)	
sgPLS-DA and LMM (n=22)			
Biological pathway	**Total**	**Selected sgPLSDA (%)**	**Selected LMM (%)**
Cell activation	57	4 (7.018)	10 (17.544)
Inflammatory response	31	2 (6.452)	2 (6.452)
Mesoderm formation	18	1 (5.556)	1 (5.556)
Regulation of phosphate metabolic process	11	1 (9.091)	2 (18.182)
Sensory perception of sound	10	2 (20)	2 (20)
Membrane protein ectodomain proteolysis	9	2 (22.222)	3 (33.333)
Immune response-activating cell surface receptor signaling pathway	8	1 (12.5)	1 (12.5)
Antigen processing and presentation of peptide antigen via MHC class II	7	3 (42.857)	4 (57.143)
Somatic diversification of immune receptors	7	2 (28.571)	2 (28.571)
Regulation of immune effector process	6	1 (16.667)	1 (16.667)
Cell recognition	5	1 (20)	1 (20)
Morphogenesis of a polarized epithelium	5	1 (20)	1 (20)
Vesicle docking during exocytosis	5	1 (20)	1 (20)
Collagen catabolic process	4	1 (25)	1 (25)
Phosphagen metabolic process	4	1 (25)	1 (25)
Regulation of cell-matrix adhesion	4	2 (50)	2 (50)
Regulation of Rho protein signal transduction	4	1 (25)	1 (25)
RNA localization	3	1 (33.333)	1 (33.333)
Iron-sulfur cluster assembly	2	1 (50)	1 (50)
Plasma membrane organization	2	1 (50)	1 (50)
Membrane fusion	1	1 (100)	1 (100)
Synaptogenesis	1	1 (100)	1 (100)

For each biological pathway, the total number of probes per pathway and the absolute and relative frequencies of probes selected per statistical approach are specified (apart from gPLS-DA where all the signals in the selected group are retained). Biological pathways that are shared across three different approaches are written in **bold** (common to sPLS-DA, sgPLS-DA and LMM) or *cursive* (common to gPLS-DA, sgPLS-DA and LMM).
sPLS: sparse Partial Least Square-Discriminant Analysis, gPLS: group Partial Least Square-Discriminant Analysis, sgPLS: sparse group Partial Least Square-Discriminant Analysis, LMM: Linear Mixed Model.

Figure 5.8: Venn diagrams representing the overlap of features and modules across the univariate and multivariate approaches for the MM (panel a) and CLL (panel b) subtypes.



For MM and proteomics (panel a) the results from gPLS-DA are not included in the diagram as all 28 inflammatory markers were retained in the optimized model.
sPLS-DA: sparse Partial Least Squares-Discriminant Analysis, gPLS-DA: group Partial Least Squares-Discriminant Analysis, sgPLS-DA: sparse group Partial Least Squares-Discriminant Analysis, LMM: Linear Mixed Model.

5.4 Discussion

The results from the regularized PLS-DA models in the proteomics dataset disclose relevant findings for the CLL and MM disease subtypes. The correct classification of samples in these sub-entities mainly relies on the cytokines sCD40L, IL4 and the chemokine Eotaxin for CLL and the growth factors TGFa, FGF2 and VEGF and the chemokines Fractalkine and MCP3 for the MM subtype. The investigated PLS-DA models failed to reveal proteins yielding an acceptable classification performance for the DLBCL and FL subtypes. On the other hand, the application of these three multivariate approaches on the transcriptomics data reveals positive results for the CLL sub-entity with a reduced set of gene expression signals driving the superior classification abilities of the statistical models. In particular, the transcripts STAP1, A_23_P500400 (ABCA6), A_23_P26854 (ARHGAP44), DNM3, WNT3, ZBTB32 are consistently highlighted as the most important to detect CLL samples. The transcripts selected for DLBCL, FL and MM subtypes were unable to yield a satisfactory classification ability. Furthermore, the analyses conducted in this chapter do not support the presence of inflammatory markers or gene expression signatures common to the four main histological subtypes being analysed, rather the findings observed for the pooled BCL study population appear to be a reflection of the favourable statistical performance for CLL and MM in proteomics and CLL only in transcriptomics.

5.4.1 Technical Assessment and Comparison of Statistical Approaches

From a methodological perspective, the analyses performed in this chapter allow for a comparative technical assessment between different statistical approaches. On the one hand, these findings shed light on the applicability and usability of the three regularized versions of PLS-DA on real-world data under different circumstances. On the other hand, a robust and detailed comparison between univariate and multivariate methods could be conducted.

In the former situation, different scenarios were contrasted: first, one where the num-

ber of observations is larger than the number of predictor variables ($n > p$, proteomics) and another where the number of features greatly exceeds the number of observations ($n << p$, transcriptomics). Second, information related to pre-existing groups of predictor variables with a similar biological function was incorporated into the analyses and systematically compared to a scenario where such information was disregarded (gPLS-DA and sgPLS-DA vs. sPLS-DA). Third, the aggregation of individual variables into groups or modules with a similar biological function was conducted in two contrasting manners, as in proteomics the grouping was performed on the basis of the information provided by the manufacturer of the omics platform (growth factors, chemokines and cytokines) while in transcriptomics such procedure was done independently by the researcher for each individual signal (using DAVID for functional annotation of probes).

Taking into consideration the points mentioned above, there seems to be a certain set of conditions that exploit the statistical performance of the investigated PLS-DA approaches. As the analyses performed on the proteomics data added valuable discoveries in relation to the CLL subtype which were undetected by the established univariate methods, it appears that the $n > p$ situation and the clear distinction of biological functions across individual features maximise the classification abilities of the models. The first point may not have played a direct role as multivariate approaches such as PLS are known to successfully accommodate the $n < p$ scenario but indirectly through the calibration procedure. The low dimension of the proteomics data may have allowed for a more precise optimization of models as opposed to the transcriptomics analysis where a limited space of parameters was assessed in the calibration process in order to reduce computational cost. This may have an impact in the final observed results: a satisfactory classification performance for FL, DLBCL and MM in transcriptomics could be possible if more complex models were explored (e.g. models with more latent variables). The second point may be responsible for the observed difference in statistical performances across regularized models: while in the proteomics analyses we consistently see the gPLS-DA and sgPLS-DA

approaches outperforming sPLS-DA, in the transcriptomics analyses the incorporation of pre-existing information regarding features with similar biological function does not yield a meaningful improvement of performance as sPLS-DA systematically outperforms the other two statistical methods. The inflammatory marker data is constituted by three clear functional groups with a comparable number of proteins per module, an attribute that is not seen in the gene expression data as most of the transcripts do not count with information about biological pathways. Furthermore, pathways are not of comparable sizes and correction for overlapping modules was needed.

In theory, the mathematical properties of the PLS technique are accommodated to adjust for potential differences in group sizes; however, in practice the application of these methods in real-world data shows they are unable to successfully incorporate information regarding biological groups under the conditions previously discussed. Consequently, these observations can lead to the recommendation that sPLS-DA may be preferred over gPLS-DA and/or sgPLS-DA where well-defined divisions of biological functions among predictor variables is unavailable. As an alternative, different functional groups of variables can be explored for gPLS-DA and sgPLS-DA at the expense of an increased computational burden. These aspects are of particular significance when considering the use of sgPLS-DA in high-dimensional data as a thorough application of the technique involves a calibration procedure where two model parameters must be optimized (in addition to the number of components). Further studies aiming at systematically comparing the three regularized methods applied on real-world data under different circumstances to the ones investigated here are required to corroborate or challenge these premises.

When comparing univariate and multivariate statistical approaches, the discoveries showcased in this chapter strongly support the use of the latter methods over the former ones with two main observations leading to such conclusion. Firstly, the regularized PLS-DA approaches were able to identify findings overlooked by the LMM (CLL subtype in proteomics) as well as to detect in a more efficient manner the most

relevant set of variables for the correct classification of samples (CLL subtype in transcriptomics). Secondly, the results from the sensitivity analysis comparing the WBC adjusted and unadjusted models reveals a substantial difference in the LMM results after WBC correction (684 vs 18 significant transcripts for the CLL subtype) which is not seen in the regularized PLS-DA; in other words, the studied multivariate approaches appear to be less sensitive to possible confounding factors. Although statistical performance and biological relevance appear to favour multivariate techniques, aspects such as computational efficiency and ability to adjust for technical-induced variability must be considered when applying these regularized PLS-DA methods.

5.4.2 Biological Relevance of Findings

For transcriptomics, the most relevant gene expression signals related to CLL also correspond to significant associations found in the LMM analyses and their biological pertinence was discussed in chapter 4. Similarly, for proteomics the possible roles of the growth factors TGFa, FGF2 and VEGF into the pathogenesis of MM were also discussed in chapter 4.

In relation to the chemokines Fractalkine and MCP3, previous studies have also shown an association between these markers and the development of this BCL subtype either in clinical or experimental settings. The former protein has been implicated in a number of pathological conditions including rheumatoid arthritis, diabetes, and some cancers where it shows pro-angiogenic and pro-inflammatory effects, which are exerted by its chemotactic activity for Natural Killer (NK) cells, Dendritic Cells (DCs) and monocytes, and mature osteoclasts [230]. Recent evidence suggests that plasma levels of Fractalkine are increased in MM patients as well as its precursor state (monoclonal gammopathy of undetermined significance [MGUS]) and the study of the concentration of this protein has been proposed as a possible new biomarker for the disease progression [230]. It is thought that its role in bone marrow vascularization is mediated through the involvement of the pro-inflammatory cytokine TNFa [230], [231], a marker that was also highlighted as a potential relevant feature by the mod-

els examined here (specifically, stability frequency analysis of sPLS-DA). On the other hand, MCP3, also known as chemokine (C-C motif) ligand 7 (CCL7), is expressed in various types of cells under physiological conditions and tumour cells under pathological conditions. It is known for being a potent chemoattractant for a variety of leukocytes and to be highly expressed in a variety of cancers including renal, gastric and colorectal [232]. It can bind to multiple transmembrane receptors and the downstream actions it exerts depend on what receptor it binds. In the particular case of MM, it has been observed that MCP3 produced by stromal cells (connective tissue cells neighbouring tumour cells) can act as a chemoattractant for human multiple myeloid cells via the interaction to the receptor CCR2 [233].

For the CLL subtype, there is some experimental evidence linking the cytokines IL4 and sCD40L with this particular BCL sub-entity [234]. As previously discussed, the phenotypic presentation of CLL is characterised by the accumulation of malignant B cells in the blood, bone marrow and secondary lymphoid organs due to the ability of the cells to escape apoptosis and it has been observed that tumour microenvironment plays a key role in the pro-survival capabilities of these malignant cells. More specifically, the tumour cells that are part of the CLL microenvironment secret several soluble factors that boost biological pathways that either activate anti-apoptosis genes or circumvent apoptotic signalling. The cytokines IL4 and sCD40L have been previously identified as most relevant secreted factors deregulating CLL pathways related to disease phenotype in terms of prolonged survival of malignant B lymphocytes. As such, the findings presented in this chapter provide some potential meta-analytical support for previously reported associations involving blood levels of inflammatory markers and specific BCL subtypes.

5.5 Conclusion

The use of multivariate techniques for sample classification purposes both confirm and improve the results obtained from the univariate statistical approaches. Com-

parable to the LMM analysis, biological markers indicative of BCL incidence were identified for MM in proteomics and CLL in transcriptomics. However, novel findings are observed as the PLS-DA analysis was successful in detecting inflammatory markers with an appropriate classification performance for CLL as well as identifying the most relevant transcripts indicative of CLL and therefore enhance sparsity in a high-dimensional setting. The exhaustive comparison between regularized approaches allows to hypothesise that the incorporation of prior biological knowledge in the form of functional groups and biological pathways introduce a valuable improvement in statistical performance and predictive accuracy, especially when a clear division between modules is available. In the following chapter I explore the interplay between transcripts and proteins by means of a two-block omics data integration approach in an attempt to unravel the biological mechanisms by which the studied molecules exert their effects.

6

Proteomics and Transcriptomics Data Integration Employing Regularized Partial Least Squares Techniques: Unravelling Complex Associations between the Two Biological Entities

6.1 Introduction

Following the results observed in the previous chapters where proteomics and transcriptomics datasets were independently analysed to identify biological markers predictive of B-cell Lymphoma (BCL), in this chapter I move towards the integration of the two omics blocks in order to identify biologically relevant information for the disease outcome under study. The same supervised Dimension Reduction Technique (DRT) previously employed in a discriminant analysis context are now used in a two-block setting whereby gene expression signals are regressed against inflammatory markers to identify co-expression patterns of interest in the two biological entities. Since the three regularized extensions of Partial Least Squares (PLS) are applied, I compare the outcome obtained from each of the statistical methods in terms of sparsity and interpretability, model performance as well as biological relevance. The consistency and robustness of the findings across techniques is also assessed. Special

emphasis is placed on the visualisation tools available for the discovery of associations between the paired omics data sets. This simultaneous analytical framework on both real-world omics datasets attempts to decipher mechanistic details in the information flow within cells that can ultimately uncover perturbations underlying the phenotype of interest. As such, the identification of alterations in the molecular biology of cells at the transcriptome and proteome levels constitute one of the many information layers needed to be explored for a comprehensive understanding of human health and disease and thus contribute to bridging the gap from environmental exposure to appearance of clinical manifestations.

6.2 Methods

The statistical methods sparse PLS (sPLS), group PLS (gPLS) and sparse group PLS (sgPLS) were applied using the "de-noised" gene expression levels as the predictor matrix and the "de-noised" immune marker concentration levels as the outcome matrix. Therefore, an asymmetric relation between datasets was assumed whereby transcripts were considered to be predictive of proteins (i.e. PLS Regression (PLS-R) was performed as opposed to the canonical mode). Statistical analyses were restricted to Chronic Lymphocytic Leukaemia (CLL) and Multiple Myeloma (MM) cases which were the two major disease subtypes for which relevant findings were identified in previous chapters (transcriptomics analyses identified signals for CLL and proteomics analyses identified markers for CLL and MM) alongside their corresponding paired control subjects. Thus, a total of 202 observations were included in the study population comprising 67 MM subtypes, 34 CLL subtypes (a total of 101 cases) and their 101 matched control individuals. The R-statistical package `mixOmics` was employed to fit the sPLS statistical model while the gPLS and sgPLS models were fitted with the R package `sgPLS`.

The definition of biological pathways and functional groups required for the gPLS and sgPLS analyses was conducted following the same criteria used in the discrim-

inant analyses context: the 29,662 individual transcripts were grouped into 850 biological pathways employing the online bioinformatics tool Database for Annotation, Visualization and Integrated Discovery (DAVID v6.8, http://david.abcc.ncifcrf.gov/) and the 28 proteins were classified into three functional groups (growth factors, chemokines and cytokines) following the information provided by the kit manufacturer. More details on the grouping of features into modules were provided in section 5.2.

6.2.1 Calibration Procedure: Defining the Optimal Value of Parameters

For the three PLS variants being conducted, the sequential strategy described in section 3.6.2.1.1 was employed to optimize the corresponding tuning parameters whereby the aim is first to define an optimal number of components followed by the optimization of the parameters defining the appropriate degree of sparsity within dimensions. In a classical PLS model (i.e. where sparsity is not imposed), the R^2 or percentage of explained variation in the outcome matrix and the Q^2 statistics were calculated to determine the optimal number of dimensions where models of up to six latent variables were explored. Taking into consideration computational cost, ability to handle the selected variables in upstream analyses and the total number of components with which the model may be fitted (which in PLS-R is determined by the total number of features in the $\mathbf{Y}$ matrix), six dimensions was deemed as an appropriate number of components to be examined.

After the appropriate number of components was determined, the additional model parameters were optimized by M-k-fold Cross Validation (CV, k = 5 and M = 50). As conducted in the previous chapter, the procedure was performed ensuring that case-control pairs were allocated to the same fold in each of the iterations in order to reduce unwanted sources of variation. The chosen metric of performance was the minimisation of the Mean Squared Error of Prediction (MSEP) averaged across the 50 CV repetitions which in turn is averaged across the 28 proteins since a prediction is

obtained for each individual feature of the outcome matrix. In the case of sPLS and gPLS, two-dimensional grids of possible values of the model parameters were investigated with each dimension seeking to define the optimal degree of sparsity in each omics matrix. On the other hand, in sgPLS a four-dimensional grid was examined as the optimal degree of sparsity in each omics matrix depends on two parameters. The specific additional tuning parameters required to be calibrated for each of the three integrative approaches are detailed as follows:

- sPLS: number of gene expression signals to retain in the predictor matrix and number of proteins to retain in the outcome matrix.
- gPLS: number of biological pathways to retain in the predictor matrix and number of functional groups to retain in the outcome matrix.
- sgPLS: number of biological pathways to retain in the predictor matrix and the degree of sparsity within the pathways (mixing parameter α_1) and number of functional groups to retain in the outcome matrix and the degree of sparsity within the groups (mixing parameter α_2).

The high-dimensional nature of the gene expression data requires a reduced sequence of values to be investigated; therefore, the same grid of values to the ones employed in the discriminant analyses situation are explored:

- sPLS: models retaining 2 to 1000 transcripts (grid resolution of 40 values).
- gPLS: models retaining 1 to 150 biological pathways (grid resolution of 28 values).
- sgPLS: models retaining 1 to 150 biological pathways (grid resolution of 28 values) and a within-group degree of sparsity (α_1) ranging from 0.05 to 0.95 (grid resolution of 11 values).

As discussed in section 5.2, the grids specified above were defined under the main assumption that the inclusion of additional transcripts or pathways would produce negligible reduction of the error of prediction and high computational cost of the

calibration procedure. In addition, it was also assumed that the ability to manage the selected variables in following analyses and extract relevant biological information would be diminished if more features were included in the models. On the other hand, given the proteomics dataset has a lower dimension, the complete model space was investigated for the calibration process of the outcome matrix:

- sPLS: models retaining 1 to 28 markers (grid resolution of 28 values).
- gPLS: models retaining 1 to 3 functional groups (grid resolution of 3 values).
- sgPLS: models retaining 1 to 3 functional groups (grid resolution of 3 values) and a within-group degree of sparsity ranging (α_2) from 0.05 to 0.95 (grid resolution of 11 values).

It is important to highlight that the grid used for the calibration of the number of modules in gPLS is the same as the one used in sgPLS (for both the predictor and outcome matrices); likewise, the same grid was employed for the mixing parameters α_1 and α_2.

6.2.2 Assessment of Calibrated Models

Once the corresponding optimal parameters were tuned for the three PLS variants and for both the predictor and outcome matrices, the prediction performance of the optimized models was evaluated by CV under the same criteria applied in the calibration procedure (k=5, M=50 and paired case-control subjects being included in the same fold). The computation of MSEP averaged across CV folds was the metric of performance of choice to compare the predictive ability of the models for each of the 28 proteins.

Pair-wise associations within and between omics datasets were investigated by examination of the visualisation outputs available for that purpose: relevance networks, Clustered Image Map (CIM)s and correlation circle plots. These tools correspond to graphical representations of a similarity matrix which is a robust approximation of the Pearson correlation coefficients of the $\mathbf{X}$ and $\mathbf{Y}$ features. In the case of correla-

tion circle plots, new coordinates of the selected features are obtained by calculating the correlation between each original variable and their associated component which allows the identification of correlated features and the direction of the association as well as subsets of variables that are important to define each component. Superimposed sample plots were also inspected to assess the possible resemblance or distinction that the three observation types being included in the study population may show in the lower dimension space spanned by the PLS components. The contribution of individual variables to the overall predictive model performance was explored by assessment of the loading coefficients of the features that were selected in the calibrated models within each component and within each omics dataset. (Details on the graphical outputs discussed here were presented in section 3.6.2.5)

Gene expression signals selected by the integrative approaches were further investigated through gene-enrichment analyses using the openly available DAVID v6.8. Finally, a quantitative and a qualitative assessment is made by comparing the findings obtained from the three PLS models at each of step of the analyses being conducted.

6.3 Results

6.3.1 Calibration Procedure

The classical PLS model fitted with up to six components shows a limited percentage of variance explained in the $\mathbf{Y}$ matrix in all models visited and a Q^2 statistics below the threshold of 0.0975 in all the situations (see Table D.1). These results support fitting the PLS regularized versions with only one dimension; however, it was considered unlikely that out of the 28 possible components only one was enough to produce models with high predictive accuracy. Therefore, three dimensions was judged as a more suitable number of latent variables to fit the regularized models as it represents a trade-off between model complexity, interpretability and computational efficiency.

Figure D.1 to Figure D.3 represent the calibration curves with the average MSEP for

all the values of the parameters tested for the three regularized models. In relation to the outcome matrix, a clear pattern can be observed for both sPLS and gPLS where the error of prediction increases as more proteins and functional groups (respectively) are included in the model for all three components. A similar pattern is observed for sgPLS as models retaining one functional group show better performance (excepting for the third component where the model with two modules yields lower error); however, the within-module sparsity of the outcome matrix does not consistently favour the model with less proteins (more specifically, the optimal vale of α_2 for the third component selecting two groups was 0.1). On the other hand, the behaviour of the prediction error in relation to the inclusion of $\mathbf{X}$ variables is consistent across all approaches and components. Low estimates of the average MSEP are observed for the first values, the highest average MSEP is seen for middle values followed by a decrease and plateau of the curves for the final set of values of the grid of parameters.

For all PLS models, the decision to define the optimal degree of sparsity was based on both the parameter value yielding the lowest average MSEP as well as visual inspection of the calibration curves in order to produce a balance between predictive power and interpretability. The final values of the tuning parameters determining the optimal degree of sparsity in both the $\mathbf{X}$ and $\mathbf{Y}$ matrices for the three dimensions and for the three regularized PLS methods as well as the corresponding number of features and/or modules selected per component are presented in Table 6.1. The retrieved cross-validated MSEP averaged across the 28 proteins is also presented in Table 6.1.

6.3.2 Assessment of Calibrated Models

The total number of features and modules selected in both the $\mathbf{X}$ and $\mathbf{Y}$ matrices across components in the final calibrated model for each of the three integrative approaches are provided in Table 6.2.

The optimized sPLS model selected 150,80 and 9 gene expression signals ($\mathbf{X}$ matrix) and two, two and three ($\mathbf{Y}$ matrix) proteins in its first, second and third components, respectively. One feature in each omics block was jointly selected across dimensions.

Table 6.1: Model parameters defining optimal degree of sparsity and number of variables and/or modules selected in both the $\mathbf{X}$ and $\mathbf{Y}$ matrices for the three dimensions and for each of the three integrative approaches.

	1st Component	2nd Component	3rd Component
sparse PLS			
N Variables $\mathbf{X}$ (transcripts)	150	80	9
N Modules $\mathbf{X}$ (biological pathways)	44	34	4
N Variables $\mathbf{Y}$ (proteins)	2	2	3
N Modules $\mathbf{Y}$ (functional groups)	2	1	2
Average MSEP	2.09	2.17	2.23
group PLS			
N Variables $\mathbf{X}$ (transcripts)	27	1	1
N Modules $\mathbf{X}$ (biological pathways)	20	1	1
N Variables $\mathbf{Y}$ (proteins)	6	18	6
N Modules $\mathbf{Y}$	1	2	1
Average MSEP	2.06	2.11	2.17
sparse group PLS			
N Variables $\mathbf{X}$ (transcripts)	48	1	1
N Modules $\mathbf{X}$ (biological pathways)	25	1	1
N Variables $\mathbf{Y}$ (proteins)	3	5	16
N Modules $\mathbf{Y}$ (functional groups)	1	1	2
Mixing Parameter α_1	0.05	0.6	0.05
Mixing Parameter α_2	0.95	0.9	0.1
Average MSEP	2.06	2.11	2.15

The resulting cross-validated MSEP averaged across the 28 proteins is also reported.
PLS: Partial Least Squares, MSEP: Mean Squared Error of Prediction.

Thus, a total of 238 unique transcripts and six unique immune markers were selected in the calibrated sPLS model which belong to 70 biological pathways and two functional groups. The vast majority of the selected transcripts (n = 138) belong to the group for which information on biological pathways was unavailable (n = 14,737); on the other hand, 54 of the chosen signals correspond to pathways containing only one transcript (i.e. one transcript correspond to an independent biological pathway). Table 6.3 shows details of the selected gene expression signals in terms of the most

Table 6.2: Total number of unique features and modules selected in both the **X** and **Y** matrices for each of the three integrative approaches.

	Transcriptomics matrix		*Proteomics matrix*	
Model type	**Transcripts**	**Biological Pathways**	**Proteins**	**Functional Groups**
sparse PLS	238	70	6	2
group PLS	29	22	18	2
sparse group PLS	50	27	16	2

PLS: Partial Least Squares.

frequent biological pathways to which they belong. In the case of the outcome matrix, the proteins retained in the calibrated model correspond to the growth factors EGF, FGF2 and VEGF which are three out of the six markers that fall into the mentioned functional group and the chemokines MCP3, MIP1a and MIP1b which are three out of the 10 markers that belong to that module. The protein MIP1a was simultaneously selected in the second and third components.

A total of 22 biological pathways were retained in the calibrated gPLS model of which 20 were selected in the first component, and one in the second and third components. These selected modules account for 29 different transcripts of which 18 correspond to one-probe pathways, as is the case for the groups chosen in the second and third components (see Table 6.3 for details). In the proteomics dataset, a total of four functional groups were selected from the calibration procedure: one group in the first and third components and two in the second component. These corresponds to the growth factor and cytokine categories accounting for a total of 18 different proteins, six of which belong to the first functional module and 12 to the second one. The growth factor markers were chosen in the first and third components while the second component selected the cytokine group in addition to the growth factor category. Consequently, an overlap of features can be observed across the **Y** components (18 unique features versus the total of 30 chosen proteins) which contrasts with the findings obtained in the predictor matrix where no overlap of transcripts across the **X** dimensions was detected.

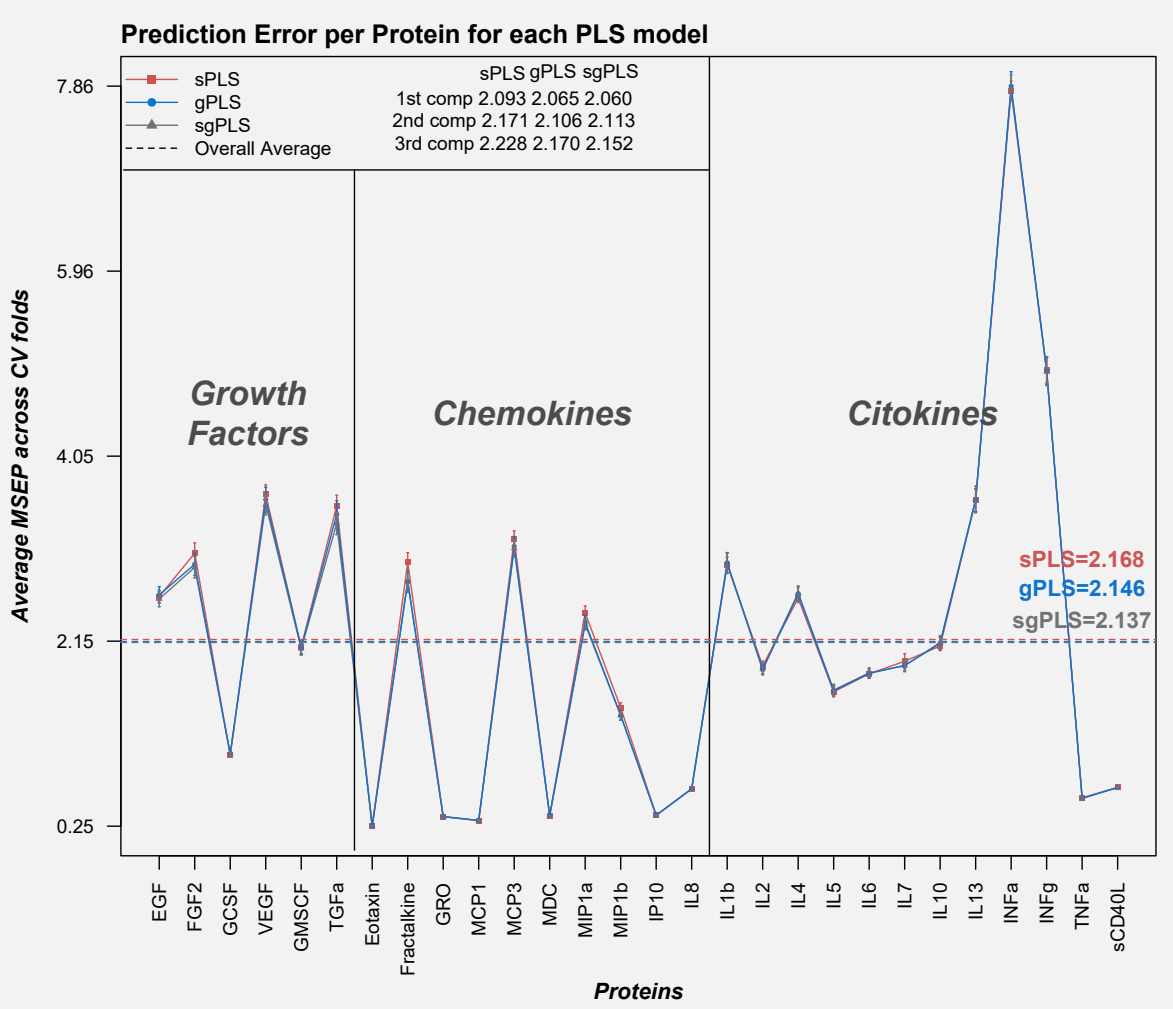
The optimized sgPLS model produced similar outputs to the gPLS model as 27 different biological pathways were chosen of which 18 correspond to one-transcript groups, with the second and third components also selecting one pathway, each of only one probe. The mixing parameter controlling the within-group sparsity in the predictor matrix (α_1) favours a full model as all the transcripts belonging to the selected groups were retained (i.e. no within-group sparsity was effectively imposed on $\mathbf{X}$). The retained modules account for a total of 50 different transcripts selected in the predictor matrix (see Table 6.3 for details). In the case of the outcome matrix, as observed in the gPLS model, the chosen functional groups were four, one in the first (growth factors) and second components (chemokines) and two in the third dimension (growth factors and chemokines). As a result of imposing within-group sparsity, the first component selected three out of the six proteins that are part of the growth factor category (FGF2, VEGF and TGFa), the second component chose five out the 10 immune markers that are part of the chemokine module (IL8, Fractalkine, MCP3, MIP1a and MIP1b) and 16 markers were retained in the final component which are the entirety of the features that belong to the chosen groups. Consequently, an overlap is observed across de $\mathbf{Y}$ components because of the 24 proteins retained in the model, 16 correspond to unique markers with the reaming eight being included in at least two components. On the other hand, there is no intersection of transcripts across the $\mathbf{X}$ components selected in the calibrated sgPLS model.

Table 6.3: Biological pathways to which the selected gene expression signals belong, total number of probes in the selected pathway and absolute and relative frequencies of the selected signals per pathway for each of the three integrative approaches.

sPLS			*gPLS*		*sgPLS*		
Biological pathway	**Total**	**Selected (%)**	**Biological pathway**	**Total**	**Biological pathway**	**Total**	**Selected (%)**
Non-annotated probes	14737	138 (0.93)	Regulation of DNA repair	4	Female pregnancy	7	7 (100)
Regulation of transcription	1254	6 (0.48)	Secretion	3	Respiratory gaseous exchange	6	6 (100)
Cation transport	240	6 (2.50)	Glial cell differentiation	2	Cytoskeletal anchoring at plasma membrane	4	4 (100)
Transcription	527	5 (0.95)	Urea transport	2	Regulation of DNA repair	4	4 (100)
G-protein coupled receptor protein signalling pathway	274	4 (1.46)			Embryonic development ending in birth or egg hatching	3	3 (100)
Phosphorus metabolic process	426	3 (0.70)			Gas transport	2	2 (100)
Gamete generation	40	3 (7.50)			Glial cell differentiation	2	2 (100)
Cell-cell adhesion	17	3 (17.65)			Positive regulation of macromolecule biosynthetic process	2	2 (100)
Macromolecule catabolic process	224	2 (0.89)			Urea transport	2	2 (100)
Cell surface receptor linked signal transduction	140	2 (1.43)					
Intracellular signalling cascade	120	2 (1.67)					
Immune response	102	2 (1.96)					
Biological adhesion	91	2 (2.20)					
Defense response	75	2 (2.67)					
Protein amino acid phosphorylation	67	2 (2.99)					
RNA export from nucleus	18	2 (11.11)					

For the three models, only the pathways containing two or more selected signals are shown corresponding to 16,4 and 9 for the sPLS, gPLS and sgPLS models, respectively. Thus, selected pathways of one probe are not displayed corresponding to 54,18 and 18 for the sPLS, gPLS and sgPLS models, respectively.
sPLS: sparse Partial Least Squares, gPLS: group Partial Least Squares, sgPLS: sparse group Partial Least Squares.

Figure 6.1: Model performance assessment for the three calibrated PLS approaches.



The overall mean (dashed lines) refers to the average MSEP across CV folds and averaged across the 28 proteins. In other words, it represents the MSEP of the optimized models. The legend box displays the average MSEP across CV folds and averaged across the 28 proteins per component.
sPLS: sparse Partial Least Squares, gPLS: group Partial Least Squares, sgPLS: sparse group Partial Least Squares, MSEP: Mean Squared Error of Prediction, CV: Cross-Validation.

6.3.3 Model Performance

The predictive ability of the three optimized models is shown in Figure 6.1 where the MSEP is displayed per each protein. The average MSEP per component and across component is also presented. Although the three calibrated models yield similar error of prediction, it can be noticed that sPLS is marginally outperformed by both gPLS and sgPLS in the three dimensions explored while the error from sgPLS was slightly

lower than gPLS for the models with one and three dimensions. Upon examination of each individual feature of the outcome matrix, it can be observed that some proteins consistently show a lower average MSEP across CV folds with Eotaxin, GRO, MCP1, MDC, IP10 (chemokines) and TNFa and sCD40L (cytokines) presenting the lowest error of prediction in the optimized PLS models.

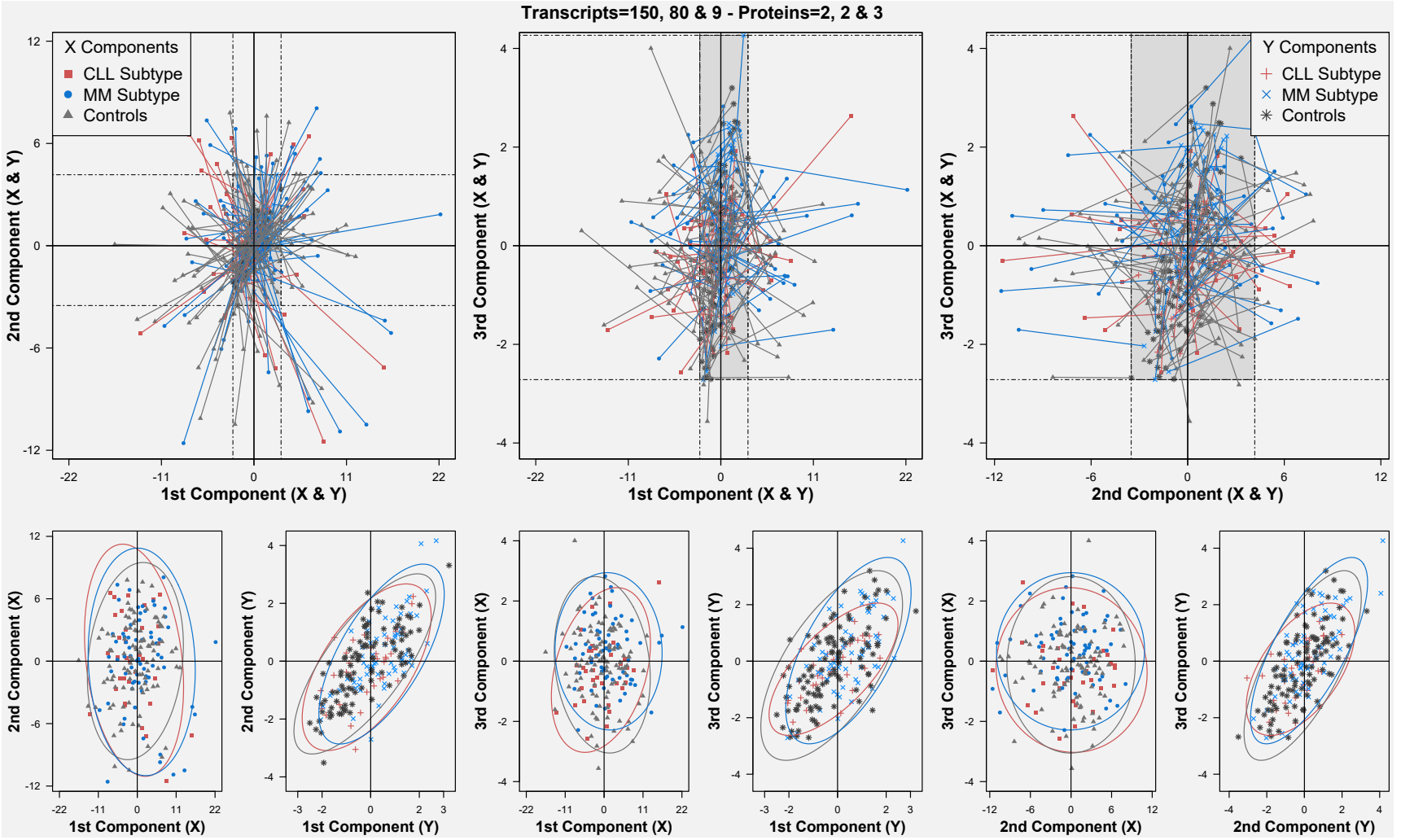
6.3.4 Assessment of Visualisation Tools

6.3.4.1 Superimposed Plots

The projection of the samples in the lower dimensional space spanned by the calibrated PLS models is displayed in Figure 6.2, Figure D.4 and Figure D.5. The three PLS models show comparable results with similar clustering patterns of samples across the combination of two dimensions used for the representation (1st vs 2nd, 1st vs 3rd and 2nd vs 3rd components). All models were unable to produce a clear distinction between the three types of observations (MM and CLL subtypes and control individuals) included in the study population across all two-dimensional spaces being examined. A marginal separation of samples can be detected in the sPLS model (2nd vs 3rd component) where the pair of second dimensions (from the $\mathbf{X}$ and $\mathbf{Y}$ matrices) tends to separate the MM observations from the other individuals as most of those samples are spread across the horizontal line (see Figure 6.2 top panel third plot from left to right). Other variables of possible biological importance (gender, smoking status, level of physical activity, median time to diagnosis) and technical influence (study of cohort) were also investigated, but a clear clustering pattern of samples did not emerge as a result (results not shown). These findings suggest that the pair-wise associations independently identified by the calibrated models between the transcriptomics and proteomics datasets are shared across the two disease subtypes as well as the control subjects. In addition, the visual representation of the samples in the lower-dimensional space allows for an assessment of the level of agreement between the two data sets according to the applied approaches. The graphics from the three PLS models show that both data sets are only slightly related as for the

vast majority of samples long arrows are generated, indicating that the position of the samples in the reduced space spanned by the $\mathbf{X}$ matrix differs considerably from their corresponding position in the $\mathbf{Y}$ matrix.

Figure 6.2: Sample representation plots displaying the location of the observations on the **X** and **Y** spaces spanned by the calibrated sPLS model (superimposed plots).



The three possible two-dimensional spaces are exhibited. The lower space spanned by the **Y** components is coloured on grey. The separate sample plots are also displayed. For each sample group, ellipses of the confidence region were drawn employing the R package `ellipse` based on the variance and mean of the matrix of the corresponding pair of components (the mean defines the location of the ellipse centre). The confidence level controlling the ellipse size was 0.95 and a total of 100 points were used. A similar output is obtained for the superimposed plots of the gPLS and sgPLS models, therefore, they are shown in the Appendix section.
sPLS: sparse Partial Least Squares, gPLS: group Partial Least Squares, sgPLS: sparse group Partial Least Squares.

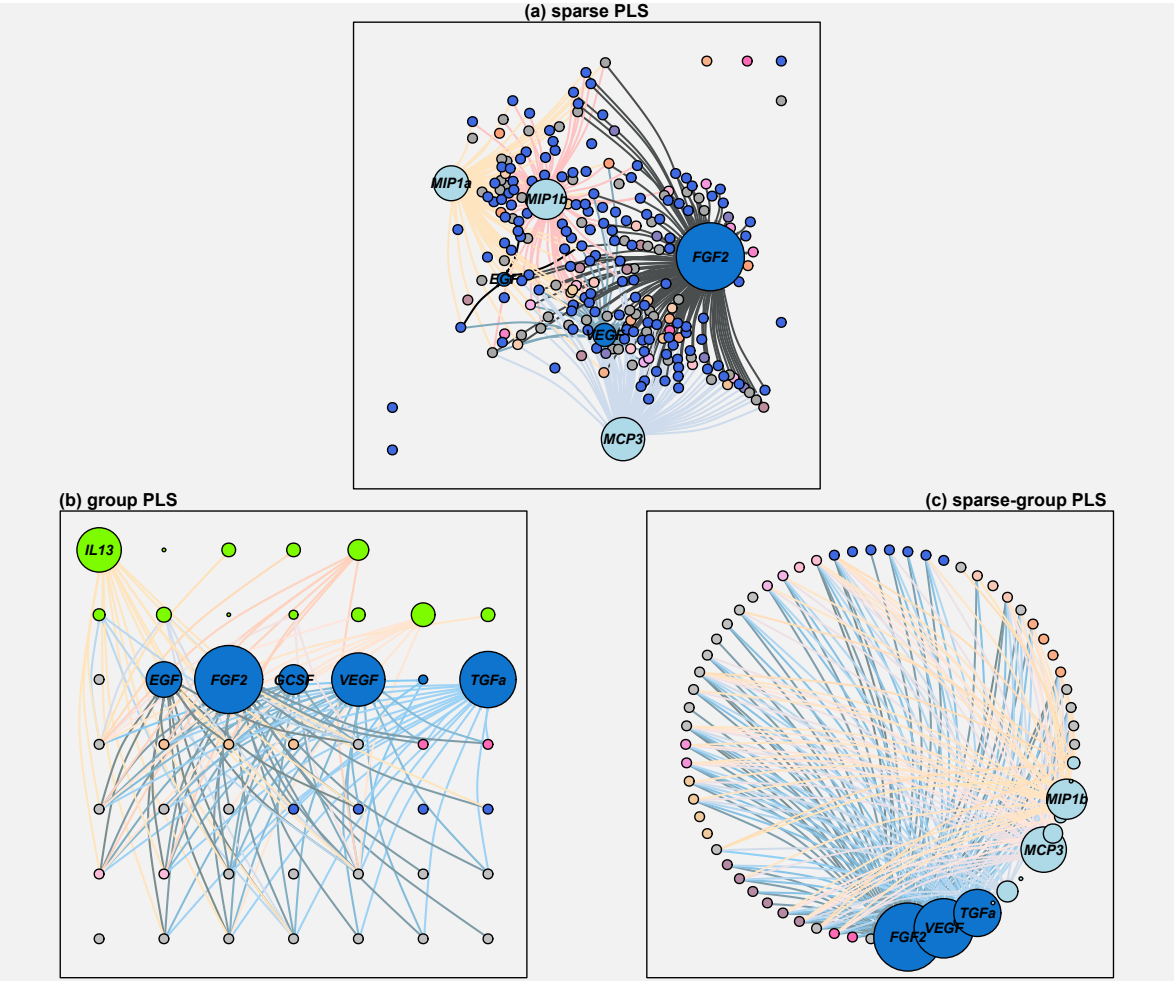
6.3.4.2 Relevance Network

The visual representation of the correlation between the selected $\mathbf{X}$ and $\mathbf{Y}$ features in the form of bipartite graphs from the three PLS approaches is displayed in Figure 6.3. Edges between nodes were drawn if the estimated correlation surpassed the 3rd quantile of the correlation distribution (in absolute value) in order to favour interpretability and remove weaker associations. In addition, in an attempt to identify possible clusters or sub-networks of subsets of variables the nodes were coloured by the biological pathway (predictor matrix) or functional group (outcome matrix) to which they belong.

Of the six immune markers selected in the sPLS approach, FGF2 (growth factor) stands out as the protein with the highest number of connections (n=115, 32.21%) followed by MCP3, MIP1b and MIP1a (chemokines) with similar number of links (n=73, 20.45%; n=68,19.05%, and n=59, 16.53%, respectively). On the other hand, considering that variables with similar connections have a close position in the network structure, inspection of the selected gene expression signals does not reveal a clear pattern of the biological pathways with the unannotated probes (n=138) dominating the graphic.

Expectedly, a more parsimonious display is exhibited by the relevance network of the gPLS model. Despite a total of 18 different proteins being selected belonging to the growth factor and cytokine categories, the majority of connections are made to the former functional group. As per sPLS, the growth factor FGF2 presents the highest number of edges drawn to it (n=23, 17.56%) followed by TGFa and VEGF (n=19,14.50% and n=18, 13.74%, respectively). The marker IL13 is highlighted as the cytokine with the most links among the other markers of its same functional group (n=15, 11.45%). In relation to the transcripts, and although most of the selected biological pathways only contain one probe (four of the 22 retained pathways correspond to modules with more than one variable), it seems that signals belonging to the same pathway display a similar connectivity pattern.

Figure 6.3: Relevance networks from the three integrative approaches.



Colours of the nodes are determined by the biological pathway or functional group to which transcripts or proteins belong (referring to the predictor and outcome matrices, respectively). All the transcripts that also constitute a single pathway (i.e. one-probe modules) are coloured in grey corresponding to 54,18 and 18 for the sPLS, gPLS and sgPLS models, respectively. Pathways containing two or more transcripts correspond to 16,4 and 9 for the sPLS, gPLS and sgPLS models, respectively. Size of the protein nodes is given by the number of connections. For the gPLS and sgPLS graphics, proteins with number of connections above the average are named. In relation to the network edges, colours are dictated by the protein to which transcripts connect and edges between nodes were drawn only if the estimated correlation exceeds the 3rd quantile of the correlation distribution (in absolute value).
PLS: Partial Least Squares.

The findings of the relevance network of the sgPLS model represent a compromise between what was observed in the sPLS and gPLS approaches. The growth factor FGF2 stands out as the protein with the greatest number of connections with the selected gene transcripts (n=39, 19.5%), similarly to the sPLS and gPLS bipartite graphs. VEGF and TGFa are growth factors that also present a substantial number of edges (n=34, 17% and n=27, 13.5%, respectively) (which is comparable to the gPLS network)

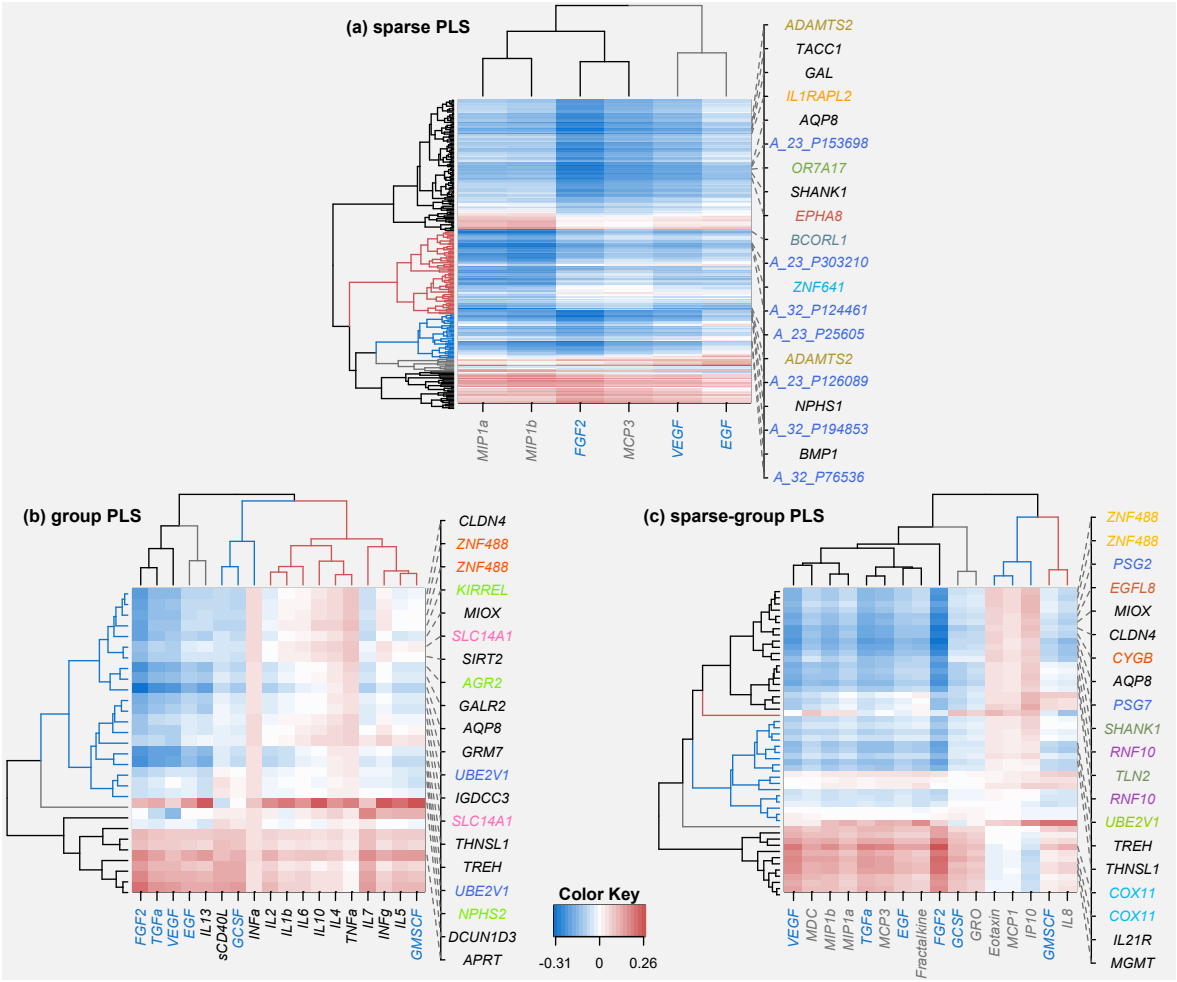
followed by the chemokines MCP3 and MIP1b (n=26, 13%, and n=23, 11.5%, respectively) (which is comparable to the sPLS network). In terms of the gene expression signals, it is observed that transcripts belonging to the same biological pathway (nine of the 27 retained modules contain more than one transcript) exhibit an equivalent connectivity with signals of a group having links to the same group of proteins.

6.3.4.3 Clustered Image Maps

The CIMs produced by the three integrative approaches are shown in Figure 6.4. The visual inspection of the heatmaps reveals that negative pair-wise associations between the $\mathbf{X}$ and $\mathbf{Y}$ features are more common and stronger than the positive correlations, which is the most apparent finding shared by the three PLS models. Of the three approaches, sPLS displays the highest relative number of negative and strong correlations followed by the sgPLS and gPLS models, with the latter showing a more balanced number of negative and positive correlation as well as weaker correlation estimates. An additional discovery shared across the all heatmaps is that the growth factor FGF2 is the marker presenting the strongest correlations (both positive and negative) with the corresponding selected gene expression signals (i.e. darkest coloured column).

When inspecting the dendrograms, the reordering of the variables given by the hierarchical clustering provide a cluster structure comparable to what was revealed in the relevance network display. The CIM from the sPLS approach shows that the proteins MIP1a, MIP1b, FGF2 and MCP3 present a high degree of similarity while the markers VEGF and EGF form another group with low dissimilarity; thus, two possible cluster of variables can be distinguished. The $\mathbf{Y}$ matrix dendrogram of gPLS shows four possible clusters containing three, two, three and 10 proteins. The first cluster (from left to right in Figure 6.4 panel b) consists of the variables with the strongest correlation estimates while the remaining clusters show weaker coefficients, with the fourth cluster mainly grouping positive weak pair-wise associations. From the immune markers dendrogram of the sgPLS model, four possible clusters can also be

Figure 6.4: Clustered Image Maps (CIMs) from the three integrative approaches.



The first 20 transcripts presenting the highest average correlation (across all selected proteins) in absolute value are named. Colours of the axis labels are determined by the biological pathway or functional group to which transcripts or proteins belong (referring to the predictor and outcome matrices, respectively). All the transcripts that constitute a single pathway (i.e. one-probe modules) are coloured in black. In relation to the dendrograms, the colours of the branches represent the clusters of features and a height of 1 and 0.2 were used as cut-off values to define the number of clusters for the predictor and outcome matrices, respectively. The Euclidian and complete methods were employed to define the distance measure and agglomeration method, respectively.
PLS: Partial Least Squares.

identified of nine, two, three and two variables; the first one (from left to right in Figure 6.4 panel c) showing the strongest correlation coefficients and third one with mainly positive estimates. On the other hand, subsets of variables can be identified from the gene expression signal dendrograms: five clusters in sPLS, two plus one outlier transcript in gPLS and three plus two outlier transcripts in sgPLS. In each of the PLS models, only one of the clusters is comprised by positive associations.

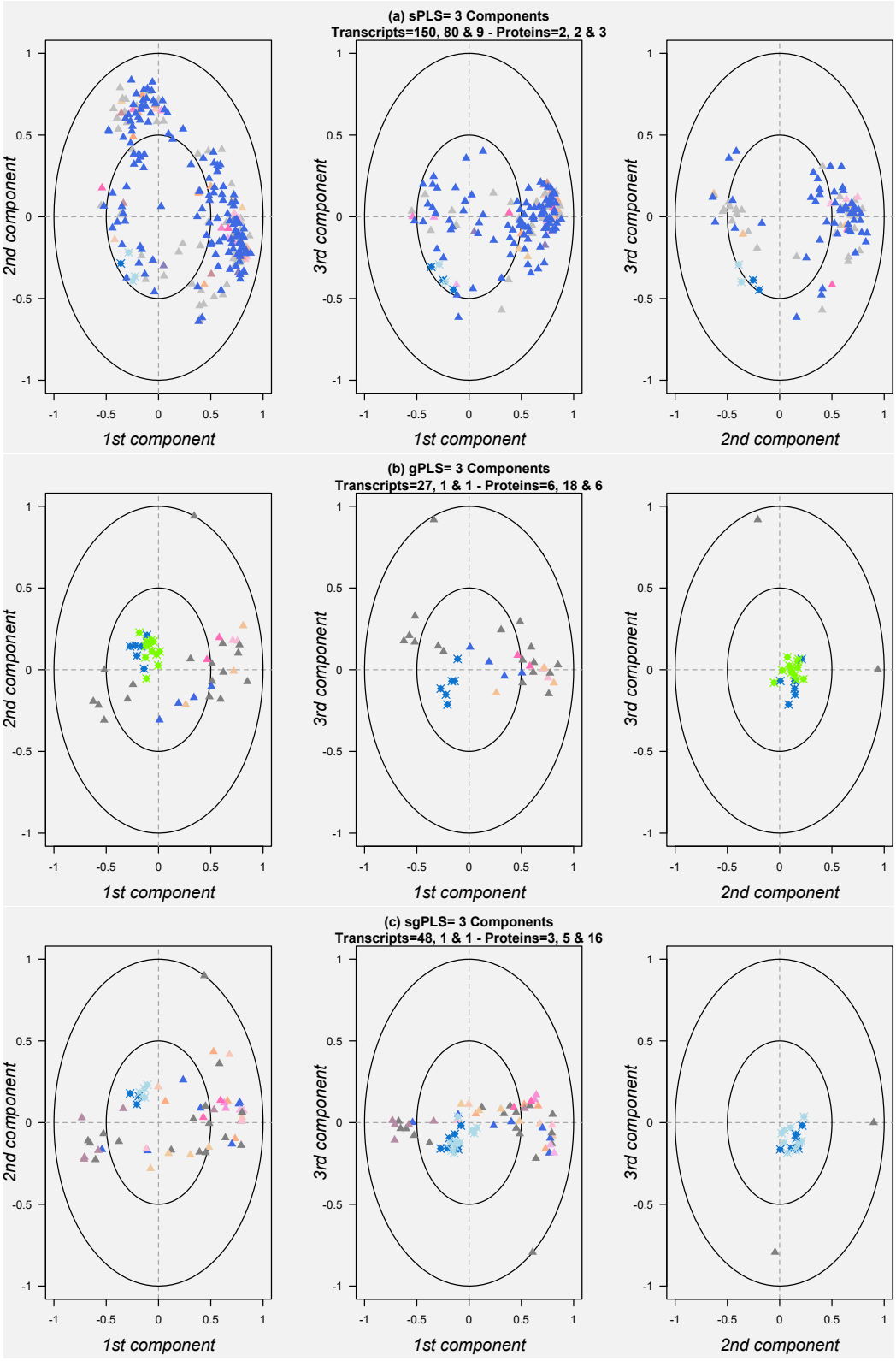
6.3.4.4 Correlation Circle Plots

The graphic illustrating the correlation circle plots of the three PLS approaches for all two-dimensional space combinations is displayed in Figure 6.5. In the three integrative approaches explored, the selected proteins are closely located to the origin (within or near the circle of radius 0.5), which is an indicator that more dimensions are needed to better visualise $\mathbf{Y}$ variable associations by means of this graphical output. Nonetheless, subsets of gene expression variables can be identified. The 1st vs 2nd component plot of the sPLS model reveals two subgroups of variables located to some extent perpendicular to each other, which suggests that variables within subgroups are highly correlated with each other while correlation between the subsets is weak or close to null. The following correlation circle plots (1st vs 3rd and 2nd vs 3rd dimensions) show one significant subset of transcripts each, mostly belonging to the group of non-annotated probes. Subsets of correlated transcripts can also be observed in the gPLS plots; however, they appear to be sparser and more disperse as well as closer to the origin, suggesting only a weak correlation pattern between variables and within groups of variables. As with the findings described in the relevance network analyses, the sgPLS model appears to offer a compromise between the sPLS and gPLS correlation circle plots as the identified subsets of variables are either denser and more compact or more spread and farther away from the origin. Finally, features within the sgPLS subgroups seem to respect the biological pathway classification as variables that belong to a given pathway are highly correlated to one another (i.e. transcripts of the same pathway are also part of the same subset).

6.3.4.5 Loading Coefficient Plots

The graphical display of the loading coefficients of the selected features in both the predictor and outcome matrices for the sgPLS model is presented in Figure 6.6 (results from the other two methods are not shown). The illustration shows that most of the retained transcripts present positive coefficients (n = 34, 68%) and the top 10 gene expression signals with the highest absolute value of loading coefficient are explicitly

Figure 6.5: Correlation circle plots from the three integrative approaches.



The three possible two-dimensional spaces are exhibited. Triangles represent gene expression signals and crossed circles represent proteins. Colours of the triangles and circles are determined by the biological pathway or functional group to which transcripts or proteins belong, respectively. All the transcripts that constitute a single pathway (i.e. one-probe modules) are coloured in grey. Two circles of radii 0.5 and 1 are drawn to better detect relevant subgroups of variables.
sPLS: sparse Partial Least Squares, gPLS: group Partial Least Squares, sgPLS: sparse group Partial Least Squares.

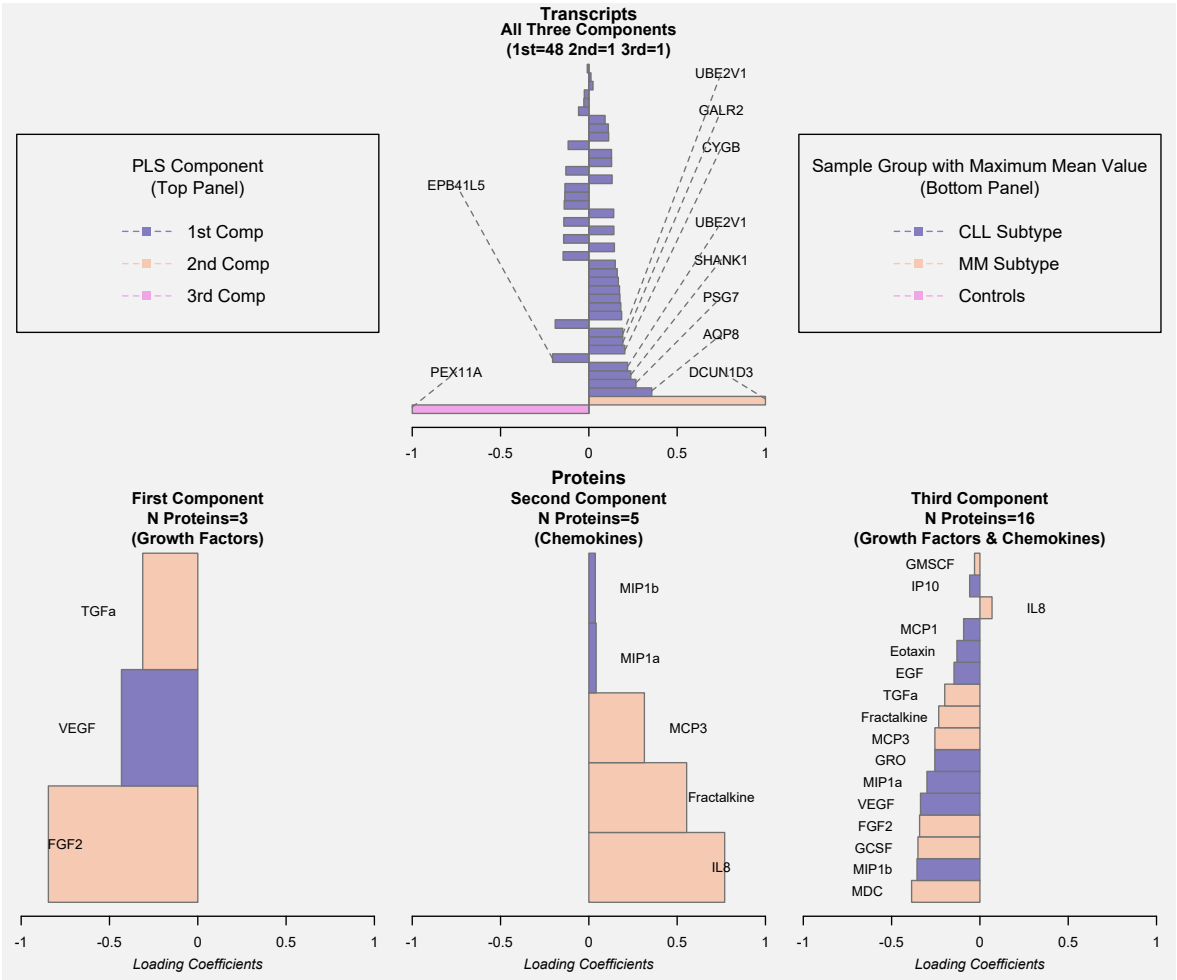
indicated. Among them, we find the single transcript retained in all three models (AQP8), the two common selected transcripts between sPLS and sgPLS (SHANK1 and PSG7) and five of the 21 common transcripts selected between gPLS and sgPLS are included (UBE2V1, GALR2, EPB41L5, UBE2V1 and DCUN1D3). More specifically, the signal AQP8 stands out as the highest contributing variable of the first component.

In relation to the proteomics data, the retained features display negative, positive, and mostly negative (n = 15, 93.75%) loading coefficients for the first, second and third components, respectively. It can be observed that the most significant variable is the growth factor FGF2 as it presents the highest absolute value in loading coefficient followed by the markers VEGF and TGFa, also growth factors. The chemokines IL8, Fractalkine and MCP3 are the proteins driving most of the variation in the second dimension while a relative even contribution is made by the 16 variables of the third component. In addition, the chosen markers retained in the calibrated sgPLS model appear to be reflection of variation patterns within MM cases and to a lesser extent CLL as these are the sample groups with the highest mean protein concentration value.

6.3.5 Biological Interpretation of the Findings

Results from the gene-enrichment analysis showed that, of the 238 different gene expression signals retained in the calibrated sPLS model, 166 (69.75%) were mapped to DAVID's database for functional annotation which were grouped into 24 different gene enriched pathways. As per gPLS and sgPLS, 24 of the 29 (82.76%) and 38 of the 50 (76%) selected transcripts were mapped to DAVID IDs being clustered in three and seven pathways, respectively. Upon further filtering (EASE score$<$0.01, setting the minimal number of probes per functional group to 5 and fold enrichment value$>$3), six, one and one biological pathways were retained. Furthermore, gene-enrichment analysis pooling all unique probes from the three integrative approaches (n=270) were matched to 209 DAVID IDs (77.41%) forming 32 functional annotation

Figure 6.6: Loading coefficients of the selected variables in both the predictor and outcome matrices for the sgPLS model.



Top panel simultaneously displays loading coefficients of the gene expression signals (predictor matrix) selected across components while bottom panels represents coefficients of the inflammatory markers (outcome matrix) separately for each of the dimensions. In the case of the transcripts, the variables presenting the 10 highest coefficient value (in absolute terms) of are named. In the case of the proteins, variables are coloured according to the sample group presenting the highest mean protein concentration value.
sgPLS: sparse group Partial Least Squares.

clusters which after the more stringent filtering were reduced to 17. The summary of the functional annotation process from the three PLS models as well as the combined selected transcripts is detailed in Table 6.4. Most of the enriched pathways relate to immune system regulation, recognition system for cell adhesion, protein transit through the cell and peptide structure, function and stability.

Table 6.4: Summary of the results from the gene-enrichment analysis from the transcripts independently selected in each of the three integrative approaches and from all unique transcripts jointly selected across the three approaches (pooled analyses).

	sparse PLS					
	Database	**Term**	**Count**	***p*-value**	**Fold Enrichment**	**Bonferroni**
1	GOTERM_MF_DIRECT	GO:0031625 ubiquitin protein ligase binding	9	0.002	4.010	0.368
2	UP_SEQ_FEATURE	domain:Ig-like C2-type 3	6	0.003	6.187	0.839
3	GOTERM_BP_DIRECT	GO:0007155 cell adhesion	11	0.003	3.049	0.935
4	UP_SEQ_FEATURE	domain:Ig-like C2-type 1	7	0.004	4.574	0.937
5	UP_SEQ_FEATURE	domain:Ig-like C2-type 2	7	0.004	4.551	0.942
6	UP_SEQ_FEATURE	short sequence motif:Cell attachment site	5	0.005	7.252	0.959
	group PLS					
	Database	**Term**	**Count**	***p*-value**	**Fold Enrichment**	**Bonferroni**
1	GOTERM_CC_DIRECT	GO:0005887 integral component of plasma membrane	7	0.007	3.756	0.390
	sparse group PLS					
	Database	**Term**	**Count**	***p*-value**	**Fold Enrichment**	**Bonferroni**
1	GOTERM_BP_DIRECT	GO:0007565 female pregnancy	5	3.405E-05	26.2	7.564E-03*

Table 6.4: Summary of the results from the gene-enrichment analysis from the transcripts selected independently in each of the three integrative approaches and from all unique transcripts jointly selected across the three PLS approaches (pooled analyses) *(cont.)*.

	Pooled analysis					
	Database	**Term**	**Count**	***p*-value**	**Fold Enrichment**	**Bonferroni**
1	UP_SEQ_FEATURE	domain:Ig-like C2-type 1	10	1.59E-04	5.093	0.114
2	UP_SEQ_FEATURE	domain:Ig-like C2-type 2	10	1.65E-04	5.068	0.118
3	UP_SEQ_FEATURE	domain:Ig-like C2-type 3	8	2.50E-04	6.430	0.174
4	UP_SEQ_FEATURE	short sequence motif:Cell attachment site	7	2.55E-04	7.914	0.177
5	GOTERM_MF_DIRECT	GO:0031625 ubiquitin protein ligase binding	11	6.98E-04	3.762	0.206
6	UP_KEYWORDS	Immunoglobulin domain	14	0.002	2.749	0.363
7	UP_KEYWORDS	Disulfide bond	49	0.002	1.530	0.365
8	UP_SEQ_FEATURE	signal peptide	49	0.003	1.507	0.857
9	UP_SEQ_FEATURE	disulfide bond	44	0.003	1.552	0.874
10	INTERPRO	IPR003598: Immunoglobulin subtype 2	9	0.003	3.754	0.719
11	GOTERM_BP_DIRECT	GO:0007155 cell adhesion	13	0.003	2.733	0.953
12	GOTERM_BP_DIRECT	GO:0007268 chemical synaptic transmission	9	0.003	3.619	0.975
13	SMART	SM00408:IGc2	9	0.005	3.416	0.406
14	UP_KEYWORDS	Signal	55	0.005	1.417	0.681
15	UP_KEYWORDS	Glycoprotein	58	0.008	1.366	0.848
16	UP_SEQ_FEATURE	metal ion-binding site:Zinc; catalytic	6	0.008	4.785	0.998
17	INTERPRO	IPR013783: Immunoglobulin-like fold	18	0.010	1.973	0.987

Biological pathways are ordered in relation to their Bonferroni adjusted *p*-values. Only one pathway reached Bonferroni significance and it is mark with a (*).
PLS: Partial Least Squares.

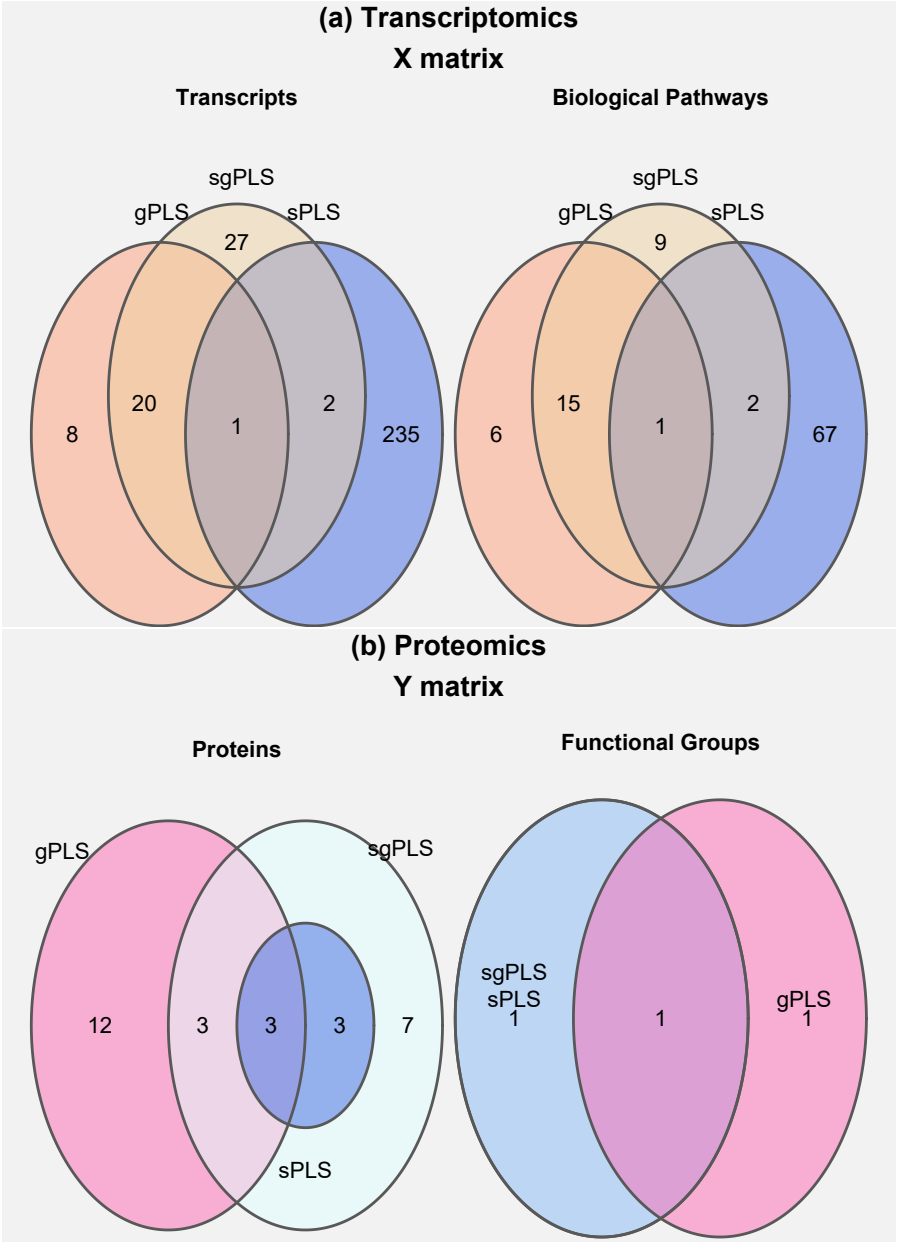
6.3.6 Comparative Assessment of Integrative Approaches

The overlap in terms of number of individual variables and modules shared across the three calibrated PLS models for both the $\mathbf{X}$ and $\mathbf{Y}$ matrices is displayed in Figure 6.7. Only one gene expression signal is common across the three statistical models (AQP8), two additional transcripts are common between the sPLS and the sgPLS models (SHANK1 and PSG7) and 20 additional probes are jointly selected in the gPLS and sgPLS methods. Inspection of the selected biological pathways reveals a similar overlapping pattern, possible because the majority of them correspond to one-feature modules. The gene AQP8 matches a single pathway and so do the two additional signals that were commonly selected by sPLS and sgPLS. As with individual probes, a greater intersection is seen between the group approaches where 16 pathways are shared between the gPLS and the sgPLS models of which 13 are composed by only one transcript. Details on the transcripts and biological pathways that are common to at least 2 of the regularized PLS models are shown in Table D.2. On the other hand, three immune markers are commonly retained across the three regularized models which correspond to EGF, FGF2 and VEGF belonging to the growth factor category. The main divergence is presented by the gPLS model which selected the cytokine group unlike the other two methods; therefore, a greater overlap is seen between sPLS and sgPLS in terms of both individual markers (six) and functional groups (two).

6.4 Discussion

The analyses conducted in this chapter constitute a comprehensive and exhaustive effort to unravel the complex associations and co-expression patters between gene expression signals and inflammatory markers that may be altered in individuals presenting BCL. Despite the limited sample size, clear findings emerge which were consistent across the three integrative approaches. The growth factors FGF2 and VEFG were consistently selected as being associated to the studied transcripts, with the former showing the strongest correlation estimates. Although less regularly across PLS

Figure 6.7: Venn diagrams representing the overlap of transcripts and biological pathways (panel a) and proteins and functional groups (panel b) shared across the three integrative approaches.



sPLS: sparse Partial Least Squares; gPLS: group Partial Least Squares; sgPLS: sparse group Partial Least Squares

models and presenting comparatively lower statistical power, the growth factor TGFa and the chemokines MCP3 and MIP1b also seem to be related to relevant gene expression signals. On the other hand, functional annotation analyses of the selected tran-

scripts exposed mainly three different group of pathways driving the connectivity pattern with the mentioned proteins, which can be categorized into intracellular and extracellular signalling, immune response and peptide signalling. Given the fact that the statistical approaches were unable to yield a clear separation between the sample types being included in the analyses (CLL and MM subtypes and control individuals), the identified pair-wise associations can be viewed as altered co-expression patterns that are commonly present in the three observation types.

6.4.1 Technical Assessment and Comparison of Integrative Approaches

The calibration process necessary to optimize a PLS model is an essential procedure for the successful application of this statistical method as aspects such as predictive accuracy, feature selection and interpretability depend on it. Despite the associated computational challenges, a through calibration procedure was conducted in this chapter to define the optimal models for the three integrative approaches and for both the predictor and outcome matrices and several observations can be drawn from this process. First, as revealed by the set of calibration curves, only marginal differences in error of prediction are observed for the different values of the model parameters being tested; a finding that is shared across the three PLS dimensions and across the three statistical methods. A possible explanation for this observation may be related to the high-dimensional nature of both the $\mathbf{X}$ and $\mathbf{Y}$ matrices and a comparatively restricted number of samples ($n <<< p+q$), which may lead to a more difficult identification of the relevant features driving the variation and to marginal differences in prediction error within dimensions. Second, for all regularized PLS models the predictive accuracy differences between components are greater to the ones observed within components, a statement that supports the exploration and optimization of models with more than one latent variable. Third, the calibration process of each approach clearly highlights which model parameters yield a better predictive accuracy, for example in sPLS this correspond to models retaining one to three proteins, in gPLS to models retaining one functional groups and in sgPLS lower values of the mixing

parameter α_1. Four, the choice to define the final model parameters invariably relies on subjective grounds as the observed minimum seldom provides an appropriate balance between statistical performance and sparsity.

In relation to the calibrated models, small differences in statistical performance can be seen across the three regularized PLS approaches. Similar to what has been demonstrated on simulated data where sPLS is outperformed by both gPLS and sgPLS, with the latter presenting the best statistical performance of the three [116], the same overall trend is viewed in this real data application; however, narrower variations in error of prediction are observed. Expectedly, marginal differences in prediction error for individual proteins are also detected across integrative approaches. As previously discussed, these differences may be limited in absolute value possibly due to the reduced number of components being explored, but they can still be significant in relative terms as they provide a clear pattern in relation to which approach provides the best model performance. Furthermore, as it was mentioned in chapter 5, the restricted improvement upon inclusion of pre-existing functional modules on the examined matrices can be explained by the grouping structure of the transcriptomics data (the majority of transcripts do not present group-annotation information, significant differences in group sizes and transcripts commonly belong to more than one pathway) as this is in attribute that only affects the gPLS and sgPLS methods. This aspect may also give explanation as to why the incorporation of the mixing parameter α_1 did not effectively impose within group sparsity in the optimized sgPLS model.

In terms of interpretability and sparsity of the optimized models, it can be observed that the sgPLS approach yields a more balanced output in comparison to those obtained from the sPLS and gPLS models as the number of features and modules selected in the predictor matrix was higher than in gPLS and lower than sPLS. Furthermore, and possibly because of the point just stated, the three main visualization tools (relevance networks, CIMs and correlation circle plots) obtained from the sgPLS model offer a compromise between those obtained from the other two approaches. These attributes facilitate the extraction of the most relevant biological features within

and between the two blocks of omics data. In contrast, the functional annotation analyses conducted on the selected gene expression signals are more informative in the case of the sPLS approach as the six selected enriched pathways appear to be more pertinent to the disease endpoint under study as opposed to the case of the group approaches were only one pathway of unspecific biological function was selected. In addition, the retained biological pathways of the sPLS model also appear to be more specific to the BCL phenotype and more aligned to the protein selection than the other integrative approaches. Thus, in terms of biological relevance the findings revealed here support a more significant contribution of the sparse statistical method over its counterparts.

As such, clear benefits and some challenging difficulties are associated with the assessed integrative methods in this real data application. Statistical performance, interpretability and sparsity are enhanced by the inclusion of pre-existing functional groups of variables with similar functions, while biological pertinence does not appear to be particularly improved.

6.4.2 Biological Relevance of Findings

Most of the selected gene expression signals and the corresponding enriched biological pathways relate to signal peptide and cell adhesion, which are molecules known to be implicated in the transportation of proteins from the cytoplasm into the extracellular space across the plasma membrane. It is an established fact in molecular biology that secretion of proteins into the extracellular space occurs when they are transported from the Endoplasmic Reticulum (ER) to the Golgi apparatus and subsequently to the plasma membrane via secretory vesicles or secretory granules, a process that depends on signal peptide-mediated translocations and molecules adhered to the plasma membrane. Once proteins are outside the cell, they exert defined extracellular functions [235]. As such, part of the biological output revealed in this chapter are indicative of this conventional protein secretion trafficking route by which the markers identified as relevant leave the cell to exert their tumour-induced angiogen-

esis and inflammatory functions.

Furthermore, it has been discovered that some growth factors and cytokines such as FGF2 and IL1b lack peptide signalling and their transportation is based on unconventional secretory pathways (i.e. ER/Golgi independent mechanisms). More specifically, in the case of the protein FGF2 there is some recent experimental evidence suggesting that the formation of disulfide bond bridges drive plasma membrane pore formation allowing the translocation of FGF2 to the cell surface [236], [237]. Judging by the findings from the functional annotation analyses and considering that FGF2 was found as the most relevant marker, the results presented in this chapter are also supportive of this unconventional mechanism for protein secretion.

Another relationship of biological significance identified in this chapter is the one established by the chemokines MCP3 and MIP1b. As discussed in chapter 5, these are proteins known to play a major role in the regulation of inflammatory processes by selectively recruiting White Blood Cells (WBCs), particularly monocytes and Natural Killer (NK) cells. Migration of these leukocyte types from the blood stream across the vascular endothelium is required for routine immunological surveillance of tissues, response to inflammation (recruitment of leukocytes to sites where an inflammation response is required) and proper functioning of the tumour microenvironment [232], [238]. More specifically, it has been described that the chemokines MCP3 and MIP1b bind to specific cell surface transmembrane receptors whose activation leads to the activation of intracellular signalling cascades that prompt migration of WBCs towards the chemokine source [232],[238]. These plasma membrane receptors are known as transmembrane domain G protein coupled receptors and the signalling cascades activated as consequence of the binding chemokine-chemokine receptors have been described to modulate not only cellular migration but also other functions such as cell survival, adhesion, invasion and proliferation. A significant group of biological pathways uncovered by the analyses conducted here are precisely related to those cellular mechanisms.

Altogether, the identified transcripts and proteins point towards two main biological

functions and the specific mechanism by which those functions are exerted, namely tumour-induced angiogenesis and inflammation. The extent to which these co-expression patterns reflect altered mechanisms leading to clinical manifestations and disease progression or biological processes occurring in circumstances where disease status is yet present remains to be investigated and constitute an unanswered interrogation opening avenues for future analyses. In particular, considering that the inflammatory markers identified in this integrative setting were also found as the most important proteins in the classification analysis for the MM subtype, it is possible that the findings discovered in this chapter are mainly driven by this specific sub-entity. Therefore, one immediate analytical step to be taken following what was revealed and discussed here is to restrict the cross-omics profiling to MM observations to confirm or refute this hypothesis.

6.5 Conclusion

The integration of the transcriptomics and proteomics EGM data by application of regularized PLS methods uncovered valuable co-expression patterns describing the interplay of the involved biomolecules and their functions. The results establish a relationship between the proteins FGF2, VEGF, TGFa, MCP3 and MIP1b and transcripts related to biological pathways such as to intracellular and extracellular signalling, cell adhesion and migration as well as peptide signalling. Such patterns appear to provide insights into the functioning of two complex biological process, namely tumour-induced angiogenesis and inflammation, which are mechanisms known to play a key role in the pathogenesis of BCL. From a methodological standpoint, most of the assessed aspects point towards a better performance of the sgPLS method over the other two integrative approaches.

7

Conclusions

Encouraged by the challenges associated to the analyses of high-throughput omics technologies and the role of environmental agents in the development human disease under the concept of the exposome, this thesis contributes to the scientific literature in the field by addressing two important research objectives: i) to assess and contrast univariate and multivariate statistical approaches on real-world omics data under different circumstances in terms of aspects such as applicability, suitability and interpretability and biological relevance of findings and ii) to identify possible biological markers indicative of B-cell Lymphoma (BCL) risk and potential underlying mechanisms leading to its onset. These two main aims were simultaneously addressed in three results chapters where the mentioned statistical approaches where applied with a focus on sample classification and omics data integration.

7.1 Summary of Findings

In chapter 4 I independently analyse proteomics and transcriptomics data from the EnviroGenoMarkers (EGM) project by means of univariate approaches to discover markers for BCL and its main histological subtypes. More specifically, considering that omics platforms are susceptible to unwanted sources of variation, I employ Linear Mixed Model (LMM)s as these are recognized statistical methods to correct for potential technical-induced noise. Predictive disease markers were identified for the subtypes Multiple Myeloma (MM) and Chronic Lymphocytic Leukaemia

(CLL) where six inflammatory markers and 684 gene expression signals respectively demonstrated to be significantly associated with disease status. For each subtype, stronger associations were consistently replicated when assessing possible limitations associated to the epidemiological study design (stratification by study cohort and analytical phase) as well as the effect of confounding factors (correction for White Blood Cell (WBC) subpopulations). Furthermore, these stronger markers showed to be statistically significant more than six years before individuals were diagnosed, a discovery suggestive of long term sub-clinical perturbations. On the other hand, results demonstrate that LMMs were efficient to characterise and correct for the main sources of technical noise and as a result to increase statistical power to detect positive associations.

The research focus shifts towards multivariate statistical approaches in chapter 5 where regularized PLS techniques are employed in a discriminant analysis context for the identification of subset of biological features able to accurately separate case and control observations. In particular, the techniques sparse, group and sparse-group Partial Least Squares Discriminant Analysis (PLS-DA) were independently applied to the proteomics and transcriptomics datasets following the same methodology undertaken in the previous chapter in order to allow for a comparative assessment between the univariate and multivariate methods. As with established statistical approaches, relevant findings were detected for the disease subtypes MM and CLL in the inflammatory marker and gene expression datasets, respectively, however, several improvements over the output from the LMM analysis emerge. First, regularized PLS methods were able to find a set of proteins displaying an optimal classification performance for CLL observations. Second, when there is a biological variability of interest the PLS methods perform better to identify the most relevant features in the corresponding predictor matrix thus enhancing sparsity when needed. Third, analytical outputs proved to be less sensitive to possible confounding factors as correction for WBC differentials did not alter findings substantially. In addition, the application of these three PLS-DA methods under different circumstances sheds light on what

conditions improve statistical performance and feature selection. It is demonstrated that the aggrupation of individual features by the similarities of their biological function does strengthen efficacy when a clear distinction between biological modules is available.

Under a regression framework, omics data integration between gene expression signals and inflammatory markers is conducted in chapter 6 using the same regularized PLS methods employed in the previous chapter but in two-block scenario. An asymmetric relationship between datasets was assumed where the purposes were to detect transcript levels predictive of protein concentration and co-expression patterns between feature types. Analyses were limited to MM and CLL subtypes as these were the observations driving the main findings in the independent omics analyses, and their corresponding matched control individuals. The identified co-expression patterns uncover two main biological mechanisms between gene expression signals and inflammatory markers, one related to tumour-induced angiogenesis and another to inflammation; the growth factors FGF2, VEGF and TGFa are linked to the first while the chemokines MCP3 and MIP1b are linked to the second. The identified transcripts are related to enriched biological pathways such as peptide signalling, cell adhesion, survival and invasion which have been proposed as the possible mechanisms by which the mentioned proteins exert their functions. From a statistical perspective, these real-world data applications show that model performance, interpretability and sparsity are superior for the sgPLS integrative approach while biological relevance appear to be enhanced in the sPLS method as more transcripts were retained in the optimized model which favours the selection of more informative enriched pathways from the functional annotation analyses.

7.2 Contribution to the State of the Art

The outputs of this thesis extend beyond the scope of each chapter and are expected to add value to the current scientific knowledge both from an epidemiological and a

statistical perspective. On the one hand, the characteristics of the EGM project allow for a valuable way of studying BCL and its main histological subtypes as most of the existing literature on this disease refers to tumour or blood samples in individuals already undergoing clinical manifestations of the disease. The prospective nature of the epidemiological studies of the EGM project allows for the discovery of biological perturbations before the onset of the clinical presentation, which avoids some of the drawbacks associated with retrospective epidemiological study designs.

On the other hand, to the best of my knowledge and apart from some of the results presented in this thesis, there is no published material known to have employed the group regularized versions of PLS in real and high-dimensional data either in a discriminatory analysis context or in a integration setting. As discussed in chapter 2, the group and sparse-group PLS methods were introduced to accommodate a disadvantage previous statistical approaches were unable to address, namely incorporation of pre-existing information in the form of group of highly correlated variables. In this regard the analyses presented in this thesis represent a valuable contribution to the existing scientific literature as two novel statistical methods were applied under different set of circumstances allowing for a comprehensive assessment of their applicability and suitability. As mentioned in the section above, it was observed that accounting for correlation structures within omics blocks does indeed reinforce and refine the results. Furthermore, the exhaustive application of sparse PLS in this thesis also constitutes a significant contribution to the literature as the widespread use of this technique to analyse omics data either with a classification purposes or in a two-block context is limited, probably because simpler univariate approaches are preferred.

7.3 Limitations

As mentioned in chapter 3, BCL is a cancer type characterised by presenting a high degree of histological heterogeneity where a substantial number of subtypes have

even been described to present sub-entities with distinguishable clinical and biological marker attributes. Thus, the reduced sample size of the study populations included here may have hindered the possibility of detecting inflammatory markers or gene expression signals that are commonly altered across the main disease subtypes as well as in individual subtypes where the analyses failed to detect markers of prediction. Furthermore, the EGM project and related studies have commonly considered MM as a type of Non-Hodgkin's Lymphoma (NHL) because it shares the cellular origin with the most common BCL subtypes (i.e. from the Germinal Centre (GC)); however, standard ICD-10 classification does not categorize it as such. As a plasma cell neoplasm, its main clinical manifestations (e.g. multiple bone lesions, production of monoclonal antibodies, hypercalcemia) differ to those seen in CLL, DLBCL and FL where painless lymphadenopathy is a common presentation across the three disorders. Considering that MM observations are the most numerous samples among all subtypes both in the proteomics and transcriptomics datasets, it is possible that the discussed findings are influenced to a certain extend by the consideration made about this plasma cell disorder.

From a statistical standpoint, it is reasonable to state that the methodologies investigated in this thesis were all applied in a thorough and exhaustive fashion. The potential drawback of multivariate approaches like PLS of them being unable to correct for technical-induced noise was overcome by means of the two-step procedure conducted in chapter 4. The other known limitation of the regularized versions of PLS is the high computational cost associated to the calibration procedure to define the model parameters. Performing this process constituted one of the biggest challenges of this PhD project, especially in a stratified analysis where statistical models are repeated over observation types (chapter 5) or in a multivariate setting (chapter 6) where the optimization of up to four different parameters was required. Although the authors of the PLS techniques employed in this thesis support the definition of the corresponding tuning parameters entirely based on subjective grounds (i.e. circumventing the calibration process), such approach may not be adequate when prior

biological knowledge is lacking or when novel findings and hypothesis-generating results are sought.

7.4 Future directions

The findings described in this thesis may serve as a foundation for further developments within the research area of statistical modelling of high-dimensional omics data for an enhancement of the biological understanding of multifactorial and complex diseases. First, as it is standard practice in epidemiological research, external validation of the observed findings in an independent population is required before they can be reliably employed in extended settings. This aspect is especially interesting for MM as the discovered biological markers seem to convert the expression levels from low to high levels when clinical presentation of the disease begins. Second, although explored in this thesis, a more robust and in-depth analysis in relation to Time to Diagnosis (TtD) is desired for a further exploitation of the prospective nature of the epidemiological study design; for example, stratification of observations per year of TtD or identifying inflammatory markers, gene expression signals or co-expression patterns predictive of TtD.

As previously discussed, PLS and its regularized extensions are not exempt of weakness, which can potentially open avenues for future work. The limitation related to their inability to accommodate categorical variables either as predictor or outcome features are commonly overcome in a two-step procedure like the one employed here: estimate the effect of a categorical variable by means of univariate methods and construct a new matrix from the residuals to which the PLS techniques are applied. However, this procedure considers the categorical variable(s) independently and their contribution cannot be measured and compared in relation to rest of the omics features. Undergoing research efforts are attempting to reduce this two-step method into one and thus to allow the modelling of categorical variables in the same manner as continuous features. Furthermore, the trade-off between orthogonality

and sparsity in the creation of the latent variables in regularized PLS methods especially in a regression setting has not been fully addressed yet, which is a factor that may have important subsequent implications in aspects such as model construction, complexity, statistical performance and ultimately biological significance.

This page intentionally left blank.

Bibliography

[1] Sophie C. Joosten, Kim M. Smits, Maureen J. Aarts, Veerle Melotte, Alexander Koch, Vivianne C. Tjan-Heijnen, and Manon Van Engeland. "Epigenetics in renal cell cancer: Mechanisms and clinical applications". *Nature Reviews Urology* **15**.7 (2018), 430–451.

[2] T. Hesman Saey. "A recount of human genes ups the number to at least 46,831". *ScienceNews* **184**.7 (2018), 5.

[3] J. Alles et al. "An estimate of the total number of true human miRNAs". *Nucleic Acids Research* **47**.7 (2019), 3353–3364.

[4] I. Ezkurdia, D. Juan, J.M. Rodriguez, A. Frankish, M. Diekhans, J. Harrow, J. Vasquez, A. Valencia, and M.L. Tress. "Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes". *Human Molecular Genetics* **23**.22 (2014), 5866–5878.

[5] International Human Genome Sequencing Consortium. "Initial sequencing and analysis of the human genome". *Nature* **409**.6822 (2001), 860–921.

[6] The 1000 Genomes Project Consortium. "A global reference for human genetic variation". *Nature* **526**.7571 (2015), 68–74.

[7] M. Kimura. *The neutral theory of molecular evolution*. Cambridge University Press, 1968.

[8] L. Kruglyak and D.A. Nickerson. "Variation is the spice of life." *Nature genetics* **27**.3 (2001), 234–236.

[9] M. Cargill et al. "Characterization of single-nucleotide polymorphisms in coding regions of human genes". *Nature genetics* **23**.3 (1999), 231–238.

[10] R. Hunt, Z.E. Sauna, S.V. Ambudkar, M.M. Gottesman, and C. Kimchi-Sarfaty. "Silent (synonymous) SNPs: should we care about them?" *Methods in Molecular Biology* **578** (2009), 23–39.

[11] R Redon et al. "Global variation in copy number in the human genome". *Nature* **444** (2006), 444–454.

[12] S.A. McCarroll and D.M. Altshuler. "Copy-number variation and association studies of human disease." *Nature genetics* **39**.(7 Suppl) (2007), S37–42.

[13] A.D. Riggs, R.A. Martienssen, and V.E.A. Russo. *Introduction. Epigenetic Mechanisms of Gene Regulation*. New York: Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 1996, pp. 1–4.

[14] A.D. Riggs and T.N. Porter. *Overview of Epigenetic Mechanisms. Epigenetic Mechanisms of Gene Regulation*. 1996, pp. 29–45.

[15] M. Kulis and M. Esteller. "DNA methylation and cancer". *Advances in Genetics* **70**.10 (2010), 27–56.

[16] Michael Stevens et al. "Estimating absolute methylation levels at single-CpG resolution from methylation enrichment and restriction enzyme sequencing methods". *Genome Research* **23**.9 (2013), 1541–1553.

[17] Mehrnaz Fatemi, Martha M. Pao, Shinwu Jeong, Einav Nili Gal-Yam, Gerda Egger, D.J. Weisenberger, and Peter A. Jones. "Footprinting of mammalian promoters: use of a CpG DNA methyltransferase revealing nucleosome positions at a single molecule level". *Nucleic Acids Research* **33**.20 (2005).

[18] Rafael A Irizarry et al. "The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores". *Nature Genetics* **41** (2009), 178–186.

[19] Matladi N. Ndlovu, Hélène Denis, and François Fuks. "Exposing the DNA methylome iceberg". *Trends in Biochemical Sciences* **36**.7 (2011), 381–387.

[20] Christoph Bock, Eleni M Tomazou, Arie B Brinkman, Fabian Müller, Femke Simmer, Hongcang Gu, Natalie Jäger, Andreas Gnirke, Hendrik G Stunnenberg, and Alexander Meissner. "Quantitative comparison of genome-wide DNA methylation mapping technologies". *Nature Biotechnology* **28** (2010), 1106–1114.

[21] P.W. Laird. "Principles and challenges of genome-wide DNA methylation analysis'". *Nature Review Genetics* **11**.3 (2010), 191–203.

[22] K Luger, AW Mäder, RK Richmond, DF Sargent, and TJ Richmond. "Crystal structure of the nucleosome core particle at 2.8 A resolution." *Nature* **389**.6648 (1997), 251–260.

[23] T Jenuwein and CD Allis. “Translating the histone code”. *Science* **293**.5532 (2001), 1074–1080.

[24] M Grunstein. “Histone acetylation in chromatin structure and transcription”. *Nature* **389**.6649 (1997), 349–352.

[25] Brian D Strahl and C David Allis. “2000 - Modificaciones Covalentes De Las Histonas.Pdf”. **403**.January (2000), 41–45.

[26] Tony Kouzarides. “Chromatin Modifications and Their Function”. *Cell* **128**.4 (2007), 693–705.

[27] AJ Bannister and T Kouzarides. “Regulation of chromatin by histone modifications”. *Cell Research* **21**.3 (2011), 381–395.

[28] A Jacquier. “The complex eukaryotic transcriptome: Unexpected pervasive transcription and novel small RNAs”. *Nature Review Genetics* **10**.12 (2009), 833–844.

[29] SA Byron, DM Van Keuren-Jensen KR, Engelthaler, JD Carpten, and DW Craig. “Translating RNA sequencing into clinical diagnostics: Opportunities and challenges”. *Nature Review Genetics* **17**.5 (2016), 257–271.

[30] Amelia Casamassimi, Antonio Federico, Monica Rienzo, Sabrina Esposito, and Alfredo Ciccodicola. “Transcriptome profiling in human diseases: New advances and perspectives”. *International Journal of Molecular Sciences* **18**.8 (2017), 1652.

[31] JS Mattick. “The central role of RNA in human development and cognition”. *FEBS Letters* **585**.11 (2011), 1600–1616.

[32] Valerio Costa, Claudia Angelini, Italia De Feis, and Alfredo Ciccodicola. “Uncovering the complexity of transcriptomes with RNA-Seq”. *Journal of Biomedicine and Biotechnology* **2010**.Article ID 853916 (2010), 19 pages.

[33] B.J. Blencowe. “Alternative splicing: New insights from global analyses”. *Cell* **126**.1 (2006), 37–47.

[34] M. Adams et al. “Complementary DNA sequencing: expressed sequence tags and human genome project”. *Science* **252**.5013 (1991), 1651–1656.

[35] MD Adams, AR Kerlavage, C Fields, and JC Venter. "3,400 new expressed sequence tags identify diversity of transcripts in human brain". *Nature Genetics* **4**.3 (1993), 256–297.

[36] Neil Shirley, Thomas Shafee, Stephen Dolan, Rohan Lowe, and Mark Bleackley. "Transcriptomics technologies". *PLOS Computational Biology* **13**.5 (2017), e1005457.

[37] Z Wang, M Gerstein, and M Snyder. "RNA-Seq: a revolutionary tool for transcriptomics". *Nature Review Genetics* **10**.1 (2009), 57–63.

[38] Z Su et al. "An investigation of biomarkers derived from legacy microarray data for their utility in the RNA-seq era". *Genome Biology* **15**.12 (2014), 523.

[39] Roman Jaksik, Marta Iwanaszko, Joanna Rzeszowska-Wolny, and Marek Kimmel. "Microarray experiments and factors which affect their reliability". *Biology Direct* **10**.1 (2015), 1–14.

[40] GA Held, G Grinstein, and Y Tu. "Relationship between gene expression and observed intensities in DNA microarrays - a modeling study". *Nucleic Acids Research* **34**.9 (2006), e70.

[41] Affymetrix. *GeneChip Expression Analysis - Technical Manual*. Tech. rep. 2009.

[42] Agilent. *User Manuals*. Tech. rep. 2015.

[43] M.F. Elshal and P. Jr. McCoy. "Multiplex Bead Array Assays: Performance Evaluation and Comparison of Sensitivity to ELISA". *Methods* **38**.4 (2006), 317–323.

[44] Patrick J. Tighe, Richard R. Ryder, Ian Todd, and Lucy C. Fairclough. "ELISA in the multiplex era: Potentials and pitfalls". *Proteomics - Clinical Applications* **9**.3-4 (2015), 406–422.

[45] Peter Mitchell. "A perspective on protein microarray". *Nature biotechnology* **20**.3 (2002), 225–229.

[46] David A. Hall, Jason Ptacek, and Michael Snyder. "Protein Microarray Technology". *Mechanisms of Ageing and Development* **128**.1 (2007), 161–167.

[47] John R Yates, Cristian I Ruse, and Aleksey Nakorchevsky. "Proteomics by mass spectrometry: approaches, advances, and applications." *Annual review of biomedical engineering* **11** (2009), 49–79.

[48] Timothy K. Toby, Luca Fornelli, and Neil L. Kelleher. “Progress in Top-Down Proteomics and the Analysis of Proteoforms”. *Annual review of biomedical engineering* **9**.1 (2016), 499–519.

[49] Fiehn O. “Metabolomics–the link between genotypes and phenotypes”. *Plant Molecular Biology* **48**.1-2 (2002), 155–171.

[50] J. K. Nicholson, J. C. Lindon, and E. Holmes. “’Metabonomics’: understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data”. *Xenobiotica* **29**.11 (1999), 1181–1189.

[51] Karan Uppal, Douglas I Walker, Ken Liu, Shuzhao Li, Young-Mi Go, and Dean P Jones. “Computational Metabolomics: A Framework for the Million Metabolome”. *Chemical Research in Toxicology* **19**.2912 (2016), 1956–1975.

[52] Warwick B. Dunn and David I. Ellis. “Metabolomics: Current analytical platforms and methodologies”. *Trends in Analytical Chemistry* **24**.4 (2005), 285–294.

[53] Roel Vermeulen, Douglas I. Walker, Gary W. Miller, Megan M. Niedzwiecki, Marc Chadeau-Hyam, and Dean P. Jones. “The Exposome: Molecules to Populations”. *Annual Review of Pharmacology and Toxicology* **59**.1 (2018), 107–127.

[54] Christopher Paul Wild. “Complementing the genome with an "exposome": The outstanding challenge of environmental exposure measurement in molecular epidemiology”. *Cancer Epidemiology Biomarkers and Prevention* **14**.8 (2005), 1847–1850.

[55] Gary W. Miller and Dean P. Jones. “The nature of nurture: Refining the definition of the exposome”. *Toxicological Sciences* **137**.1 (2014), 1–2.

[56] Stephen M. Rappaport and Martyn T. Smith. “Environment and Disease Risks”. *Science* **330**.6003 (2010), 460–461.

[57] Christopher Paul Wild. “The exposome: from concept to utility”. *International Journal of Epidemiology* **41**.1 (2012), 24–32.

[58] Paolo Vineis, Aneire E. Khan, Jelle Vlaanderen, and Roel Vermeulen. “The impact of new research technologies on our understanding of environmental causes of disease: The concept of clinical vulnerability”. *Environmental Health* **8**.1 (2009), 1–10.

[59] Paolo Vineis, Karin VanVeldhoven, Marc Chadeau-Hyam, and Toby J. Athersuch. "Advancing the Application of Omics-Based Biomarkers in Environmental Epidemiology". *Environmental and Molecular Mutagenesis* **54**.7 (2013), 461–467.

[60] Paolo Vineis. "Exposomics: Mathematics meets biology". *Mutagenesis* **30**.6 (2015), 719–722.

[61] Florence Guida et al. "Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation". *Human Molecular Genetics* **24**.8 (2015), 2349–2359.

[62] Federica Saletta, Giuseppe Matullo, Maurizio Manuguerra, Sabrina Arena, Alberto Bardelli, and Paolo Vineis. "Exposure to the tobacco smoke constituent 4-aminobiphenyl induces chromosomal instability in human cancer cells". *Cancer Research* **67**.15 (2007), 7088–7094.

[63] Helmut Schöllnberger, Niko Beerenwinkel, Rudolf Hoogenveen, and Paolo Vineis. "Cell Selection as Driving Force in Lung and Colon". *Cancer Research* **70**.17 (2010), 6797–6803.

[64] Paolo Vineis, Arthur Schatzkin, and John D. Potter. "Models of carcinogenesis: An overview". *Carcinogenesis* **31**.10 (2010), 1703–1709.

[65] Paolo Vineis and Frederica Perera. "Molecular epidemiology and biomarkers in etiologic cancer research: The new in light of the old". *Cancer Epidemiology Biomarkers and Prevention* **16**.10 (2007), 1954–1965.

[66] Iain M. Johnstone and D. Michael Titterington. "Statistical challenges of high-dimensional data". *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **367**.1906 (2009), 4237–4253.

[67] J.A. Nelder and R.W.M. Wedderburn. "Generalized linear models". *Journal of the Royal Statistical Society. Series A* **135**.3 (1972), 370–384.

[68] T. Hastie and T. Robert. "Generalized additive models". *Statistical Science* **1**.3 (1986), 297–310.

[69] T.J. Hastie and R.J. Tibshirani. *Generalized Additive Models*. Chapman, Hall/CRC. Monographs on Statistics, and Applied Probability 43, 1990.

[70] J. Johnson and J. Dinardo. *Econometrics methods*. 4th Editio. McGraw-Hill, 1997.

[71] Y. Benjamini and Y Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". *Journal of the Royal Statistical Society, Series B* **57**.1 (1995), 289–300.

[72] J.D. Storey. "A direct approach to false discovery rates John". *Journal of the Royal Statistical Society, Series B* **64**.3 (2002), 479–498.

[73] J.D. Storey. "The positive false discovery rate: a Bayesian interpretation and the q-value". *The Annals of Statistics* **31**.6 (2003), 2013–2035.

[74] Sture Holm. "Board of the Foundation of the Scandinavian Journal of Statistics A Simple Sequentially Rejective Multiple Test Procedure Author (s): Sture Holm Published by : Wiley on behalf of Board of the Foundation of the Scandinavian Journal of Statistics Stable U". *Scandinavian Journal of Statistics* **6**.2 (1978), 65–70.

[75] Y Hochberg. "A sharper Bonferroni procedure for multiple tests of significance". *Biometrica* **75**.4 (1988), 800–802.

[76] Y. Benjamini and D. Yekuteli. "The Control of the False Discovery Rate in Multiple Testing under Dependency". *The Annals of Statistics* **29**.4 (2001), 1165–1188.

[77] P.H. Westfall and S.S. Young. *Resampling-based multiple testing: examples and methods for p-value adjustment.* Wiley, 1993.

[78] Peter H. Westfall and James F. Troendle. "Multiple testing with minimal assumptions". *Biometrical Journal* **50**.5 (2008), 745–755.

[79] M. Chadeau-Hyam, G. Campanella, T. Jombart, L. Bottolo, L. Portengen, P. Vineis, B. Liquet, and RC.H. Vermeulen. "Deciphering the Complex: Methodological Overviewof Statistical Models to DeriveOMICS-Based Biomarkers". *Environmental and Molecular Mutagenesis* **54**.1 (2013), 542–557.

[80] A. Hoerl and R. Kennard. "Ridge Regression". *In Encyclopedia of Statistic Sciences* **8** (1988), 129–136.

[81] Robert Tibshirani. "Regression Shrinkage and Selection via the Lasso". *Journal of the Royal Statistical Society Series B* **58**.1 (1996), 267–288.

[82] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*. Vol. 103. 2013, pp. 175–184.

[83] M. Yuan and Y. Lin. "Model selection and estimation in regression with grouped variables". *Journal of the Royal Statistical Society Series B* **68**.1 (2006), 49–67.

[84] J. Friedman, T. Hastie, and R. Tibshirani. "Regularization Paths for Generalized Linear Models via Coordinate Descent". *Journal of Statistical Software* **33**.1 (2010), 1–22.

[85] J. Friedman, T. Hastie, and R. Tibshirani. "A note on the group lasso and a sparse group lasso" (2010), 1–8.

[86] N. Simon, J Friedman, T.J. Hastie, and R. Tibshirani. "A sparse-group lasso". *Journal of Computational and Graphical Statistics* **22**.2 (2013), 231–245.

[87] A. Puig, A. Wiesel, and A. Hero. "A multidimensional shrinkage-thresholding operator". In: *statistical signal processing, in 'SSP '09. IEEE/SP 15th Workshop on Statistical Signal Processing'*. Vol. 2122. 1. 2009, pp. 113–116.

[88] Y. Li, B. Nan, and J Zhu. "Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure." *Biometrics* **71**.2 (2015), 354–363.

[89] Ian T. Jollife. *Principal Component Analysis*. New York: Springer Verlag, 1986.

[90] J. Cadima and I.T. Jollife. "Loadings and correlations in the interpretation of principal components". *Journal of Applied Statistics* **22**.2 (1995), 203–214.

[91] Ian T. Jolliffe, Nickolay T. Trendafilov, and Mudassir Uddin. "A Modified Principal Component Technique Based on the LASSO". *Journal of Computational and Graphical Statistics* **12**.3 (2003), 531–547.

[92] Hui Zou, Hastie Trevor, and Tibshirani Robert. "Sparse Principal Component Analysis". *Pattern Recognition* **61**.2 (2017), 524–536.

[93] Qian Yang and Hongying Liu. "Sparse principal component analysis via regularized rank-k matrix approximation". *Beijing Hangkong Hangtian Daxue Xuebao/Journal of Beijing University of Aeronautics and Astronautics* **43**.6 (2017), 1239–1246.

[94] Daniela M. Witten, Robert Tibshirani, and Trevor Hastie. "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis". *Biostatistics* **10**.3 (2009), 515–534.

[95] H. Wold. "Estimation of principal components and related models by iterative least squares". *Multivariate Analysis* Academic Press, New York, Wiley. (1966), 391–420.

[96] H. Wold. "Partial Least Squares". *Encyclopedia of the Statistical Sciences* In Samuel Kotz and Norman L.Johnson, editors. Wiley. (1985), 581–591.

[97] F.L Bookstein, P.D. Sampson, A.P Streissguth, and H.M. Barr. "Exploiting redundant measurement of dose and developmental outcome: New methods from the behavioral teratology of alcohol". *Developmental Psychology* **32**.3 (1996), 404–415.

[98] F.J. Rohlf and M Corti. "Use of two-block partial least-squares to study covariation in shape". *Systematic Biology* **49**.4 (2000), 740–753.

[99] Anjali Krishnan, Lynne J. Williams, Anthony Randal McIntosh, and Hervé Abdi. "Partial Least Squares (PLS) methods for neuroimaging: A tutorial and review". *NeuroImage* **56**.2 (2011), 455–475.

[100] H. Abdi and L.J. Williams. "Partial Least Squares Methods: Partial Least Squares Correlation and Partial Least Square Regression". *Methods in Molecular Biology* **930** (2013), 549–579.

[101] J.A. Wegelin. *A survey of partial least squares (pls) methods, with emphasis on the two-block case. (Technical Report).* Tech. rep. University of Washington, 2000.

[102] V. Vinzi, L. Trinchera, and S. Amato. "Pls path modeling: from foundations to recent developments and open issues for model assessment and improvement". *Handbook of Partial Least Squares* (2010), 47–82.

[103] A.D. Cak, E.F. Moran, R. de O. Figueiredo, D. Lu, G. Li, and S. Hetrick. "Urbanization and small household agricultural land use choices in the brazilian amazon and the role for the water chemistry of small streams". *Journal of Land Use Science* **11**.2 (2016), 203–221.

[104] D.R. Hardoon, S Szedmak, and J Shawe-Taylor. "Canonical correlation analysis: an overview with application to learning methods". *Neural Computation* **16**.12 (2004), 2639–2664.

[105] G. Guo and G. Mu. "Joint estimation of age, gender and ethnicity: Cca vs. pls". In: *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. 2013, pp. 1–6.

[106] P. Geladi and B.R. Kowalski. "Partial least-squares regression: a tutorial". *Analytica Chimica Acta* **10**.9 (1986), 1–17.

[107] Svante Wold, Michael Sjöström, and Lennart Eriksson. "PLS-regression: a basic tool of chemometrics". *Chemometrics and Intelligent Laboratory Systems* **58**.2 (2001), 109–130.

[108] R. Rosipal and N Kramer. "Overview and recent advances in partial least squares". *Subspace, Latent Structure and Feature Selection: Statistical and Optimization Perspectives Workshop* (2006), 34–51.

[109] H. Hotelling. "Relations between two sets of variates". *Biometrika* **28**.3-4 (1936), 321–371.

[110] Pierre Lafaye de Micheaux, Benoit Liquet, and Matthew Sutton. "A Unified Parallel Algorithm for Regularized Group PLS Scalable to Big Data". **Preprint a** (2017), 1–20.

[111] K.-A. Le Cao, D. Rossouw, C. Robert-Granie, and P. Besse. "A Sparse PLS for Variable Selection when Integrating Omics Data". *Statistical Applications in Genetics and Molecular Biology* **7**.1 (2008), 35.

[112] S. Wold, H. Martens, and H. Wold. "The Multivariate Calibration Problem in Chemistry Solved by the PLS Method". *Conference Proceeding Matrix pencils* (1983), 286–293.

[113] S. Wold, A. Ruhe, H. Wold, and W.J. Dunn. "The collinearity problem in linear regression. The partial least squares (pls) approach to generalized inverses". *SIAM Journal on Scientificic and Statistical Computing* **5**.3 (1984), 735–743.

[114] S. De Jong. "SIMPLS: an alternative approach to partial least squares regression". *Chemometrics and intelligent laboratory systems* **18** (1993), 251–263.

[115] Hyonho Chun and Sündüz Kele. "Sparse partial least squares regression for simultaneous dimension reduction and variable selection". *J. R. Statist. Soc. B* **72**.1 (2010), 3–25.

[116] Benoît Liquet, Pierre Lafaye De Micheaux, Boris P. Hejblum, and Rodolphe Thiébaut. "Group and sparse group partial least square approaches applied in genomics context". *Bioinformatics* **32**.1 (2016), 35–42.

[117] Danh V. Nguyen and David M. Rocke. "Tumor classification by partial least squares using microarray gene expression data". *Bioinformatics* **18**.1 (2002), 39–50.

[118] Danh V Nguyen and David M Rocke. "Multi-class cancer classification via partial least squares with gene expression profiles". *Bioinformatics* **18**.9 (2002), 1216–1226.

[119] Richard G. Brereton and Gavin R. Lloyd. "Partial least squares discriminant analysis: Taking the magic away". *Journal of Chemometrics* **28**.4 (2014), 213–225.

[120] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. New York, 2001.

[121] Richard G. Brereton. *Chemometrics: Data Analysis for the Laboratory and Chemical Plant*. 2003.

[122] R.A. Fisher. "The use of multiple measurement in taxonomic problems". *Annals of Eugenics* **7** (1936), 179–188.

[123] K.V. Mardia, J.T. Kent, and J.M Bibby. "Multivariate Analysis". *Academic Press, London* (1979).

[124] P.C. Mahalanobis. "On the generalised distance in statistics". *In Proceedings National Institute of Science, India* **2**.1 (1936), 49–55.

[125] J.H. Friedman. "Regularized discriminant-analysis". *Journal of the American Statistical Association* **84**.405 (1987), 165–175.

[126] S. Dixon and B. Richard. "Comparison of performance of five common classifiers represented as boundary methods: Euclidean Distance to Centroids, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Learning Vector Quantization and Support Vector Machines, as dependent on". *Chemometrics and intelligent laboratory systems* **95**.1 (2009), 1–17.

[127] Anne Laure Boulesteix and Korbinian Strimmer. "Partial least squares: A versatile tool for the analysis of high-dimensional genomic data". *Briefings in Bioinformatics* **8**.1 (2007), 32–44.

[128] Kim Anh Lê Cao, Simon Boitard, and Philippe Besse. "Sparse PLS discriminant analysis: Biologically relevant feature selection and graphical displays for multiclass problems". *BMC Bioinformatics* **12**.1 (2011), 253.

[129] B. Liquet, Pi. L. De Micheaux, and C. Broc. *sgPLS: Sparse Group Partial Least Square Methods. R package version 1.7.* 2017.

[130] E. Riboli and R. Kaaks. "The EPIC Project: rationale and study design." *International Journal of Epidemiology* **26**.1 (1997), 6–14.

[131] Domenico Palli et al. "A Molecular Epidemiology Project on Diet and Cancer: The Epic-Italy Prospective Study. Design and Baseline Characteristics of Participants". *Tumori Journal* **89**.6 (2003), 586–593.

[132] E Riboli et al. "European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection". *Public Health Nutrition* **5**.6b (2002), 1113–1124.

[133] Göran Hallmans et al. "Cardiovascular disease and diabetes in the Northern Sweden Health and Disease Study Cohort- evaluation of risk factors and their interactions". *Scandinavian Journal of Public Health* **31**.61 (2003), 18–24.

[134] FD Groves, MS Linet, LB Travis, and SS Devesa. "Cancer Surveillance Series: Non-Hodgkin's Lymphoma Incidence by Histologic Subtype in the United States From 1978 Through 1995". *Journal of the National Cancer Institute* **92**.15 (2000), 1240–1251.

[135] WHO. *Tumours of Haematopoietic and Lymphoid Tissues*. World Health Organisation: Lyon, 2001.

[136] A. Fritz, C. Percy, A. Jack, K. Shanmugaratnam, L. Sobin, D.M. Parkin, and S. Whelan. *International Classification of Diseases for Oncology (ICD-O-3)*. Ed. by World Health Organization 2000. 3rd Editio. Geneva, Switzerland, 2000.

[137] Steven H. Swerdlow et al. "The 2016 revision of the World Health Organization classification of lymphoid neoplasms." *Blood* **127** (2016), 2375–2390.

[138] Katia Basso and Riccardo Dalla-Favera. "Germinal centres and B cell lymphomagenesis". *Nature Reviews Immunology* **15**.3 (2015), 172–184.

[139] J Ferlay, I Soerjomataram, M Ervik, R Dikshit, S Eser, C Mathers, M Rebelo, DM Parkin, D Forman, and F Bray. "GLOBOCAN 2012: Estimated Cancer Incidence, Mortality and Prevalence Worldwide in 2012 v1.0" (2012).

[140] Surveillance Epidemiology and End Results (SEER) Program. *National Cancer Institute, DCCPS, Cancer Statistics Branch*.

[141] Dai Chihara, Loretta J. Nastoupil, Jessica N. Williams, Paul Lee, Jean L. Koff, and Christopher R. Flowers. "New insights into the epidemiology of non-Hodgkin lymphoma and implications for therapy". *Expert Review of Anticancer Therapy* **15**.5 (2015), 531–544.

[142] Cancer-Research-UK. *Non-Hodgkin lymphoma incidence statistics*.

[143] Kate R Shankland, James O Armitage, and Barry W Hancock. "Non-Hodgkin lymphoma". *The Lancet* **380**.9844 (2012), 848–857.

[144] James O Armitage, Randy D Gascoyne, Matthew A Lunning, and Franco Cavalli. "Non-Hodgkin lymphoma". *The Lancet* **390**.10091 (2017), 298–310.

[145] P. Boffetta. "I. Epidemiology of adult non-Hodgkin lymphoma". *Annals of Oncology* **22**.4 (2011), 27–31.

[146] R Cartwright et al. "The rise in incidence of lymphomas in Europe 1985-1992." *European Journal of Cancer* **35**.4 (1999), 627–633.

[147] TR Holford, T Zheng, ST Mayne, and LA McKay. "Time trends of non-Hodgkin's lymphoma: are they real? What do they mean?" *Cancer Research* **52**.19 (1992), 5443s–5446s.

[148] A. Smith, D. Howell, R. Patmore, A. Jack, and E. Roman. "Incidence of haematological malignancy by sub-type: A report from the Haematological Malignancy Research Network". *British Journal of Cancer* **105**.11 (2011), 1684–1692.

[149] L Hardell and M Eriksson. "Is the decline of the increasing incidence of non-Hodgkin lymphoma in Sweden and other countries a result of cancer preventive measures?" *Environmental Health Perspectives* **111**.14 (2003), 1704–1706.

[150] Yawei Zhang et al. "Personal use of hair dye and the risk of certain subtypes of non-Hodgkin lymphoma". *American Journal of Epidemiology* **167**.11 (2008), 1321–1331.

[151] S. A M van de Schans, D. E. Issa, O. Visser, P. Nooijen, P. C. Huijgens, H. E. Karim-Kos, M. L G Janssen-Heijnen, and J. W W Coebergh. "Diverging trends in incidence and mortality, and improved survival of non-Hodgkin's lymphoma, in the Netherlands, 1989-2007". *Annals of Oncology* **23**.1 (2012), 171–182.

[152] A Carbone and A Gloghini. "AIDS-related lymphomas: from pathogenesis to pathology." *British Journal of Haematology* **130**.5 (2005), 662–670.

[153] L Kinlen. "Immunosuppressive therapy and acquired immunological disorders". *Cancer Research* **52**.19 (1992), 5474s–5476s.

[154] E Bayerdörffer, A Neubauer, B Rudolph, C Thiede, N Lehn, S Eidt, and M Stolte. "Regression of primary gastric lymphoma of mucosa-associate lymphoid tissue type after cure of Helicobacter pylori infection." *Lancet* **345**.8965 (1995), 1591–1594.

[155] TP Giordano, L Henderson, O Landgren, EY Chiao, JR Kramer, H El-Serag, and EA Engels. "Risk of non-Hodgkin lymphoma and lymphoproliferative precursor diseases in US veterans with hepatitis C virus". *JAMA* **297**.18 (2007), 2010–2017.

[156] S de Sanjose et al. "Hepatitis C and non-Hodgkin lymphoma among 4784 cases and 6269 controls from the International Lymphoma Epidemiology Consortium." *Clinical Gastroenterology and Hepatology* **6**.4 (2008), 451–458.

[157] C Colli, B Leinweber, R Müllegger, A Chott, H Kerl, and L Cerroni. "Borrelia burgdorferi-associated lymphocytoma cutis: clinicopathologic, immunophenotypic, and molecular study of 106 cases". *Journal of Cutaneous Pathology* **31**.3 (2004), 232–240.

[158] AJ Ferreri et al. "Chlamydophila psittaci eradication with doxycycline as first-line targeted therapy for ocular adnexae lymphoma: final results of an international phase II trial". *Journal of Clinical Oncology* **30**.24 (2012), 2988–2994.

[159] C Melenotte et al. "B-cell non-Hodgkin lymphoma linked to Coxiella burnetii". *Blood* **127**.1 (2016), 113–121.

[160] A Altieri, JL Bermejo, and K Hemminki. "Familial risk for non-Hodgkin lymphoma and other lymphoproliferative malignancies by histopathologic subtype: the Swedish Family-Cancer Database". *Blood* **106**.2 (2005), 668–672.

[161] SS Wang et al. "Family history of hematopoietic malignancies and risk of non-Hodgkin lymphoma (NHL): a pooled analysis of 10 211 cases and 11 905 controls from the International Lymphoma Epidemiology Consortium (InterLymph)". *Blood* **109**.8 (2007), 3479–3488.

[162] BA Bassig, Q Lan, N Rothman, Y Zhang, and T Zheng. "Current understanding of lifestyle and environmental factors and risk of non-hodgkin lymphoma: an epidemiological update". *Journal of Cancer Epidemiology* (2012), 2012:978930.

[163] A Nieters et al. "Smoking and lymphoma risk in the European prospective investigation into cancer and nutrition". *American Journal of Epidemiology* **167**.9 (2008), 1081–1089.

[164] JJ Castillo, RR Ingham, JL Reagan, M Furman, S Dalia, and J Mitri. "Obesity is associate with increase relative risk of diffuse large B-cell lymphoma: a meta-analysis of observational studies". *Clinical Lymphoma, Myeloma & Leukemia* **14**.2 (2014), 122–130.

[165] G Lenz and LM Staudt. "Aggressive lymphomas". *The New England Journal of Medicine* **362**.15 (2010), 1417–1429.

[166] Ralf Küppers. "Mechanisms of B-cell lymphoma pathogenesis". *Nature Reviews Cancer* **5**.4 (2005), 251–262.

[167] Marc Seifert, René Scholtysik, and Ralf Küppers. *Origin and Pathogenesis of B Cell Lymphomas*. 2019, pp. 1–18.

[168] Jay H Lubin, Joanne S Colt, David Camann, Scott Davis, James R Cerhan, Richard K Severson, Leslie Bernstein, and Patricia Hartge. "Epidemiologic evaluation of measurement data in the presence of detection limits". *Environmental Health Perspectives* **112**.17 (2004), 1691–1696.

[169] Dennie G A J Hebels et al. "Performance in omics analyses of blood samples in long-term storage: Opportunities for the exploitation of existing biobanks in environmental health research". *Environmental Health Perspectives* **121**.4 (2013), 480–487.

[170] Francesco Marabita et al. "An evaluation of analysis pipelines for DNA methylation profiling using the Illumina HumanMethylation450 BeadChip platform". *Epigenetics* **8**.3 (2013), 333–346.

[171] Jie Liu and Kimberly D. Siegmund. "An evaluation of processing methods for HumanMethylation450 BeadChip data". *BMC Genomics* **17**.1 (2016), 1–11.

[172] Yu Jia Shiah, Michael Fraser, Robert G. Bristow, and Paul C. Boutros. "Comparison of pre-processing methods for Infinium HumanMethylation450 BeadChip array". *Bioinformatics* **33**.20 (2017), 3151–3157.

[173] Claudia Sala, Pietro Di Lena, Danielle Fernandes Durso, Andrea Prodi, Gastone Castellani, and Christine Nardini. "Evaluation of pre-processing on the meta-analysis of DNA methylation data from the Illumina HumanMethylation450 BeadChip platform". *PLoS ONE* **15**.3 (2020), 1–15.

[174] Andrew E. Teschendorff, Charles E. Breeze, Shijie C. Zheng, and Stephan Beck. "A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies". *BMC Bioinformatics* **18**.105 (2017), 1511–4.

[175] EA Houseman, WP Accomando, DC Koestler, BC Christensen, CJ Marsit, HH Nelson, JK Wiencke, and KT Kelsey. "DNA methylation arrays as surrogate measures of cell mixture distribution." *BMC bioinformatics* **13**.86 (2012), 1–16.

[176] Alexander J. Titus, Rachel M. Gallimore, Lucas A. Salas, and Brock C. Christensen. "Cell-type deconvolution from DNA methylation: A review of recent applications". *Human Molecular Genetics* **26**.2 (2017), 216–224.

[177] Lovisa E. Reinius, Nathalie Acevedo, Maaike Joerink, Göran Pershagen, Sven Erik Dahlén, Dario Greco, Cilla Söderhäll, Annika Scheynius, and Juha Kere. "Differential DNA methylation in purified human blood cells: Implications for cell lineage and studies on disease susceptibility". *PLoS ONE* **7**.7 (2012), e41361.

[178] JT Leek, RB Scharpf, HC Bravo, D Simcha, B Langmead, WE Johnson, D Geman, K Baggerly, and RA Irizarry. "Tackling the widespread and critical impact of batch effects in high-throughput data". *Nature Review Genetics* **11**.10 (2010), 733–739.

[179] Almudena Espín-Pérez, Chris Portier, Marc Chadeau-Hyam, Karin van Veldhoven, Jos C.S. Kleinjans, and Theo M.C.M. de Kok. "Comparison of statistical methods and the use of quality control samples for batch effect correction in human transcriptome data". *PLoS ONE* **13**.8 (2018), 1–19.

[180] Cliona M McHale et al. "Global gene expression profiling of a population exposed to a range of benzene levels". *Environmental Health Perspectives* **119**.5 (2011), 628–634.

[181] M. Chadeau-Hyam et al. "Prediagnostic transcriptomic markers of Chronic lymphocytic leukemia reveal perturbations 10 years before diagnosis". *Annals of Oncology* **25**.5 (2014), 1065–1072.

[182] Roel Vermeulen et al. "Pre-diagnostic blood immune markers, incidence and progression of B-cell lymphoma and multiple myeloma: Univariate and functionally informed multivariate analyses". *International Journal of Cancer* **143**.6 (2018), 1335–1347.

[183] WE Johnson, C Li, and A Rabinovic. "Adjusting batch effects in microarray expression data using empirical Bayes methods". *Biostatistics* **8**.1 (2007), 118–127.

[184] Jeffrey T. Leek and John D. Storey. "Capturing heterogeneity in gene expression studies by surrogate variable analysis". *PLoS Genetics* **3**.9 (2007), 1724–1735.

[185] AE. Teschendorff, J Zhuang, and M Widschwendter. "Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies". *Bioinformatics* **27**.11 (2011), 1496–1505.

[186] S Franceschi et al. "Infection with Hepatitis B and C Viruses and Risk of Lymphoid Malignancies in the European Prospective Investigation into Cancer and Nutrition (EPIC)". *Cancer Epidemiology Biomarkers & Prevention* **20**.1 (2011), 208–214.

[187] Catharina M. Van Veldhoven et al. "Physical activity and lymphoid neoplasms in the European Prospective Investigation into Cancer and nutrition (EPIC)". *European Journal of Cancer* **47**.5 (2011), 748–760.

[188] M Chadeau-Hyam et al. "Prediagnostic transcriptomic markers of Chronic lymphocytic leukemia reveal perturbations 10 years before diagnosis." *Annals of Oncology* **25**.5 (2014), 1065–1072.

[189] Florian Rohart et al. "Supplemental Files An online resource for the Molecular Classification of Human Mesenchymal Stromal Cells". *PeerJ* **4** (2016), 1–28.

[190] F. Rohart, B. Gautier, A. Singh, and K-A. Lê Cao. "S1 Supplementary Information mixOmics: An R package for 'omics feature selection and multiple data integration". *PLoS Computational Biology* **13**.11 (2017), 1–14.

[191] F. Rohart, B. Gautier, A. Singh, and K-A. Lê Cao. "mixOmics: An R package for 'omics feature selection and multiple data integration". *PLoS Computational Biology* **13**.11 (2017), 1–14.

[192] Mevik Bjørn-Helge and Ron Wehrens. "The pls Package: Principal Component and Partial Least Squares Regression in R". *Journal of Statistical Software* **18**.2 (2007).

[193] Tenenhaus M. *La regression PLS: theorie et pratique.* Editions T. 1998.

[194] A. Umetri. *SIMCA-P for windows, Graphical Software for Multivariate Process Modeling.* Umea, Sweden, 1996.

[195] Johan A. Westerhuis, Huub C.J. Hoefsloot, Suzanne Smit, Daniel J. Vis, Age K. Smilde, Ewoud J.J. Velzen, John P.M. Duijnhoven, and Ferdi A. Dorsten. "Assessment of PLSDA cross validation". *Metabolomics* **4**.1 (2008), 81–89.

[196] Loong Chuen Lee, Choong Yeun Liong, and Abdul Aziz Jemain. "Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: A review of contemporary practice strategies and knowledge gaps". *Analyst* **143**.15 (2018), 3526–3539.

[197] Ewa Szymańska, Edoardo Saccenti, Age K. Smilde, and Johan A. Westerhuis. "Double-check: Validation of diagnostic statistics for PLS-DA models in metabolomics studies". *Metabolomics* **8**.1 (2012), 3–16.

[198] Tahir Mehmood, Kristian Hovde Liland, Lars Snipen, and Solve Sæbø. "A review of variable selection methods in Partial Least Squares Regression". *Chemometrics and Intelligent Laboratory Systems* **118** (2012), 62–69.

[199] L. Eriksson, E. Johansson, N. Kettaneh-Wold, and S. Wold. "Multi-and megavariate data analysis" ().

[200] Il Gyo Chong and Chi Hyuck Jun. "Performance of some variable selection methods when multicollinearity is present". *Chemometrics and Intelligent Laboratory Systems* **78**.1 (2005), 103–112.

[201] Ryan Gosselin, Denis Rodrigue, and Carl Duchesne. "A Bootstrap-VIP approach for selecting wavelength intervals in spectral imaging applications". *Chemometrics and Intelligent Laboratory Systems* **100**.1 (2010), 12–21.

[202] Kim Anh Lê Cao, Pascal G.P. Martin, Christèle Robert-Granié, and Philippe Besse. "Sparse canonical methods for biological data integration: Application to a cross-platform study". *BMC Bioinformatics* **10**.34 (2009), 1–17.

[203] Ignacio González, Kim Anh Lê Cao, Melissa J. Davis, and Sébastien Déjean. "Visualising associations between paired 'omics' data sets". *BioData Mining* **5**.1 (2012), 1–23.

[204] Neil Pearce. "Analysis of matched case-control studies". *BMJ (Online)* **352** (2016), 1–4.

[205] Chia-Ling Kuo, Yinghui Duan, and James Grady. "Unconditional or Conditional Logistic Regression Model for Age-Matched Case–Control Data?" *Frontiers in Public Health* **6**.March (2018), 6–8.

[206] DW Hosmer, S Lemeshow, and RX Sturdivant. "Applied Logistic Regression". In: 3rd. 2013. Chap. 4th, pp. 89 –153.

[207] Mathias Uhlén et al. "Tissue-based map of the human proteome". *Science* **347**.6220 (2015).

[208] Neha Korde, Sigurdur Y. Kristinsson, and Ola Landgren. "Monoclonal gammopathy of undetermined significance (MGUS) and smoldering multiple myeloma (SMM): Novel biological insights and development of early treatment strategies". *Blood* **117**.21 (2011), 5573–5581.

[209] RA Kyle, TM Therneau, SV Rajkumar, DR Larson, MF Plevak, JR Offord, A Dispenzieri, JA Katzmann, and LJ Melton. "Prevalence of monoclonal gammopathy of undetermined significance". *The New England Journal of Medicine* **354**.13 (2006), 1362–1369.

[210] Paolo Ghia and Federico Caligaris-Cappio. "Monoclonal B-cell lymphocytosis: Right track or red herring?" *Blood* **119**.19 (2012), 4358–4362.

[211] MC Lanasa et al. "Immunophenotypic and gene expression analysis of monoclonal B-cell lymphocytosis shows biologic characteristics associated with good prognosis CLL". *Leukemia* **25**.9 (2011), 1459–1466.

[212] M Gkotzamanidou, D Christoulas, VL Souliotis, A Papatheodorou, MA Dimopoulos, and E Terpos. "Angiogenic cytokines profile in smoldering multiple myeloma: No difference compared to MGUS but altered compared to symptomatic myeloma". *Medical Science Monitor* **19** (2013), 1188–1194.

[213] C Greco, G Vitelli, G Vercillo, R Vona, D Giannarelli, I Sperduti, F Pisani, E Capoluongo, MC Petti, and F Ameglio. "Reduction of serum IGF-I levels in patients affected with monoclonal gammopathies of undetermined significance or multiple myeloma. Comparison with bFGF, VEGF and K-ras gene mutation". *Journal of Experimental and Clinical Cancer Research* **28**.1 (2009), 28–35.

[214] AC Ng et al. "Bone microstructural changes revealed by high-resolution peripheral quantitative computed tomography imaging and elevated DKK1 and MIP-1α levels in patients with MGUS". *Blood* **118**.25 (2011), 6529–6534.

[215] V. Scudla, T. Pika, M. Budikova, J. Petrova, J. Minarik, J. Bacovsky, K. Langova, and Zivna J. "The importance of serum levels of selected biological parameters in the diagnosis, staging and prognosis of multiple myeloma". *Neoplasma* **57**.2 (2010), 102–110.

[216] D Pulte, MT Redaniel, H Brenner, L Jansen, and M Jeffreys. "Recent improvement in survival of patients with multiple myeloma: variation by ethnicity". *Leukemia & Lymphoma* **55**.5 (2014), 1083–1089.

[217] DF Jelinek, RC Tschumper, GA Stolovitzky, SJ Iturria, Y Tu, J Lepre, N Shah, and NE Kay. "Identification of a global gene expression signature of B-chronic lymphocytic leukemia." *Molecular Cancer Research* **1**.5 (2003), 346–361.

[218] Daruka Mahadevan et al. "Gene expression and serum cytokine profiling of low stage CLL identify WNT/PCP, Flt-3L/Flt-3 and CXCL9/CXCR3 as regulators of cell proliferation, survival and migration." *Human Genomics and Proteomics* 2009: 453634 (2009), doi: 10.4061/2009/453634.

[219] P Salven, A Orpana, L Teerenhovi, and H Joensuu. "Simultaneous elevation in the serum concentrations of the angiogenic growth factors VEGF and bFGF is an independent predictor of poor prognosis in non-Hodgkin lymphoma: a single-institution study of 200 patients." *Blood* **96**.12 (2000), 3712–3718.

[220] P Salven, L Teerenhovi, and H Joensuu. "A high pretreatment serum basic fibroblast growth factor concentration is an independent predictor of poor prognosis in non-Hodgkin's lymphoma." *Blood* **94**.10 (1999), 3334–3339.

[221] N Sato et al. "Elevated Level of Plasma Basic Fibroblast Growth Factor in Multiple Myeloma Correlates with Increased Disease Activity". *Japanese Journal of Cancer Research* **93** (2002), 459–466.

[222] O Sezer, C Jakob, J Eucker, K Niemöller, F Gatz, K Wernecke, and K Possinger. "Serum levels of the angiogenic cytokines basic fibroblast growth factor (bFGF), vascular endothelial growth factor (VEGF) and hepatocyte growth factor (HGF) in multiple myeloma." *European Journal of Haematology* **66**.2 (2001), 83–88.

[223] FT Wu, MO Stefanini, F Mac Gabhann, CD Kontos, BH Annex, and AS Popel. "A systems biology perspective on sVEGFR1: its biological function, pathogenic role and therapeutic use." *Journal of Cellular and Molecular Medicine* **14**.3 (2010), 528–552.

[224] E Balcan, F Demirkiran, Y Aydin, C Sanioglu, T Bese, M Arvas, T Akçay, and T Cift. "Serum levels of epidermal growth factor, transforming growth factor, and c-erbB2 in ovarian cancer". *International Journal of Gynecological Cancer* **22**.7 (2012), 1138–1142.

[225] IO Kara, B Sahin, R Gunesacar, and C Unsal. "Clinical significance of hepatocyte growth factor, plateletderived growth factor-AB, and transforming growth factor-alpha in bone marrow and peripheral blood of patients with multiple myeloma." *Advances in Therapy* **23**.4 (2006), 635–645.

[226] VL Pannain, JR Morais, and V Damasceno-Ribeiro, O Avancini-Alves. "Transforming growth factor α immunoreactivity. A study in hepatocellular carcinoma and in non-neoplastic liver tissue". *Annals of Hepatology* **11**.4 (2012), 495–499.

[227] Patrick De Boever, Britt Wens, Anyiawung Chiara Forcheh, Hans Reynders, Vera Nelen, Jos Kleinjans, Nicolas Van Larebeke, Geert Verbeke, Dirk Valkenborg, and Greet Schoeters. "Characterization of the peripheral blood transcriptome in a repeated measures design using a panel of healthy individuals". *Genomics* **103**.1 (2014), 31–39.

[228] HL Wong, RM Pfeiffer, TR Fears, R Vermeulen, S Ji, and CS Rabkin. "Reproducibility and correlations of multiplex cytokine levels in asymptomatic persons." *Cancer Epidemiology, Biomarkers & Prevention* **17**.12 (2008), 3450–3456.

[229] SL Navarro, TM Brasky, Y Schwarz, X Song, CY Wang, AR Kristal, M Kratz, E White, and JW Lampe. "Reliability of Serum Biomarkers of Inflammation from Repeated Measures in Healthy Individuals". *Cancer Epidemiology, Biomarkers & Prevention* **21**.7 (2012), 1167–70.

[230] Valentina Marchica et al. "Bone marrow CX3CL1/Fractalkine is a new player of the pro-angiogenic microenvironment in multiple myeloma patients". *Cancers* **11**.3 (2019).

[231] J Delgado-Calle et al. "Bidirectional Notch signaling and osteocyte-derived factors in the bone marrow microenvironment promote tumor cell proliferation and bone destruction in multiple myeloma". *Cancer Research* **76**.5 (2016), 1089–1100.

[232] YS Lee and YB Cho. *Tumor Microenvironment: The Role of Chemokines - Part A*. 2020, pp. 33–43.

[233] I. Vande Broek, K. Asosingh, K. Vanderkerken, N. Straetmans, B. Van Camp, and I. Van Riet. "Chemokine receptor CCR2 is expressed by human multiple myeloma cells and mediates migration to bone marrow stromal cell-produced monocyte chemotactic proteins MCP-1, -2 and -3". *British Journal of Cancer* **88**.6 (2003), 855–862.

[234] Nupur Bhattacharya et al. "Loss of cooperativity of secreted CD40L and increased dose-response to IL4 on CLL cell viability correlates with enhanced activation of NF-kB and STAT6". *International Journal of Cancer* **136**.1 (2015), 65–73.

[235] TA Rapoport. "Protein translocation across the eukaryotic endoplasmic reticulum and bacterial plasma membranes". *Nature* **450**.663-669 (2007).

[236] Giuseppe La Venuta, Marcel Zeitler, Julia P. Steringer, Hans Michael Müller, and Walter Nickel. "The startling properties of fibroblast growth factor 2: How to exit mammalian cells without a signal peptide at hand". *Journal of Biological Chemistry* **290**.45 (2015), 27015–27020.

[237] Hans Michael Müller et al. "Formation of disulfide bridges drives oligomerization, membrane pore formation, and translocation of fibroblast growth factor 2 to cell surfaces". *Journal of Biological Chemistry* **290**.14 (2015), 8925–8937.

[238] N Mukaida, S Sasaki, and T Baba. *CCL4 Signaling in the Tumor Microenvironment*. 2020, pp. 23–32.

[239] BM Bolstad, RA Irizarry, M. Astrand, and TP Speed. "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias". *Bioinformatics* **19**.2 (2003), 185–193.

[240] Pierre Comon. "Independent Component Analysis , a new concept?" *Signal Processing* **36**.94 (1994), 287–314.

[241] Aapo Hyvarinen, Erkki Oja, and Karhunen Juha. *Independent Component Analysis*. Wiley. New York, 2001.

[242] Sandra Waaijenborg, Philip C. Verselewel De Witt Hamer, and Aeilko H. Zwinderman. "Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis". *Statistical Applications in Genetics and Molecular Biology* **7**.1 (2008).

[243] Elena Parkhomenko, David Tritchler, and Joseph Beyene. "Sparse canonical correlation analysis with application to genomic data integration". *Statistical Applications in Genetics and Molecular Biology* **8**.1 (2009).

[244] Anastasia Lykou and Joe Whittaker. "Sparse CCA using a lasso with positivity constraints". *Computational Statistics and Data Analysis* **54**.12 (2010), 3144–3157.

[245] Endre Anderssen, Knut Dyrstad, Frank Westad, and Harald Martens. "Reducing overoptimism in variable selection by cross-model validation". *Chemometrics and Intelligent Laboratory Systems* **84**.1-2 (2006), 69–74.

[246] Richard G. Brereton. "Consequences of sample size, variable selection, and model validation and optimisation, for predicting classification ability from analytical data". *Trends in Analytical Chemistry* **25**.11 (2006), 1103–1111.

Appendices

A

Supplementary Material for Chapter 3

A.1 Omics Data Pre-processing Steps

LOcally WEighted Scatterplot Smoothing (LOESS) and A-quantile methods are within- and between-array normalization approaches (respectively) commonly applied to gene expression data from RNA microarray platforms. The former employs a locally weighted polynomial regression of the intensity scatterplot of each array in order to obtain a calibration factor while the later seeks to make the empirical distribution of probe intensities for each array in a set of arrays the same. In LOESS the smoothing curve is fitted to the M versus A plot, where M is the difference in log expression values and A is the average of the log expression values. On the other hand, the A-quantile method ensures that the A-values have the same distribution across arrays leaving the M-values unchanged [239].

As discussed in section 3.3, pre-processing steps are not as well-established for DNA methylation (DNAm) as for RNA microarray data [170], [171], [172], [173]. The procedure is hindered by several factors, mainly driven by the nature and the construction of the microarray chips which include: the use of two different probe types (Infinium I and Infinium II), the presence of non-specific probes (oligonucleotides that ambiguously map multiple genomic locations) and SNPs probes (HM450 arrays are based on Cytosine/Thymine SNPs introduced after bisulfite conversion, thus if a SNP is present it may be possible that the array measures a difference in genotype rather than a difference in DNAm).

A.2 Reference-based Deconvolution Algorithm

Broadly speaking, the Houseman algorithm estimates proportions of cell types present in the reference DNAm database using a technique known as linear constrained projection, a method that assumes that the methylation profile of a given sample is a weighted linear sum of the reference profiles present in the database. The inference of the weights proceeds by means of least-squares minimization subject to non-negativity and normalization constraints. That is, the sum of inferred weights must add to 1 or to a number that is less than or equal to 1 (in order to allow for the possibility that the reference database does not contain all relevant cell subtypes). The reference DNAm database employed for the analysis was made available by Reinius et al. [177].

A.3 Approaches for Batch Effect Removal

In this section I provide a brief description on three standard statistical methods for the removal of batch effects: Combatting Batch Effects (ComBat), SVA and ISVA. As discussed in section 3.5, the first approach assumes that the main sources responsible for the technical variation are known while the last two approaches model potentially unknown confounding variables. In the former case, the general formulation of the model is as follows:

$$\boldsymbol{y}_{ijg} = \boldsymbol{\beta}_{0g} + \mathbf{X}\boldsymbol{\beta}_g + \boldsymbol{\gamma}_{ig} + \boldsymbol{\delta}_{ig}\boldsymbol{\epsilon}_{ijg}, \tag{A.1}$$

where $\boldsymbol{y}_{ijg}$ represents the expression value for feature g for sample j from batch i, $\boldsymbol{\beta}_{0g}$ is the overall feature expression level, $\mathbf{X}$ is a design matrix for variables and covariates of interest, $\boldsymbol{\beta}_g$ is the vector of regression coefficients associated to $\mathbf{X}$ and $\boldsymbol{\epsilon}_{ijg}$ is the error term assumed to be normally distributed. The terms $\boldsymbol{\gamma}_{ig}$ and $\boldsymbol{\delta}_{ig}$ represent the additive and multiplicative batch effects of batch i for feature g, respectively. These batch effect parameters (also referred to as location and scale parameters representing

mean and variance) are estimated by the empirical Bayes method.

SVA or ISVA are more appropriate to account for technically induced variation [178] in situations where the true sources of technical variation are unknown or cannot be adequately identified or modelled. Briefly, the predictor matrix $\mathbf{X}_{ij}$, with i ($i = 1, \ldots, p$) labelling the predictor variables and j ($j = 1, \ldots, n$) labelling the samples, is modelled as a function of the vector $\boldsymbol{y} = (\boldsymbol{y}_1, ..., \boldsymbol{y}_n)$ representing the phenotype of interest

$$\mathbf{X}_{ij} = f_i(\boldsymbol{y}_j) + \boldsymbol{\epsilon}_{ij}, \tag{A.2}$$

from which the residual matrix is obtained

$$\mathbf{R}_{ij} = \mathbf{X}_{ij} - f_i \boldsymbol{y}_j, \tag{A.3}$$

This residual matrix is decomposed using SVD/PCA in the case of SVA or Independent Component Analysis (ICA) [240], [241] in the case of ISVA. The aim is then to identify subset of features driving this orthogonal residual variation which are used to construct (independent) surrogate variables; the significant (independent) surrogate variables can be included as covariates in subsequent regression analyses. The use of ICA allows the technical confounders to be uncorrelated in a non-linear fashion and thus to be modelled as statistically independent variables (a stronger condition than the linear uncorrelatedness imposed by an SVD/PCA in SVA).

A.4 Metrics of Performance for Canonical Mode

To the best of my knowledge, the overwhelming majority of literature in which Partial Least Squares (PLS) methods are applied to the analysis of high dimensional data focuses on regression and prediction where biological information indicates one omics data is expected to explain the other. However, when the purpose is to conduct an exploratory analysis where there is no assumption on the relationship between the two sets of variables or when a reciprocal relationship between the two data sets is

expected, the analysis relies on canonical correlation-based methods [202]. In that scenario, Canonical Correlation Analysis (CCA) and its regularized version (rCCA) are usually preferred over PLS canonical mode. These approaches are statistically difficult to assess mainly because they do not fit into a regression and prediction framework where the estimation of the prediction error is employed as a metric of performance to evaluate the quality of the model.

One adopted approach is to arbitrarily determine the model specifications: for example, number of dimensions $H = 3$ and number of selected variables per component to 100 [202]. A higher H makes the visual representation of the results more cumbersome while the selection size of 100 has been deemed as small enough to allow for the identification of individual relevant features and large enough to reveal important functional categories or pathways. After the value of the parameters has been pre-specified, the biological applications of the results are examined and the model parameters are modified accordingly (e.g. if a clustering effect is lost as the model is made more complex, the simpler model is preferred).

For CCA and its regularized versions, the choice of the optimal parameters has been primarily based on the estimation of either the correlation between latent variables or the proportion of variance in $\mathbf{X}$ and $\mathbf{Y}$ that can be explained by the latent variables. (In CCA the components are more commonly known as canonical variates). In the former case, a CV[1] procedure is conducted and the optimal parameters are chosen on the grounds of the computation of the test sample correlation; two criteria have been proposed that are calculated for different combinations of sparseness parameters, the first one being [242]:

$$\frac{1}{k}\sum_{j=1}^{k}\left|\left|Cor(\mathbf{X}_{-j}\boldsymbol{u}^{-j},\mathbf{Y}_{-j}\boldsymbol{v}^{-j})\right| - \left|Cor(\mathbf{X}_{j}\boldsymbol{u}^{-j},\mathbf{Y}_{j}\boldsymbol{v}^{-j})\right|\right|, \tag{A.4}$$

[1] As described in section 3.6.2.1.1, CV is normally considered as a procedure employed in regression (i.e. for prediction purposes); however, the authors proposing the metrics of performance for canonical mode discussed in this chapter also refer to CV as a method to estimate correlation or covariance. Therefore, in this thesis the concept CV is employed in both regression and canonical contexts.

where k is the number of folds while $\boldsymbol{u}^{-j}$ (resp. $\boldsymbol{v}^{-j}$) are the canonical loading vectors estimated from the training set $\mathbf{X}_{-j}$ (resp. $\mathbf{Y}_{-j}$) in which subset j was removed. Thus, this criterion seeks to minimize the mean difference between canonical correlation in the training and test sets. A simplified measure is to maximize the test sample correlation [243]:

$$\frac{1}{k}\sum_{j=1}^{k}\left|Cor(\mathbf{X}_j\boldsymbol{u}^{-j},\, \boldsymbol{Y}_j\boldsymbol{v}^{-j})\right| \tag{A.5}$$

On the other hand, the measure of explained variability in each data set is the diagnostic statistics employed to tune the optimal parameters in CCA and rCCA, which can be estimated in relation to the "same" or "opposite" canonical variates $(\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_h)$ and $(\boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_h)$ [244].The Redundacy (Rd) criterion, or part of explained variance by each component in relation with its associated data set, is computed as follows:

$$Rd(\mathbf{X}|\boldsymbol{\xi}_h) \;=\; \frac{tr\left[var(E(\mathbf{X}|\boldsymbol{\xi}_h))\right]}{tr\left[var(\mathbf{X})\right]} \;=\; \frac{1}{p}\sum_{j=1}^{p} cor^2(\boldsymbol{x}_j, \boldsymbol{\xi}_h) \tag{A.6}$$

where $cor^2(\boldsymbol{x}_j, \boldsymbol{\xi}_h)$ is the vector of the correlations between the h^{th} $\mathbf{X}$ variate and the j^{th} column of $\mathbf{X}$ and the term $E(\mathbf{X}|\boldsymbol{\xi}_h)$ denotes the regression coefficients from the linear regression of each column of $\mathbf{X}$ on $\boldsymbol{\xi}_h$. The corresponding proportions for the dataset $\mathbf{Y}$ in relation to $\boldsymbol{\omega}_h$ of are computed likewise. Similarly, Rd can be computed in relation to the "opposite" variate:

$$Rd(\mathbf{X}|\boldsymbol{\omega}_h) \;=\; \frac{tr\left[var(E(\mathbf{X}|\boldsymbol{\omega}_h))\right]}{tr\left[var(\mathbf{X})\right]} \tag{A.7}$$

Again, the corresponding proportions for the dataset $\mathbf{Y}$ in relation to $\boldsymbol{\xi}_h$ are computed likewise. Equation A.6 and Equation A.7 can be employed to determine the optimal number of dimensions in CCA and rCCA; however, when the aim is to introduce sparsity in the canonical loading vectors, these expressions can only be used to define the optimal tuning parameters for the first dimension. As discussed in section 2.2.2.4, the regularized versions of dimension reduction techniques lose the orthogonality property of the loading weights, meaning that information contained in dimension

$h+1$ is partially shared by dimension h. Thus, for dimensions $h = 2$ in rCCA, the Rd criteria to compute proportion of variability in the matrix $\mathbf{X}$ in relation to its "same" and "opposite" canonical variate is adapted as follows (equations for the $\mathbf{Y}$ matrix set follow likewise and are therefore omitted):

$$Rd(\mathbf{X}|\boldsymbol{\xi}_2) = \frac{tr\left[var(E(\mathbf{X} - E(\mathbf{X}|\boldsymbol{\xi}_1, \boldsymbol{\omega}_1)|\boldsymbol{\xi}_2])\right)}{tr\left[var(\mathbf{X})\right]} \tag{A.8}$$

$$Rd(\mathbf{X}|\boldsymbol{\omega}_2) = \frac{tr\left[var(E(\mathbf{X} - E(\mathbf{X}|\boldsymbol{\xi}_1, \boldsymbol{\omega}_1)|\boldsymbol{\omega}_2])\right)}{tr\left[var(\mathbf{X})\right]} \tag{A.9}$$

Following a similar principle, it has been suggested to extend the criteria discussed above for CCA to PLS canonical mode and base the selection of both the number of dimensions and the parsimony per dimension on the computation of the covariance between the $\mathbf{X}$ and $\mathbf{Y}$ scores (as opposed to the correlation), which is the optimization criterion maximized in two-block PLS $Cov(\mathbf{X}_{h-1}\boldsymbol{u}_h, \mathbf{Y}_{h-1}\boldsymbol{v}_h)$ [116]. It is worth highlighting that correlation is defined as
$Cor(\mathbf{X}_{h-1}\boldsymbol{u}_h, \mathbf{Y}_{h-1}\boldsymbol{v}_h) = Cov(\mathbf{X}_{h-1}\boldsymbol{u}_h, \mathbf{Y}_{h-1}\boldsymbol{v}_h)/\sqrt{Var(\mathbf{X}_{h-1}\boldsymbol{u}_h)}\sqrt{Var\mathbf{Y}_{h-1}\boldsymbol{v}_h}$; therefore, the aim of CCA is to maximize the covariance between $\mathbf{X}$ and $\mathbf{Y}$ variates while simultaneously minimizing the individual variances of the latent variables.

A.5 Cross-model Validation

It has been observed that for a CV procedure to give reliable error rate estimates, the complete optimization modelling process must be cross-validated [245], [246]. That is, the predicted observation should in no way be used in the development of the model. To circumvent this issue, Cross-Model Validation (CMV), also known as Double Cross-Validation (2CV) can be conducted [245]. As done in a single CV approach, CMV divides the complete data set into 2 parts, namely test set and rest set. The test set is set aside and repeated CV is performed using the rest set: the data is split into training and validation (sometimes called optimization) sets. This is the inner CV loop (CV1) and it is used to select the optimal value of a tuning parameter based on

the performance of a chosen diagnostic statistics. After the calibrated model has been defined, all observation in the rest set are used to fit a model with the optimal parameters and the performance is assessed by means of the test set that was initially left aside. This is referred as the outer CV loop (CV2). Typically, both CV1 and CV2 procedures are repeated until each sample has been included in the validation set (and test set) once and only once. Thus, CMV simultaneously optimize model complexity (CV1) while assessing final model quality (CV2). However, computational aspects and sample size restrict the wide application of this approach.

This page intentionally left blank.

B

Supplementary Material for Chapter 4

B.1 Overlapping and Improvements in Relation to Cited Papers

The analysis of gene expression data employing univariate statistical approaches was originally published in *"Prediagnostic transcriptomic markers of Chronic lymphocytic leukemia reveal perturbations 10 years before diagnosis"* [181] while the analysis of inflammatory markers data employing univariate statistical approaches was originally published in *"Pre-diagnostic blood immune markers, incidence and progression of B-cell lymphoma and multiple myeloma: Univariate and functionally informed multivariate analyses"* [182]. Therefore, there is a partial overlap between the results presented in chapter 4 and the ones published in these articles, which include the following:

- Assessment of significant associations between biological markers and disease status stratified by the major four histological subtypes using Linear Mixed Models (LMM)s (for both transcriptomics and proteomics).
- Predictive performance assessment of the statistically significant variables from the LMM by logistic regression conducted on the "de-noised" concentration levels (for both transcriptomics and proteomics).
- Time to Diagnosis (TtD) analysis stratified by median time elapsed between recruitment of cohort participants and clinical diagnosis (6 years) (transcriptomics

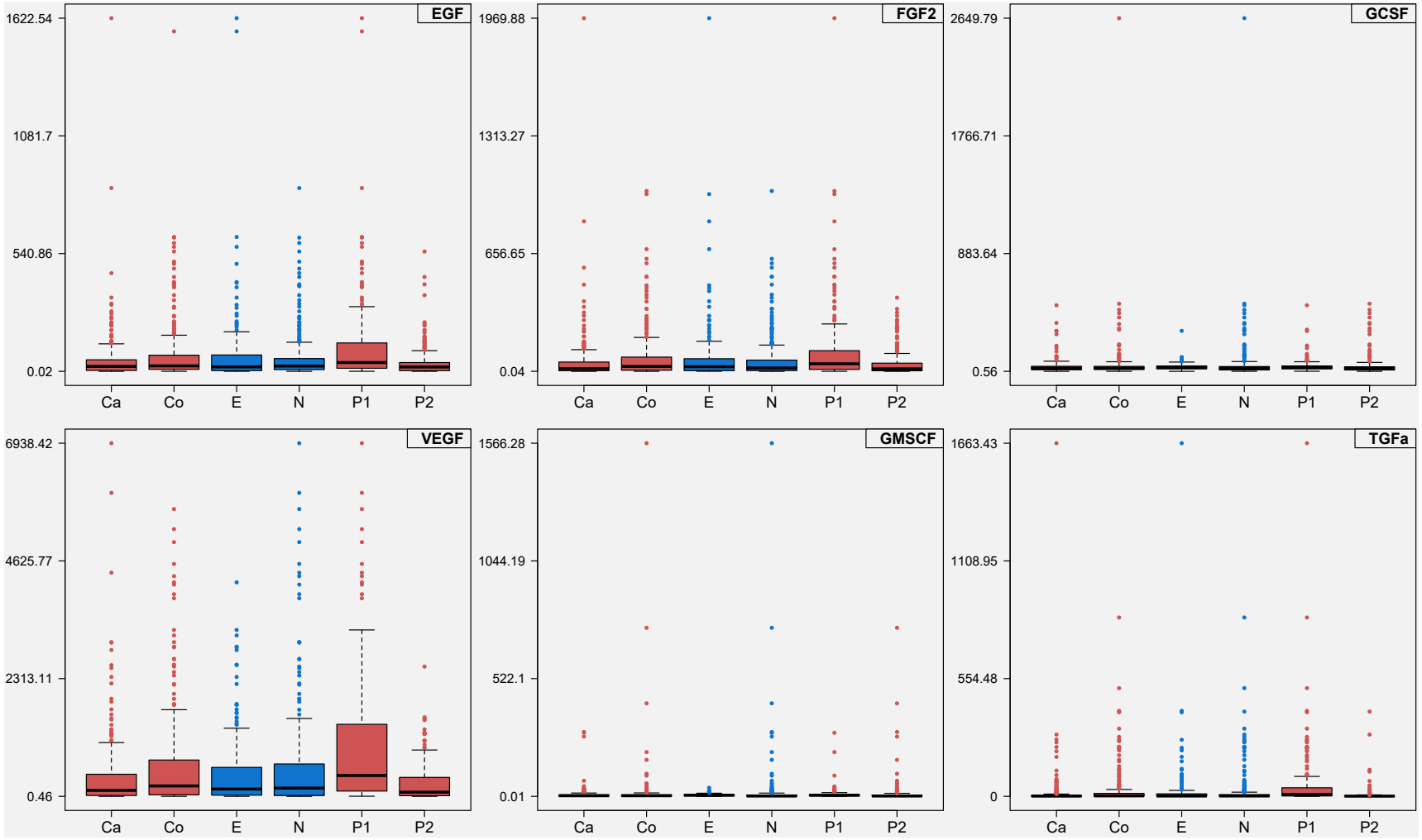
only).

- Functional annotation analyses using the openly available Database for Annotation, Visualization and Integrated Discovery (DAVID v6.8, http://david.abcc.ncifcrf.gov/) (transcriptomics only, it does not apply to proteomics).
- Study design validation and technical replication of the results in terms of epidemiological cohorts (EPIC-Italy and NSHDS) and experimental analytical phase (phase 1 and 2) (for both transcriptomics and proteomics).
- Assessment of technical-induced noise (transcriptomics only).
- Full adjustment for White Blood Cell (WBC) differentials in the LMMs (proteomics only).

Overall, the overlapping results between chapter 4 and the cited papers show small variations in the final outputs which may come down to differences in population sizes. The remaining of analyses presented in the chapter which were not listed above correspond to novel incorporation introduced in this thesis.

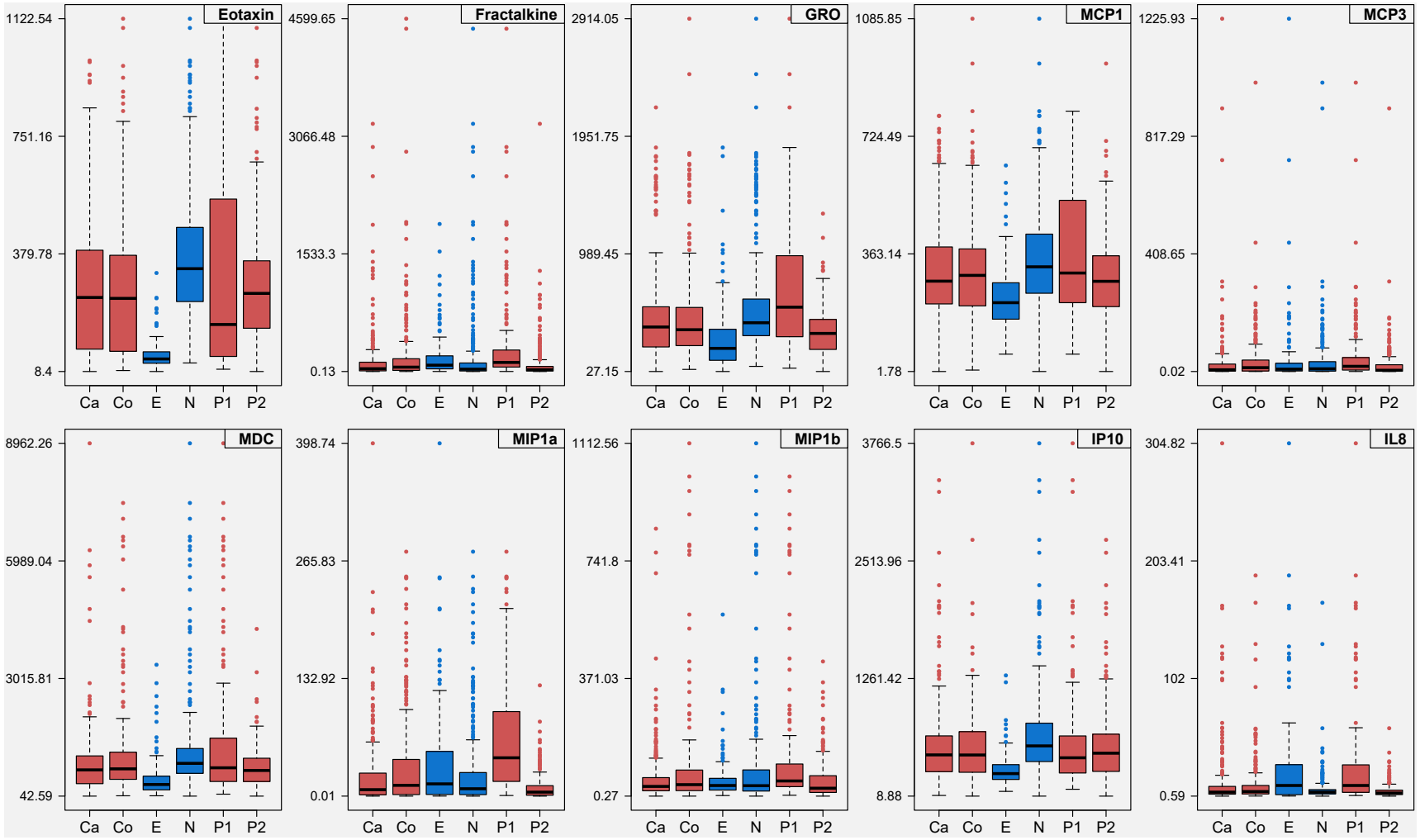
Supplementary Figures

Figure B.1: Box-and-whisker plot summarizing the concentration levels for inflammatory markers stratified by case-control status, study cohort and experimental phase for proteins belonging to the group growth factors (n=6).



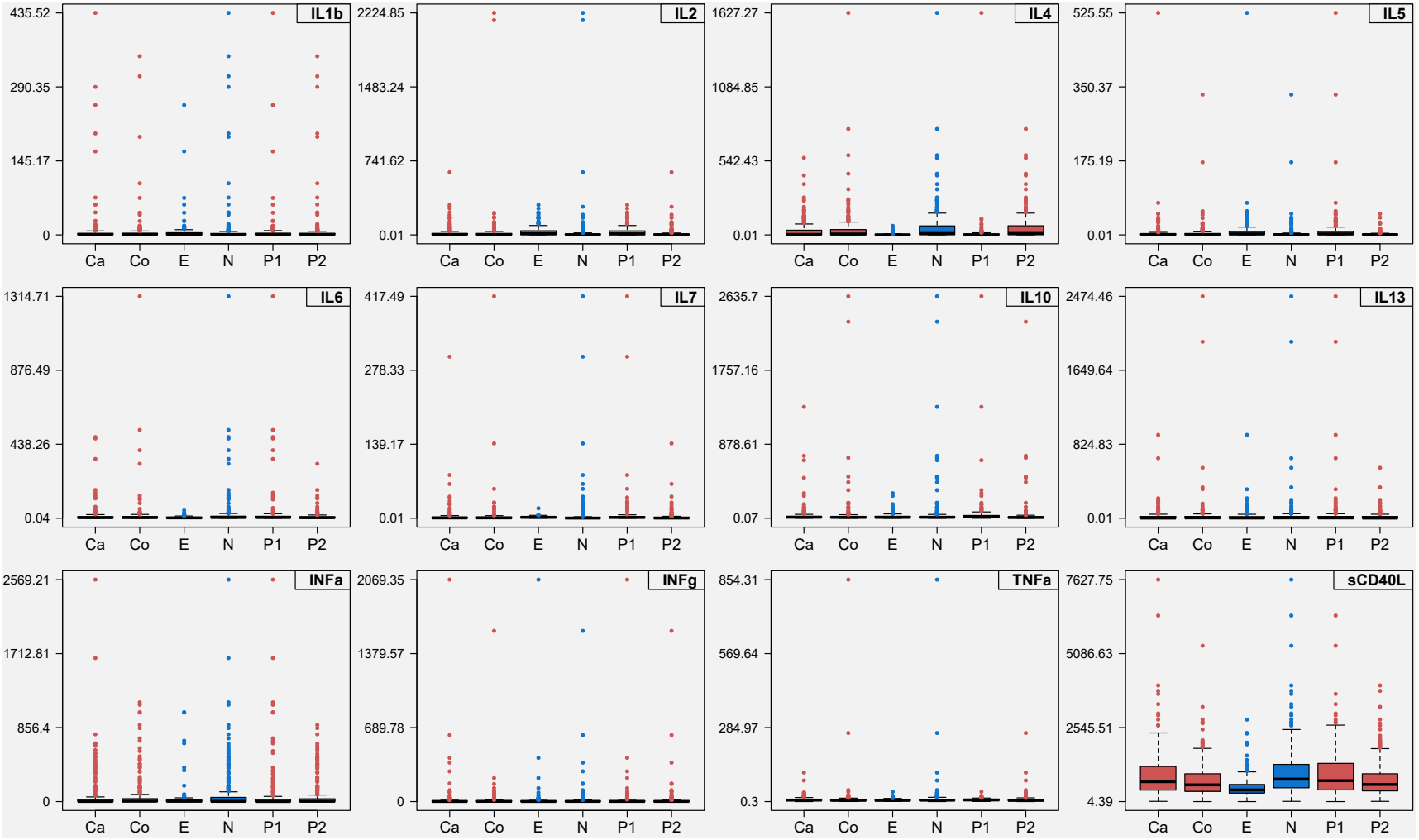
Ca:Cases, Co:Controls, E:EPIC: European Prospective Investigation into Cancer and Nutrition-Italy (EPIC-Italy), N:Northern Sweden Health and Disease Study (NSHDS), P1: Study Phase 1, P2: Study Phase 2.

Figure B.2: Box-and-whisker plot summarizing the concentration levels for inflammatory markers stratified by case-control status, study cohort and experimental phase for proteins belonging to the group chemokines (n=10).



Ca:Cases, Co:Controls, E:EPIC: European Prospective Investigation into Cancer and Nutrition-Italy (EPIC-Italy), N:Northern Sweden Health and Disease Study (NSHDS), P1: Study Phase 1, P2: Study Phase 2.

Figure B.3: Box-and-whisker plot summarizing the concentration levels for inflammatory markers stratified by case-control status, study cohort and experimental phase for proteins belonging to the group cytokines (n=12).



Ca:Cases, Co:Controls, E:EPIC: European Prospective Investigation into Cancer and Nutrition-Italy (EPIC-Italy), N:Northern Sweden Health and Disease Study (NSHDS), P1: Study Phase 1, P2: Study Phase 2.

Figure B.4: Histograms displaying the relative frequency distribution of the concentration levels of the 28 proteins under study before and after logarithmic transformation (top and bottom panel, respectively).

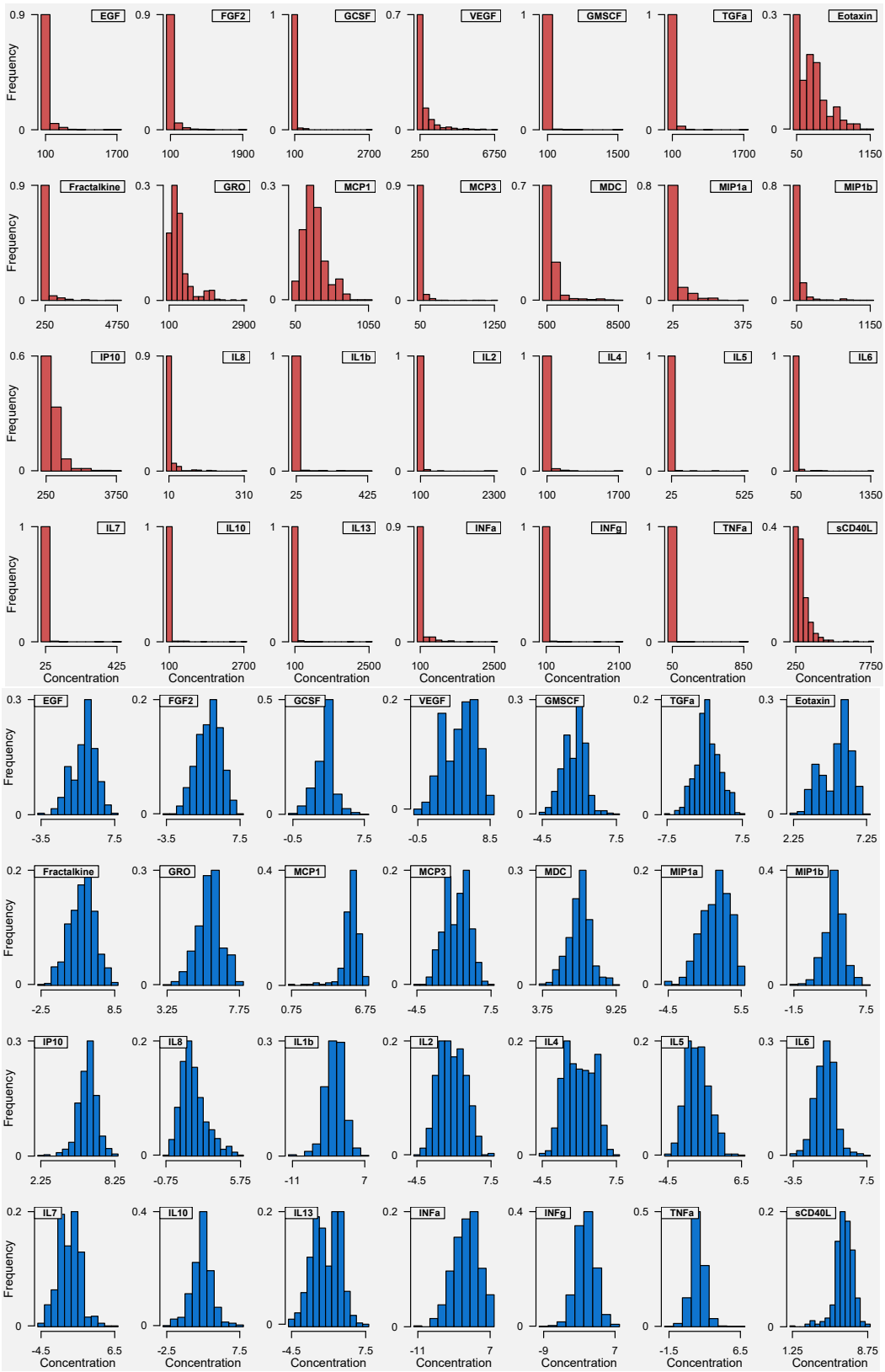
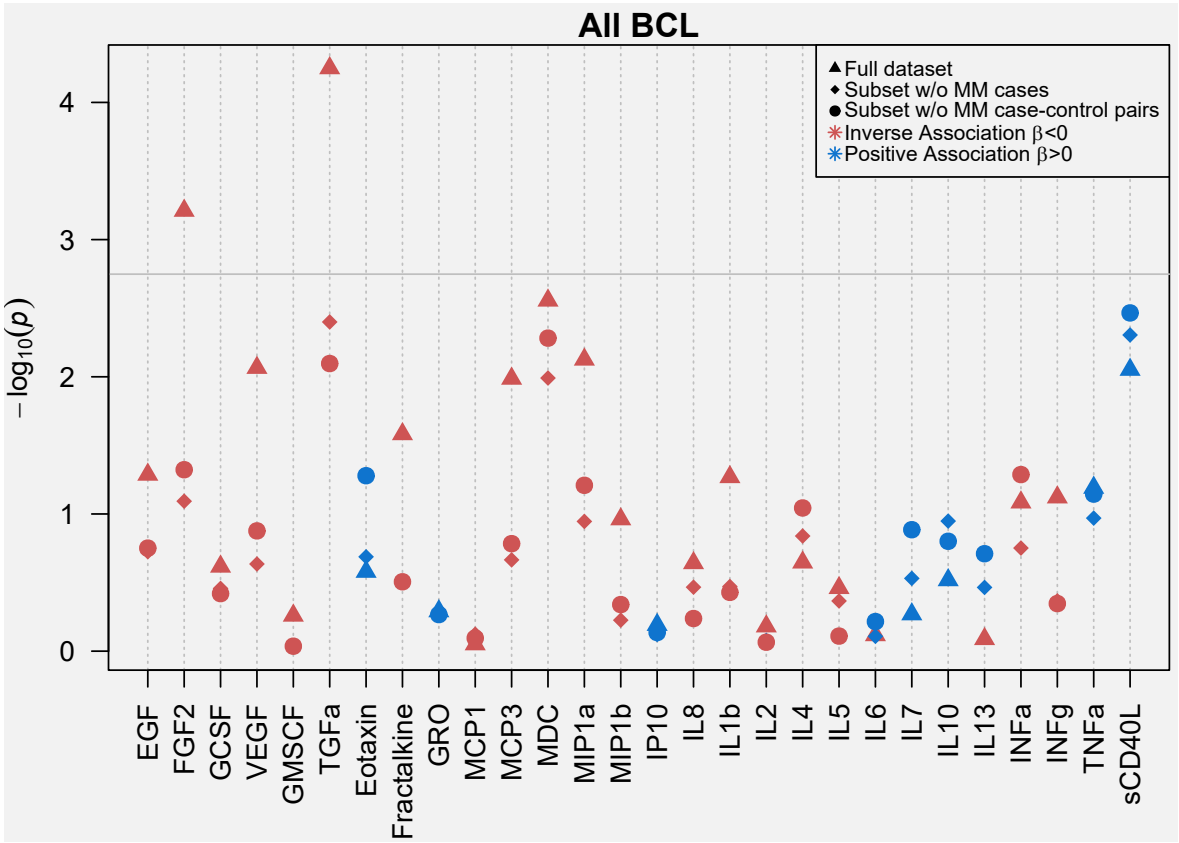
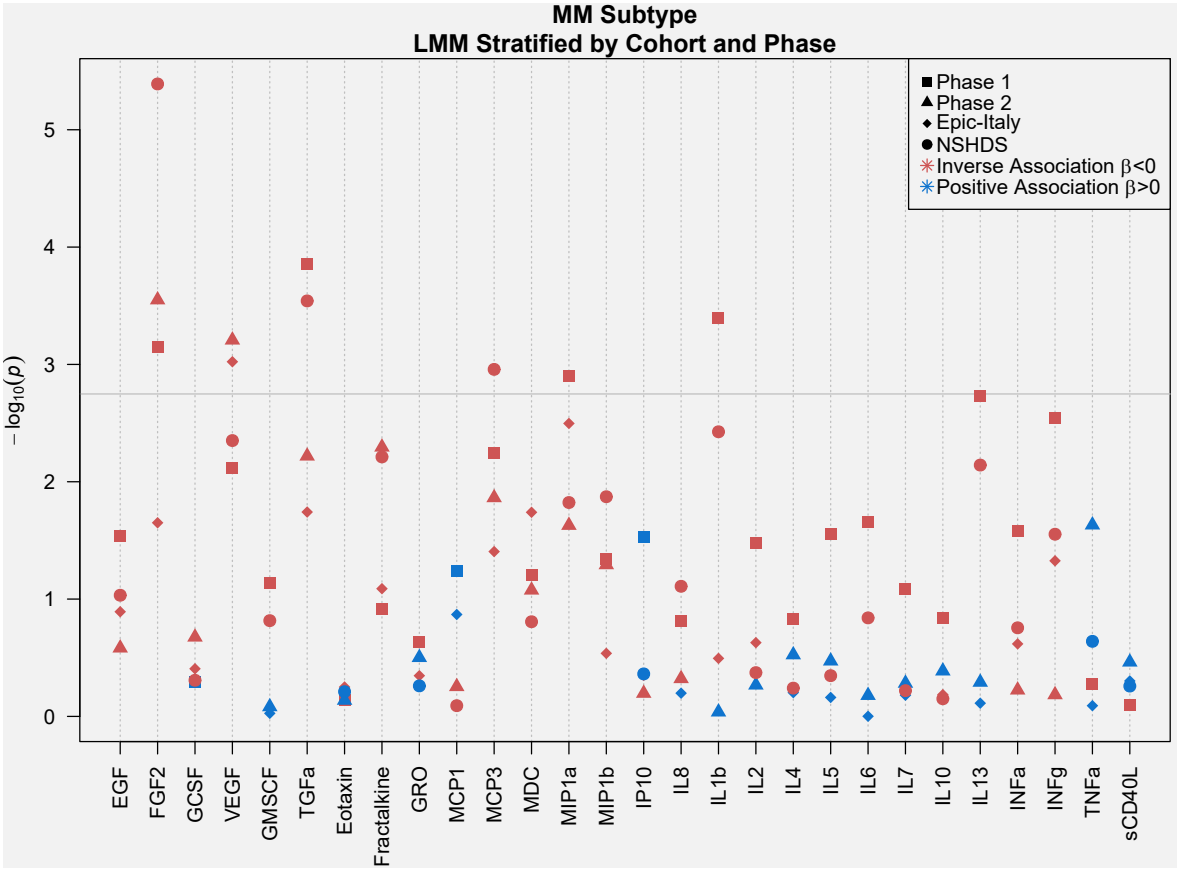


Figure B.5: Results of the LMM analyses between log-transformed values of proteins and BCL case-control status.



Results are displayed for BCL observations excluding MM subtype samples (76 cases and 152 case-control pairs). Strength of association (Y-axis) is measured by -$\log_{10}$ transformed p-values and the grey horizontal line represents the Bonferroni corrected per-test significance level ensuring a FWER control at 5%. Direction of the association is represented in red and blue for the negative and positive regression coefficients, respectively. Results are presented for the pooled BCL population (triangles), for the population subset excluding MM cases (diamonds) and for the population subset excluding MM cases-control pairs (circles). LMM: Linear Mixed Model.

Figure B.6: Results of the LMM analyses between log-transformed values of proteins and MM case–control status stratified by study cohort and analytical phase.



Strength of association (Y-axis) is measured by -$\log_{10}$ transformed p-values and the grey horizontal line represents the Bonferroni cut-off value ensuring a control of the FWER below 0.05. Direction of the association is represented in red and blue for the negative and positive regression coefficients, respectively. Results are presented for the pooled BCL population (triangles), for the population subset analysed in experimental phase 1 (square), for the population subset analysed in experimental phase 2 (triangle) for the population subset drawn from the Epic-Italy cohort (diamonds) and for the population subset drawn from the NSHDS cohort (circles). The corresponding number of cases and controls are 21 and 84 for Epic-Italy, 55 and 184 for NSHDS, 28 and 96 for phase 1 and 48 and 172 for phase 2.)
LMM: Linear Mixed Model, EPIC: European Prospective Investigation into Cancer and Nutrition, NSHDS: Northern Sweden Health and Disease Study.

Supplementary Tables

Table B.1: Distribution of BCL subtypes and genders across phases and cohorts with successfully analysed proteomics samples.

	Phase 1		*Phase 2*		*Total*	
Disease Subtype	**Epic-Italy n (%)**	**NHSDS n (%)**	**Epic-Italy n (%)**	**NHSDS n (%)**	**Epic-Italy n (%)**	**NHSDS n (%)**
CLL	5 (10)	7 (15.22)	6 (17.65)	24 (17.39)	11 (13.10)	31 (16.85)
DLBCL	10 (20)	9 (19.57)	1 (2.94)	24 (17.39)	11 (13.10)	33 (17.93)
FL	14 (28)	5 (10.87)	6 (17.65)	14 (10.14)	20 (23.80)	19 (10.33)
MM	10 (20)	18 (39.13)	11 (32.35)	37 (26.81)	21 (25)	55 (29.89)
Others	11 (22)	7 (15.22)	10 (29.41)	39 (28.26)	21 (25)	46 (25)
Total	50 (100)	46 (100)	34 (100)	138 (100)	84 (100)	184 (100)
Sex						
Female	60 (60)	38 (41.30)	40 (58.82)	134 (48.55)	100 (59.52)	172 (46.74)
Male	40 (40)	54 (58.70)	28 (41.18)	142 (51.45)	68 (40.48)	196 (53.26)
Total	100 (100)	92 (100)	68 (100)	276 (100)	168 (100)	368 (100)

Numbers refer to only cases in the top panel and cases plus controls in the bottom panel.

Table B.2: Distribution of BCL subtypes and genders across phases and cohorts with successfully analysed proteomics and epigenetics samples.

	Phase 1		*Phase 2*		*Total*	
Disease Subtype	**Epic-Italy n (%)**	**NHSDS n (%)**	**Epic-Italy n (%)**	**NHSDS n (%)**	**Epic-Italy n (%)**	**NHSDS n (%)**
CLL	4 (8.89)	5 (11.90)	3 (10.71)	12 (14.29)	7 (9.59)	17 (13.49)
DLBCL	9 (20)	9 (21.43)	1 (3.57)	16 (19.05)	10 (13.70)	25 (19.84)
FL	12 (26.67)	5 (11.90)	5 (17.86)	7 (8.33)	17 (23.29)	12 (9.52)
MM	10 (22.22)	17 (40.48)	11 (39.29)	24 (28.57)	21 (28.77)	41 (32.54)
Others	10 (22.22)	6 (14.29)	8 (28.57)	25 (29.76)	18 (24.66)	31 (24.60)
Total	45 (100)	42 (100)	28 (100)	84 (100)	73 (100)	126 (100)
Sex						
Female	52 (57.78)	38 (45.24)	30 (53.57)	72 (42.86)	82 (56.16)	110 (43.65)
Male	38 (42.22)	46 (54.76)	26 (46.43)	96 (57.14)	64 (43.84)	142 (56.35)
Total	90 (100)	84 (100)	56 (100)	168 (100)	146 (100)	252 (100)

Numbers refer to only cases in the top panel and cases plus controls in the bottom panel.

Table B.3: Distribution of BCL subtypes and genders across phases and cohorts with successfully analysed transcriptomics samples.

	Phase 1		*Phase 2*		*Total*	
Disease Subtype	**Epic-Italy n (%)**	**NHSDS n (%)**	**Epic-Italy n (%)**	**NHSDS n (%)**	**Epic-Italy n (%)**	**NHSDS n (%)**
CLL	5 (10.64)	4 (10)	6 (20.69)	19 (16.38)	11 (14.47)	23 (14.74)
DLBCL	8 (17.02)	8 (20)	0 (0)	21 (18.1)	8 (10.53)	29 (18.59)
FL	14 (29.79)	5 (12.5)	5 (17.24)	13 (11.21)	19 (25)	18 (11.54)
MM	10 (21.28)	17 (42.5)	10 (34.48)	30 (25.86)	20 (26.32)	47 (30.13)
Others	10 (21.28)	6 (15)	8 (27.59)	33 (28.45)	18 (23.68)	39 (25)
Total	47 (100)	40 (100)	29 (100)	116 (100)	76 (100)	156 (100)
Sex						
Female	56 (59.57)	32 (40)	32 (55.17)	104 (44.83)	88 (57.89)	136 (43.59)
Male	38 (40.43)	48 (60)	26 (44.83)	128 (55.17)	64 (42.11)	176 (56.41)
Total	94 (100)	80 (100)	58 (100)	232 (100)	152 (100)	312 (100)

Numbers refer to only cases in the top panel and cases plus controls in the bottom panel.

Table B.4: Distribution of BCL subtypes and genders across phases and cohorts with successfully analysed transcriptomics and epigenetics samples.

	Phase 1		*Phase 2*		*Total*	
Disease Subtype	**Epic-Italy n (%)**	**NHSDS n (%)**	**Epic-Italy n (%)**	**NHSDS n (%)**	**Epic-Italy n (%)**	**NHSDS n (%)**
CLL	4 (9.3)	4 (10.26)	3 (12)	7 (10.14)	7 (10.29)	11 (10.19)
DLBCL	8 (18.6)	8 (20.51)	0 (0)	14 (20.29)	8 (11.76)	22 (20.37)
FL	12 (27.91)	5 (12.82)	5 (20)	7 (10.14)	17 (25)	12 (11.11)
MM	10 (23.26)	17 (43.59)	10 (40)	20 (28.99)	20 (29.41)	37 (34.26)
Others	9 (20.93)	5 (12.82)	7 (28)	21 (30.43)	16 (23.53)	26 (24.07)
Total	43 (100)	39 (100)	25 (100)	69 (100)	68 (100)	108 (100)
Sex						
Female	50 (58.14)	32 (41.03)	26 (52)	52 (37.68)	76 (55.88)	84 (38.89)
Male	36 (41.86)	46 (58.97)	24 (48)	86 (62.32)	60 (44.12)	132 (61.11)
Total	86 (100)	78 (100)	50 (100)	138 (100)	136 (100)	216 (100)

Numbers refer to only cases in the top panel and cases plus controls in the bottom panel.

Table B.5: Median (minimum - maximum) values of immune markers stratified by BCL subtypes for participants of the EPIC-Italy cohort.

Protein	*CLL (n=11)*	*DLBL (n=11)*	*FL (n=20)*	*MM (n=21)*
EGF	25.95 (1.47-265.24)	27.58 (0.35-305.56)	28.25 (0.25-211.41)	11.57 (0.15-338.79)
FGF2	23.76 (1-836.47)	73.94 (4.27-222.42)	19.95 (0.91-359.63)	8.83 (0.99-390.82)
GCSF	31.99 (6.61-101.51)	31.1 (18.17-54.55)	30.09 (4.2-304.12)	28.29 (0.56-104.8)
VEGF	240.55 (3.27-1245.17)	203.09 (8.33-2875.48)	231.38 (1.76-2346.19)	23.37 (0.81-2578.85)
GMSCF	1.91 (0.11-10.47)	4.71 (0.81-12.18)	5.32 (0.15-20.45)	5.96 (0.19-24.01)
TGFa	0.39 (0.06-17.06)	2.55 (0.6-81.12)	2.46 (0-1663.43)	0.78 (0.04-185.92)
Eotaxin	42.22 (35.96-114.76)	53.16 (17.28-241.08)	54.15 (8.4-118.72)	46.51 (22.17-111.48)
Fractalkine	71.64 (12.16-373.32)	117.74 (39.75-870.06)	100.41 (6.69-525.23)	71.38 (0.98-1564.23)
GRO	145.04 (27.15-1789.58)	313.71 (115.07-524.73)	269.72 (81.05-1343.18)	199.54 (61.54-569.37)
MCP1	216.8 (154.03-454.23)	184.53 (127.87-634.85)	197.22 (98.66-351.39)	223.55 (140.45-549.76)
MCP3	5.27 (0.6-734.54)	39.28 (1.79-192.25)	10.47 (0.14-165.72)	5.86 (0.09-83.15)
MDC	232.3 (94.77-1252.65)	355.73 (91.4-2142.34)	328.66 (80.47-1065.02)	245.16 (103.41-2570.29)
MIP1a	8.72 (0.27-82.86)	20.31 (1.25-398.74)	26.55 (0.27-140.62)	2.64 (0.14-210.96)
MIP1b	26.73 (2.61-336.15)	31.5 (20.67-58.48)	37.45 (10.57-174.02)	23.18 (1.99-119.93)
IP10	231.11 (195.79-818.51)	395.74 (111.74-1294.32)	237 (74.09-449.82)	231.22 (86.73-459.02)
IL8	2.93 (0.72-123.57)	16.34 (1.5-304.82)	11.18 (1.12-162.49)	4.01 (1.51-120.25)
IL1b	0.37 (0.02-10.24)	4.34 (0.03-163.78)	1.33 (0-17.46)	1.35 (0.05-7.73)
IL2	1.07 (0.09-166.8)	16.67 (1.2-69.11)	19.63 (0.13-224.77)	3.44 (0.09-83.17)
IL4	0.47 (0.23-4.23)	0.38 (0.08-3.16)	0.75 (0.07-22.22)	1.07 (0.17-59.3)
IL5	0.33 (0.08-525.55)	3.19 (0.17-75.84)	4.91 (0.1-39.69)	1.34 (0.11-49.96)
IL6	1.22 (0.28-27.67)	2.26 (0.19-45.99)	3.31 (0.23-29.17)	2.1 (0.13-10.44)
IL7	1.21 (0.22-18.76)	1.94 (0.74-4.86)	2.21 (0.14-6.46)	1.4 (0.34-5.5)
IL10	9.24 (0.86-297.55)	4.7 (0.18-266.23)	13.39 (0.18-161.99)	11.39 (0.26-88.63)
IL13	1.57 (0.23-929.4)	10.61 (0.53-33.85)	3.48 (0.02-140.31)	1.52 (0.07-44.11)
INFa	13.9 (0.25-362.53)	0.28 (0-394.49)	3.95 (0-673.94)	3.79 (0.01-81.92)
INFg	0.54 (0.11-2069.35)	1.99 (0.04-57.33)	1.48 (0-34.1)	0.86 (0.01-25.99)
TNFa	6.2 (2.44-12.44)	4.71 (0.81-12.18)	6.02 (1.64-20.45)	5.56 (1.33-10.18)
sCD40L	383.16 (110.35-1429.72)	398.44 (143.41-2819.65)	390.52 (282.82-2339.77)	431.83 (155.27-2027.04)

EPIC: European Prospective Investigation into Cancer and Nutrition.

Table B.6: Median (minimum - maximum) values of immune markers stratified by BCL subtypes for participants of the NSHDS cohort.

Protein	*CLL (n=31)*	*DLBL (n=33)*	*FL (n=19)*	*MM (n=55)*
EGF	23.16 (1.24-275.71)	18.42 (0.02-129.72)	23.32 (0.67-241.93)	19.24 (0.25-842.01)
FGF2	21.9 (0.58-195.91)	20.08 (0.39-236.19)	23.59 (0.3-309.99)	7.6 (0.04-483.71)
GCSF	23.77 (1.45-363.57)	17.79 (1.08-496)	23.23 (2.96-57.51)	24.4 (1.16-303)
VEGF	261.49 (4.98-2520.86)	118.53 (2.99-2218.52)	234.04 (5.37-5962.78)	84.1 (0.46-6938.42)
GMSCF	0.64 (0.05-281.84)	3.13 (0.11-265.24)	1.12 (0.07-44.77)	1.48 (0.02-33.61)
TGFa	1.16 (0.06-100.19)	0.85 (0-290.88)	2.23 (0.01-267.82)	0.98 (0-254.11)
Eotaxin	373.04 (192.63-813.23)	351.77 (142.96-985.52)	329.76 (152.89-990.58)	341.53 (95.61-944.76)
Fractalkine	22.85 (2.15-2927)	35.54 (0.58-1260.43)	42.9 (0.66-1347.33)	19.22 (0.13-3229.59)
GRO	442.67 (138.76-1534.08)	435.02 (134.68-2188.12)	358.68 (102.64-1782.63)	475.63 (78.81-1748.74)
MCP1	355.69 (70.58-787.6)	325.7 (17.17-689.56)	306.84 (2.2-747.54)	346.09 (1.78-786.86)
MCP3	7.92 (0.37-121.48)	9.43 (0.08-914.3)	10.69 (0.36-70.88)	2.3 (0.02-312.82)
MDC	904.71 (160.02-8962.26)	921.91 (42.59-2493.68)	873.91 (214.98-5878.03)	889.71 (170.88-6259.24)
MIP1a	10.83 (0.2-100.73)	8.87 (0.05-124.7)	7.66 (0.01-230.61)	5.88 (0.01-207.93)
MIP1b	36.69 (2.53-285.75)	37 (2.52-273.28)	28.61 (1.07-702.97)	30.85 (0.27-843.45)
IP10	535.4 (192.06-1648.62)	532.43 (151-2021.13)	534.5 (38.07-1959.53)	588.06 (15.86-3374.05)
IL8	3.1 (0.59-29.57)	4.22 (0.89-59.12)	2.81 (0.69-9.8)	3.83 (0.64-41.85)
IL1b	0.18 (0.01-435.52)	0.41 (0.01-199.02)	0.24 (0.01-14.17)	0.28 (0-21.24)
IL2	0.85 (0.04-201.86)	1.96 (0.24-282.42)	2.33 (0.2-47.92)	1.04 (0.07-123.28)
IL4	1.19 (0.12-435.84)	9.14 (0.02-274.25)	6.7 (0.53-184.39)	8.67 (0.02-372.43)
IL5	0.25 (0.03-49.89)	0.72 (0.03-40.95)	0.27 (0.06-35.55)	0.63 (0.01-29.25)
IL6	2.33 (0.09-350.99)	4.93 (0.04-133.18)	2.63 (0.38-480.8)	4.64 (0.04-46.45)
IL7	0.29 (0.03-81.26)	0.36 (0.01-64.42)	0.51 (0.02-303.91)	0.35 (0.05-39.07)
IL10	7.44 (0.1-479.88)	13.59 (1.5-1322.5)	11.57 (2.15-688.93)	10.61 (0.26-130.1)
IL13	1.19 (0.09-669.04)	4.05 (0.12-204.7)	0.88 (0.02-178.38)	1.09 (0.01-173.84)
INFa	2.39 (0-670.65)	2.67 (0-778.15)	2.97 (0.01-2569.21)	1.54 (0-645.33)
INFg	1.87 (0.02-281.84)	2.37 (0.02-619.75)	1.74 (0.28-106.09)	1.26 (0-168.82)
TNFa	7.44 (1-35.58)	6.56 (0.85-81.53)	4.98 (0.96-34.57)	6.9 (1.22-24.74)
sCD40L	937.92 (189.09-3994.34)	709.77 (22.96-6393.33)	859.35 (92.92-1857.87)	858.07 (32.53-3807.63)

NSHDS: Northern Sweden Health and Disease Study.

Table B.7: Results of the LMM analyses between log-transformed values of proteins and case-control status.

Protein	*All BCL (n=536)*		*CLL (n=310)*		*DLBCL (n=312)*		*FL (n=307)*		*MM (n=344)*	
	β	***p*-value**	β	***p*-value**	β	***p*-value**	β	***p*-value**	β	***p*-value**
EGF	-0 .28	5 .16 E-02	-0 .05	8 .51 E-01	-0 .55	4 .96 E-02	-0 .19	5 .18 E-01	-0 .53	1 .66 E-02
FGF2	-0 .5	6 .15 E-04	-0 .19	5 .06 E-01	-0 .10	7 .22 E-01	-0 .31	2 .61 E-01	-1 .11	4 .85 E-07
GCSF	-0 .10	2 .42 E-01	-0 .15	3 .97 E-01	-0 .08	6 .44 E-01	0 .03	8 .62 E-01	-0 .13	3 .20 E-01
VEGF	-0 .42	8 .60 E-03	0 .12	6 .86 E-01	-0 .33	2 .72 E-01	-0 .09	7 .81 E-01	-1 .00	4 .23 E-05
GMSCF	-0 .07	5 .50 E-01	-0 .32	1 .85 E-01	0 .23	3 .44 E-01	0 .08	7 .41 E-01	-0 .22	2 .56 E-01
TGFa	-0 .68	5 .62 E-05	-0 .68	3 .08 E-02	-0 .5	1 .43 E-01	-0 .43	2 .02 E-01	-1 .08	2 .78 E-05
Eotaxin	0 .05	2 .64 E-01	0 .12	1 .57 E-01	0 .13	1 .46 E-01	0 .02	8 .41 E-01	0 .03	n6 .59 E-01
Fractalkine	-0 .31	2 .62 E-02	-0 .20	4 .60 E-01	-0 .03	9 .43 E-01	0 .03	9 .31 E-01	-0 .72	9 .14 E-04
GRO	0 .03	5 .11 E-01	0 .02	8 .46 E-01	0 .07	4 .11 E-01	-0 .04	6 .51 E-01	0 .02	7 .29 E-01
MCP1	-0 .01	8 .90 E-01	0 .14	1 .60 E-01	0 .04	7 .30 E-01	-0 .09	3 .92 E-01	0 .01	8 .57 E-01
MCP3	-0 .4	1 .03 E-02	-0 .33	2 .46 E-01	-0 .07	8 .55 E-01	-0 .26	3 .73 E-01	-0 .91	1 .09 E-04
MDC	-0 .16	2 .78 E-03	-0 .08	3 .89 E-01	-0 .26	1 .44 E-02	-0 .08	4 .43 E-01	-0 .18	1 .66 E-02
MIP1a	-0 .35	7 .48 E-03	0 .09	7 .04 E-01	-0 .16	5 .53 E-01	-0 .28	2 .71 E-01	-0 .72	3 .57 E-04
MIP1b	-0 .16	1 .09 E-01	0 .03	8 .75 E-01	-0 .13	5 .30 E-01	-0 .02	9 .21 E-01	-0 .44	5 .25 E-03
IP10	0 .03	6 .43 E-01	0 .12	2 .46 E-01	0 .04	6 .71 E-01	-0 .03	7 .24 E-01	0 .05	5 .68 E-01
IL8	-0 .08	2 .28 E-01	-0 .15	2 .22 E-01	-0 .02	8 .81 E-01	-0 .08	5 .52 E-01	-0 .12	2 .24 E-01
IL1b	-0 .27	5 .38 E-02	-0 .50	5 .50 E-02	0 .31	2 .27 E-01	-0 .27	3 .26 E-01	-0 .59	3 .88 E-03
IL2	-0 .05	6 .59 E-01	-0 .09	6 .92 E-01	0 .02	9 .42 E-01	0 .23	3 .43 E-01	-0 .23	2 .12 E-01
IL4	-0 .15	2 .26 E-01	-0 .70	7 .04 E-03	-0 .36	1 .77 E-01	0 .07	8 .02 E-01	-0 .01	9 .45 E-01
IL5	-0 .10	3 .46 E-01	-0 .24	2 .21 E-01	0 .11	5 .91 E-01	0 .01	9 .94 E-01	-0 .09	5 .78 E-01
IL6	-0 .03	7 .64 E-01	-0 .15	4 .66 E-01	-0 .19	4 .06 E-01	0 .01	9 .83 E-01	-0 .22	1 .98 E-01
IL7	0 .07	5 .38 E-01	-0 .22	3 .46 E-01	0 .08	7 .51 E-01	0 .31	2 .26 E-01	-0 .07	7 .27 E-01
IL10	0 .12	3 .03 E-01	-0 .10	6 .75 E-01	0 .25	3 .07 E-01	0 .30	2 .41 E-01	-0 .06	7 .49 E-01
IL13	-0 .03	8 .15 E-01	-0 .16	5 .82 E-01	0 .17	5 .38 E-01	0 .18	5 .55 E-01	-0 .53	2 .79 E-02
INFa	-0 .41	8 .23 E-02	-0 .07	8 .75 E-01	-0 .81	8 .92 E-02	0 .08	8 .67 E-01	-0 .68	6 .10 E-02
INFg	-0 .32	7 .56 E-02	-0 .18	5 .81 E-01	-0 .01	9 .97 E-01	0 .02	9 .65 E-01	-0 .74	6 .39 E-03
TNFa	0 .12	6 .43 E-02	0 .13	3 .02 E-01	0 .12	3 .30 E-01	0 .10	4 .53 E-01	0 .12	1 .79 E-01
sCD40L	0 .19	8 .88 E-03	0 .35	1 .42 E-02	0 .19	1 .86 E-01	0 .25	7 .27 E-02	0 .11	3 .32 E-01

Results are displayed separately for all BCL observations and the four main histological subtypes including cases and all controls subjects.
LMM: Linear Mixed Model

Table B.8: Results of the LMM analyses between log-transformed values of proteins and case-control status adjusted for the estimated WBC (n=199 case-control pairs).

Protein	*All BCL (n=398)*		*CLL (n=223)*		*DLBCL (n=234)*		*FL (n=228)*		*MM (n=261)*	
	β	***p*-value**	β	***p*-value**	β	***p*-value**	β	***p*-value**	β	***p*-value**
EGF	-0.29	9.19 E-02	-0.04	9.03 E-01	-0.36	2.49 E-01	-0.09	7.85 E-01	-0.58	2.10 E-02
FGF2	-0.49	3.56 E-03	0.28	5.04 E-01	0.02	9.29 E-01	-0.22	4.74 E-01	-1.12	3.91 E-06
GCSF	-0.12	2.11 E-01	-0.28	2.93 E-01	-0.04	8.53 E-01	0.07	7.52 E-01	-0.20	1.79 E-01
VEGF	-0.43	1.84 E-02	0.63	1.86 E-01	-0.22	5.31 E-01	-0.02	9.53 E-01	-1.16	2.27 E-05
GMSCF	-0.09	5.20 E-01	0.24	5.50 E-01	0.27	3.30 E-01	0.13	6.65 E-01	-0.33	1.20 E-01
TGFa	-0.65	1.05 E-03	-0.86	8.50 E-02	-0.16	6.68 E-01	-0.52	1.80 E-01	-1.03	2.54 E-04
Eotaxin	0.03	5.70 E-01	0.12	3.76 E-01	0.14	1.34 E-01	-0.02	8.23 E-01	0.03	6.85 E-01
Fractalkine	-0.49	3.01 E-03	-0.3	4.93 E-01	-0.04	9.25 E-01	0.17	6.12 E-01	-1.07	7.34 E-06
GRO	0.05	4.06 E-01	0.09	5.81 E-01	0.15	1.35 E-01	-0.01	9.15 E-01	0.05	5.33 E-01
MCP1	0.1	7.97 E-02	0.15	3.74 E-01	0.18	1.37 E-01	0.03	8.47 E-01	0.15	1.12 E-01
MCP3	-0.41	1.97 E-02	-0.24	6.14 E-01	0.21	4.91 E-01	-0.29	3.90 E-01	-1.06	3.98 E-05
MDC	-0.15	1.94 E-02	-0.03	8.68 E-01	-0.15	2.33 E-01	-0.10	4.20 E-01	-0.20	2.62 E-02
MIP1a	-0.34	2.18 E-02	0.07	8.63 E-01	0.02	9.25 E-01	-0.14	6.18 E-01	-0.69	1.75 E-03
MIP1b	-0.13	2.53 E-01	0.24	4.66 E-01	-0.05	8.27 E-01	0.03	9.15 E-01	-0.45	1.19 E-02
IP10	0.05	3.83 E-01	0.01	9.70 E-01	0.04	7.27 E-01	0.02	8.58 E-01	0.11	2.38 E-01
IL8	-0.08	3.10 E-01	-0.22	2.74 E-01	0.07	6.29 E-01	-0.10	5.54 E-01	-0.08	4.71 E-01
IL1b	-0.39	2.64 E-02	-0.32	4.68 E-01	0.38	1.87 E-01	-0.54	1.05 E-01	-0.67	3.07 E-03
IL2	-0.22	1.25 E-01	-0.11	7.70 E-01	-0.06	8.41 E-01	0.12	6.67 E-01	-0.43	4.28 E-02
IL4	-0.22	1.29 E-01	-1.07	9.71 E-03	-0.34	2.53 E-01	0.04	9.09 E-01	-0.14	5.41 E-01
IL5	-0.12	3.25 E-01	-0.49	1.30 E-01	0.33	1.57 E-01	-0.08	7.47 E-01	-0.19	2.74 E-01
IL6	-0.06	6.59 E-01	-0.58	9.62 E-02	-0.07	7.96 E-01	0.13	6.52 E-01	-0.30	1.22 E-01
IL7	0.09	5.02 E-01	-0.23	5.26 E-01	0.30	2.43 E-01	0.56	5.31 E-02	-0.35	7.25 E-02
IL10	0.08	5.67 E-01	-0.43	2.52 E-01	0.33	2.18 E-01	0.34	2.51 E-01	-0.08	7.26 E-01
IL13	-0.15	4.13 E-01	-0.3	5.47 E-01	0.26	4.15 E-01	0.10	7.85 E-01	-0.76	5.95 E-03
INFa	-0.52	7.42 E-02	-0.16	8.38 E-01	-0.83	1.42 E-01	0.20	7.33 E-01	-0.90	3.65 E-02
INFg	-0.44	3.53 E-02	0.08	8.84 E-01	0.21	5.27 E-01	-0.08	7.86 E-01	-0.83	5.90 E-03
TNFa	0.12	1.03 E-01	-0.18	3.25 E-01	0.2	1.40 E-01	0.09	5.80 E-01	0.11	2.96 E-01
sCD40L	0.23	9.65 E-03	0.43	7.46 E-02	0.37	2.49 E-02	0.31	8.77 E-02	0.04	7.10 E-01

Results are displayed separately for all BCL observations and the four main histological subtypes including cases and all controls subjects. Models are adjusted for estimated blood proportions of CD8, CD4, NK cells, B cells, and monocytes.
LMM: Linear Mixed Model, WBC: White Blood Cell, NK: Natural Killer.

Table B.9: Results of the LMM analyses between log-transformed values of proteins and case-control status for the participants with WBC estimates available (n=199 case-control pairs).

Protein	*All BCL (n =398)*		*CLL (n =223)*		*DLBCL (n =234)*		*FL (n =228)*		*MM (n =261)*	
	β	***p*-value**	β	***p*-value**	β	***p*-value**	β	***p*-value**	β	***p*-value**
EGF	-0.31	6.79 E-02	0.03	9.44 E-01	-0.36	2.38 E-01	-0.05	8.76 E-01	-0.63	1.09 E-02
FGF2	-0.47	4.09 E-03	-0.06	8.88 E-01	0.06	8.47 E-01	-0.18	5.31 E-01	-1.14	2.05 E-06
GCSF	-0.12	1.96 E-01	-0.27	2.15 E-01	-0.03	8.63 E-01	0.07	7.33 E-01	-0.16	2.69 E-01
VEGF	-0.42	2.12 E-02	0.37	3.53 E-01	-0.21	5.37 E-01	0.01	9.77 E-01	-1.12	3.51 E-05
GMSCF	-0.07	5.97 E-01	0.05	8.86 E-01	0.33	2.37 E-01	0.11	7.09 E-01	-0.35	9.71 E-02
TGFa	-0.66	7.21 E-04	-1.02	1.29 E-02	-0.12	7.53 E-01	-0.44	2.48 E-01	-1.06	1.40 E-04
Eotaxin	0.02	6.08 E-01	0.03	7.95 E-01	0.15	1.15 E-01	-0.06	5.67 E-01	0.02	6.93 E-01
Fractalkine	-0.47	3.63 E-03	-0.3	4.04 E-01	-0.02	9.83 E-01	0.14	6.83 E-01	-1.00	2.55 E-05
GRO	0.03	6.20 E-01	-0.01	9.32 E-01	0.14	1.65 E-01	-0.04	7.54 E-01	0.04	6.28 E-01
MCP1	0.09	1.29 E-01	0.15	2.94 E-01	0.18	1.42 E-01	0.05	6.89 E-01	0.17	6.38 E-02
MCP3	-0.36	4.01 E-02	-0.27	4.87 E-01	0.32	3.18 E-01	-0.21	5.15 E-01	-1.09	2.14 E-05
MDC	-0.16	9.50 E-03	-0.12	3.69 E-01	-0.14	2.55 E-01	-0.08	4.89 E-01	-0.20	2.17 E-02
MIP1a	-0.33	2.64 E-02	-0.01	9.64 E-01	0.05	8.62 E-01	-0.13	6.40 E-01	-0.68	1.88 E-03
MIP1b	-0.13	2.36 E-01	0.02	9.13 E-01	-0.04	8.64 E-01	0.03	9.21 E-01	-0.43	1.46 E-02
IP10	0.04	5.61 E-01	0.03	8.21 E-01	0.04	7.08 E-01	0.03	8.26 E-01	0.12	1.60 E-01
IL8	-0.05	5.39 E-01	-0.08	6.16 E-01	0.11	4.44 E-01	-0.06	7.03 E-01	-0.11	3.13 E-01
IL1b	-0.36	3.77 E-02	-0.34	3.39 E-01	0.48	1.07 E-01	-0.5	1.29 E-01	-0.74	1.23 E-03
IL2	-0.15	2.83 E-01	0.12	7.21 E-01	0.04	8.75 E-01	0.12	6.81 E-01	-0.41	5.00 E-02
IL4	-0.23	1.15 E-01	-0.94	5.98 E-03	-0.3	3.13 E-01	-0.03	9.08 E-01	-0.15	5.06 E-01
IL5	-0.1	4.28 E-01	-0.27	3.12 E-01	0.4	9.73 E-02	-0.03	9.01 E-01	-0.24	1.69 E-01
IL6	-0.02	8.96 E-01	-0.26	3.65 E-01	0.01	9.65 E-01	0.09	7.34 E-01	-0.32	9.24 E-02
IL7	0.11	4.22 E-01	-0.25	4.09 E-01	0.35	1.95 E-01	0.54	5.94 E-02	-0.26	1.81 E-01
IL10	0.11	4.39 E-01	-0.31	3.17 E-01	0.4	1.56 E-01	0.29	3.41 E-01	-0.09	7.03 E-01
IL13	-0.06	7.47 E-01	0.03	9.42 E-01	0.38	2.50 E-01	0.09	8.03 E-01	-0.77	4.95 E-03
INFa	-0.44	1.25 E-01	0.15	8.32 E-01	-0.63	2.80 E-01	0.26	6.57 E-01	-0.92	3.04 E-02
INFg	-0.39	5.84 E-02	-0.06	9.10 E-01	0.27	4.24 E-01	0.03	9.86 E-01	-0.95	1.74 E-03
TNFa	0.14	6.35 E-02	0.08	6.82 E-01	0.24	1.00 E-01	0.15	3.58 E-01	0.09	4.10 E-01
sCD40L	0.22	1.36 E-02	0.43	3.30 E-02	0.37	2.85 E-02	0.3	8.32 E-02	0.08	4.85 E-01

Results are displayed separately for all BCL observations and the four main histological subtypes including cases and all controls subjects.
LMM: Linear Mixed Model, WBC: White Blood Cell.

Table B.10: Strength of association and effect size of the significant association identified in the partial WBC adjustment from the LMM including all BCL cases and controls (n=398).

	Unadjusted		*Partial WBC Adjustment*												*Full WBC*	
	LMM		**CD8**		**CD4**		**NK cells**		**B cells**		**Monocytes**		**Granulocytes**		*LMM*	
Protein	β	***p*-value**	β	***p*-value**	β	***p*-value**	β	***p*-value**	β	***p*-value**	β	***p*-value**	β	***p*-value**	β	***p*-value**
TGFa	0.68	5.62E-05	0.65	7.91E-04	0.66	6.50E-04	0.66	7.05E-04	0.64	1.11E-03	0.66	6.25E-04	0.66	7.45E-04	-0.65	1.05E-03

The corresponding results from the WBC unadjusted LMM (n=398) and full WBC adjustment model (n=536) are also displayed.
WBC: White Blood Cell, LMM: Linear Mixed Model.

Table B.11: Strength of association and effect size of the significant associations identified in the partial WBC adjustment from the LMM including MM cases and all controls (n=261).

	Unadjusted		*Partial WBC Adjustment*												*Full WBC*	
	LMM		**CD8**		**CD4**		**NK cells**		**B cells**		**Monocytes**		**Granulocytes**		*LMM*	
Protein	β	***p*-value**	β	***p*-value**	β	***p*-value**	β	***p*-value**	β	***p*-value**	β	***p*-value**	β	***p*-value**	β	***p*-value**
FGF2	1 .11	4 .85 E-07	1 .16	1 .56E-06	1 .11	3 .68E-06	1 .16	2 .10E-06	1 .11	4 .11E-06	1 .14	2 .07E-06	1 .18	8 .58E-07	1 .12	3 .91 E-06
VEGF	-1	4 .23 E-05	1 .15	2 .13E-05	1 .08	6 .15E-05	1 .19	1 .58E-05	1 .09	5 .40E-05	1 .12	3 .55E-05	1 .17	1 .24E-05	1 .16	2 .27 E-05
TGFa	1 .08	2 .78 E-05	1 .07	1 .16E-04	1 .04	1 .90E-04	1 .07	1 .67E-04	1 .01	2 .46E-04	1 .06	1 .40E-04	1 .09	8 .66E-05	1 .03	2 .54 E-04
Fractalkine	0 .72	9 .14 E-04	1 .03	1 .22E-05	-1	2 .95E-05	1 .09	6 .97E-06	0 .97	4 .18E-05	-1	2 .53E-05	1 .04	9 .46E-06	1 .07	7 .34 E-06
MCP3	0 .91	1 .09 E-04	1 .09	2 .19E-05	1 .06	3 .31E-05	1 .11	2 .16E-05	1 .04	4 .64E-05	-1 .1	1 .92E-05	1 .12	1 .13E-05	1 .06	3 .98 E-05
MIP1a	0 .72	3 .57 E-04	-0 .7	1 .49E-03	—	—	0 .72	1 .21E-03	—	—	—	—	0 .71	1 .03E-03	0 .69	1 .75 E-03
IL1b	—	—	0 .74	1 .18E-03	0 .73	1 .46E-03	—	—	—	—	0 .74	1 .13E-03	0 .76	8 .92E-04	—	—
INFg	—	—	0 .96	1 .58E-03	—	—	—	—	—	—	0 .95	1 .72E-03	0 .97	1 .27E-03	—	—

The corresponding results from the WBC unadjusted LMM (n=344) and full WBC adjustment model (n=261) are also displayed. Proteins are ordered in relation to their corresponding strength of association with MM case/control status from the results of the WBC unadjusted model. Strength of association and effect size of the proteins that did not reach statistical significance in the corresponding models are not shown.
LMM: Linear Mixed Model, WBC: White Blood Cell, NK: Natural Killer.

Table B.12: Results of the LMM analyses between log-transformed values of proteins and case-control status for all BCL and MM observations stratified by median TtD.

Protein	*All BCL (n=536)*				*MM Subtype (n=344)*			
	TtD<6 years (n=402)		**TtD>6years (n=402)**		**TtD<6 years (n=313)**		**TtD>6years (n=299)**	
	β	***p*-value**	β	***p*-value**	β	***p*-value**	β	***p*-value**
EGF	-0.46	1.29 E-02	-0.14	4.02 E-01	-0.72	9.40 E-03	-0.29	3.54 E-01
FGF2	-0.61	1.11 E-03	-0.44	1.33 E-02	-1.07	8.21 E-05	-1.18	2.48 E-04
GCSF	-0.18	8.98 E-02	-0.02	8.05 E-01	-0.15	3.63 E-01	-0.11	5.75 E-01
VEGF	-0.52	8.36 E-03	-0.38	5.27 E-02	-1.06	4.17 E-04	-0.93	8.85 E-03
GMSCF	0.18	2.31 E-01	-0.31	4.11 E-02	-0.09	6.89 E-01	-0.35	2.24 E-01
TGFa	-0.77	4.70 E-04	-0.63	2.09 E-03	-1.28	7.84 E-05	-0.72	5.24 E-02
Eotaxin	0.04	5.14 E-01	0.06	2.67 E-01	0.05	4.99 E-01	-0.02	8.60 E-01
Fractalkine	-0.44	1.26 E-02	-0.21	2.17 E-01	-0.83	1.52 E-03	-0.60	6.19 E-02
GRO	0.05	4.29 E-01	0.00	9.47 E-01	0.07	3.96 E-01	-0.04	6.75 E-01
MCP1	-0.04	5.69 E-01	0.01	8.90 E-01	-0.02	8.47 E-01	0.06	6.11 E-01
MCP3	-0.54	4.81 E-03	-0.32	9.75 E-02	-0.95	7.94 E-04	-0.94	6.69 E-03
MDC	-0.20	2.66 E-03	-0.13	3.74 E-02	-0.23	1.94 E-02	-0.12	2.87 E-01
MIP1a	-0.52	1.28 E-03	-0.24	1.33 E-01	-0.95	1.19 E-04	-0.38	1.95 E-01
MIP1b	-0.13	3.06 E-01	-0.24	5.53 E-02	-0.51	1.02 E-02	-0.36	1.11 E-01
IP10	0.07	2.75 E-01	-0.01	8.40 E-01	0.07	4.35 E-01	0.01	9.26 E-01
IL8	-0.09	2.69 E-01	-0.07	4.09 E-01	-0.23	5.85 E-02	0.06	6.89 E-01
IL1b	-0.26	1.44 E-01	-0.27	1.02 E-01	-0.54	2.90 E-02	-0.66	3.00 E-02
IL2	-0.02	8.72 E-01	-0.06	6.80 E-01	-0.31	1.82 E-01	-0.06	8.30 E-01
IL4	-0.07	6.95 E-01	-0.20	2.14 E-01	-0.01	9.67 E-01	0.04	9.14 E-01
IL5	-0.06	6.34 E-01	-0.11	3.82 E-01	-0.08	6.85 E-01	-0.08	7.27 E-01
IL6	0.01	9.22 E-01	-0.08	5.59 E-01	-0.18	3.91 E-01	-0.26	2.87 E-01
IL7	0.14	3.50 E-01	0.01	9.27 E-01	-0.09	7.02 E-01	-0.02	9.53 E-01
IL10	0.12	4.46 E-01	0.17	2.35 E-01	-0.17	4.64 E-01	0.17	5.23 E-01
IL13	0.07	7.14 E-01	-0.12	4.98 E-01	-0.53	7.15 E-02	-0.48	1.61 E-01
INFa	-0.51	9.26 E-02	-0.42	1.49 E-01	-0.96	3.78 E-02	-0.33	5.23 E-01
INFg	-0.39	6.85 E-02	-0.26	2.27 E-01	-0.62	5.37 E-02	-0.91	2.09 E-02
TNFa	0.13	8.48 E-02	0.07	3.42 E-01	0.04	7.37 E-01	0.24	7.98 E-02
sCD40L	0.18	4.00 E-02	0.19	3.72 E-02	0.04	7.43 E-01	0.20	2.10 E-01

Results for CLL, DLBCL and FL subtypes did not provide significant findings and are not shown.
LMM: Linear Mixed Model, TtD: Time to Diagnosis.

Table B.13: Results of the LMM analyses between log-transformed values of proteins and case-control status for all BCL and MM observations stratified by study cohort.

Protein	*Epic-Italy*				*NSHDS*			
	All BCL (n=168)		**MM (n=105)**		**All BCL (n=368)**		**MM (n=239)**	
	β	***p*-value**	β	***p*-value**	β	***p*-value**	β	***p*-value**
EGF	-0.62	1.9 E-02	-0.68	1.3 E-01	-0.10	5.3 E-01	-0.43	9.3 E-02
FGF2	-0.73	2.1 E-03	-0.90	2.2 E-02	-0.41	2.0 E-02	-1.19	4.1 E-06
GCSF	0.02	8.2 E-01	-0.15	3.9 E-01	-0.13	2.3 E-01	-0.12	4.9 E-01
VEGF	-0.80	2.4 E-03	-1.49	9.5 E-04	-0.27	1.7 E-01	-0.82	4.5 E-03
GMSCF	-0.08	5.3 E-01	0.02	9.4 E-01	-0.12	4.7 E-01	-0.35	1.5 E-01
TGFa	-1.16	1.3 E-04	-1.12	1.8 E-02	-0.48	1.9 E-02	-1.09	2.9 E-04
Eotaxin	-0.02	8.4 E-01	-0.08	5.6 E-01	0.07	1.5 E-01	0.04	6.2 E-01
Fractalkine	-0.28	1.1 E-01	-0.51	8.2 E-02	-0.35	6.3 E-02	-0.75	6.1 E-03
GRO	-0.04	7.0 E-01	-0.12	4.5 E-01	0.05	2.9 E-01	0.04	5.5 E-01
MCP1	0.00	9.8 E-01	0.12	1.4 E-01	0.00	9.8 E-01	-0.03	8.1 E-01
MCP3	-0.59	1.7 E-02	-0.82	3.9 E-02	-0.35	7.8 E-02	-0.94	1.1 E-03
MDC	-0.25	1.3 E-02	-0.37	1.8 E-02	-0.12	5.3 E-02	-0.11	1.6 E-01
MIP1a	-0.70	3.9 E-04	-0.96	3.2 E-03	-0.17	2.9 E-01	-0.59	1.5 E-02
MIP1b	-0.06	6.2 E-01	-0.24	2.9 E-01	-0.18	1.7 E-01	-0.49	1.3 E-02
IP10	0.04	5.8 E-01	-0.09	4.5 E-01	0.03	7.2 E-01	0.09	4.3 E-01
IL8	0.17	1.8 E-01	0.08	6.3 E-01	-0.21	5.8 E-03	-0.20	7.8 E-02
IL1b	-0.36	1.9 E-01	-0.37	3.2 E-01	-0.27	9.1 E-02	-0.70	3.7 E-03
IL2	-0.18	3.7 E-01	-0.40	2.4 E-01	-0.04	7.9 E-01	-0.18	4.2 E-01
IL4	-0.09	6.1 E-01	0.14	6.3 E-01	-0.22	1.8 E-01	-0.14	5.7 E-01
IL5	0.06	7.1 E-01	0.10	6.9 E-01	-0.19	1.7 E-01	-0.15	4.5 E-01
IL6	0.14	4.2 E-01	0.00	1.0 E+00	-0.13	3.5 E-01	-0.30	1.4 E-01
IL7	0.03	7.7 E-01	0.07	6.6 E-01	0.09	6.1 E-01	-0.13	6.0 E-01
IL10	0.07	7.7 E-01	-0.16	6.5 E-01	0.12	3.5 E-01	-0.08	7.1 E-01
IL13	0.44	8.3 E-02	0.12	7.7 E-01	-0.27	1.4 E-01	-0.76	7.2 E-03
INFa	-0.46	2.9 E-01	-0.83	2.4 E-01	-0.39	1.6 E-01	-0.56	1.8 E-01
INFg	-0.50	1.2 E-01	-0.98	4.7 E-02	-0.25	2.3 E-01	-0.70	2.8 E-02
TNFa	0.10	1.8 E-01	0.03	8.1 E-01	0.10	2.2 E-01	0.14	2.3 E-01
sCD40L	0.11	3.3 E-01	0.12	5.0 E-01	0.23	1.3 E-02	0.07	5.5 E-01

Results for CLL, DLBCL and FL subtypes did not provide significant findings and are not shown.
LMM: Linear Mixed Model, EPIC: European Prospective Investigation into Cancer and Nutrition, NSHDS: Northern Sweden Health and Disease Study.

Table B.14: Results of the LMM analyses between log-transformed values of proteins and case-control status for all BCL and MM observations stratified by analytical phase.

Protein	*Phase 1*				*Phase 2*			
	All BCL (n=192)		**MM (n=124)**		**All BCL (n=344)**		**MM (n=220)**	
	β	***p*-value**	β	***p*-value**	β	***p*-value**	β	***p*-value**
EGF	-0.63	1.25 E-02	-0.86	2.92 E-02	-0.11	5.45 E-01	-0.32	2.61 E-01
FGF2	-0.69	2.52 E-03	-1.23	7.16 E-04	-0.41	3.08 E-02	-1.05	2.82 E-04
GCSF	0.07	5.12 E-01	0.12	5.06 E-01	-0.19	1.14 E-01	-0.24	2.11 E-01
VEGF	-0.57	2.95 E-02	-1.13	7.68 E-03	-0.38	5.92 E-02	-1.07	6.20 E-04
GMSCF	0.04	8.24 E-01	-0.56	7.36 E-02	-0.12	4.46 E-01	0.05	8.27 E-01
TGFa	-1.17	4.62 E-05	-1.73	1.41 E-04	-0.44	3.80 E-02	-0.89	6.04 E-03
Eotaxin	0.01	8.67 E-01	-0.04	7.26 E-01	0.06	2.38 E-01	0.02	7.29 E-01
Fractalkine	-0.25	2.09 E-01	-0.49	1.20 E-01	-0.37	4.89 E-02	-0.83	5.07 E-03
GRO	0.01	9.91 E-01	-0.15	2.33 E-01	0.03	5.74 E-01	0.08	3.15 E-01
MCP1	0.05	3.68 E-01	0.16	5.78 E-02	-0.04	5.84 E-01	-0.08	5.57 E-01
MCP3	-0.61	8.13 E-03	-1.00	5.70 E-03	-0.30	1.51 E-01	-0.79	1.37 E-02
MDC	-0.23	1.93 E-02	-0.28	6.22 E-02	-0.13	3.86 E-02	-0.15	8.38 E-02
MIP1a	-0.64	4.24 E-04	-0.90	1.26 E-03	-0.20	2.39 E-01	-0.63	2.35 E-02
MIP1b	-0.17	2.45 E-01	-0.50	4.52 E-02	-0.17	2.00 E-01	-0.42	5.10 E-02
IP10	0.10	2.01 E-01	0.26	2.96 E-02	-0.03	7.03 E-01	-0.06	6.36 E-01
IL8	0.02	8.96 E-01	-0.28	1.53 E-01	-0.15	3.55 E-02	-0.08	4.76 E-01
IL1b	-0.49	1.17 E-01	-1.53	4.05 E-04	-0.14	2.86 E-01	0.02	9.17 E-01
IL2	-0.01	9.59 E-01	-0.71	3.34 E-02	-0.10	4.65 E-01	0.13	5.41 E-01
IL4	-0.36	1.24 E-01	-0.53	1.49 E-01	-0.06	7.06 E-01	0.26	2.99 E-01
IL5	-0.13	5.24 E-01	-0.67	2.78 E-02	-0.10	3.74 E-01	0.17	3.37 E-01
IL6	-0.09	7.01 E-01	-0.82	2.19 E-02	-0.04	7.30 E-01	0.07	6.63 E-01
IL7	0.10	6.36 E-01	-0.52	8.14 E-02	0.05	7.25 E-01	0.15	5.22 E-01
IL10	0.22	4.05 E-01	-0.55	1.46 E-01	0.07	5.07 E-01	0.15	4.10 E-01
IL13	-0.01	9.79 E-01	-1.62	1.85 E-03	-0.09	4.89 E-01	0.15	5.12 E-01
INFa	-0.66	1.72 E-01	-1.71	2.65 E-02	-0.28	2.43 E-01	-0.20	5.96 E-01
INFg	-0.54	1.51 E-01	-1.81	2.89 E-03	-0.21	2.24 E-01	-0.12	6.54 E-01
TNFa	0.13	1.12 E-01	-0.07	5.28 E-01	0.12	1.46 E-01	0.29	2.34 E-02
sCD40L	0.27	3.19 E-02	-0.06	8.02 E-01	0.16	7.72 E-02	0.12	3.44 E-01

Results for CLL, DLBCL and FL subtypes did not provide significant findings and are not shown.
LMM: Linear Mixed Model.

Table B.15: Results of the ULR model relating MM subtype case-control status and quantile categories of log-transformed protein concentration levels stratified by study cohort.

	EPIC-Italy (n=105)						*NSHDS (n=239)*					
Protein	**Quantiles Limits**	**Cases/Controls (n)**	**OR**	**low CI**	**High CI**	**P-trend**	**Quantiles Limits**	**Cases/Controls (n)**	**OR**	**low CI**	**High CI**	**P-trend**
FGF2	Q1=< 2.9	8/23	Ref	Ref	Ref	0.001	Q1=< 2.9	26/44	Ref	Ref	Ref	0.0015
	Q2=2.9 - 4.26	9/13	3.82	0.8	21.29		Q2=2.9 - 4.26	16/54	0.45	0.2	0.96	
	Q3=4.26 - 5.35	2/27	0.14	0.01	0.99		Q3=4.26 - 5.35	9/40	0.36	0.14	0.86	
	Q4=> 5.35	2/21	0.21	0.02	1.27		Q4=> 5.35	4/46	0.14	0.04	0.4	
VEGF	Q1=< 4.44	10/18	Ref	Ref	Ref	0.0255	Q1=< 4.44	22/49	Ref	Ref	Ref	0.0206
	Q2=4.44 - 6.15	7/22	0.6	0.15	2.32		Q2=4.44 - 6.15	13/45	0.54	0.23	1.24	
	Q3=6.15 - 7.43	2/20	0.16	0.02	0.83		Q3=6.15 - 7.43	16/47	0.71	0.32	1.55	
	Q4=> 7.43	2/24	0.11	0.01	0.58		Q4=> 7.43	4/43	0.18	0.05	0.52	
TGFa	Q1=< 1.27	8/17	Ref	Ref	Ref	0.0312	Q1=< 1.27	23/50	Ref	Ref	Ref	0.0113
	Q2=1.27 - 2.65	7/27	0.23	0.04	1.07		Q2=1.27 - 2.65	16/40	0.77	0.33	1.73	
	Q3=2.65 - 3.91	5/19	0.43	0.08	2.07		Q3=2.65 - 3.91	12/48	0.51	0.22	1.16	
	Q4=> 3.91	1/21	0.05	0	0.38		Q4=> 3.91	4/46	0.17	0.05	0.5	
Fractalkine	Q1=< 4.11	4/16	Ref	Ref	Ref	0.1087	Q1=< 4.11	24/51	Ref	Ref	Ref	0.0462
	Q2= 4.11 - 5.12	12/27	2.41	0.54	12.93		Q2= 4.11 - 5.12	15/40	0.75	0.33	1.66	
	Q3= 5.12 - 6.21	3/27	0.42	0.06	2.78		Q3=5.12 - 6.21	8/40	0.43	0.16	1.09	
	Q4=> 6.21	2/14	0.61	0.05	5.75		Q4=> 6.21	8/53	0.31	0.12	0.75	
MCP3	Q1=< 0.97	7/16	Ref	Ref	Ref	0.2103	Q1=< 0.97	28/51	Ref	Ref	Ref	0.0118
	Q2=0.97 - 2.48	7/26	0.76	0.18	3.26		Q2=0.97 - 2.48	10/41	0.4	0.16	0.93	
	Q3=2.48 - 3.7	5/27	0.31	0.06	1.47		Q3=2.48 - 3.7	7/40	0.29	0.1	0.73	
	Q4=> 3.7	2/15	0.15	0.01	1.1		Q4=> 3.7	10/52	0.36	0.15	0.81	
MIP1a	Q1=< 1.58	8/17	Ref	Ref	Ref	0.0088	Q1=< 1.58	19/50	Ref	Ref	Ref	0.0186
	Q2= 1.58 - 2.84	11/24	1.09	0.28	4.26		Q2= 1.58 - 2.84	21/43	1.25	0.57	2.77	
	Q3= 2.84 - 3.64	0/20					Q3=2.84 - 3.64	8/47	0.41	0.15	1.04	
	Q4=> 3.64	2/23	0.16	0.02	0.89		Q4=> 3.64	7/44	0.36	0.12	0.94	

OR: Odds Ratio, CI: Confidence Interval, ULR: Unconditional Logistic Regression, EPIC: European Prospective Investigation into Cancer and Nutrition, NSHDS: Northern Sweden Health and Disease Study.

Table B.16: Results of the CLR model relating MM subtype case-control status and quantile categories of log-transformed protein concentration levels.

Protein	*Quantiles Limits*	*Cases/Controls (n)*	*OR*	*low CI*	*High CI*	*P-trend*
FGF2	Q1=< 2.67	33/19	Ref	Ref	Ref	0.0093*
	Q2=2.67 - 4.23	25/19	0.78	0.32	1.9	
	Q3=4.23 - 5.29	12/19	0.36	0.14	0.94	
	Q4=> 5.29	6/19	0.19	0.06	0.56	
VEGF	Q1=< 4.25	30/19	Ref	Ref	Ref	0.0917
	Q2=4.25 - 5.92	21/19	0.66	0.26	1.67	
	Q3=5.92 - 7.38	17/19	0.59	0.23	1.53	
	Q4=> 7.38	8/19	0.26	0.09	0.74	
TGFa	Q1=< 1.27	31/19	Ref	Ref	Ref	0.0028*
	Q2=1.27 - 2.75	26/19	0.68	0.27	1.72	
	Q3=2.75 - 3.96	15/19	0.43	0.16	1.1	
	Q4=> 3.96	4/19	0.11	0.03	0.36	
Fractalkine	Q1=< 4.07	28/19	Ref	Ref	Ref	0.0581
	Q2=4.07 - 5.16	27/19	1.1	0.45	2.72	
	Q3=5.16 - 6.28	12/19	0.42	0.15	1.15	
	Q4=> 6.28	9/19	0.36	0.12	0.97	
MCP3	Q1=< 0.88	33/19	Ref	Ref	Ref	0.1341
	Q2=0.88 - 2.44	19/19	0.55	0.21	1.41	
	Q3=2.44 - 3.47	11/19	0.33	0.12	0.88	
	Q4=> 3.47	13/19	0.45	0.17	1.16	
MIP1a	Q1=< 1.6	28/19	Ref	Ref	Ref	0.0521
	Q2=1.6 - 2.73	28/19	1.1	0.46	2.63	
	Q3=2.73 - 3.59	9/19	0.35	0.11	1	
	Q4=> 3.59	11/19	0.4	0.14	1.08	

(*) represents statistically significant associations.
OR: Odds Ratio, CI: Confidence Interval, CLR: Conditional Logistic Regression.

Table B.17: Common transcripts differentially expressed between the WBC unadjusted LMM (n=266), the B-cells proportion adjusted LMM (n=194) and the full WBC adjustment LMM (n=194) from the CLL-specific analysis.

			WBC Unadjusted LMM		*B cells Adjustment*		*Full WBC Adjustment*	
	Agilent ID	**Gene Name**	***f***	***p*-value**	***f***	***p*-value**	***f***	***p*-value**
1	A_23_P500400	ABCA6	17.40	1.04E-62	6.92	4.57E-24	6.45	3.65E-22
2	A_32_P53234	—	5.28	1.18E-44	2.11	7.65E-12	2.09	1.45E-11
3	A_23_P26854	ARHGAP44	24.49	1.30E-43	6.94	2.38E-13	6.54	2.99E-12
4	A_23_P27332	TCF4	3.83	7.14E-41	2.08	1.35E-12	2.08	1.72E-12
5	A_24_P29733	CDK14	3.96	4.43E-39	2.16	1.40E-10	2.04	2.37E-09
6	A_23_P131024	ZBTB32	5.27	5.47E-39	2.58	2.81E-10	2.85	9.28E-12
7	A_24_P691826	—	5.84	1.46E-35	2.63	4.72E-09	2.72	2.64E-09
8	A_23_P130158	WNT3	13.75	5.67E-35	5.06	7.37E-11	4.45	2.29E-09
9	A_24_P931428	TCF4	3.79	3.58E-34	1.94	1.08E-07	1.94	2.36E-07
10	A_23_P56553	METTL8	2.71	8.26E-29	1.65	9.88E-07	1.69	4.74E-07
11	A_23_P201211	FCRL5	5.20	2.74E-28	2.45	1.64E-06	2.55	8.92E-07
12	A_32_P44394	AIM2	2.63	3.78E-24	1.97	2.07E-07	1.96	4.59E-07
13	A_24_P184803	COCH	3.91	6.57E-24	2.32	2.21E-07	2.42	1.32E-07

Gene expression signals are ordered in relation to their corresponding strength of association with CLL case/control status from the WBC unadjusted LMM. Fold change (f) estimates derived from the regression coefficient (β) obtained from the LMM.
LMM: Linear Mixed Model, WBC: White Blood Cell.

Table B.18: Significant associations identified in the analysis stratified by median TtD for the population including all BCL cases and controls.

	Pooled Analysis (n=464)					*TtD<6 years (n=348)*					*TtD>6 years (n=348)*			
	Agilent ID	**Gene Name**	***f***	***p*-value**		**Agilent ID**	**Gene Name**	***f***	***p*-value**		**Agilent ID**	**Gene Name**	***f***	***p*-value**
1	A_23_P500400	ABCA6	1.68	9.52E-09	1	A_23_P26854	ARHGAP44	2.00	9.36E-08	1	A_23_P500400	ABCA6	1.67	8.96E-09
2	A_23_P26854	ARHGAP44	1.86	3.28E-08	2	A_24_P204574	—	0.71	6.49E-07	2	A_23_P310931	CNR2	1.22	8.63E-07
3	A_23_P210581	KCNG1	0.76	4.78E-07	3	A_23_P500400	ABCA6	1.69	3.33E-07	3	A_23_P145889	CDK14	1.32	6.39E-07
4	A_32_P44394	AIM2	1.26	9.36E-07	4	A_24_P484904	—	0.74	3.16E-07	4	A_32_P44394	AIM2	1.32	3.08E-07
5	A_23_P145889	CDK14	1.28	1.54E-06	5	A_23_P210581	KCNG1	0.71	2.03E-07					

The corresponding results of the pooled analysis are also displayed. Transcripts are ordered in relation to their corresponding strength of association with BCL case/control status from the results of the TtD pooled analysis. Fold change (f) estimates derived from the regression coefficient (β) obtained from the Linear Mixed Model.
TtD: Time to Diagnosis.

Table B.19: Significant associations identified in the analysis stratified by median TtD for the population including CLL cases and all controls.

	Pooled Analysis (n=464)					*TtD<6 years (n=348)*					*TtD>6 years (n=348)*			
	Agilent ID	**Gene Name**	***f***	***p*-value**		**Agilent ID**	**Gene Name**	***f***	***p*-value**		**Agilent ID**	**Gene Name**	***f***	***p*-value**
1	A_23_P500400	ABCA6	1.68	9.52E-09	1	A_23_P26854	ARHGAP44	2.00	9.36E-08	1	A_23_P500400	ABCA6	1.67	8.96E-09
2	A_23_P26854	ARHGAP44	1.86	3.28E-08	2	A_24_P204574	—	0.71	6.49E-07	2	A_23_P310931	CNR2	1.22	8.63E-07
3	A_23_P210581	KCNG1	0.76	4.78E-07	3	A_23_P500400	ABCA6	1.69	3.33E-07	3	A_23_P145889	CDK14	1.32	6.39E-07
4	A_32_P44394	AIM2	1.26	9.36E-07	4	A_24_P484904	—	0.74	3.16E-07	4	A_32_P44394	AIM2	1.32	3.08E-07
5	A_23_P145889	CDK14	1.28	1.54E-06	5	A_23_P210581	KCNG1	0.71	2.03E-07					

Transcripts are ordered in relation to their corresponding strength of association with MM case/control status from the results of the TtD pooled analysis. Fold change (f) estimates derived from the regression coefficient (β) obtained from the Linear Mixed Model.
TtD: Time to Diagnosis.

C

Supplementary Material for Chapter 5

C.1 Overlapping and Improvements in Relation to Cited Papers

The analysis of inflammatory markers data employing multivariate statistical approaches was originally published in *"Pre-diagnostic blood immune markers, incidence and progression of B-cell lymphoma and multiple myeloma: Univariate and functionally informed multivariate analyses"* [182]. Therefore, there is a partial overlap between the results presented in chapter 5 and the ones published in this article, which include the following:

– Identification of proteins indicative of future risk of B-cell Lymphoma (BCL) and its main histological subtypes employing sparse and sparse group Partial Least Squares- Discriminant Analysis (sPLS-DA and sgPLS-DA).

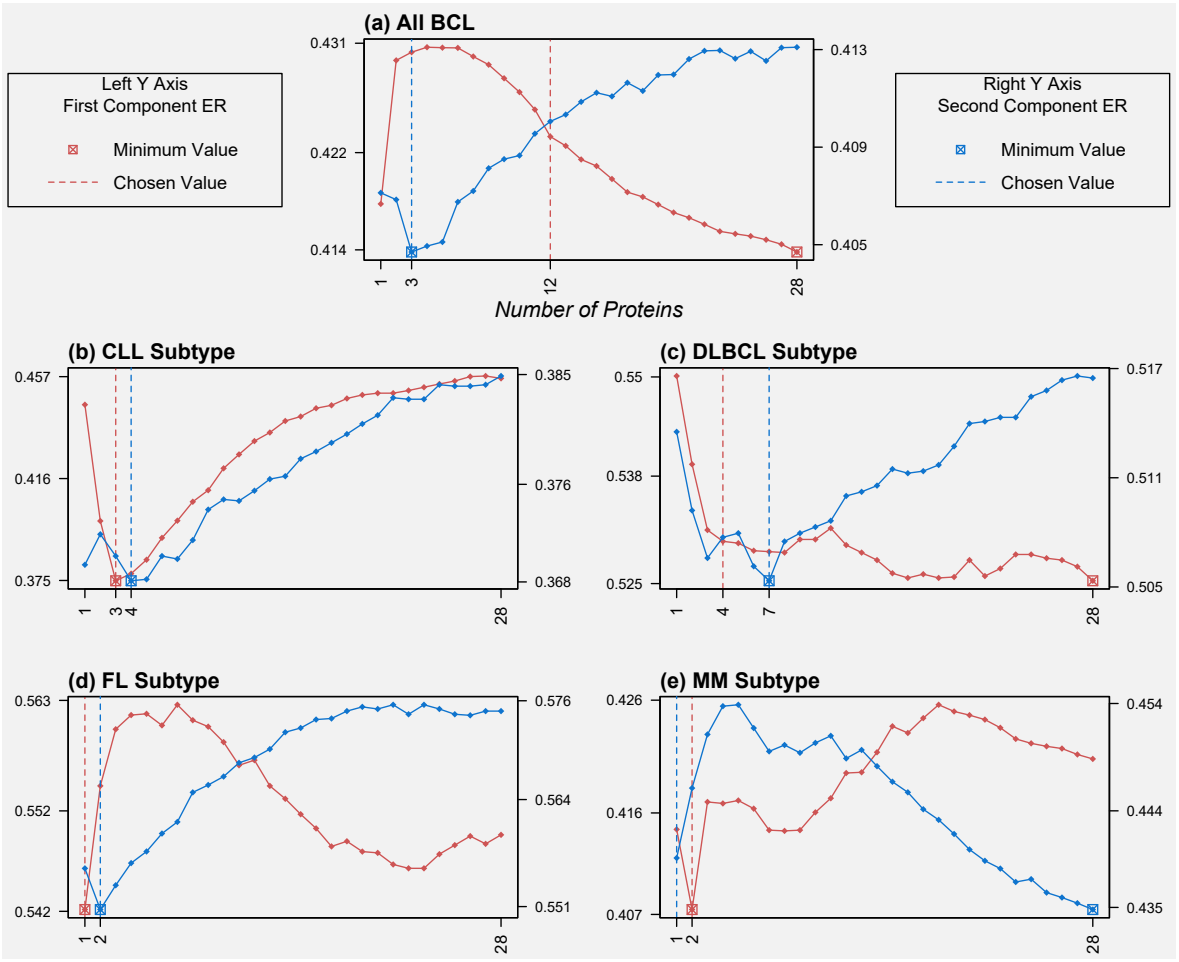
The overlapping results between chapter 5 and the cited paper show significant variations in the final outputs which may come down to differences in the approach taken for the calibration procedure. The paper explored only one dimension for the model fitting procedure of sPLS-DA and sgPLS-DA following the $H = min(p, G - 1)$ criteria, where p is the total number of predictor variables and G is the total number of classes [128]. As a consequence, visualization tools such as sample representation plots were not investigated in the paper and correspond to valuable incorporations added in this thesis. This methodological difference between the two works can also be responsi-

ble for the positive finding identified for the Chronic Lymphocytic Leukaemia (CLL) subtype and proteomics in chapter 5 which were not discovered in the published article.

The remaining of analyses presented in this chapter which were not listed above correspond to novel incorporation introduced in this thesis.

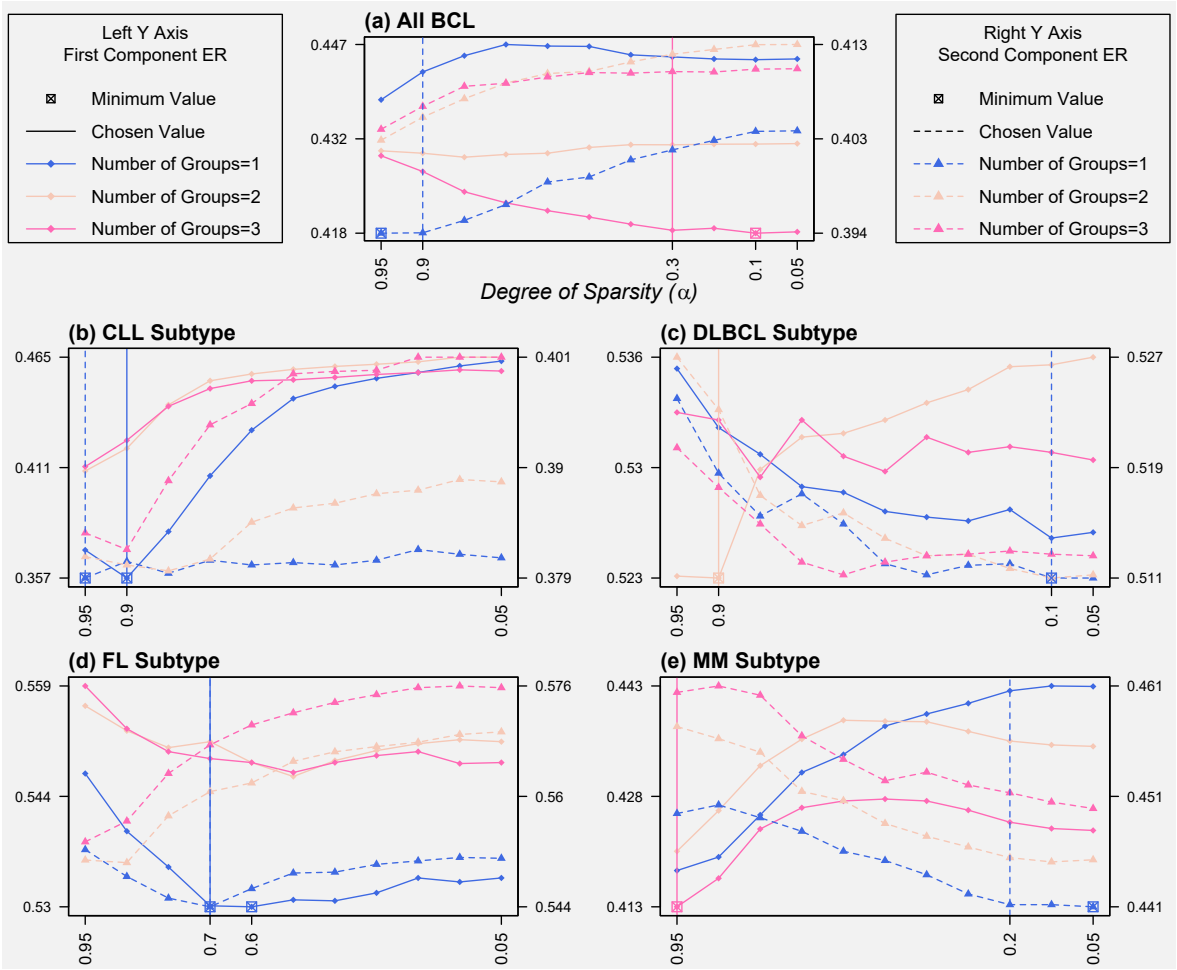
Supplementary Figures

Figure C.1: Average overall misclassification ER from the calibration procedure of the sPLS-DA model for the two dimensions and for the five study populations (proteomics dataset).



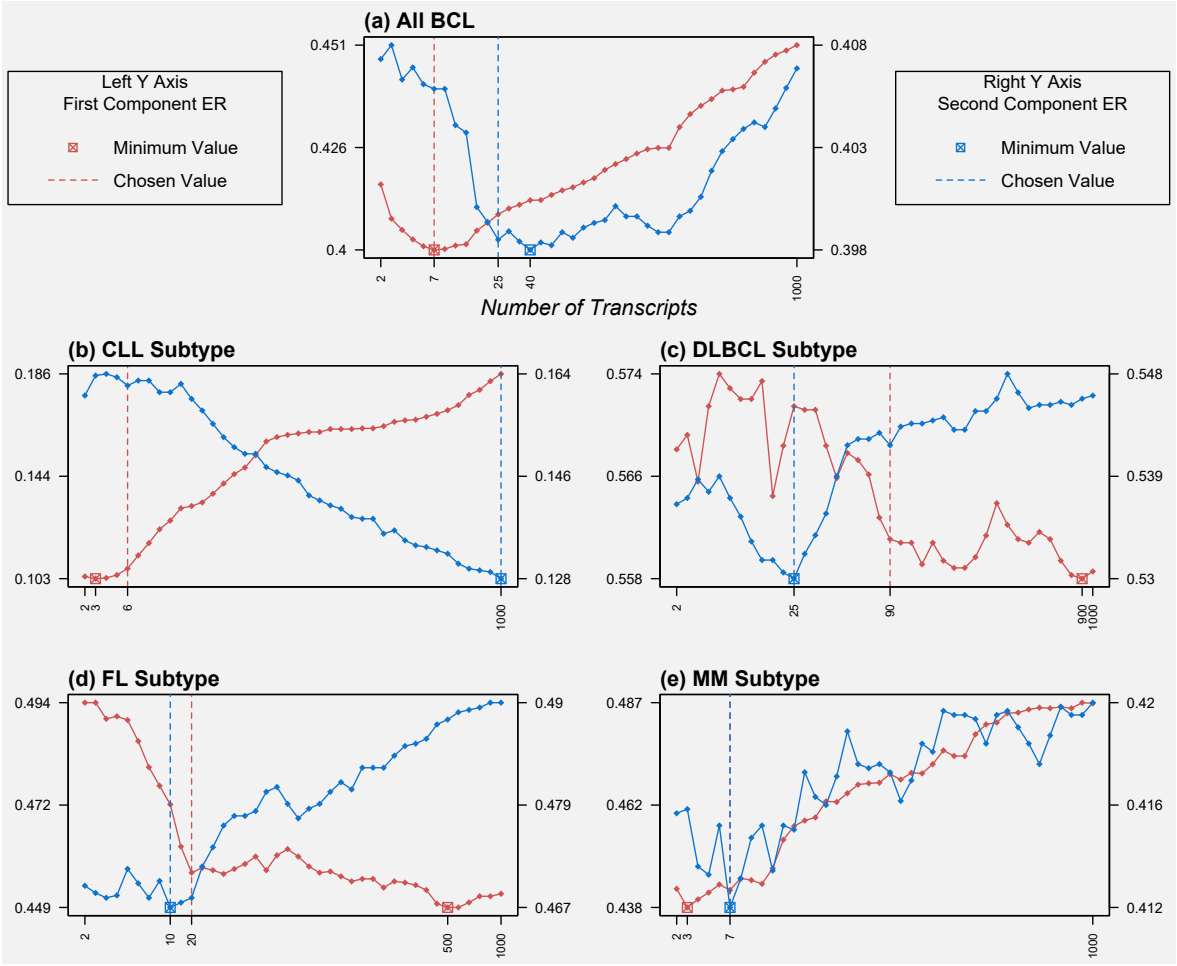
In each graphic, calibration curves represent models fitted with 1 to 28 proteins to retain in the predictor matrix and the two dimensions are shown simultaneously. For each component, the observed minimum value and the chosen value to retain in the model are represented by a cross and a dashed vertical line, respectively. Calibration of the second component was conducted retaining in the previous dimension the chosen number of variables.
ER: Error Rate; sPLS-DA: sparse Partial Least Squares-Discriminant Analysis.

Figure C.2: Average overall misclassification ER from the calibration procedure of the sgPLS-DA model for the two dimensions and for the five study populations (proteomics dataset).



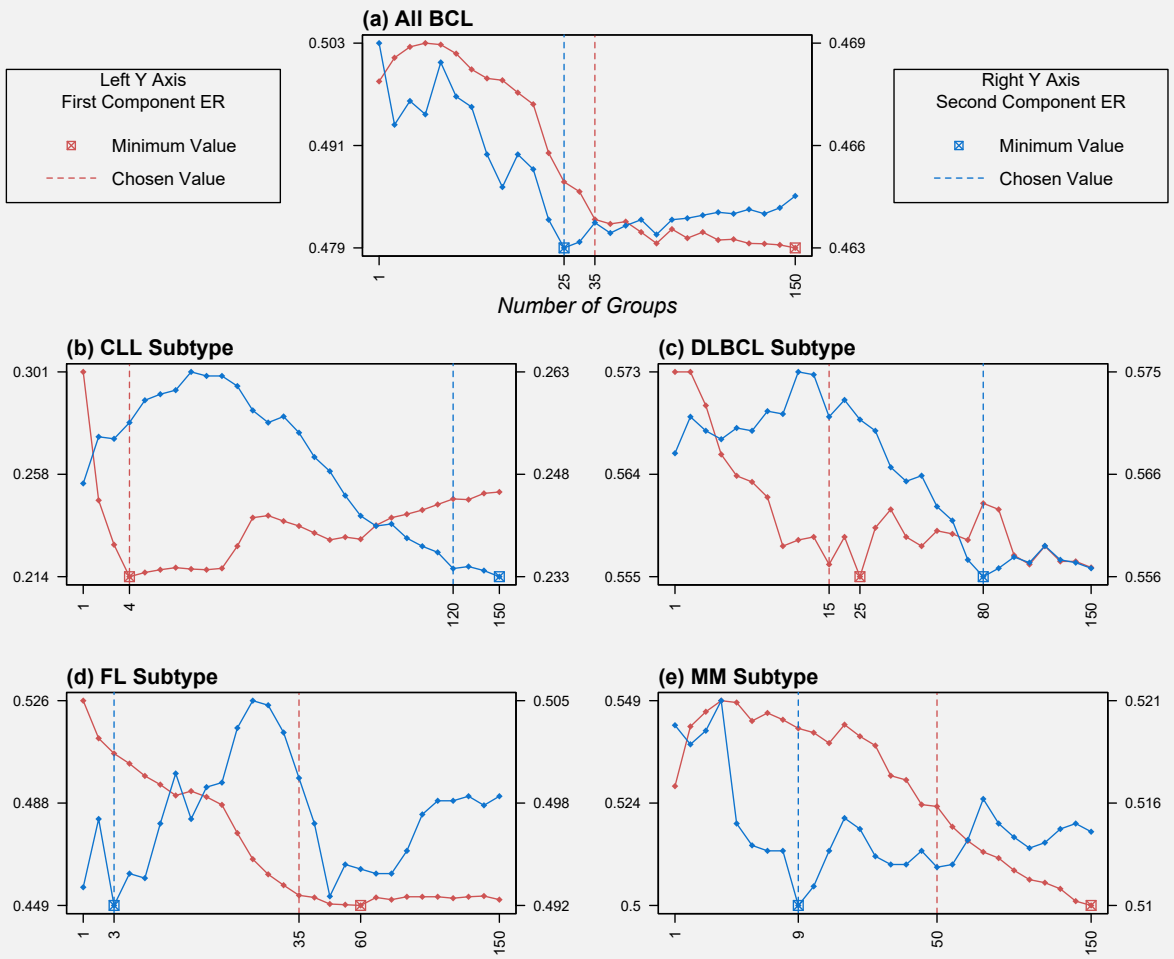
A single calibration curve represents models fitted with 11 different values of the mixing parameter α_1 defining the within-group sparsity for models retaining either 1, 2 or 3 functional groups (blue, light orange and pink, respectively). Since the two PLS dimensions are represented simultaneously, each graphic contains six calibration curves representing models retaining 1 to 3 functional groups for the first and second component (solid and dashed lines, left and right axes, respectively). For each component, the observed minimum value and the chosen value of the model parameters are represented by a cross and a vertical line, respectively; colours are given by the number of groups yielding the lowest ER. Calibration of the second component was conducted retaining in the previous dimension the chosen number of functional groups and α_1.
ER: Error Rate; sgPLS-DA: sparse group Partial Least Squares-Discriminant Analysis.

Figure C.3: Average overall misclassification ER from the calibration procedure of the sPLS-DA model for the two dimensions and for the five study populations (transcriptomics dataset).



In each graphic, calibration curves represent models fitted with 2 to 1000 transcripts (grid resolution of 40 values) to retain in the predictor matrix and the two dimensions are shown simultaneously. For each component, the observed minimum value and the chosen value to retain in the model are represented by a cross and a dashed vertical line, respectively. Calibration of the second component was conducted retaining in the previous dimension the chosen number of variables.
ER: Error Rate; sPLS-DA: sparse Partial Least Squares-Discriminant Analysis.

Figure C.4: Average overall misclassification ER from the calibration procedure of the gPLS-DA model for the two dimensions and for the five study populations (transcriptomics dataset).



In each graphic, calibration curves represent models fitted with 1 to 150 biological pathways (grid resolution of 28 values) to retain in the predictor matrix and the two dimensions are shown simultaneously. For each component, the observed minimum value and the chosen value to retain in the model are represented by a cross and a dashed vertical line, respectively. Calibration of the second component was conducted retaining in the previous dimension the chosen number of biological pathways.
ER: Error Rate; gPLS-DA: group Partial Least Squares-Discriminant Analysis.

Figure C.5: Average overall misclassification ER from the calibration procedure of the sgPLS-DA model for the two dimensions and for the five study populations (transcriptomics dataset).

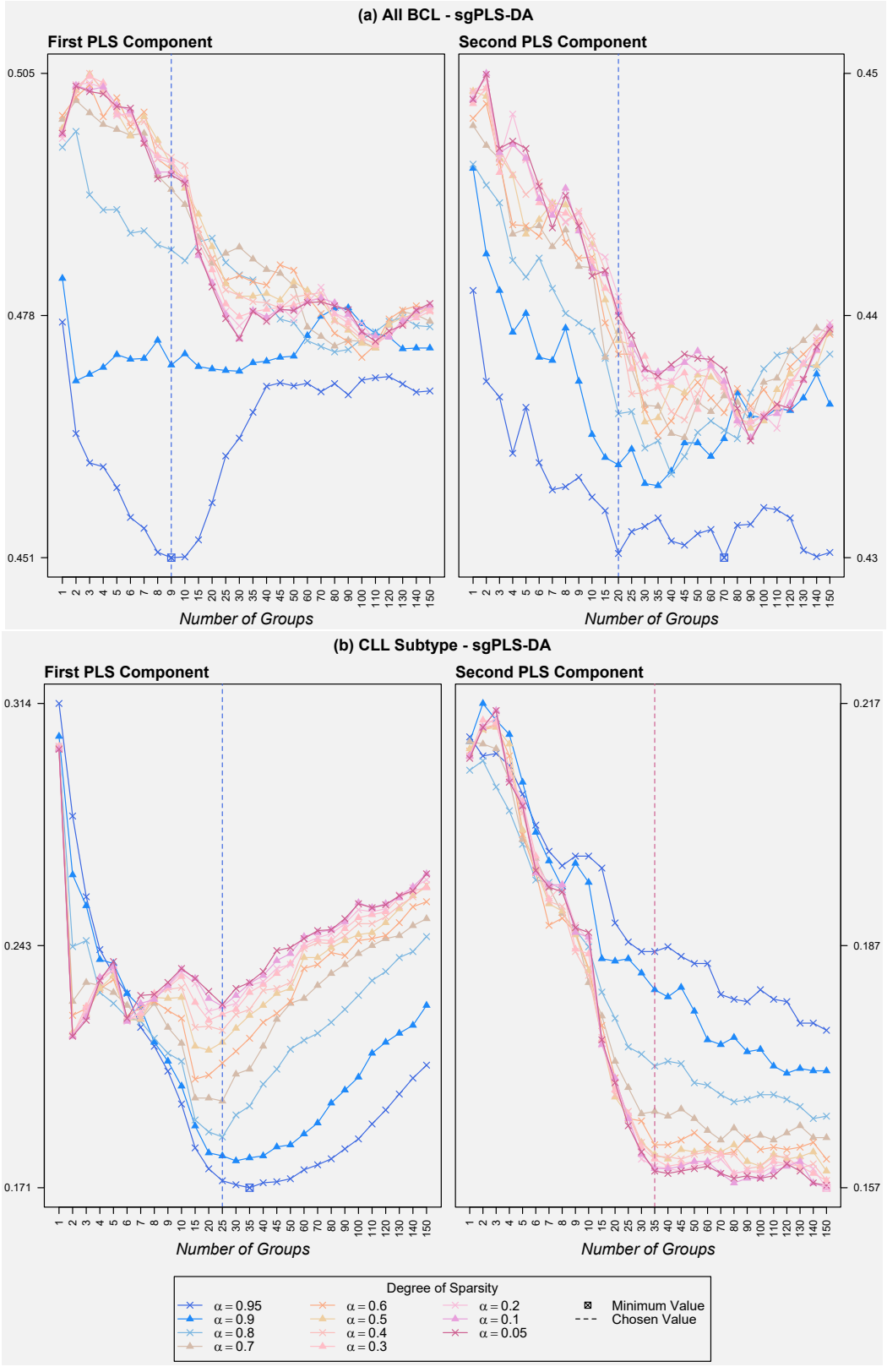


Figure C.5: Average overall misclassification ER from the calibration procedure of the sgPLS-DA model for the two dimensions and for the five study populations (transcriptomics dataset) *(cont.)*.

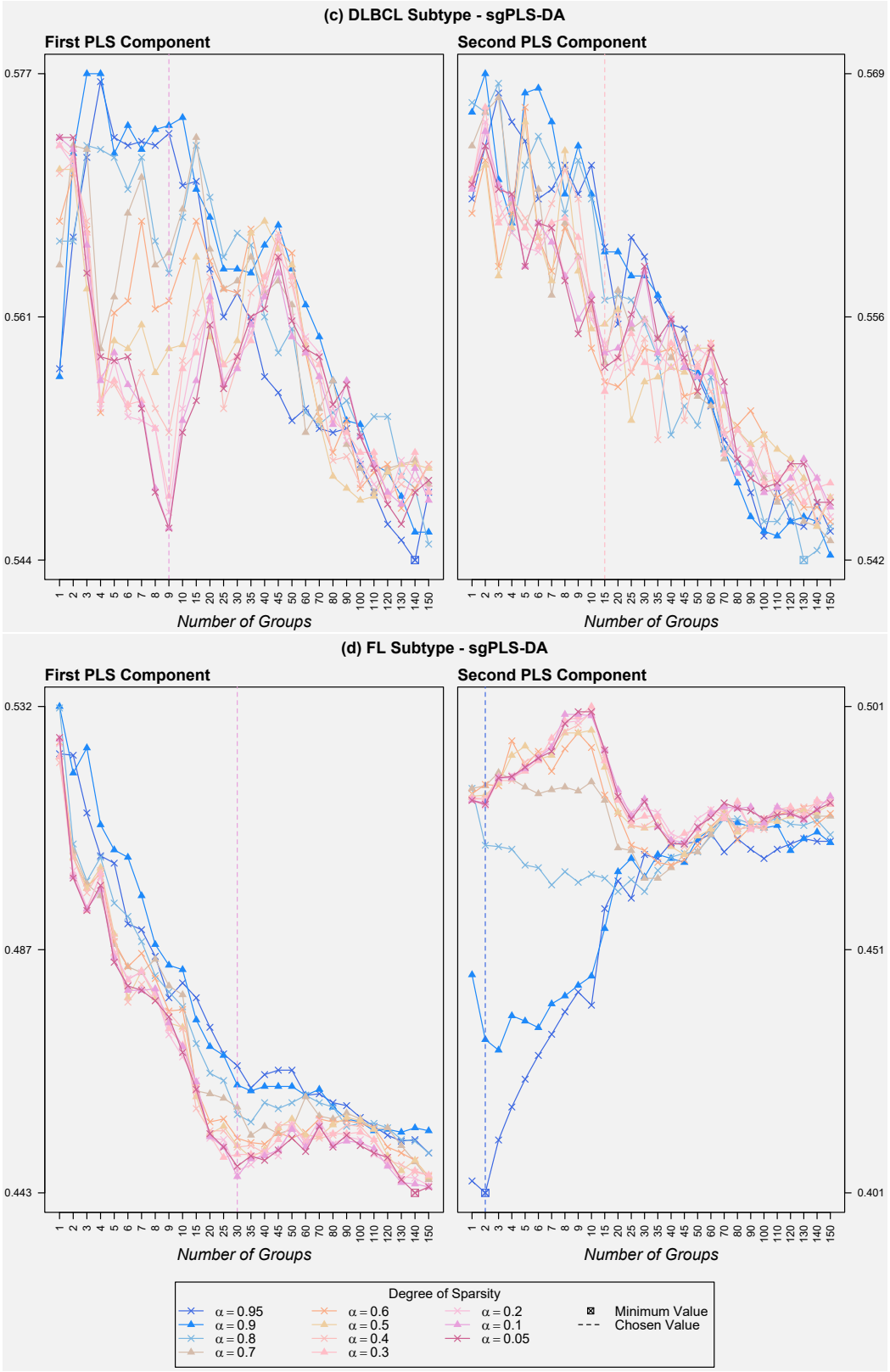
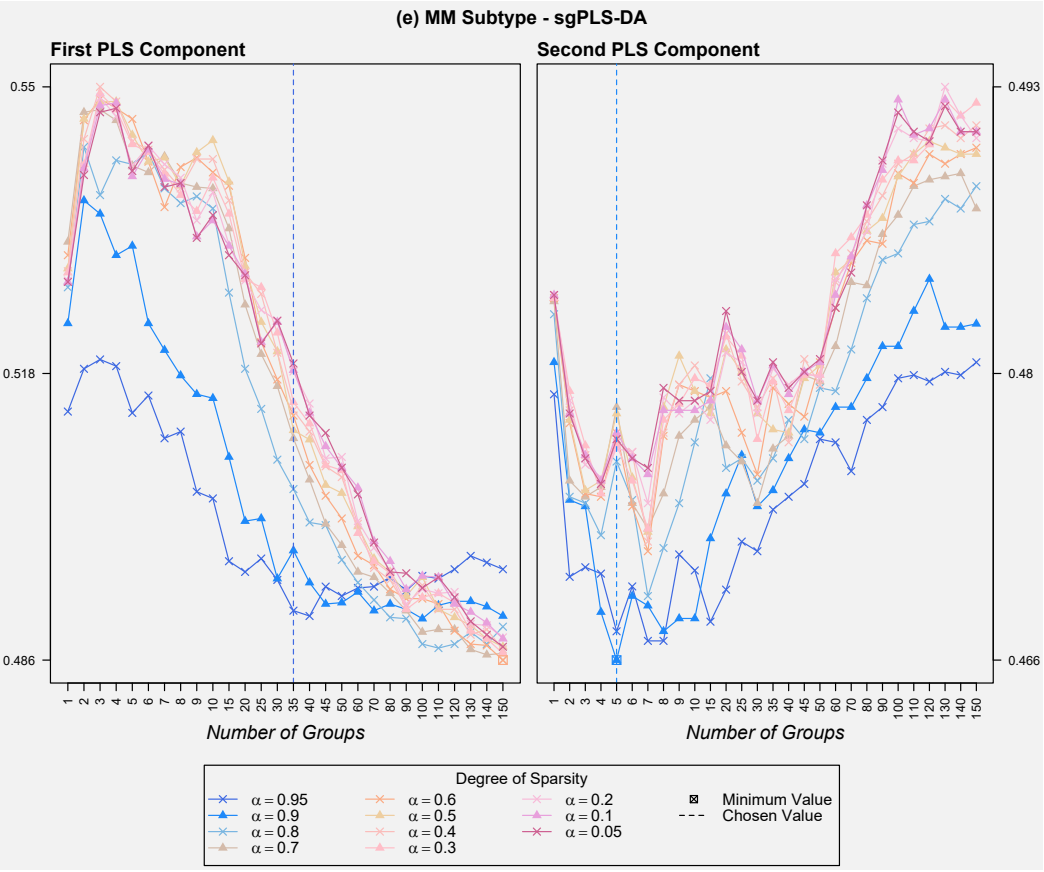


Figure C.5: Average overall misclassification ER from the calibration procedure of the sgPLS-DA model for the two dimensions and for the five study populations (transcriptomics dataset) *(cont.)*.



A single calibration curve represents models fitted with 28 different numbers of biological pathways to retain in the predictor matrix. In each plot, the set of calibration curves represents models fitted with 11 different values of the mixing parameter α_1 defining the within-group sparsity. For each component, the observed minimum value and the chosen value of the model parameters are represented by a cross and a dashed vertical line, respectively; colours are given by the value of α_1 yielding the lowest ER. Calibration of the second component was conducted retaining in the previous dimension the chosen number of biological pathways and α_1.
ER: Error Rate; sgPLS-DA: sparse group Partial Least Squares-Discriminant Analysis.

Supplementary Tables

Table C.1: Most common biological pathways to which individual transcripts were allocated.

	Biological Pathway	*N Probes*	*%*
1	Non-annotated	14737	49.683
2	Regulation of transcription	1254	4.228
3	Transcription	527	1.777
4	Phosphorus metabolic process	426	1.436
5	G-protein coupled receptor protein signalling pathway	274	0.924
6	Cation transport	240	0.809
7	Mitotic cell cycle	236	0.796
8	Macromolecule catabolic process	224	0.755
9	Via transesterification reactions	198	0.668
10	DNA-dependent	184	0.620

Of the 29,662 transcripts, 14,925 have information related to biological pathways and were grouped into a total of 849 different modules. The top 10 pathways containing the highest number of probes are reported.

Table C.2: Percentage of variance explained in the outcome matrix (R^2) and mean and standard deviation of the Discriminant Q^2 (DQ^2) statistics of PLS-DA models fitted with one to two dimensions for the five study populations (proteomics dataset).

	R^2		DQ^2	
	1 Component	*2 Components*	*1 Component*	*2 Components*
All BCL	1.994	4.030	0.304 (0.009)	0.248 (0.007)
CLL	11.530	17.080	0.444 (0.060)	0.273 (0.061)
DLBCL	4.428	6.877	0.365 (0.076)	0.243 (0.036)
FL	2.451	6.979	0.444 (0.055)	0.365 (0.040)
MM	3.603	8.103	0.463 (0.016)	0.316 (0.029)

Results of the calibration procedure for the number of components. R^2 refers to the cumulative percentage of $\mathbf{Y}$ variance explained by the $\mathbf{X}$ components.
PLS-DA: Partial Least Squares-Discriminant Analysis.

Table C.3: Average overall missclassification ER from the calibration procedure of the gPLS-DA model for the two dimensions and for the five study populations (proteomics dataset).

	1st Component			*2nd Component*		
	N Group=1	**N Group=2**	**N Group=3**	**N Group=1**	**N Group=2**	**N Group=3**
All BCL	0.445	0.435	0.414	0.401	0.407	0.405
CLL	0.463	0.462	0.456	0.419	0.409	0.408
DLBCL	0.525	0.535	0.525	0.515	0.506	0.505
FL	0.534	0.549	0.550	0.549	0.560	0.581
MM	0.443	0.438	0.421	0.415	0.433	0.445

Calibration of the second component was conducted retaining in the previous dimension the number of functional groups yielding the minimum ER (see Table 5.1).
ER: Error Rate; gPLS-DA: group Partial Least Squares-Discriminant Analysis.

Table C.4: Classification performances of the three calibrated PLS-DA models for the BCL pooled populations excluding CLL and MM observations (proteomics dataset).

	sparse PLS-DA				*group PLS-DA*				*sparse group PLS-DA*			
	Overall ER	**ER Controls**	**ER Cases**	**AUC**	**Overall ER**	**ER Controls**	**ER Cases**	**AUC**	**Overall ER**	**ER Controls**	**ER Cases**	**AUC**
All BCL	0.406	0.419	0.393	0.594	0.414	0.426	0.402	0.587	0.394	0.402	0.387	0.606
w/o CLL	0.429	0.428	0.431	0.572	0.424	0.430	0.418	0.577	0.407	0.403	0.411	0.593
w/o MM	0.398	0.403	0.392	0.604	0.424	0.428	0.420	0.578	0.398	0.401	0.394	0.603

The classification performance of the population pooling all BCL case-control pairs is shown to facilitate comparison. These results are part of the sensitivity analysis process.
PLS-DA: Partial Least Squares-Discriminant Analysis, ER: Error Rate, AUC: Area Under the Curve.

Table C.5: Classification performances of the three calibrated PLS-DA models for the for the five study populations conducted on the proteomics dataset after correction for WBC populations.

	sparse PLS-DA				*group PLS-DA*				*sparse group PLS-DA*			
	Overall ER	**ER Control**	**ER Cases**	**AUC**	**Overall ER**	**ER Control**	**ER Cases**	**AUC**	**Overall ER**	**ER Control**	**ER Cases**	**AUC**
All BCL	0.413	0.427	0.399	0.588	0.403	0.422	0.383	0.597	0.380	0.400	0.360	0.620
CLL	0.348	0.413	0.284	0.685	0.424	0.460	0.388	0.647	0.319	0.338	0.300	0.721
DLBCL	0.500	0.491	0.509	0.585	0.473	0.474	0.472	0.594	0.477	0.472	0.483	0.591
FL	0.500	0.460	0.540	0.580	0.538	0.539	0.538	0.585	0.483	0.447	0.518	0.587
MM	0.412	0.451	0.372	0.599	0.415	0.469	0.360	0.593	0.417	0.459	0.376	0.594

These results are part of the sensitivity analysis process.
PLS-DA: Partial Least Square-Discriminant Analysis, WBC: White Blood Cells; ER: Error Rate; AUC: Area Under the Curve.

Table C.6: Percentage of variance explained in the outcome matrix (R^2) and mean and standard deviation of the Discriminant Q^2 (DQ^2) statistics of PLS-DA models fitted with one to two dimensions for the five study populations (transcriptomics dataset).

	R^2		DQ^2	
	1 component	***2 components***	***1 component***	***2 components***
All BCL	4.206	7.858	0.365 (0.034)	0.267 (0.005)
CLL	23.441	33.964	0.547 (0.051)	-0.029 (0.099)
DLBCL	9.447	20.231	0.405 (0.079)	0.459 (0.022)
FL	5.801	20.632	0.454 (0.022)	0.458 (0.040)
MM	2.888	17.291	0.419 (0.032)	0.559 (0.023)

Results of the calibration procedure for the number of components. R^2 refers to the cumulative percentage of $\mathbf{Y}$ variance explained by the $\mathbf{X}$ components.
PLS-DA: Partial Least Squares-Discriminant Analysis.

Table C.7: Biological pathways to which the selected gene expression signals belong, total number of probes in those selected pathway and absolute and relative frequencies of the selected signals per pathway for the study population including all BCL case-control pairs and for each of the three regularized approaches.

sparse PLS-DA		
Biological Pathways	**Total**	**Selected (%)**
Non-annotated probes	14737	3 (0.020)
Phosphorus metabolic process	426	1 (0.235)
Cation transport	240	1 (0.417)
Cell activation	57	1 (1.754)
Mesoderm formation	18	1 (5.556)
group PLS-DA		
Biological Pathways	**Total**	
Lymphocyte homeostasis	4	
Regulation of DNA repair	4	
Regulation of Rho protein signal transduction	4	
Acute inflammatory response	3	
Myeloid dendritic cell activation	3	
Positive regulation of neurotransmitter secretion	3	
RNA localization	3	
Telomere maintenance	3	
Water transport	3	
Cell-cell junction assembly	2	
sparse group PLS-DA		
Biological Pathways	**Total**	**Selected (%)**
Immune system development	46	7 (15.217)
Cell activation	57	5 (8.772)
Protein amino acid dephosphorylation	21	3 (14.286)
Proline biosynthetic process	9	3 (33.333)
Inflammatory response	31	2 (6.451)
DNA modification	12	2 (16.667)
Kidney development	6	2 (33.333)
Negative regulation of macromolecule biosynthetic process	6	2 (33.333)
Regulation of Rho protein signal transduction	4	2 (50)
Positive regulation of cell death	3	2 (66.667)

For the sparse PLS-DA model, the total number of chosen biological pathways are displayed (five) while for group and sparse group models the first 10 pathways are displayed, which are ordered in terms of number of total probes per module and number of selected probes per module, respectively.
PLS-DA: Partial Least Squares-Discriminant Analysis

Table C.8: Classification performances of the three calibrated PLS-DA models for the BCL pooled populations excluding CLL observations (transcriptomics dataset).

	sparse PLS-DA				*group PLS-DA*				*sparse group PLS-DA*			
	Overall ER	**ER Control**	**ER Cases**	**AUC**	**Overall ER**	**ER Control**	**ER Cases**	**AUC**	**Overall ER**	**ER Control**	**ER Cases**	**AUC**
All BCL	0.400	0.259	0.541	0.601	0.463	0.403	0.524	0.541	0.430	0.346	0.513	0.572
w/o CLL	0.417	0.396	0.438	0.584	0.496	0.472	0.519	0.530	0.447	0.439	0.455	0.557

The classification performance of the population pooling all BCL case-control pairs is shown to facilitate comparison. These results are part of the sensitivity analysis process.
PLS-DA: Partial Least Squares-Discriminant Analysis, ER: Error Rate; AUC: Area Under the Curve.

Table C.9: Classification performances of the three calibrated PLS-DA models for the five study populations conducted on the transcriptomics dataset after correction for WBC populations.

	sparse PLS-DA				*group PLS-DA*				*sparse group PLS-DA*			
	Overall ER	**ER Control**	**ER Cases**	**AUC**	**Overall ER**	**ER Control**	**ER Cases**	**AUC**	**Overall ER**	**ER Control**	**ER Cases**	**AUC**
All BCL	0.422	0.392	0.451	0.581	0.430	0.414	0.445	0.573	0.418	0.398	0.437	0.585
CLL	0.209	0.196	0.222	0.818	0.467	0.474	0.460	0.625	0.459	0.483	0.436	0.631
DLBCL	0.491	0.509	0.473	0.584	0.506	0.479	0.533	0.592	0.504	0.485	0.523	0.597
FL	0.413	0.415	0.411	0.618	0.482	0.493	0.471	0.592	0.475	0.482	0.468	0.585
MM	0.454	0.436	0.472	0.575	0.458	0.423	0.493	0.568	0.453	0.420	0.486	0.574

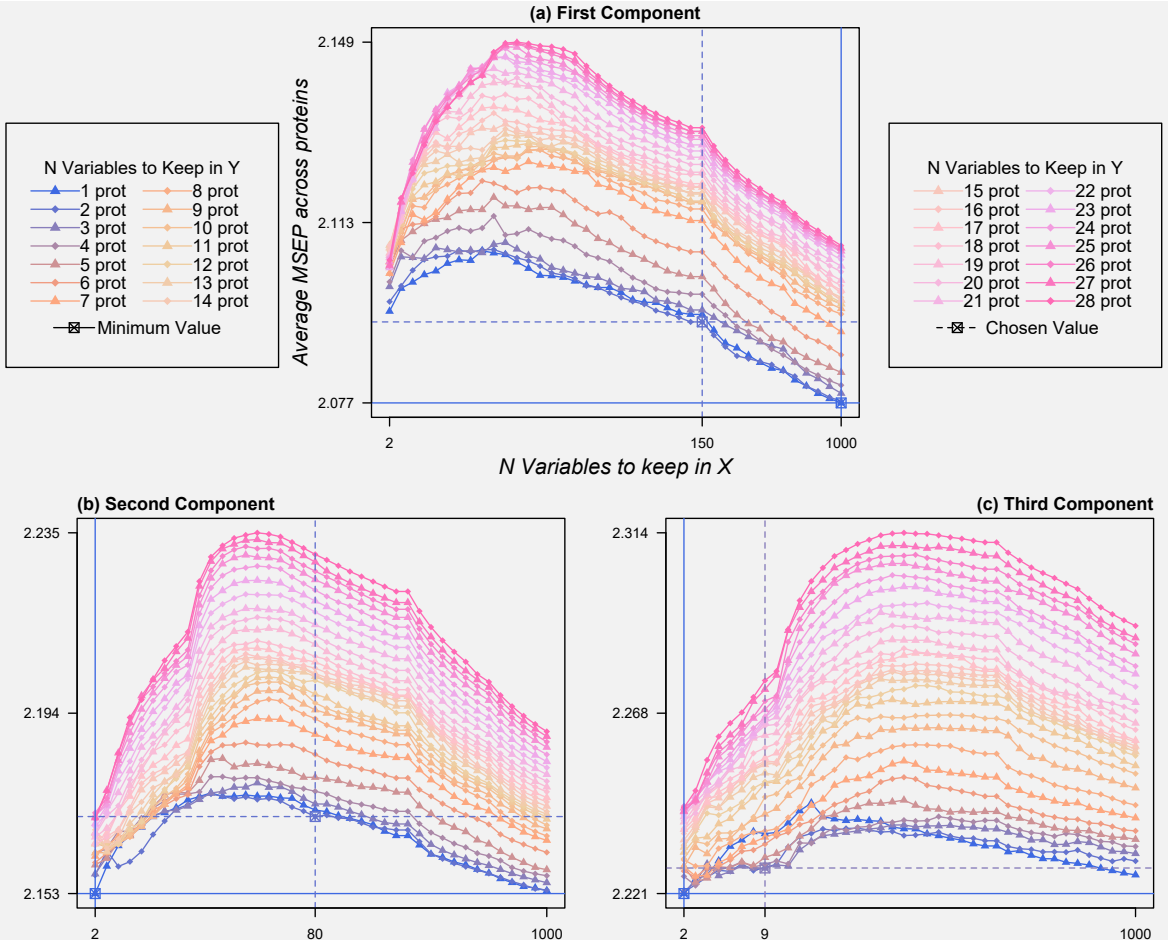
These results are part of the sensitivity analysis process.
PLS-DA: Partial Least Squares-Discriminant Analysis, WBC: White Blood Cells; ER: Error Rate; AUC: Area Under the Curve.

D

Supplementary Material for Chapter 6

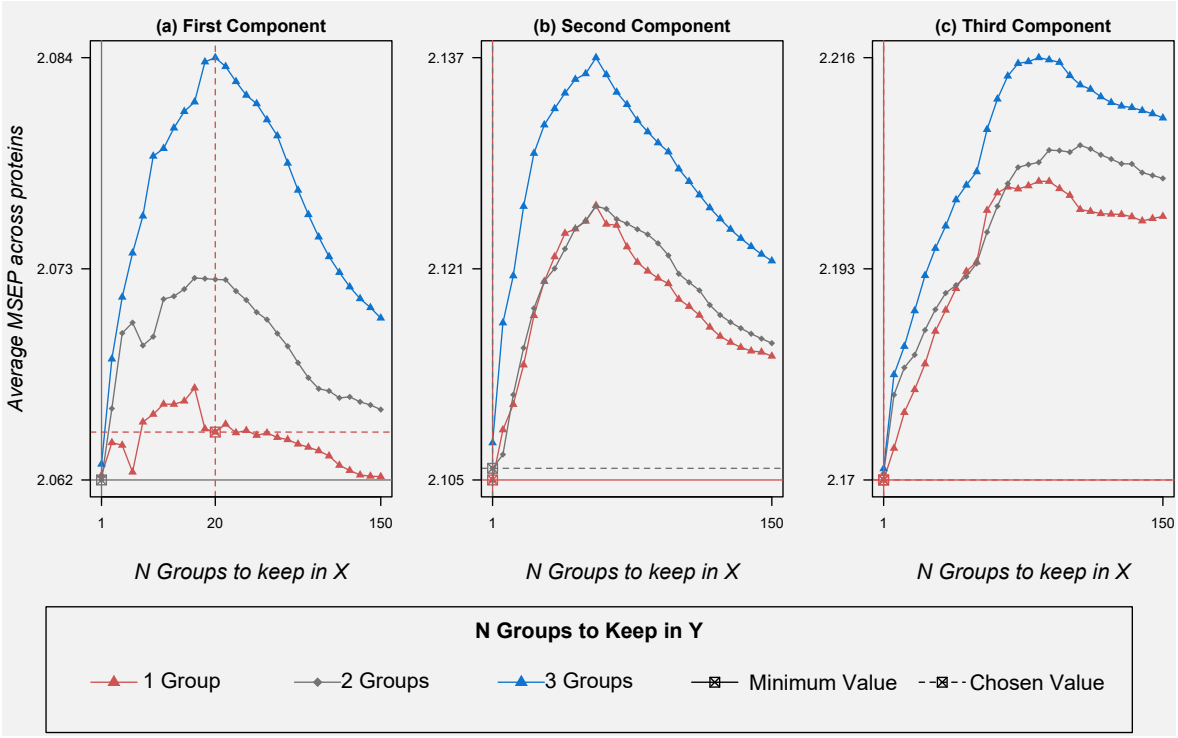
Supplementary Figures

Figure D.1: Average MSEP from the calibration procedure of the sPLS model for each of the three dimensions.



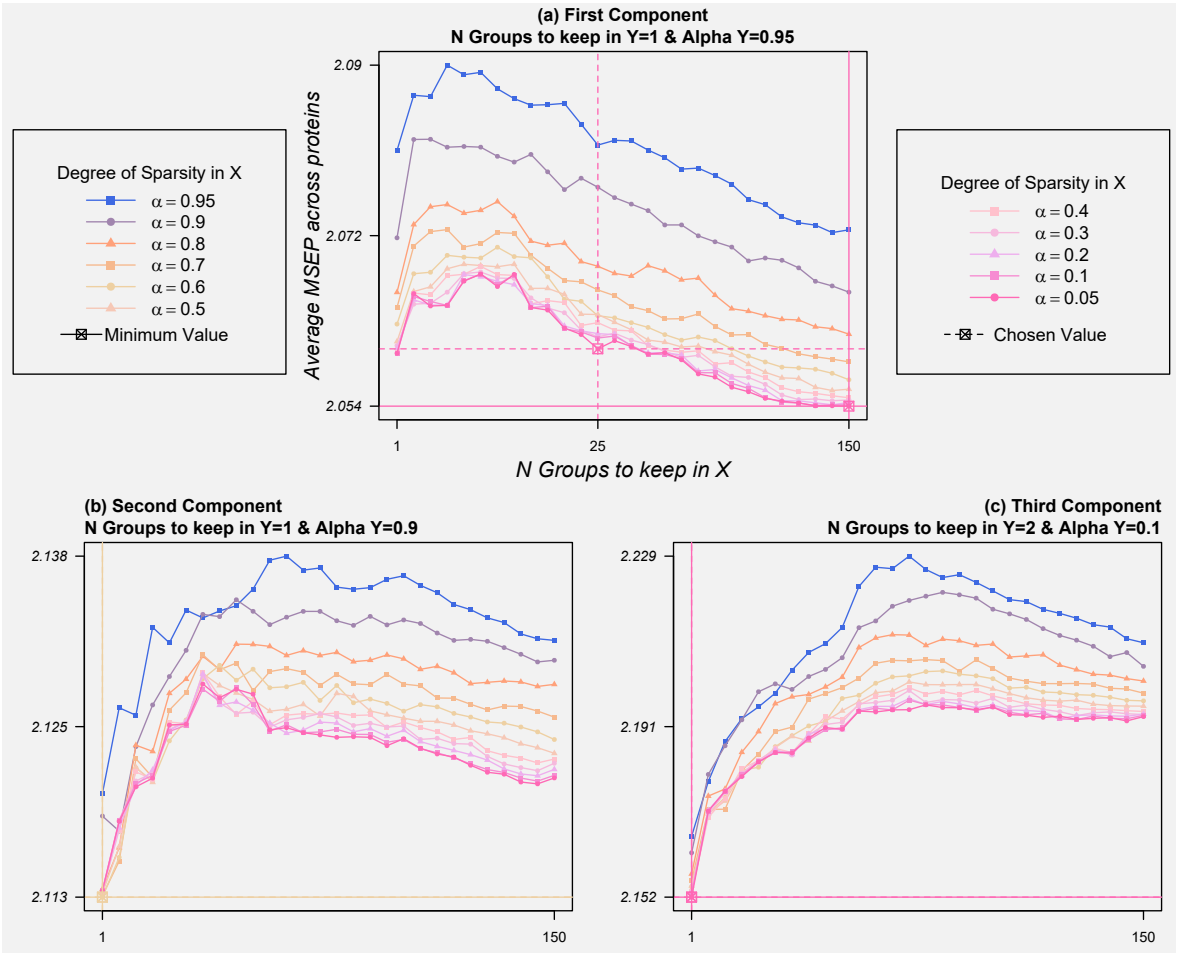
In each graphic, calibration curves represent models fitted with 1 to 28 proteins to retain in the outcome matrix. Solid and dashed lines represent the number of variables yielding the minimum average MSEP and the chosen number of variables (vertical lines), respectively and their corresponding MSEP (horizontal lines). Colour of the lines is dictated by the number of **Y** variables. Calibration of the second and third components was conducted retaining in the previous dimension(s) the chosen number of variables.
MSEP: Mean Squared Error of Prediction; sPLS: sparse Partial Least Squares.

Figure D.2: Average MSEP from the calibration procedure of the gPLS model for each of the three dimensions.



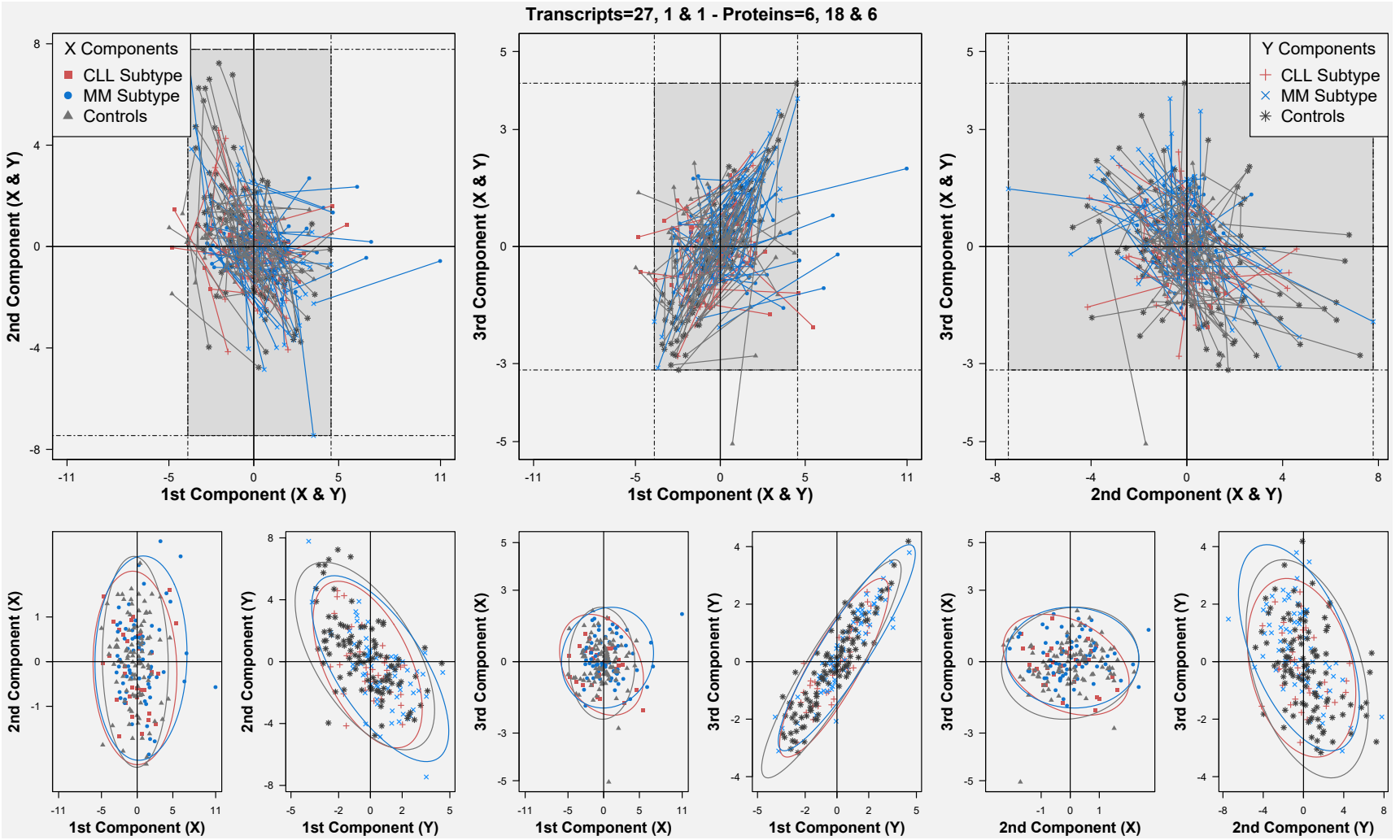
In each graphic, calibration curves represent models fitted with 1 to 3 functional group to retain in the outcome matrix. Solid and dashed lines represent the number of modules yielding the minimum average MSEP and the chosen number of modules (vertical lines), respectively and their corresponding MSEP (horizontal lines). Colour of the lines is dictated by the number of $\mathbf{Y}$ modules. Calibration of the second and third components was conducted retaining in the previous dimension(s) the chosen number of modules.
MSEP: Mean Squared Error of Prediction; gPLS: group Partial Least Squares.

Figure D.3: Average MSEP from the calibration procedure of the sgPLS model for each of the three dimensions.



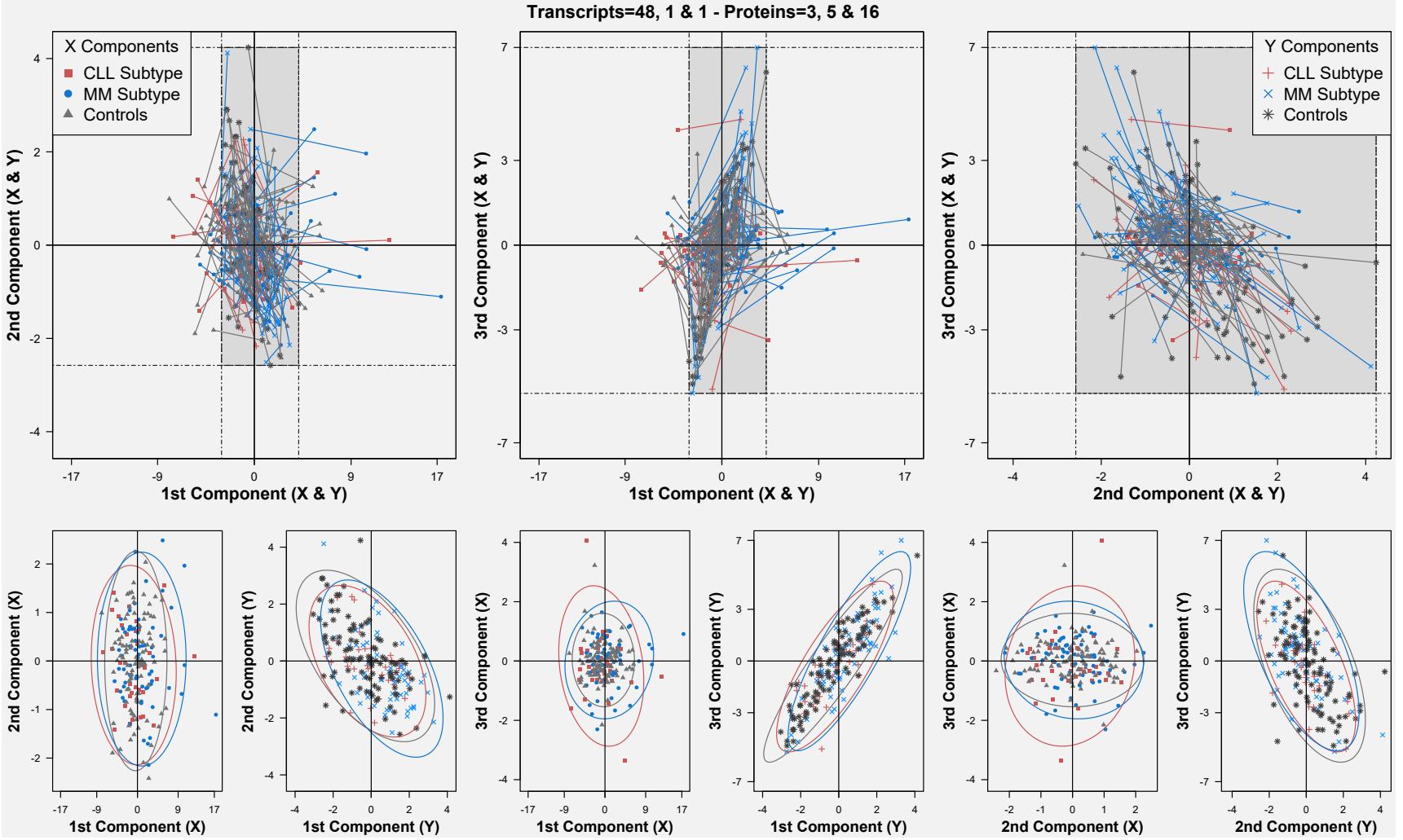
In each graphic, calibration curves represent models fitted with 11 different values of the mixing parameter α_1 defining the degree of sparsity of the predictor matrix. Solid and dashed lines represent the value of the model parameters of $\mathbf{X}$ yielding the minimum MSEP and the chosen value of model parameters of $\mathbf{X}$ (vertical lines), respectively and their corresponding average MSEP (horizontal lines). Colour of the lines is dictated by the value of the mixing parameter α_1. Calibration of the second and third components was conducted retaining in the previous dimension(s) the chosen number of functional groups and the chosen value of α_1. Each panel displays the calibration curves for one value of number of functional groups to keep in $\mathbf{Y}$ and for one value of the mixing parameter α_2 defining the degree of sparsity of the outcome matrix. The values of these model parameters are specified in the title of each panel and were defined based on the minimization of the average MSEP. Thus, the set of calibration curves for the other possible values of number of groups to keep in $\mathbf{Y}$ and the other possible values of α_2 are not shown.
MSEP: Mean Squared Error of Prediction; sgPLS: sparse group Partial Least Squares.

Figure D.4: Sample representation plots displaying the location of the observations on the **X** and **Y** spaces spanned by the calibrated gPLS model (superimposed plots).



The three possible two-dimensional spaces are exhibited. The lower sample space is coloured on grey. The separate sample plots are also displayed. For each sample group, ellipses of the confidence region were drawn employing the R package `ellipse` based on the variance and mean of the matrix of the corresponding pair of components (the mean defines the location of the ellipse centre). The confidence level controlling the ellipse size was 0.95 and a total of 100 points were used.
gPLS: group Partial Least Squares.

Figure D.5: Sample representation plots displaying the location of the observations on the **X** and **Y** spaces spanned by the calibrated sgPLS model (superimposed plots).



The three possible two-dimensional spaces are exhibited. The lower sample space is coloured on grey. The separate sample plots are also displayed. For each sample group, ellipses of the confidence region were drawn employing the R package `ellipse` based on the variance and mean of the matrix of the corresponding pair of components (the mean defines the location of the ellipse centre). The confidence level controlling the ellipse size was 0.95 and a total of 100 points were used.
sgPLS: sparse group Partial Least Square.

Supplementary Tables

Table D.1: Percentage of variance explained in the outcome matrix (R^2) and mean and standard deviation of the Q^2 statistics of PLS models fitted with one to six dimensions.

Model Size	R^2	Q^2
1 component	0.214	-0.023 (-0.009)
2 components	0.858	0.004 (0.004)
3 components	1.594	-0.022 (0.008)
4 components	2.643	0.002 (0.007)
5 components	3.670	0.005 (0.007)
6 component	4.521	0.005 (0.005)

Results of the calibration procedure for the number of components. R^2 refers to the cumulative percentage of **Y** variance explained by the **X** components.
PLS: Partial Least Squares.

Table D.2: Transcripts and biological pathways that are common to at least two of three integrative approaches.

	sparse PLS and group PLS		
	Agilent ID	**Gene Name**	**Biological Pathway**
1	A_23_P26522	AQP8	monocarboxylic acid transport
	sparse PLS and sparse group PLS		
	Agilent ID	**Gene Name**	**Biological Pathway**
1	A_23_P208636	SHANK1	cytoskeletal anchoring at plasma membrane *
2	A_24_P331761	PSG7	female pregnancy *
3	A_23_P26522	AQP8	monocarboxylic acid transport
	group PLS and sparse group PLS		
	Agilent ID	**Gene Name**	**Biological Pathway**
1	A_24_P944640	EPB41L5	axial mesoderm development
2	A_23_P68998	MIOX	carbohydrate catabolic process
3	A_23_P127002	THNSL1	cellular amino acid biosynthetic process
4	A_23_P60016	PTTG3P	chromosome organization
5	A_23_P104323	MGMT	DNA ligation
6	A_23_P77756	GALR2	inositol metabolic process
7	A_23_P8754	AASS	lysine metabolic process
8	A_23_P26522	AQP8	monocarboxylic acid transport
9	A_24_P227927	IL21R	natural killer cell activation
10	A_23_P3274	IGDCC3	neuromuscular process controlling balance
11	A_24_P115183	CLDN4	pathogenesis
12	A_32_P104000	DCUN1D3	regulation of S phase of mitotic cell cycle
13	A_23_P104819	TREH	trehalose metabolic process
14	A_23_P23966	ZNF488	glial cell differentiation
15	A_24_P398210	ZNF488	glial cell differentiation
16	A_23_P55616	SLC14A1	urea transport
17	A_24_P926507	SLC14A1	urea transport
18	A_23_P11461	UBE2V1	regulation of DNA repair
19	A_23_P218685	UBE2V1	regulation of DNA repair
20	A_23_P501996	UBE2V1	regulation of DNA repair
21	A_24_P5935	UBE2V1	regulation of DNA repair
	Across 3 models		
	Agilent ID	**Gene Name**	**Biological Pathway**
1	A_23_P26522	AQP8	monocarboxylic acid transport

The two pathways marked with (*) correspond to modules containing seven and four transcripts, respectively (in order of appearance). The sPLS model selected one transcript in each of those while the sgPLS selected all signals. For the rest of the pathways, the totality of the transcripts that belong to that module were retained in the models being compared.
PLS: Partial Least Squares