Alma Mater Studiorum - Università di Bologna

DOTTORATO DI RICERCA IN

FISICA

Ciclo 35

Settore Concorsuale: 02/D1 - FISICA APPLICATA, DIDATTICA E STORIA DELLA FISICA

Settore Scientifico Disciplinare: FIS/07 - FISICA APPLICATA A BENI CULTURALI, AMBIENTALI, BIOLOGIA E MEDICINA

DEVELOPMENT OF MACHINE LEARNING METHODS FOR MULTI-MODAL BIOMARKERS DETECTION AND INTEGRATION

Presentata da: Daniele Dall'Olio

Coordinatore Dottorato	**Supervisore**
Michele Cicoli	Gastone Castellani

Esame finale anno 2023

Abstract

Innovation comes through understanding. In medicine, innovation depends on a better knowledge of the mechanism of the human body, which represents a complex system of multi-scale constituents. The health of individuals stems from the maintenace of organizing biological principles at any scale, from the molecular level to the functional level. Diseases manifest a perturbation of the organizing principles which damages the healthy circuit of life. Unraveling the complexity underneath diseases proves to be challenging. A deep understanding of the inner workings comes with dealing with many heterogeneous information. Fortunately, the outstanding advancement in biotechnology boosted the availability of massive amount of heterogeneous data, which measure the status of the underlying biological organizations. Exploring the molecular status and the organization of genes, proteins, metabolites provides insights on what is driving a disease in all its aspects, from aggressiveness to curability. Molecular constituents, however, are only the building blocks of the whole human body and cannot currently tell the whole story of diseases. This is why nowadays attention is growing towards the contemporary exploitation of information used in clinics, like imaging data (e.g., biopsy images) and non-invasive analysis (e.g., blood test), and molecular statuses (e.g. genomic alterations). The former are closer to the phenotype and represents the health macroscopic status, whereas the latter examine what is driving the cells to malfunction. Modelling data to reproduce the scale-up from the inner workings of cells to disease phenotypes requires multi-disciplinary involvement. Thus, sophisticated holistic methods are nowadays drawing much interest to address the problem of integrating heterogeneous data. The heterogeneity may derive from the diversity across data types, i.e., multiple types of different natures, and from the diversity within diseases, i.e., variability among individuals. Data integration should then ideally tackle all sources of heterogeneity. Several approaches of data integration were introduced in the last two decades to help dealing with heterogeneous data and this thesis contributes in that direction. Here, four studies conducted data integration using customly designed workflows that implement new methods and views to tackle the heterogeneous characterization of diseases. All studies shared the same motivation: bringing personalized medicine closer to reality. To this end the first study devoted to determine shared gene regulatory signatures for onco-hematology and it showed partial co-regulation across blood-related diseases. The second study considered only one of such hematological malignancy, Acute Myeloid Leukemia, and refined the unsupervised integration of genomic mutations and karyotype aberrations, which turned out to better resemble clinical practice. To observe the impact of data integration on additional diseases the third and fourth studies focused respectively on artherosclerosis and breast cancer. Network integration for artherosclerosis demonstrated, as a proof of concept, the impact of network intelligibilty when it comes to model heterogeneous data, which showed to accelerate the identification of new potential pharmaceutical targets. On another note, the fourth study introduced a new method to integrate multiple data types in a unique latent heterogeneous-representation that facilitated the selection of important data types to predict the tumour stage of invasive ductal carcinoma. The results of these four studies laid the groundwork to ease the detection of new biomarkers ultimately beneficial to medical practice and to the ever-growing field of Personalized Medicine.

Contents

1 Introduction 2
1.1 Omics and holistic complexity 4
1.2 Multi-omics and data integration 5
1.3 Network support for integration 6

2 Transcription Factors Bi-Clustering for Knowledge Discovery on Onco-Hematological Diseases 7
2.1 Introduction 7
2.2 Methods 8
2.2.1 Dataset pre-processing 8
2.2.2 Negative controls estimation 9
2.2.3 Batch correction 10
2.2.4 Gene summarization 10
2.2.5 Regulatory Networks generation 11
2.2.6 Adjusted δ-trimax for bi-clustering 11
2.2.7 Signature extraction 14
2.3 Results 14
2.3.1 Controlling systematic unwanted effects 14
2.3.2 Clustering diseases hierarchically 14
2.3.3 Bi-clusters detection 14
2.3.4 Biological signatures 15
2.3.5 Diseases similarity across TFs pools 17
2.4 Discussion 18

3 Automatic Molecular Driver Identification and Classification of Acute Myeloid Leukemia via Non-Central Hypergeometric Refinement 23
3.1 Introduction 23
3.2 Methods 24
3.2.1 Database 24
3.2.2 Pre-Processing 24
3.2.3 Imputation 26

3.2.4 AML as a Hierachical Dirichlet Mixture Model 27
3.2.5 HDMM fit and imputation control 29
3.2.6 Convergence of single HDP chains 30
3.2.7 Convergence across all DP chains 30
3.2.8 Multinomial parameters estimation 32
3.2.9 Non-Central Hypergeometric Refinement 33
3.2.10 Automatic molecular assignment 33
3.3 Results 34
3.3.1 AML components 34
3.3.2 Correspondence between automatic classification and clinical classifications 34
3.3.3 The case of t(6;9) and t(15;17) 37
3.3.4 Consistency of automatic classifiers on survival predictions 37
3.3.5 Survival predictions over ambiguous and un-classified subjects . . 39
3.4 Discussion 40

4 Empowering In-depth definition of Simvastatin and Ezetimibe Effects in Humans by Intelligible Heterogeneous Networks 44
4.1 Introduction 44
4.2 Methods 46
4.2.1 Dataset pre-processing 46
4.2.2 Human protein-protein interaction network 46
4.2.3 Gene Expression Profiling 46
4.2.4 Genes – biochemical parameters association 47
4.2.5 DNA methylation analysis 47
4.2.6 Building treatment-specific heterogeneous networks 48
4.2.7 Gene prioritization and gene module analysis 48
4.2.8 Disease classes exploration and GO enrichment analysis 49
4.3 Results 49
4.3.1 Treatment-specific heterogeneous networks 49
4.3.2 Prioritizing genes by different types 50
4.3.3 Identification of treatment-unique modules 55
4.3.4 Experimental validation of putative target genes 57
4.4 Discussion 59

5 Multi-modal integration via infomax-trained Neural Networks 65
5.1 Introduction 65
5.2 Methods 67
5.2.1 Data filtering 67
5.2.2 Data selection 67
5.2.3 Modalities pre-processing 68

5.2.4 Tumour stage classifier . . . 69
5.2.5 Multi-modal integration method . . . 69
5.2.6 Training . . . 74
5.2.7 Testing on simulated data . . . 75
5.3 Results . . . 76
5.3.1 TCGA breast cancer data . . . 76
5.3.2 Simulated data . . . 76
5.4 Discussion . . . 78

6 Methodologies 82
6.1 Methodology behind Transcription Factors Bi-Clustering . . . 82
6.1.1 RMA . . . 82
6.1.2 fRMA . . . 85
6.1.3 Microarray quality control . . . 87
6.1.4 Removal of Unwanted Variation . . . 88
6.1.5 Estimating negative controls . . . 90
6.1.6 PANDA . . . 90
6.1.7 Bi-clustering and tri-clustering with δ-max . . . 92
6.2 Methodology behind Automatic Molecular Driver Identification . . . 95
6.2.1 Multiple Imputation by Chained Equations . . . 95
6.2.2 Dirichlet Processes . . . 96
6.2.3 Multivariate non-central hypergeometric distributions . . . 100
6.2.4 Survival Analysis . . . 100
6.3 Methodology behind Intelligible Heterogeneous Networks . . . 103
6.3.1 RNA-seq quantification . . . 103
6.3.2 Modelling RNA-seq with DESeq . . . 104
6.3.3 Combining p-values . . . 105
6.3.4 Network Centralities . . . 106
6.3.5 Network Propagation . . . 107
6.4 Methodology behind infomax-based multi-modal integration . . . 108
6.4.1 Convolutional Neural Networks . . . 108
6.4.2 DenseNet . . . 109
6.4.3 Attention Modules . . . 110
6.4.4 LassoNet . . . 110
6.4.5 Mutual Information . . . 113
6.4.6 Deep InfoMax . . . 115

7 Conclusion 117

A Supplementary for Transcription Factors Bi-Clustering 119
A.1 List of GEO datasets 119
A.2 Systematic effects after RMA 119
A.3 Bi-clusters overlap 119
A.4 Bi-clusters according to partitions 121

B Supplementary for Automatic Molecular Driver Identification 127
B.1 Potential available molecular variables 127
B.2 AML components details 130
B.3 Known classifications of AML 131
B.4 Automatic classifiers survival curves 132

C Supplementary for Intelligible Heterogeneous Networks 140
C.1 Biologically known seeds 140
C.2 Identification of treatment-related differentially expressed genes 140
C.3 Association of clinical parameters to the treatments 141
C.4 Detection of differentially methylated regions depending on treatment . . 142

D Supplementary for infomax-based multi-modal integration 144
D.1 Attention Module 144
D.2 Optimal Network for WSI 144
D.3 Extended-LassoNet fine tuning 145
D.4 Integrated simulated signature with noise 146

Chapter 1

Introduction

Approximately thirty years ago the international scientific community set an ambitious challenge: to completely sequence the human DNA. To achieve a breakthrough of such magnitude, numerous scientists with different background and competences committed to this challenge, whose name was the Human Genome Project (HGP) [1]. Experts in biological fields, like biologists and genetists, joined forces with physicists, mathematicians and other professionals to design and realize a successful program. The necessity of an interdisciplinary team was due to the need of exploring, studying and analysing a complex biological system with advanced technologies, sophisticated techniques and accurate methods. In 2003 the HGP accomplished the sequencing of roughly 85% of the whole human genome[1] and its success was recognized by many. The HGP gauged much interest around biology from the new generation of scientists and laid the groundwork for the biomedical discoveries of the new century [3]. At the project conclusion, though, a practical revolution was still lacking. In fact, the experiments to perform DNA sequencing were expensive and lengthy, which indicated a poor scalability to large masses. The solution did not take long to show and after a few years from the conclusion of the HGP new advanced laboratory techniques were introduced under the name of Next Generation Sequencing (NGS) [4]. High-throughput was then feasible for DNA sequencing and the research community started to acquire and to stock tons of data. As the time went by NGS became progressively cheaper and in the next years the cost for a single genome will drop down $1000 [5], which will make it more affordable for hospitals and healthcare centers. Thereafter, the global scientific community faced challenges similar to the HGP. The research questions were mostly induced by medical demands that urged, once again, a transversal commitment. The need of a interdisciplinary effort was due in particular to the analysis of the huge amount of data produced by the NGS experiments. Competences in modeling, statistics and informatics were crucial, alongside biological knowledge, to take full advantage of the NGS data and deliver relevant answers to the

[1] The lingering gaps in the human DNA sequencing were filled in the following decade until 2021, when the complete sequencing was announced [2].

research questions.

Distinct questions shared the same desire of better understanding the etiology of a severe condition (or disease) to discover innovative strategies to prevent and counterattack it. Following the HGP and the introduction of NGS, the research attention shifted towards genomics and ultimately paid off. Mutations highly associated with diseases, especially tumors, were discovered and led to the concept of target-therapy [6]. Although genomics studies enhanced the molecular characterization of several diseases, they soon showed that the DNA is a single piece of a much broader puzzle. Hence, throughout the last decade, several types of molecular biological entities were explored alternatively to the genome owing to a fast-paced technological advancement.

Each entity portraits a vast collection of biological constituents and its name typically features the suffix *-ome*. In analogy, the suffix *-omics* characterizes research areas focusing on such entities. Nowadays, omics data, i.e., data derived from one of the omics area, are frequently reported in literature and multiple datasets from the same omics are publicly available now. This availability growth offers the opportunity to develop and employ models on huge amount of data (if such datasets are collected accordingly). Furthermore, since many omics data are becoming progressively affordable, the so-called *multi-omics* studies are significantly piling up [7]. Multi-omics researches can ideally complete the biological puzzle by pooling together different entities that are expected to harbour interdependent and complementary processes. Although it is yet to be thoroughly explored, the multi-omics field is expected to thrive if the combination of single-omics data truly captures their organic whole.

Therefore, today, the main interdisciplinary urgency in the biomedical area calls on physicists, mathematicians and data professionals to develop, apply and deploy methods that can successfully perform data integration on omics and multi-omics data. Driven by the motivation of providing practical insights to clinicians and biomedical personnel, this thesis deals with four different data integration works carrying potential medical and pharmaceutical implications. The widespread leading intuition of these works, commonly promoted by the biomedical community, is that at a certain time point a complex biological internal change occurs and produces a dysfunctional and non self-reparable defect that causes either a condition or a disease. Efforts were then devoted to unveil biomarkers for such internal change.

The herein thesis reports four studies focused on data integration that could bring pharmaceutical and clinical insights. This introductory chapter covers essential omics and biological concepts in the first section and it briefly surveys the current state-of-the-art of data integration in the second section. Each of the next four chapters describes the body of work behind one study in all its aspects. Besides, additional detailed methodology and information are provided for all studies respectively in the methods chapter and in the supplementary material.

1.1 Omics and holistic complexity

As mentioned, the simple suffixes -ome and -omics currently trend in many research field. Their origin is believed to come from either Greek or Sanskrit, and they refer to the concept of wholeness [8]. Thus, in biology, many words endowed with the -ome suffix are introduced to better express an entity in its totality, from its building blocks to the phenomena it rules over. Similarly, -omics ending terms relate to the complete family of studies and researches about a specific entity.

To understand the role of each existing omics field, a broad description of cells inner mechanism is necessary. As first stated by Crick in 1954, the DNA is the fundamental constituent of human life and encodes the basic instructions for the human organism. The long double-helix structure of the DNA contains sequences of nucleotides that consist of sugar molecules binded to a phosphate group and either one between 4 nucleobases: adenine and guanine, known as the purines, alongside cytosine and thymine, known as the pyrimidines. Two joint strands of nucleotides form the DNA. Namely, each purine binds to a pyrimidine (adenine−thymine and guanine−cytosine) and their bounds keep together the double helix. To pass on the information contained within a DNA region, i.e., a partial DNA sequence, an enzyme known as RNA polymerase temporarily splits the double helix, reads one strand and reproduces the corresponding other[2]. This transcription process ends up with an RNA strand, known as messenger RNA (mRNA), that is specular to the DNA region processed by the RNA polymerase, usually referred to as gene. It follows that the so-called genome is the collection of genes in the DNA. The mRNA strand is then used to code a protein; yet, not all its sequences partecipate to the translation process. In fact, only some regions within the mRNA (exons) are retained and others (introns) are removed. The RNA transcript that actually builds a protein is then the concatenation of exons. The transcriptome represents, consequently, the collection of the RNA transcripts as well as the proteome accounts for all the overall proteins. In short, genes are the core entities of the DNA and mRNA transcripts process their sequences in order to eventually translate them to proteins [9]. Additionally, the transcription process is affected by a range of environmental phenomena that are commonly referred to as epigenetics [10]. The binding of the RNA polymerase to the DNA sequence is affected by phenomena that change either the biochemistry (e.g., methylation) or organization (e.g., histone modifications) of the double-helix. As for the RNA transcription, such changes are catalyzed by enzymes and are responsible of making genes accessible or inaccessible to the RNA polymerase, which results in active and inactive genes. Yet, all enzymes are proteins, which leads to the mind-bending knowledge that the translation of proteins depends on proteins themselves. Except being catalyzing enzymes, proteins carry out many different functions for a cell, like signaling, and they cover most of a

[2] Differently, though, RNA polymerase replaces thymine with another pyrimidine molecule called uracyl.

cell's activity [11, 12]. They also cooperate to build metabolic processes that produce small molecules (e.g., ATP, fatty acid, etc.) providing essential regulation mechanisms for the cell [13, 14]. The metabolome is then the collection of such molecules, known as metabolites.

Genome, transcriptome, epigenome and metabolome organize holistically so that each cell of the human organism has its own functions. That is, genes, RNAs, proteins and metabolites combine in recurrent patterns which results in pathways and motifs. At the same time, pathway and motifs cooperate to form modules of recurrent patterns that specialize a cell function, which exerts according to the cell's organization of all functions [15]. Further, the functionality of a cell's module can depend on other cell's organizations and, on tissue and organs specializations, which scale the landscape of biological interactions up to the interplay between different cell types [16].

In summary, an organism can be depicted as a system of interacting biological processes operating at different scales. At any scale there is no unique or static pattern of biological processes, rather multiple dynamic patterns are expected to interplay [17].

1.2 Multi-omics and data integration

In the last decade reductionist studies, which analyse omics individually, showed limitations [18–21]. In turn, thanks to the massive hetereogeneous data provided by new modern biotechnologies, the recent multi-omics studies promised improvement both in terms of better modelling and in terms of biological comprehension [22, 23]. The field of multi-omics is strongly involved in the topic of data integration. The ability to integrate data from different sources in biology and medicine may indeed capture the true inner interplay between the molecular constituents.

Data integration is a general concept with no specific definition. Yet, three approaches of data integration are commonly acknowledged [24]: early integration, intermediate integration and late integration. The first approach, early integration, takes all input types and simply concatenates them before running any type of model. That is, all types are considered as a unique large source with no emphasis on within-type information. The second approach, intermediate integration, capitalizes on within-type information by integrating the two types halfway through a model. Doing so, the source origin impacts on the integration but the model cross-intersects them at some point before yielding an output. The third approach, late integration, independently models each type and only afterwards combines their outcomes. This last approach privileges the single types and presumes the integration is captured by an aggregation of the transformed sources.

The three mentioned approaches describe when integration should be performed in a model. Alternatively, data integration can be considered not only in terms of when but also in terms of what is actually to be integrated [25]. As a matter of fact, suppose multiple datasets are collected and endowed with multiple heterogeneous types. Integration can then occur in potentially three directions: across datasets, across types and both. The first case of integration, called horizontal integration, aims at joint modelling all datasets per type. This might be necessary for scenarios where data come from different providers (e.g., laboratories, hospitals, research centers, etc.) or were acquired by different technologies. Also, horizontal integration can be used to borrow information from multiple conditions and diseases [26]. The second case, called vertical integration, looks for the integration of data types for the same dataset. This is where multi-modal studies comes down. Lastly, the third case ideally moves through both directions to perform integration in order to account for whole the underlying heterogeneity.

1.3 Network support for integration

Data integration was declined in many fields [27] and, among them, the network field showed to be one of the most promising [28]. A network is a collection of nodes joined by edges that can be directed or undirected. In particular, real networks are known to be scale-free. Scale-free networks feature a degree distribution that follows a power law, where the degree quantifies the number of connections of a node. Conversely, the degree distribution of random networks is described by the Poisson distribution. Real networks can be imagined as the product of evolution: few original essential nodes existed and then the sequential addition of other nodes progresses the structure of the network [29]. Notably, scale-free networks highlight groups of densely connected nodes, usually referred to as hubs. Especially in the recent field of Network Medicine hubs embody essential biological constituents, like essential genes, whose place in networks is central [16].

Graph theory from physics and mathematics offers a versatile framework to implement the heterogeneous and multiscale interconnections of a living organism [17,30–33]. Patterns are well represented by networks and the whole system of interacting biological processes becomes a web of multiple dynamic networks [29]. Networks have been determined both experimentally and computationally at each level of molecular biology and, despite their still incomplete characterization, they are able to provide valuable information. Functional pathways, likewise KEGG [34, 35] and REACTOME [36], are commonly used networks of genes in system's biology researches due to their ability to map single genes information to higher level knowledge. Similarly, gene regulatory networks [37, 38] describe how genes program each other, which potentially can explain high-order mechanisms like disease initiation and patophenotype emergence.

Integration is intuitive when exerted on networks because the graph-like architecture reflects the concept of interactions and interconnections expected in reality. Also Bayesian Networks [39] revealed to be a platform for data integration [40]. These graphical models represents statistical causal relationships between variables and they can adapt to all three data integration approaches.

Chapter 2

Transcription Factors Bi-Clustering for Knowledge Discovery on Onco-Hematological Diseases

2.1 Introduction

Diseases have different etiologies and clinical implications, which reproduces in the urgent need of many distinct pharmaceutical countermeasurements. Discovering therapies with efficient outcomes over multiple diseases is therefore a tough challenge but of groundbreaking potentiality. In order to determine whether a drug can be exploited across more than one disease, a deep knowledge of the molecular biology is crucial. Targeted gene therapies are the most representative example of what can be accomplished by understanding which biological factors characterize a disease.

Genes are the most studied entities and an anomaly on a gene can be detected in different ways. A gene can be mutated, which entails changes in its nucleotides sequence, or differentially expressed, which implies either an over-representation or under-representation of its sequence. Such anomalies indicate a disorder which can potentially cause harmful consequences for the human organism.

Oncology is currently the medical field that most invest on molecular biology research and in the last decades genetics have been the primary focus. The distinct specializations within oncology are at different stages in the actual knowledge and application of techniques related to molecular biology, and among all the specializations hematology shines out. The so-called onco-hematological sector is recognized as the main driver sector in medicine for the usage of the molecular biology in clinics. Indeed, the breakthrough of targeted therapy was achieved with Imatinib, a drug invented to inhibit the fusion protein BCR-ABL for Philadelphia chromosome positive subjects suffering from the Chronic Myeloid Leukemia, which is a rare blood disease [41]. Thus, hematology is at the core of medical innovation.

Onco-hematology can be described generally as the field of blood and blood-related cancers. Several pathologies exist and are classified in: leukemia, lymphoma, myeloma, myelodysplatics syndrome and myeloproliferative diseases. Such pathologies differentiate based on the type of blood cell they affect. Understanding how similar these onco-hematological disorders are and how they can be classified based only on molecular information is the main interest of this study.

Using gene expression levels, hematological disorders were grouped together to extract common biological signatures over transcription factors (TFs), i.e., proteins that promote the transcription of genes. This research is motivated by the ambition of discovering new common biomarkers, in terms of TFs, for well-distinct hematological disorders through the usage of an adjusted bi-clustering algorithm. With the application of such algorithm the aim is also to provide insights to promote drug repurposing.

2.2 Methods

2.2.1 Dataset pre-processing

As a consequence of the leading role of onco-hematology in medicine, there are many publicly available datasets on the Internet. After a thorough assessment on data type and laboratory platform we collected microarray data from 36 datasets available on the Gene Expression Omnibus (GEO). All microarrays were acquired by the Affymetrix Human Genome U133 Plus 2.0 platform [42]. Two blood-related conditions, myelofibrosis (AMM) and monoclonal gammopathy of undetermined significance (MGUS), were apriori excluded upon clinical indications. Besides, three types of lymphoma, i.e., mucosa-associated lymphoid tissue (MALT), splenic marginal zone lymphoma (SMZL) and marginal zone lymphoma (MZL) , were joint together to form a single type of lymphoma (MZLs). Then, healthy subjects were filtered out alongside pediatric subjects, i.e., younger than 18, and exclusively untreated subjects took part in the analyses. No subject had replicas in the dataset.

Quality control and filtering were carried out first via image assessment and then via thresholding on GNUSE values [43] in two consecutive stages (more details in section 6.1.3). First, corrupted images, i.e., with scratches, hazes, unusual spots, were removed upon observation. This stage took out 1563 subjects with the exclusion of an entire dataset (GSE31312) dedicated to DLBCL. Afterwards, the fRMA approach [44] was used to perform background-correction, normalization and summarization. Details can be found in section 6.1.2. Then the second step of filtering removed all subjects with GNUSE values below 1.25 that were 404, along with another whole dataset taken out (GSE79533). This recommended GNUSE threshold indicates that the variability of a subject intensity values could not exceed the size of the first quartile of the median subject intensity distribution. Eventually, a total of 5442 subjects over 34 GSE was considered of good quality. The complete collection of these datasets is reported in Supplementary Table A.1. Totally 13 onco-hematological diseases were covered by the good-quality cohort (Table 2.1).

Disease	Acronym	Subjects
Acute Lymphocytic Leukemia	ALL	730
Acute Myeloid Leukemia	AML	1230
Burkitt Lymphoma	BL	38
Chronic Lymphocytic Leukemia	CLL	862
Chronic Myeloid Leukemia	CML	115
Diffuse Large B-Cell Lymphoma	DLBCL	417
Follicular Lymphoma	FL	452
Hodgkin Lymphoma	HL	98
Mantle Cell Lymphoma	MCL	158
Myelo-Dysplastic Syndromes	MDS	338
Multiple Myeloma	MM	595
Marginal Zone Lymphomas	MZLs	138
Peripheral T-Cell Lymphoma	PTCL	271

Table 2.1: Dataset overview across diseases. Totally 5442 subjects were included in the dataset for a group of 13 onco-hematological diseases.

2.2.2 Negative controls estimation

The Least Variant Set, or LVS, approach was utilized to drive the correction of systematic variability due to the data providers. This approach, section 6.1.5, targets the probesets playing no role with the biological effect and whose variability depends exclusively on the systematic variability. To this end, after the summarization step, it first estimates each probeset subjects variability and each probeset logarithmic standard error of the residuals. Then it performs a quantile regression over these two quantities and it considers as negative controls those probesets below the fitted curve. The main idea is that the least variable set of probesets are the ones below a certain quantile of the subjects variability distribution. Though, such variability distribution necessarily depends on the residuals of each probesets model. Therefore, the quantile regression is employed to account for the change of the quantile value per logarithmic standard error of residuals. In fact, the fitted curve passes through the estimated quantile values.

Both the logarithmic standard error of residuals and the subjects variability was computed for each probeset using the fRMA outcomes, i.e. the weights of the M-estimator, the probes residuals and the subjects espression estimates. Then the 60% quantile was set as target and the quantile regression was run accordingly. In other words, the probesets belonging to the subjects variability distribution below the 60% quantile was determined to be the negative controls.

Negative control probesets were estimated with this approach separately for every batch. Then only the intersection across all batches was kept and potential interesting probesets were filtered out. Eventually, the total estimated negative control probesets were 9381 out of 54675.

2.2.3 Batch correction

The correction of batch effect was performed according to the RUV approach [45]. On the negative control probesets, RUV estimated the batch signal and then removed it from the whole data (more details in section 6.1.4). To establish an optimal number of unwanted factors and an optimal penalty parameter, which controls the amount of total removed signal, a grid search was created. The number of unwanted factors was tested in the interval $[1, 100]$ and for each of these values the penalty parameter was tested for $\{10^x\}_{x\in[0,10]}$. Hence, this grid search was based on a total of 1100 sets of parameters.

The silhoutte score was the metric used to assist in determining optimal parameters. For each correction, data was divided per disease and PCA was run individually for every disease. Batches were considered as clusters and the silhoutte score was set to use the Euclidean distance over the two first principal components space. Then, the silhoutte scores (one for each cluster per disease) were computed and their mean squared was defined as the driving metric for optimization. The parameters with the lowest mean squared silhoutte score were 99 and 1 respectively for the number of unwanted factors and the Ridge parameter. Ultimately, the correction was performed using these parameters.

2.2.4 Gene summarization

The annotation of the Affymetrix platform contributed to define a criterion on how to yield expression values for genes. Given the g-th gene associated with a set of probesets $\rho(g)$ its expression value was determined by:

$$y(g) = \sum_{p\in\rho(g)} w_p x_p \ , \tag{2.1}$$

were x_p is the batch-corrected expression value for the p-th probeset and w_p is its weight. The weights of probesets were estimated as the inverse of the number of genes a probesets is associated with according to the Affymetrix annotation.

2.2.5 Regulatory Networks generation

After gene summarization, the PANDA [46] algorithm was employed to generate regulatory networks. Beforehand, to set up PANDA, the position weight matrices (PWSs) of 1149 transcription factors (TFs) were downloaded from the CIS-BP database [47] and the promoter regions of genes were downloaded from the most recent human genome release provided by UCSC [48]. Next, each TF was scanned along every region by the FIMO tool from the MEME suite [49] to estimate the probability of a motif in the promoters for a TF. These probabilities were then thresholded using value 10^{-5} to yield the initial agnostic binary regulatory network. Moreover, the TFs were mapped onto the STRING Protein Protein Interaction network (PPI) [50] to generate the initial TF-to-TF network. Though, even such network was disease-agnostic. The initial network feeding PANDA with diverse biological input was the gene-to-gene network that is approximated by the genes correlation matrix of each disease. Given these three networks, the PANDA algorithm was able to find an agreement for them (see details in section 6.1.6). To be noted, negative control probesets were converted to negative control genes (4262), and were removed prior to PANDA. Also genes and TFs that did not appear initially in both expression data and regulatory network were excluded.

2.2.6 Adjusted δ-trimax for bi-clustering

The 13 regulatory networks estimated by PANDA were binded together to compose a tri-dimensional object R with size $D \times T \times G$, where D stand for the total number of diseases, T for all TFs and G for the collection of genes. In this work the PANDA outcomes yielded a $13 \times 18315 \times 1010$ object. Then, to form bi-clusters of diseases and TFs, an adaptation of the δ-trimax [51] method was developed (section 6.1.7 reports more details). Since the goal is to discover TFs that behave similarly across diseases, the Square Residue (SR) is formulated as,

$$\mathrm{SR}(r_{dtg}) = \left(r_{dtg} - \frac{1}{D_{bic}} \sum_{d \in D_{bic}} r_{dtg} \right)^2 \tag{2.2}$$

and the mean SR (MSR) is accordingly $\frac{1}{D_{bic}} \frac{1}{T_{bic}} \frac{1}{G_{bic}} \sum_{d \in D_{bic}} \sum_{t \in T_{bic}} \sum_{g \in G_{bic}} \mathrm{SR}(r_{dtg})$. The terms D_{bic}, T_{bic} and G_{bic} respectively indicate the number of diseases, TFs and genes contained by the bic-th bi-cluster. Herein δ-trimax algorithm is adjusted to either delete or add diseases and TFs until a bi-cluster meets the criterion: $\mathrm{MSR}(R_{bic}) < \delta$. Therefore, $G_{bic} = G$ always holds.

To determine bi-clusters the canonic procedure of the δ-trimax algorithm [51] or, more generally, of the CC algorithm [52] was adapted and utilized. This procedure iters four operations to identify bi-clusters one by one: multiple node deletion, single node deletion, node addition and masking. Each of these steps mimics what happens in the two-dimensional version of the algorithm, since the main interest lies only on diseases and TFs. A neat difference between the known methods and the one presented here is the masking technique. Usually, once a bi-cluster is determined, each of its belonging nodes, i.e., diseases and TFs, is masked in the original three-dimensional object to prevent the same identified pattern to be re-clustered or mixed with others. However, random filling out, or replacing as missing values, does not prevent cases where nodes are assigned multiple times, which complicates the post-processing and the understanding of the final bi-clusters. A rough alternative is to stop considering diseases and TFs once they are clustered, which solves the previous problem of masking but could prevent total clustering, i.e., not all the nodes of R are eventually assigned to a bi-cluster.

Here, a new masking procedure, Algorithm 1, is introduced before the four-operations iteration is executed. This procedure is a preparation step applied on the entire matrix that yields a completely clean sub-matrix to be passed on to the standard δ-trimax procedure. Hence, at the end of Algorithm 1, no element (node) of the sub-matrix belongs to a previous determined bi-cluster. Such sub-matrix does not result randomly or roughly by removing all diseases and TFs that were previously involved in some bi-clusters. Instead, it is obtained by schematically deleting the already-assigned nodes with the largest average SR, updating both the average SRs at every deletion. This scheme drives the δ-algorithm to work only on clean sub-matrices of R and, thus eventually, each element of R belongs to a single bi-cluster.

Algorithm 1 Newly introduced propedeutic masking procedure for the adjusted δ-trimax algorithm.

```
procedure REPEAT(until the submatrix is free from already-assigned nodes)
    Compute the average SR for only diseases and TFs with at least one node already assigned
    Remove the node (either disease or TF) with the highest average SR
end procedure
```

Along with the new masking, a last final removal step is added to ease further post-processing steps and manipulation of the final bi-clusters. In fact, even after a bi-cluster has MSR less than δ it can contain nodes (diseases or TFs) with average squared residue greater than δ. This fact may limit to consistently split bi-clusters in post-processing operation, which can be handy when comparing different bi-clusters sharing only a subset of either diseases or TFs. To address this possible issue a refinement procedure (Algorithm 2) that iteratively takes out the nodes with average squared residue higher than δ is run.

With all innovation implemented, the adjusted δ-trimax algorithm for bi-clustering becomes Algorithm 3.

Algorithm 2 Newly introduced node refinement technique for the adjusted δ-trimax algorithm.

```
procedure Repeat(until the submatrix no node (disease/TF) has average SR > $\delta$)
    Compute the average SR for all diseases and TFs in the bi-cluster
    Remove the node with the highest average SR greater than $\delta$
end procedure
```

Algorithm 3 Overview of the adjusted δ-trimax algorithm.

```
Set threshold $\delta$
Set multiple node deletion positive parameter $\alpha$
Set $A = R$, with $D$ and $T$ being the number of diseases and the number of TFs respectively
procedure Repeat(until all nodes have been assigned)
    Preparation of $A$: cleaning $A$ from already-assigned nodes
    Multiple node deletion: removal of $d$s and $t$s with average residue $> \alpha\mathrm{MSR}(A)$
    Single node deletion: iterative removal of a single $d$ or $t$ with the largest average residue until $\mathrm{MSR}(A) < \delta$
    Node addition: addition of $\{d|d \notin D_A\}$ and $\{j|j \notin T_A\}$ with average residue $< \mathrm{MSR}(A)$
    Node refinement: iterative removal of a single $d$ or $t$ with average residue higher than $\delta$
end procedure
```

2.2.7 Signature extraction

All discovered bi-clusters intrinsically carried a biological signature. Indeed, the list of TFs they include expresses which biological entities drove the adjusted δ-trimax method to pool a subset of diseases together. Hence, there was no need of any specific signature extraction algorithm. Nevertheless, the TFs of any bi-clusters could be sorted to point out the TFs with the lowest MSR.

2.3 Results

2.3.1 Controlling systematic unwanted effects

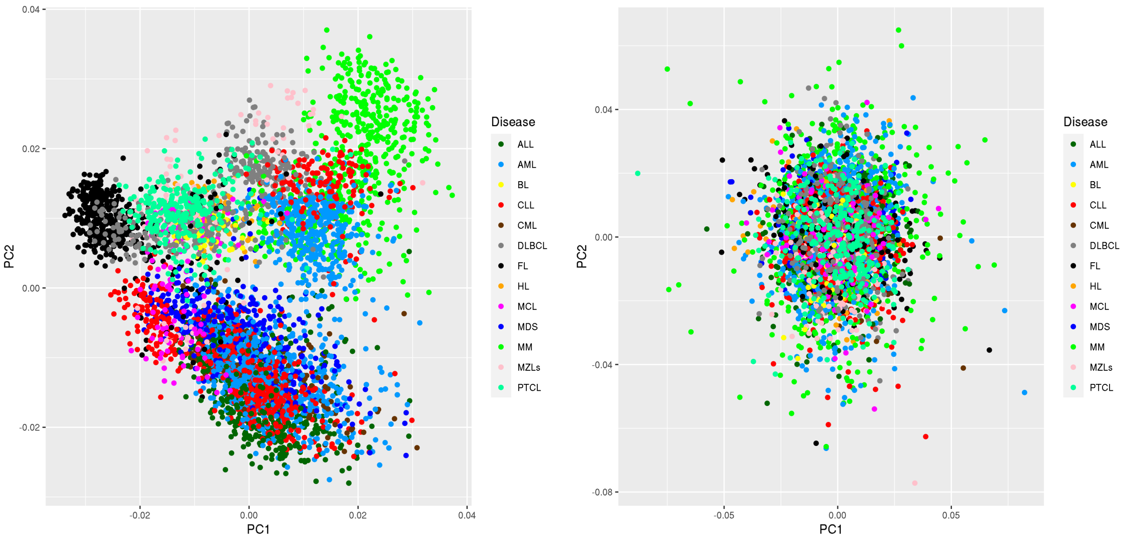
Upon completing the fRMA procedure and having filtered out low-quality subjects, the PCA performed on the whole data scattered the variability of diseases as shown in Figure 2.1a. Diseases generally form overlapping clusters along the first two principal components but, observing the plot individually per disease, systematic effect was clear for AML, CLL, CML, DLBCL, FL, MCL, MDS, MM and MZLs. Figure 2.2a illustrates how the seven batches influence the projection of subjects suffering from MZLs on the two principal components. RUV was then performed to subtract this influence with customly determined optimal parameters. The result over all diseases showed (Figure 2.1b) the lack of clusters previously observable. Plus, Figure 2.2b depicted the representative effect of RUV at the disease level.

2.3.2 Clustering diseases hierarchically

The PANDA algorithm generated a regulatory network estimate for each disease. To first observe how diseases cluster based on the overall regulations determined by PANDA hierarchical clustering was exploited. The result delineated the occurence of two singleton-clusters, BL and FL, three clusters of two diseases (MCL plus PTCL, MZLs plus MM, DLBCL plus HL) and a single larger cluster with the remaining diseases. Figure 2.3 shows that although BL and FL are singletons, the former is included in the hierarchy while the latter is totally sidelined.

2.3.3 Bi-clusters detection

The mathematical entity to perform the bi-clustering on was a three dimensional tensor, whose sizes were: number of diseases (13), number of genes (18315) and number of TFs (1010). Diseases and TFs were chosen to be the output dimensions of any bi-cluster, letting the method use the large number of genes to determine whether diseases are similarly regulated by some common TFs. The similarity criterion of the bi-clustering method depended on the δ parameter, that controls the granularity of the clustering.



(a) (b)

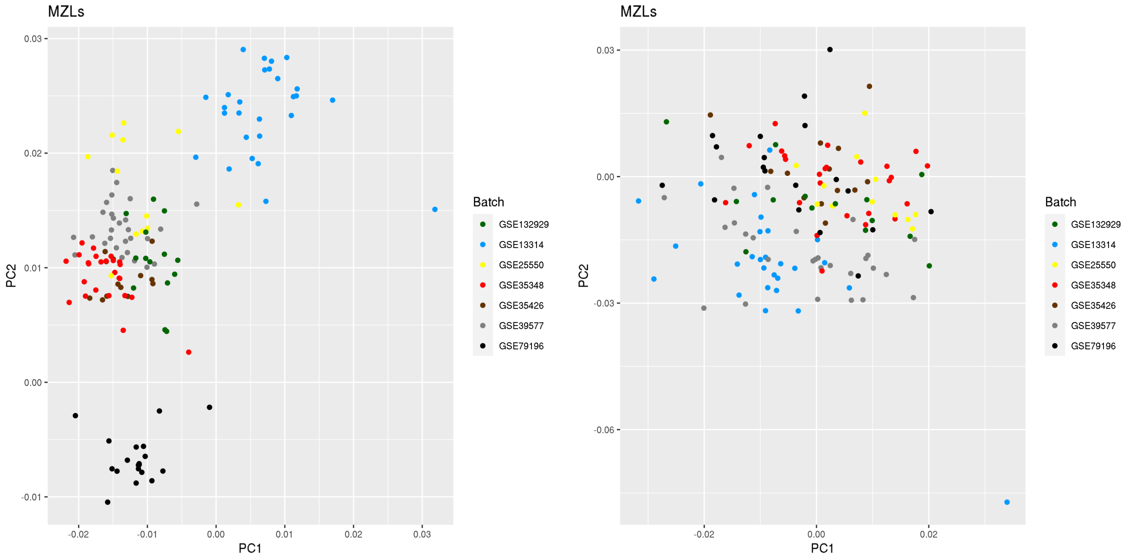
Figure 2.1: Principal components run over all diseases prior to the RUV batch correction (on left) and afterwards (on right).

The selected δ was determined as the largest value forcing the method to generate a number of clusters greater than one. Therefore this δ can be considered as the first level of granulation. With δ set as explained, 25 bi-clusters resulted, rendered in Figure 2.4 and summarized in Table 2.2.

There are no overlap between the 25 bi-clusters, since the method was specifically designed to prevent such scenario. Though, several subsets of diseases occur in more than one bi-clusters. The evaluation of the similarity between a subset of diseases account for every bi-clusters where such subset appear. Doing so, the subset of diseases with the largest number of similar TFs is the pair ALL plus AML, with 971 TFs. All identified subsets of diseases with their associated number of similar TFs are reported in Supplementary Table A.2. Notably, FL remained a singleton on all TFs as well as BL except only for those in the large twelve diseases bi-cluster.

2.3.4 Biological signatures

Out of all 25 bi-clusters, only 12 had more than one disease. Among the disease-singleton bi-clusters AML showed the shortest signature, with 36 TFs, followed by ALL (37), CLL (98), MDS (114), MCL (142), MZLs (146), MM (147), CML (152), PTCL (185), HL (188), DLBCL (188), BL (224) and to end with FL, which had all TFs. Over 35 TFs, the



(a) (b)

Figure 2.2: Projection of subjects suffering from MZLs on the two first principal components, run over all diseases, prior to the RUV batch correction (on left) and afterwards (on right). The effects of the correction on MZLs are representative of what occurs in all the other batch-affected diseases (AML, CLL, CML, DLBCL, FL, MCL, MDS, MM and MZLs).

13 diseases were spread on each of their singleton bi-clusters. According to PCA the top ten edges causing a separation along the first principal component indicated differences in the regulation of genes CYP3A43, NLRP8 and C8A. On the the second principal component the top five edges highlighted differences in regulation of USP47, DHFRL1 and PHF3. Bi-clusters with multiple diseases showed diverse signatures once their TFs were sorted based on MSR. The top three most similar TFs were considered as a short signature. The 12-diseases bi-cluster had TCF23, PPARD and ZBTB49, whereas the bi-cluster of ALL, AML, CLL, CML, MCL, MDS, MM and MZLs had TCF23, RXRB and ETV7. The bi-cluster with ALL, AML, CLL and MDS reported RXRB, PPARD and NR2F1. The following four bi-clusters with three diseases emerged: ALL, AML plus CLL, AML, CLL plus MCL, MCL, MM plus MZLs and DLBCL, HL plus PTCL. Their short signature was respectively NR2F1, ETV7 and RXRB; MBNL2, ZBTB49 and TLX1; ETV7, ZBTB7C and TCF23; TCF23, ZBTB43 and TLX1. Further, MCL plus PTCL reported ZNF778, ETV7 and TLX1 whereas MCL plus MZLs had HES4, ETV7 and TCF23, and MM plus MZLs had ZBTB7C, ETV7 and HES2. The remaining two bi-clusters of two diseases, ALL plus AML and ALL plus MDS highlighted ETV7, RXRB

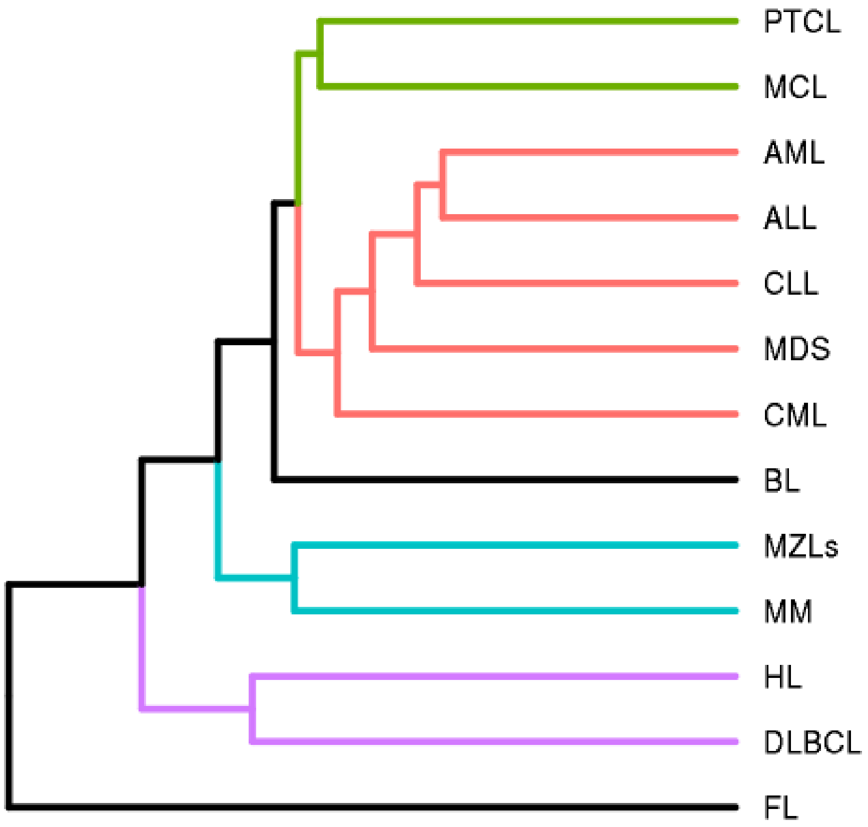


Figure 2.3: Hierarchical clustering based on the overall regulation. The PANDA algorithm estimated regulatory networks for each disease and then Euclidean distance was utilized to calculate the distance between all pairs of diseases.

and NR2F1, and RXRB, PPARD and ZBTB48 respectively. Redundancy occured across short signatures since several genes revealed to be really close in expression.

2.3.5 Diseases similarity across TFs pools

The landscape of diseases similarities was not the same across all TFs. As a matter of fact, diseases organized in 16 different partitions over diverse sets of TFs (Supplementary Figures A.2, A.3, A.4 and A.5). Over 786 TFs the diseases clustered together except for FL. The bi-cluster of eight diseases (ALL, AML, CLL, CML, MCL, MDS, MM and MZLs) occurred alongside either singletons (BL, DLBCL, FL, HL and PTCL) or DLBCL, HL plus PTCL and two singletons (BL and FL). Across the 16 partitions the bi-cluster of ALL, AML, CLL plus MDS appeared five times with the remaining diseases clustered in different ways. Only seven partitions emerged over more than five TFs, whereas some of

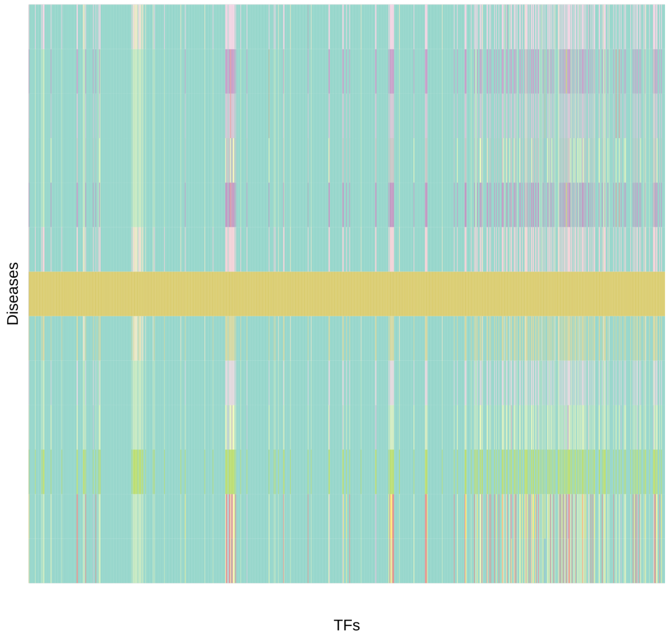


Figure 2.4: Portrait of adjusted δ-trimax result. Colors represent different bi-clusters, for a total of 25. Bi-clusters consists of submatrices whose Mean Square Residue (MSR) is lower than δ. The MSR is calculated with respect to the average TFs values across diseases. Bi-clusters can then be imagined as sub-matrices with common values along diseases for every TF.

the remainings occured only on a single TF, namely for ZNF555, NEUROD1, ZNF449, KFL3 and ZNF35. No more than two non-singleton bi-clusters appeared within the partitions and six of them had only one non-singleton bi-cluster.

2.4 Discussion

This work shows an approach to study the etiology of onco-hematological diseases and to explore in-depth similarities in terms of genes regulations. Also it stresses the importance of handling the presence of low-quality subjects and of systematic noise.

Initially the large collected dataset was extremely prone to systematic effect because it was intentionally composed of datasets whose laboratory source and year of generation were different. This trait of the cohort hindered the capability to compare subjects from distinct batches and to establish whether the differences were actually biological and not systematic. To control this trait fRMA and RUV approaches were combined, since they were purposedly designed to tackle data normalization. The normalized data by fRMA was extremely efficient in closing the gap between batches, especially when

Diseases	**Number of TFs**
ALL, AML, BL, CLL, CML, DLBCL, HL, MCL, MDS, MM, MZLs, PTCL	786
ALL, AML, CLL, CML, MCL, MDS, MM, MZLs	72
ALL, AML, CLL, MDS	36
DLBCL, HL, PTCL	36
ALL, AML, CLL	15
AML, CLL, MCL	3
MCL, MM, MZLs	3
ALL, AML	62
MCL, PTCL	3
ALL, MDS	2
MM, MZLs	2
MCL, MZLs	1
FL	1010
BL	224
DLBCL	188
HL	188
PTCL	185
CML	152
MM	147
MZLs	146
MCL	142
MDS	114
CLL	98
ALL	37
AML	36

Table 2.2: Bi-clusters resulting from the adjusted δ-trimax method. The 25 bi-clusters do not overlap. The largest cluster comprehends all 13 diseases. Table is sorted first by the number of diseases and, if tied, by the number of transcription factors (TFs) included in a bi-cluster.

compared to RMA (Supplementary Figure A.1). In fact, based on PCA, subjects were close and overlapping without any strong outlier cluster. Yet several diseases were not compact and were scattered around in what seemed separate groups unrelated to all observed factors but the laboratory source. At this point RUV was performed under the assumption that, provided a list of negative controls, the biological signal was sufficiently uncorrelated from the systematic signal. After RUV, the batch effects were notably reduced and subjects showed to shrink and overlap on the first two principal components regardless of their disease. Although no clusters were clearly observable after RUV application, this is theoretically plausible since the variation between cancer diseases is not expected to emerge as a major modification of all genes regulation. Besides, the hierarchical clusters pointed out that there were still enough differences in the data to characterize the diseases up to four levels of similiarities. As expected, leukemias plus MDS clustered together since the latter is known especially to progress into AML. Lymphomas conversely showed diverse similarities. Interestingly, MM was found overall similar in regulation to MZLs, which could suggest why these two conditions may coexist in rare occasions [53]. Across the non-Hodgkin lymphomas, DLBCL was found to be the most similar to HL and even in this case it could explain why coexisting cases emerged [54]. The completely non-Hodgkin lymphoma cluster was PTCL with MCL, which tackle two different cells, respectively T-cells and B-cells, but have shown mutual involvement [55]. As mentioned, FL and BL did not cluster with any other disease suggesting their uniqueness in terms of overall regulation. Given this comprehensive landscape of the onco-hematological diseases, another step analysis would be required to determine the common signatures within each cluster. Besides, the standard hierarchical clustering gave an overall picture of the regulation, which might not manifest subtle similar regulations. In other words, diseases could have very similar (or different) regulations only on a small set of TFs and genes, and have very different (or similar) regulations on all the remainings. This hypothesis could not be tested by the standard hierarchical clustering. In literature bi-clustering techniques are well-known to tackle the clustering of two factors contemporarily. This is even more true for microarray array expression data, which several techniques were invented for. Though, the 13 regulatory networks formed a three dimensional object (diseases, TFs and genes) that is not commonly addressed by bi-clustering techniques but rather by a few tri-clustering approaches. Since the number of edges in the regulatory networks was significantly high and the goal of this work was to provide intelligible signatures for groups of diseases, signatures of TFs were targeted. By doing so, diseases were clustered based on how similarly their TFs overall regulate the genes. This helped to focus on 1100 TFs instead of 18315 genes (or millions of edges), which implicitly reduced the dimensionality of the clustering.

The δ-trimax algorithm was then adjusted to be a technique working on a three-dimensional object but generating bi-clusters. Totally 25 bi-clusters were obtained and 16 partitions resulted. As expected, the landscape became more fragmented and subtle than the one provided by the standard hierarchical clustering. Several similarities were confirmed especially between leukemias plus MDS and between lymphomas. Unlike previously, though, FL was the only singleton bi-cluster and BL showed to be similar to other eleven diseases over 786 TFs. The presence of such a large bi-cluster indicates that there is a strong backbone of TFs that commonly regulate genes, which aligns with the assumption that only a few lethal are carried by a disease. The independence of FL could be explained then in terms of a radical modification of several genes regulations or the presence of an unwanted factor that overcame the normalization and batch-correction procedure. All diseases appeared as singletons over a set of TFs. Except for FL, BL was the most solitary disease while ALL and AML were the least ones. Even from the standard clustering ALL and AML revealed to belong to the deeper level of the hieararchy, which makes them the most active in sharing regulations (only 37 and 36 TFs respectively unshared). Interestingly, over 35 TFs, diseases did not cluster at all likely because too much differences occur in how TFs overall regulate the 18315 genes. Based on the PCA outcomes, genes CYP3A43, NLRP8 and C8A could particularly be responsibile for the lack of clusters. Gene CYP3A43 could indicate that the metabolism of several drugs is regulated very differently across diseases and should be accounted for [56]. The other two genes, NLRP8 and C8A, respectively may point out that subtle inflammation modifications and immunodeficiencies should be always addressed specificially per disease. Such differences could be easily explored after the δ-trimax was run since there was no need of further advanced techniques to extract signatures and list of regulated genes. Short TFs signatures were as well extracted to observe the similarities between diseases in the same bi-clusters. The largest bi-cluster composed by twelve diseases reported TCF23 to be the most similar regulator, which also appeared to be one of the most similar regulators for several other bi-clusters of diseases. Yet, the strong similarity could imply also that TCF23 regulate genes involved in the basic functionality of the organism. This might also apply for recurrent similar TFs like PPARD, RXRB and ETV7. Of greater interest might be genes like HES4 and HES2 that resulted strongly similar only for MZLs respectively with MCL and MM.

To make further progress in analysing common TFs, bi-clusters were organized in 16 partitions to observe over the same set of TFs how diseases grouped. Among the 34 TFs where all leukemias, MDS, MM and MZLs cluster together and DLBCL, HL and PTCL formed the other non-singleton bi-cluster, NFATC1 was discovered. This TF, known to play role in hematopoietic cell transformation [57], marked the difference (along with other 33 TFs) between a subset of lymphomas and a leukemia-driven group. Besides, NFATC1 is also drug-targeted according to DrugBank [58], and it might be worth to explore whether it can contribute to aid either monitoring or prevention or treatment of the clustered diseases. This applies also for the DNMT1, which is drug-targeted and it belongs to the list of the 59 TFs that clustered AML and ALL together and left all the others as singletons. Besides, DNMT1 is already studied in hematology [59] due to its crucial regulation of DNA methylation. Yet not only drug-targeted TF helped to separate diseases and were already known to take part in hematopoietic processes. For instance, only over KLF3 [60] diseases divided in ALL plus MDS and AML, CLL plus MCL. Also in such case all remainings diseases appeared as singletons. The presence of such many singletons across partition could be explained by the choice of the δ of δ-trimax algorithm. In fact all results suggested that the one largest bi-cluster absorbed most consistent signal and left only the highly discording signal to be captured by smaller bi-clusters. Alternatively, multiple decrementing δ could be utilized to look how the bi-clusters transform as the imposed level of similarity becomes stricter.

This work reported a strategy to deal with diseases similarities and differences at a subtle level that leads to determine simple but intelligible signatures, i.e., biomarkers. These biomarkers could potentially play a role in discovering new regulatory mechanisms shared by diseases and in highlighting new common molecular targets for drug repurposing.

Chapter 3

Automatic Molecular Driver Identification and Classification of Acute Myeloid Leukemia via Non-Central Hypergeometric Refinement

3.1 Introduction

Great interest lies in the omics characterization of the Acute Myeloid Leukemia (AML) and in the identification of its molecular subtypes. Deep knowledge on subtypes can enhance the medical practice in terms of precise diagnosis and accurate prognosis, which heavily affect the choice and timing of treatment. The World Health Organization (WHO) published updated guidelines on AML subtypes in 2017, declaring several genomic subtypes [61]. Much effort is then required to obtain detailed molecular data about a AML subject, since they may reveal paramount information for its health. Despite the current WHO classification, other subtypes may be discovered in the next years, especially new rare ones, which are the most difficult to discover in a typical medium size cohort of subjects.

The HARMONY Alliance was designed to tackle the characterization of several hematological malignancies from different angles. The Alliance promoted collaborations of numerous academic institutes around Europe and encouraged private companies to take part into the research of onco-hematological diseases. With regard to AML, a large cohort has been assembled according to the OMOP common data model with the intention to capture even rare properties thanks to the power of big data.

With this study, AML subtypes were modelled as a Hierarchical Dirichlet Mixture

Model (HDMM) [62] with a large cohort and with a methodological approach to better control the effects of imputation values and statistical fluctuation. To enhance the identification of non-trivial driver genes, a new refinement step was introduced based on a non-central Hypergeometric distribution. Such distribution was also employed to define an automatic classification approach able to assign subjects that did not take part in the HDMM fitting. The unbiased nature of the automatic classificator was originally designed to support clinical practice when dealing with molecular alterations such as genomic mutations and karyotypic aberrations. All results were compared with the WHO official guidelines and with the clinical expectations. Namely, overall survival (OS) was considered as the main clinical quantity to assess the goodness of the newly introduced assignment method.

3.2 Methods

3.2.1 Database

The HARMONY alliance project provided the AML data through a privacy-secured platform based on the OMOP common data model. A total of 4160 subjects were recruited by six providers, whose identity was restricted according to the HARMONY project policy. The distribution of subjects across providers is shown in Table 3.1.

	DP_{12}	DP_1	DP_2	DP_4	DP_5	$TCGA$
Number of subjects	144	1542	660	185	636	993

Table 3.1: Number of subjects in the HARMONY Alliance database per data provider. The label DP stands for Data Provider and each DP is numbered chronologically

The cohort was well balanced in gender, with 53% males (2200) and 47% females (1960), and also evenly distributed along age between sexes (Figure 3.1). Age was defined to be the age at the time of diagnosis.

3.2.2 Pre-Processing

The richness of the HARMONY database was a valuable resource to achieve an accurate description of the molecular landscape of AML. Though, when data come from multiple centers or laboratories the lack of overlap between available data usually come into play. Besides, the database should be representative according to the clinical questions. Therefore several pre-processing steps were necessary before performing the main analyses.

Firstly, children and adolescents could interfere throughout the characterization of AML since pediatric diseases are tipycally different from adult ones. Hence, subjects

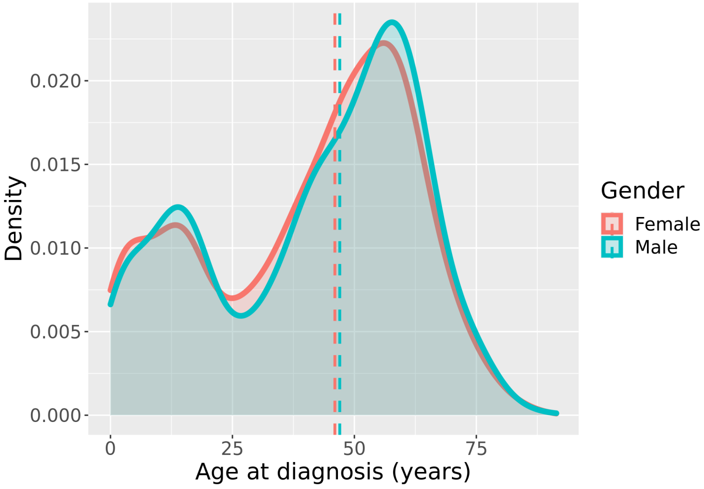


Figure 3.1: Density plot of age at diagnosis for males and females. The vertical dashed lines point to the median age for each gender.

younger than 16 years old, about 19% (793) of the total, were filtered out. Secondly, only genomic and karyotypic alterations were selected to characterize the disease and within HARMONY these were a total of 122 alterations that spanned from single point mutations to chromosomic aberrations. The entire list of such alterations is reported in Supplementary Table B.1 and counts 31 karyotypic aberrations and 91 gene mutation statuses.

Third, the expected subtypes of the AML landscape were assumed to be representative of at least 10 subjects, i.e. any clustering algorithm should ideally assign at least 10 subjects to all subtypes. Consequently, all alterations with less than 10 occurrences were taken out, which implied the removal of 35 of them. Mainly they were genomic mutations (PTEN, CBLB, CUX1, CBLC, MPL, CALR, HRAS, GATA1, SF3A1, CDKN2A, JAK3, JAK1, ABCG2, U2AF2, PTPRT, ATRX, GNAS, RB1, MLL5, ABL1, PRPF40B, SF1, SH2B3, VHL, TERC, ASXL2, DCK, DCLK1, WAC, ABCB1, DIS3, BRINP3), but $t(9;22)$, $t(10;11)$, $t(3;5)$ were removed as well. Fourth, unavailable data had to be taken care of and initially subjects missing all data (7 subjects) were excluded. After that, missingness covered about 23% fraction of the dataset and several patterns of missingness were revealed to be recurrent (Figure 3.2), i.e. missingness did not occur at random (MNAR). To move towards a familiar missing at random (MAR) scenario, only subjects sharing random missingness should be retained but this would eventually end up in a small unrepresentative cohort. Yet, in the context of AML, few genes per subject are usually targeted and the pre-dominance of unmutated genes mitigates the issue. Then the fraction of missing data was used to get rid of both alterations lacking the most amount of data [63] and subjects with very few available data. The strategy was to

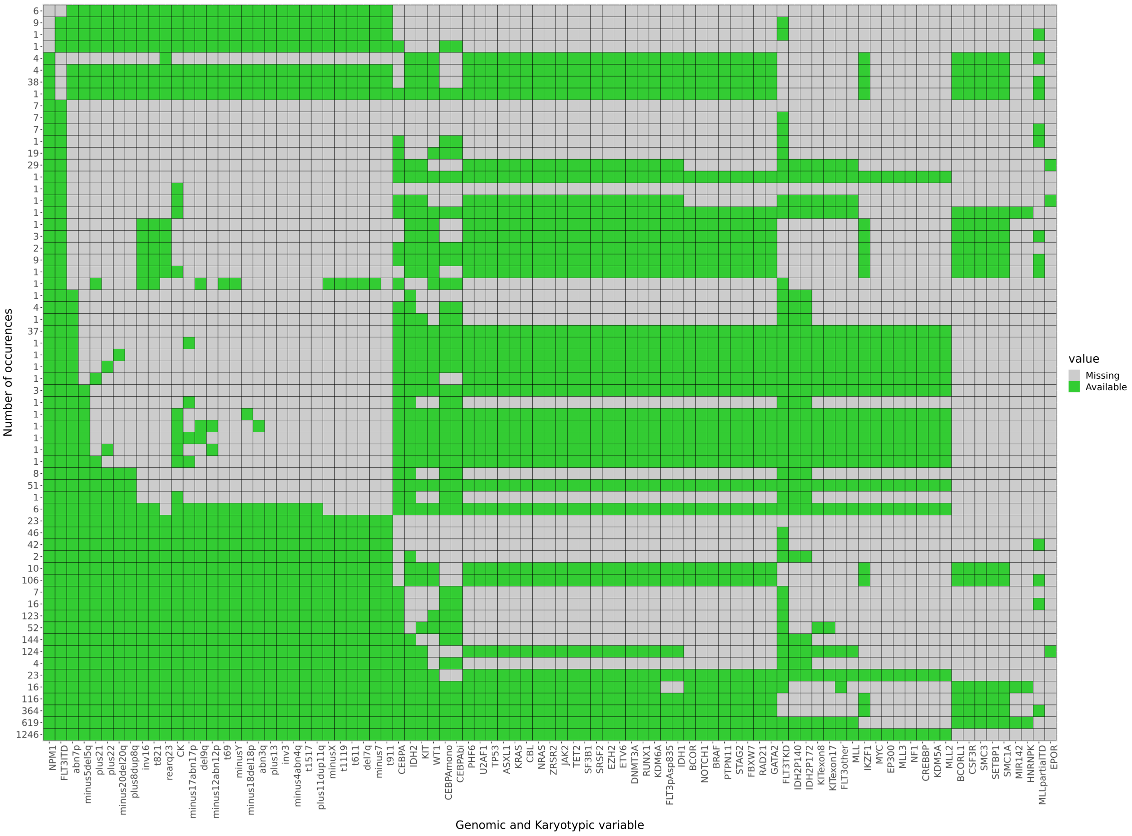


Figure 3.2: Patterns of available/missing data across genomic and karyotypic alterations. The number of occurrences of a pattern is reported on the y-axis.

impute at most the 10% of the whole dataset, which is a good general tradeoff when it comes to imputation. The empirical iterative approach that was utilized to perform filtering on both subjects and alterations is illustrated in Figure 3.3.

The approach cut off 215 subjects and 17 genomic variables (EPOR, MLL^{PTD}, MIR142, HNRNPK, BCORL1, CSF3R, SMC3, SETBP1, SMC1A, MYC, EP300, MLL3, NF1, CREBBP, KDM5A, MLL2, IKZF1) and left a percentage of missingness equal to 9.78%. The fifth and last step of pre-processing was to neglect from the analysis the subjects with no alterations at all (97). Eventually the data consisted of 3048 subjects and 70 alterations: 42 genomic mutation statues and 28 karyotypic aberrations.

3.2.3 Imputation

After pre-processing the ratio of available to missing data was significantly improved, alongside the patterns of recurrent missingness (Figure 3.4). Since completely unaltered

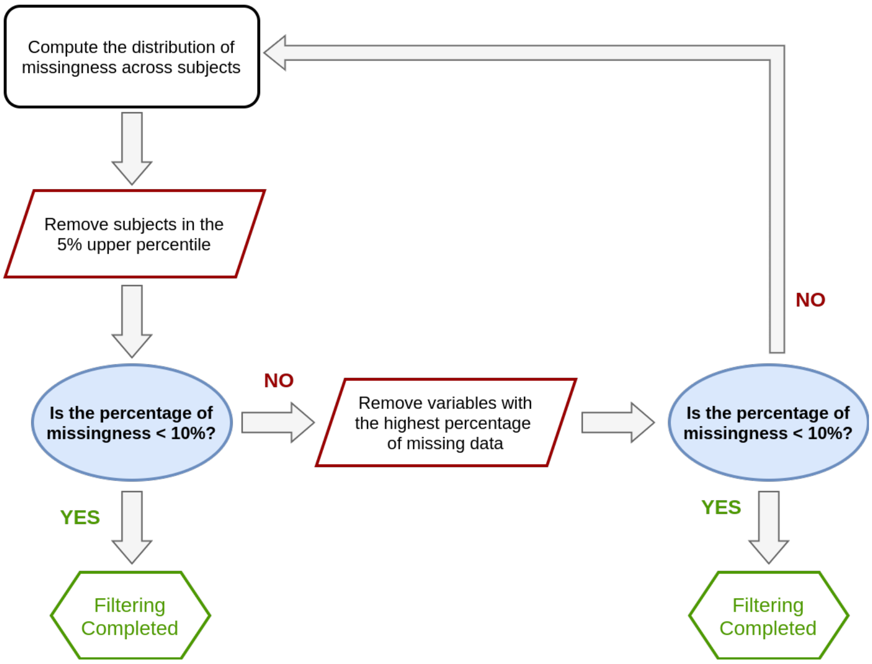

Compute the distribution of missingness across subjects
Remove subjects in the 5% upper percentile
Is the percentage of missingness < 10%?
NO
Remove variables with the highest percentage of missing data
Is the percentage of missingness < 10%?
NO
YES
Filtering Completed
YES
Filtering Completed

Figure 3.3: Schematic illustration of the empirical iterative approach defined to filter out poor alterations and subjects. The pre-set threshold of imputable data was 10%.

subjects were removed, the fraction of data downsized to 8.78%. Since the scenario remained MNAR, the strategy was then to simply control that any final outcomes would have turned out invariant with respect to any imputed values. To this end, missingness was imputed 10 times: 9 times using Multiple Imputation by Chained Equations (MICE[1]) [64] and once setting all unavailable data to zero. MICE imputations were chosen to capture relationships between variables and to potentially predict values more accurately. Gender, age and quality of life were fed to MICE as covariates hoping to enhance imputation. In contrast, the single imputation with all zeroes was used since it is the most conservative way of treating the unknown in AML. As mentioned, mutations and aberrations happen only on a few genes and chromosomes in AML, and that makes even a single alteration extremely relevant.

3.2.4 AML as a Hierachical Dirichlet Mixture Model

The problem of modelling AML can be depicted as follows. A new subject suffering from AML comes to a hospital and is taken to a room, with all people in the room having its same AML genomic and karyotypic subtype among K biological subtypes. There also might be multiple rooms for the same subtype. Further, a new subject always has the

[1] See section 6.2.1

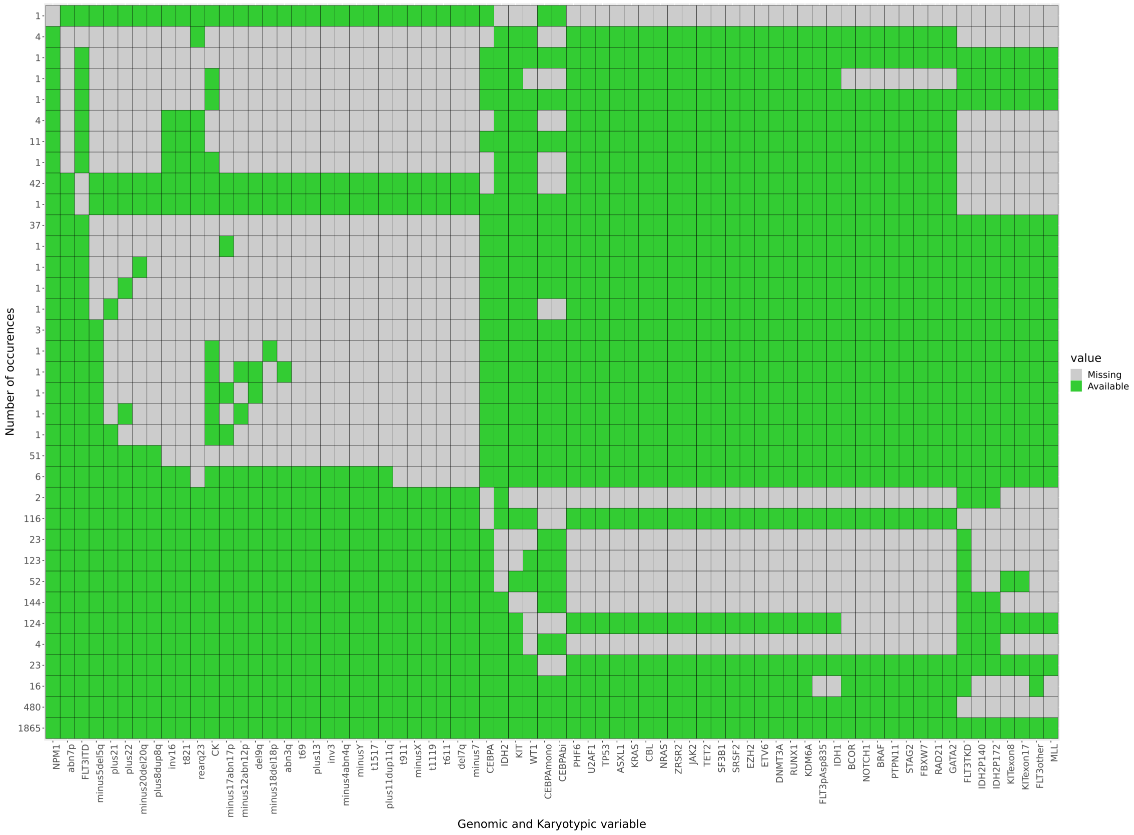


Figure 3.4: Patterns of available/missing data across genomic and karyotypic alterations after empirical filtering. The number of occurrences of a pattern is reported on the y-axis.

chance to suffer from a new subtype of the disease and therefore being taken to a room still empty. Though, a new subject is more likely to be affected by an already known subtype and, more popular is a subtype, greater is the chance that a new patient suffers from it. This type of problem can be represented by the so-called Dirichlet Process (DP). The problem can be also posed as follows: the joint probability of alterations occurrences can be described by a DP (section 6.2.2).

Thus, the foremost assumption is that subjects suffering from AML cluster in several biological subtypes. In this work every subtype is supposed to be a multinomial distribution of 70 molecular alterations and the difference across subtypes lies in the probability of occurrence of such alterations. The expectation is then that the cohort could be represented as a mixture of multinomial distributions with each subject being a realization according to a DP. The non-parametric nature of a DP does not need a fixed number of subtypes (i.e. multinomial distributions) since there is always the possibility of new subtype to emerge. To make further progress and to achieve a more detailed mixture of multinomial we added a layer to the DP, defining a Hierarchical Dirichlet Mixture Model (HDMM). Basically, by adding an intermediate layer, each subject becomes a collection of alterations and each is assigned at every iteration of the DP. Doing so, a single alteration is assigned by the HDP based both on the subject it emerged from and on the assignments of all other alterations of its kind. Eventually this leads to yield several multinomials but also to remark the need to classify an entire subject to a single one.

3.2.5 HDMM fit and imputation control

To estimate the joint posterior distribution of the HDMM a Gibbs sampler was employed. A total of 5000 samplings, called burn-in samples, were set to wait for the target distribution to stabilize. The Gibbs sampler was a Markov chain Monte Carlo (MCMC) and, as such, its bias depended on how it was initialized [65]. The starting point, called seed, of a MCMC chain strongly influences the samplings at the beginning of the chain and can also prevent an accurate exploration of the targeted distribution. Therefore, differently initialized multiple chains were run to achieve better unbiased estimations. After the independent chains topped convergence, the average of the parameters of the multinomial distributions across chains was used to produce the final outcome [66, 67]. Ideally, averaging multiple chain would also mean bias equal to zero but this is not always true.

To account for missing data the following assumptions were made. Since a small part of the data was unknown, the underlying real HDMM could not be exactly captured even theoretically. Only the known data could drive towards the most reliable estimate of the HDMM. Imputing data was then analog to initialize a chain randomly, since the imputed values biased the sampler's walk. Upon this scenario, the missingness could be dealt with similarly to what explained for different random chain initializations. That is, several chains were run both for multiple random seeds and for multiple imputation values. Only afterwards, samplings were pooled together, as if they were a unique long chain, to estimate the final HDMM. The expectations is that the pool-and-average procedure should highlight the most robust structure of the HDMM, minimizing the influences of random starts and imputation values. To control for the Gibbs sampler random starts the HDMM was estimated 10 times starting from 10 distinct random values. Additionally, to control for imputation values each of such HDMMs was fitted for all 10 sets of imputed values. That is, 100 HDMM chains were totally run and since for each chain 10^4 posterior samples were collected, 10^6 posterior samples were eventually gathered. Everytime a posterior sample was extracted, the parameters of the multinomials, i.e., the subtypes, were estimated, which implies that 10^6 sets of multinomials were collected.

3.2.6 Convergence of single HDP chains

Two criteria were adopted to establish whether a single HDMM chain converged. The first one was based on empirical evidence and expertise with the HDMM fitting procedure: a chain reaches convergence when at least 60% of posterior samplings share the same number of components[2]. The number of components throughout samplings was expected to differ because the HDMM always leaves the possibility of adding new components and this holds even when the theoretical true number of components is achieved. In other words, upon convergence, the true number of components was expected to be stable through the MCMC samplings and to be the smallest one.

The second criterion focused on the parameters of the components, i.e., the multinomials. The parameters of a multinomial distribution are known to be probabilities that sum up to one. After averaging their values across all posterior samples (that shared the converged number of components) and calculating their 95% confidence interval (CI), using the highest posterior density (HPD) interval, only parameters whose CI did not overlap with zero were considered non-null, whereas the others were zeroed out. It follows that, upon convergence, although their sum should be approximately one, it could not be so. Therefore, to set a consistent threshold, the lower end of the CI of the distribution across chains of the smallest sum among components was chosen. Empirically, the second criterion eventually was: a chain converged when all its components had parameters whose sum was greater than about 0.694.

All chains that do not respect both criteria were removed from the subsequent analysis.

3.2.7 Convergence across all DP chains

Convergence was also expected across different chains in order to prove their independence from the random seed and from the different imputed values. Firstly, for each chain, the number of components was defined as the most frequent number of components across the samplings. Next, only the samplings with such number were preserved and the parameters of the components were estimated by averaging. In addition, parameters whose 95% CI overlapped with zero were considered unsignificant and zeroed out. At the end of this first step a list of K_n components $\{\vec{w}_{kn}\}_{k=1}^{K_n}$ for each n-th chain was obtained.

[2] It was noticed on all chains that running the MCMC to get 10^4 posterior samples multiple times, the most frequent number of components was always the same when it was shared by roughly the 60% of the posterior samplings.

Components Definition

Ideally all chains should have the same number of components ($K_n = K, \forall n$) with approximately the same parameters ($\{\vec{w}_{kn}\}_{k=1}^{K_n} = \{\vec{w}_k\}_{k=1}^{K}, \forall n$). This ideal expectation cannot be achievable when the HDMM assumptions are not completely met or if the simulation finishes before reaching an extremely accurate estimation of the HDMM. However, a feasible and intuitive expectation upon convergence is: every component from a chain points uniquely to a component from another chain. In other words, a component from a chain corresponds univocally to a component from another chain, and this is true for all components. In fact, if real strong components actually exist, they should turn out for each chain and they should have the same signature across chains. With N chains we would expect N replicas for each component. This rationale can be represented with graphs, where each vertex is a component from a chain. Upon convergence, a real component should be characterized by N fully connected verteces where the edges represent correspondences across chains. With no convergence, it would not be viable either to determine N replicas or to achieve fully connection between them, because they would be significantly different from chain to chain.

To evaluate the similarity between components the cosine similarity was used:

$$d(\vec{w}_1, \vec{w}_2) = \frac{\vec{w}_1 \cdot \vec{w}_2}{||\vec{w}_1||\,||\vec{w}_2||}\ . \tag{3.1}$$

All components from a chain were compared to each component from all other chains. A component from a chain connects to the component of another chain with the maximum cosine distance. Multiple components with the same maximum cosine distance value are not expected upon convergence, and, if they occur, it might suggest a lack of convergence. Therefore, all chains with components having non univocally connections was considered as unconverged, and excluded from further analysis. At the end, a graph of components was built to visualize which components always emerged and how many replicas they had across chains.

Merging components

Maximal cliques, i.e., fully connected sub-graphs with maximal size, with N nodes were obtained to determine the targeted components. The clique Cl_k for the k-th components represented all components across all chains that were found to harbour the same signature. When $K_n = K, \forall n$, the totally unconnected maximal cliques are K, one for each targeted component. Else, when K_ns differ, multiple components from the same chain point to a single component from another chain. That is, overlapping maximal cliques emerged. This event can occur because one chain estimates a single component as one, whereas another chain estimates it, but split it in multiple similar parts. In other words, the same signature is found in both chains but represented through a different number of components. It is not trivial to say if the overlapping maximal cliques represents a single targeted component or are actually multiple ones. The following procedure was then exploited.

Suppose L overlapping maximal cliques compose a compound. Now, count the number of chains x_l that estimate the compound in l different parts, for all $l \in [1, L]$. Then a multinomial test can be run on the L counts to control whether the compound splits in l uniformly. If it is, then the compound is found to be non-enriched in any particular l and is kept together. If it is not, $l = \text{argmin}_{l'} x_{l'}$ is removed and the multinomial test is re-run with the remaining ls. The procedure stop either when the only one l is left or when multiple ls are equally represented. By doing so, ambiguous components across chains were sorted out.

3.2.8 Multinomial parameters estimation

So far, the estimation of parameters was taken for granted but to understand how they were actually estimated, a deeper look at the fitting procedure is required. What the HDMM samplings actually showed is how each alteration was assigned to one of the components. As mentioned, indeed, with the hierachical

structure the alterations are the objects assigned to the components. Therefore if the total number of alterations in the data is T, then T objects are assigned for every HDMM sampling. This means that a component at one sampling was a cluster of alterations. The more a particular molecular alteration was assigned to a component, the higher its probability was in the multinomial it was supposed to come from. The quickest way to estimate the parameters of a component was, thus, to normalize the counts for each molecular alteration for the total number of alterations assigned. That is, if z_{kj} is the number of the j-th molecular alterations assigned to the component k and the vector $(z_{k1}, .., z_{kM})$ gathers the counts for all M alterations, then the parameters of the assumed multinomials are estimated as $(\frac{z_{k1}}{\sum_{j=1}^{M} z_{kj}}, .., \frac{z_{kM}}{\sum_{j=1}^{M} z_{kj}})$.

Now, to understand how the final estimates of the parameters were determined across all samplings of all chains, assume to add to the counts z_{kj} another two indeces z_{kjsh}, one for the sampling (s) and one for the chain h. The count z_{kjsh} is consequently the number of times the j-th molecular alteration is assigned to the k-th component at the s-th sampling of the h-th chain. After the targeted components signatures is determined by the graph-based approach, for each chain h and component k all $z_{k'jsh}$ are summed over all k'-th components connected to the targeted k-th component, i.e., $z_{kjsh} = \sum_{k'|k' \in Cl_k} z_{k'jsh}$. As mentioned, the connections are derived directly from the maximal clique Cl_k for the k-th targeted component. Next, the maximum a posterior (MAP) estimate over all samplings z_{kjsh} is taken, which means that z_{kjh}^{MAP} is the most frequent z_{kjsh} across all s. It follows that the multinomial parameters for each k-th components are estimated as previously explained: $\vec{w}_{kh}^{Multi} = (\frac{z_{k1h}^{MAP}}{\sum_{j=1}^{M} z_{kjh}^{MAP}}, .., \frac{z_{kMh}^{MAP}}{\sum_{j=1}^{M} z_{kjh}^{MAP}})$.

Ultimately, the median across all h-th chains were taken and normalized to yield the final parameter estimates.

3.2.9 Non-Central Hypergeometric Refinement

The refinement approach proposed in this work adjusts the final part of the multinomial parameters estimation. Instead of using the z_{kjsh}^{MAP} to approximate the parameters of multinomials a Multivariate Fisher's Non-Central Hypergeometric Distribution (MFNCHD) was fitted (see details in section 6.2.3). The MFNCHD models scenarios where the balls from an urn are drawn with a bias. In other words, suppose each molecular alteration is a ball with a certain color and there are as many colors as molecular alterations. Then the sum of the MAP estimates $\sum_j z_{kjsh}^{MAP}$ is the number of draws and the MAP estimates themselves are the expected number of times the colors are drawn. Compared with the number of times a molecular alteration appears in the data, which defines the total number of balls for a specific color, the non-central distribution fit determines the weight for each color, i.e., molecular alteration, in the component. In this way, if a rare alteration in the data is frequently assigned to a single component, then its weight will be very high only for such component and very low for the others. Conversely, a frequent molecular alteration in the data is more likely to be extracted in several components and, if it is not particulary enriched in any components, it will not have a high weight despite being popular.

It follows that after the weights of the MFNCHDs are estimated for each k-th component for every chain, $\vec{w}_{kh}^{MFNCHD}$, the final parameters for each component are computed by taking the median across all h-th chains.

3.2.10 Automatic molecular assignment

Once the K components and their weights, $\vec{w}_k^{Multi}$ or $\vec{w}_k^{MFNCHD}$, were estimated, an automatic approach to classify subjects was designed. With the multinomials scenario the assignment approach for each i-th subject can be formulated as:

$$\kappa_i = \underset{k}{\operatorname{argmax}} \frac{\Gamma(\sum_j x_{ij} + 1)}{\prod_j \Gamma(x_{ij} + 1)} \prod_{j=1}^{M=70} \left(w_{jk}^{Multi}\right)^{x_{ij}} \quad , \tag{3.2}$$

where κ_i is the component the i-th subject is assigned to. The x_{ij} stands for the j-th molecular alteration status of the i-th subject. Therefore, the assignment approach leverages the probability mass function of the multinomial distribution to deduce which component has the higher probability to yield a subject.

The same principle applies for MFNCHD, where the assignment formula becomes:

$$\kappa_i = \underset{k}{\operatorname{argmax}} \frac{1}{P_0} \prod_{j=1}^{M=70} \binom{m_{ij}}{x_{ij}} w_{jk}^{x_{ij}} \quad , \tag{3.3}$$

where P_0 is the partition function and m_{ij} is the maximum number of times the j-th alteration could potentially occur for the i-th sample. Since all alterations are binary, i.e., zero or one, all m_{ik} are set to one, and formula 3.3 simplifies to:

$$\kappa_i = \underset{k}{\text{argmax}} \frac{1}{P_0} \prod_{j|x_{ij}=1} w_{jk} \ , \tag{3.4}$$

with

$$P_0 = \sum_{\vec{y}|\sum_j y_j = \sum_j x_{ij}} \left(\prod_{j|y_j=1} w_{jk} \right) \ . \tag{3.5}$$

When multiple ks are associated to the maximum value, the subject is considered ambiguous.

3.3 Results

3.3.1 AML components

Of the 100 HDMM chains, 7 were found to be unconverged and were removed. Their lack of convergence was not related to any particular imputed values. After all 93 HDMM chains were compared through the graph-based approach, a total of 12 components were discovered. Each component was first estimated as a multinomial distribution and second as a MFNCHD. The top molecular composition of the components for both distributions are showed in Table 3.2.

The complete signatures of molecular alterations for both distributions are reported in Supplementary Tables B.2-B.13. Several overlaps could be noticed between the two version of the components but what stood out the most is the difference in ordering. The molecular alterations are sorted in descending order based on their probabilities for the multinomial case and based on their weights for the MFNCHD case.

3.3.2 Correspondence between automatic classification and clinical classifications

Both automatic classifiers were run to assign the subjects to an AML component. All 3048 subjects were classified and no ambiguous cases resulted. The NPM1-driven compponenent was eventually the most popular for either classifier. In contrast, though, subjects distributed significantly different as showed in Table 3.3. As a matter of fact, the NPM1-driven component received more than 700 subjects with the multinomial-based classifier than with the MFNCHD-based one. Such differences, of course in diverse proportions, were also neat for most components.

Multinomial-based components	**MFNCHD-based components**
NPM1, DNMT3A, FLT3$^{\text{ITD}}$	NPM1, DNMT3A, FLT3$^{\text{ITD}}$
Complex Karyotype, TP53, -5/del(5q)	Complex Karyotype, TP53, -5/del(5q)
RUNX1, ASXL1, SRSF2	RUNX1, ASXL1, SRSF2
IDH2, IDH2$^{\text{p140}}$, NPM1	IDH2$^{\text{p140}}$, IDH2, NPM1
IDH2, IDH2$^{\text{p172}}$, IDH2$^{\text{p140}}$	IDH2$^{\text{p172}}$, IDH2, IDH2$^{\text{p140}}$
t(8;21), -Y, KIT	t(8;21), -Y, KIT
CEBPA, CEBPA$^{\text{bi-allelic}}$, GATA2	CEBPA$^{\text{bi-allelic}}$, CEBPA, GATA2
-7, NRAS, inv(3)	inv(3), -7, abn(3q)
inv(16), NRAS, KIT	inv(16), KIT, KRAS
FLT3$^{\text{ITD}}$, FLT3$^{\text{other}}$, t(15;17)	t(15;17), t(6;9), WT1
t(x;11q23), t(9;11), NRAS	t(x;11q23), t(9;11), t(6;11)
CEBPA, CEBPA$^{\text{mono-allelic}}$, KIT	CEBPA$^{\text{mono-allelic}}$, CEBPA, KIT

Table 3.2: Top three molecular alterations of the 12 components of Acute Myeloid Leukemia when estimating the final parameters as from a multinomial distribution (on the left) and as from a MFNCHD.

To observe how well the automatic classifiers behaved with respect to clinically-oriented AML classification systems, the classes assigned to the subjects were compared to the classes expected by the WHO and by Pappaemmanuil et al., which is a popular

Multinomial-based classification		**MFNCHD-based classification**	
NPM1	1660	NPM1	899
Complex Karyotype	219	Complex Karyotype	356
RUNX1	265	RUNX1	424
IDH2 - IDH2^{p140}	98	IDH2^{p140}	210
IDH2 - IDH2^{p172}	89	IDH2^{p172}	146
t(8;21)	114	t(8;21)	144
CEBPA - CEBPA$^{\text{bi-allelic}}$	83	CEBPA$^{\text{bi-allelic}}$	143
-7	107	inv(3)	154
inv(16)	125	inv(16)	172
FLT3^{ITD}	152	t(15;17)	145
t(x;11q23)	84	t(x;11q23)	114
CEBPA - CEBPA$^{\text{mono-allelic}}$	52	CEBPA$^{\text{mono-allelic}}$	141

Table 3.3: Number of subjects assigned to each components for both automatic classification approaches: one based on the multinomial distribution, the other based on the MFNCHD.

HDMM-derived system. To establish a correspondence between the classifier assignments and the expected clinical classes, the driver molecular alterations were compared. Only subjects that could be classified by the reference systems were considered at this stage.

Respectively, the WHO system and the Pappaemmanuil et al. system could classify 2012 and 2308 subjects out of the total cohort (3048 subjects). As reported by Table 3.4 the MFNCHD-based classifier outperformed the multinomial-based classifier based on accuracy with respect to both reference systems.

Automatic classifier	**Accuracy w.r.t. WHO (2012 subjects)**	**Accuracy w.r.t. Pappaemmanuil et al. (2308 subjects)**
Multinomial-based	71%	66%
MFNCHD-based	82%	78%

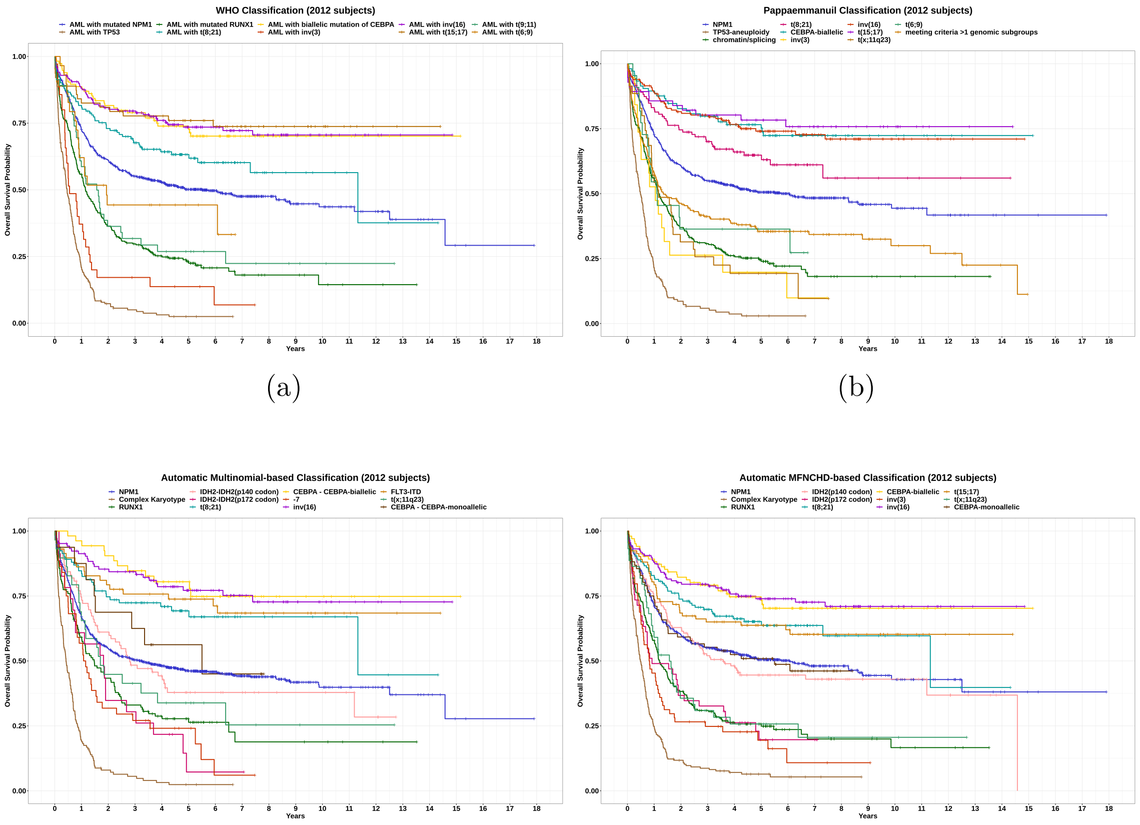
Table 3.4: Overview of how accurate the automatic classifiers reproduced two widespread clinically-oriented classification systems. The WHO system stratifies AML in molecular classes according to the clinical and biological knowledge, whereas the Pappaemmanuil et al. adopts a clinically driven decision system suggested by a HDMM with underlying multinomials. The comparison was run over subjects that could actually be classified by the known classification systems, i.e., no-ambiguous or un-classified subjects were neglected.

3.3.3 The case of t(6;9) and t(15;17)

In terms of AML components, the biggest difference between both the WHO and Pappaemmanuil et al. reference systems was the absence of a t(6;9)-driven component. The multinomial-based classifier split the subjects with t(6;9) between the NPM1-driven and the FLT3$^{\mathrm{ITD}}$-driven components. Differently, according to the MFNCHD-based classifier, they were mostly assigned to the t(15;17)-driven component. In theory, the FLT3$^{\mathrm{ITD}}$-driven component for the multinomial-based classifier and the t(15;17)-driven component for the MFNCHD-based classifier represent the same signature in different ways. It follows that subjects with either t(6;9) or t(15;17) fell both under the same component due to their strong co-occurence with the FLT3$^{\mathrm{ITD}}$ alteration. Notably, though, it was the MFNCHD-based approach to showcase the role of t(6;9) and t(15;17) in the component.

3.3.4 Consistency of automatic classifiers on survival predictions

Overall survival probability, section 6.2.4, is a well-known clinical measurement to determine a prognosis and, here, it was used to observe how molecular alterations impact on the risk of death of the subjects. Exclusively subjects that could be classified by WHO were considered due to its application in clinical practice. The automatic classifiers and the reference systems (WHO and Pappaemmanuil et al.) were all used to predict survival and then were all compared via Kaplan-Meier curves (Figures 3.5).



(a)

(b)

(c)

(d)

Figure 3.5: Overlook of the Kaplan-Meier survival curves over the subjects that could be classified by WHO. Each plot was generated using the classes from WHO itself (3.5a), Pappaemmanuil et al. (3.5b), the automatic multinomial-based classifier (3.5c) and the automatic MFNCHD-based classifier (3.5d).

The KM curves qualitatively highlighted that both automatic classifiers and Pappaemmanuil et al. were able to capture the main risks expectation along the years. Slight changes were noticed due to the presence of additional AML components with respect to the WHO reference system.

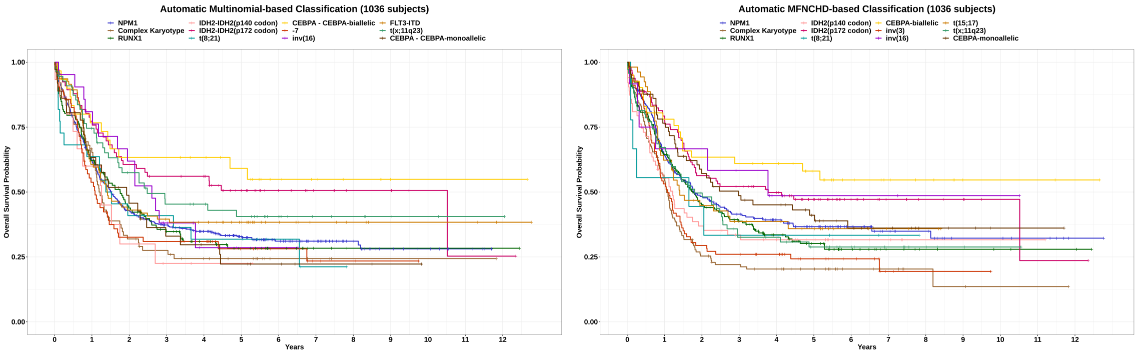
To make further progress, the comparison were also performed by fitting Cox Proportional Hazards (CPH) models. The classes determined by the automatic classifiers were used as predictors in the CPH models while controlling for age and gender. This was also performed for the classes expected by the two reference systems in order to compare the survival predictions. The concordance index showed that all classification systems performed similarly but the WHO was still the best to predict survival outcomes.

Classification system	**Concordance w.r.t. WHO classified subjects (totally 2012)**
WHO	0.705
Pappaemmanuil et al.	0.699
Multinomial-based	0.685
MFNCHD-based	0.699

Table 3.5: Performances, in terms of concordance index, of the Cox Proportional Harzards (CPH) models that were fitted for each classification systems. The classes were considered as predictors and age plus gender were added it in as covariates.

3.3.5 Survival predictions over ambiguous and un-classified subjects

The automatic classifier did not show limitations in assigning subjects. Based on the likelihood of the two distributions, subjects with several molecular alterations, which were ambiguous or left un-classified by WHO and Pappaemmanuil et al., were assigned smoothly. The KM survival curves for WHO un-assigned subjects (a total of 1036) are illustrated in Figures 3.6. The Pappaemmanuil curve was not reported because more than 500 subjects were left un-classified as well.



(a) (b)

Figure 3.6: Kaplan Meier survival curves for the WHO-unassigned subjects

Differently from as previously seen, the KM curves tended to clutter and overlap (especially for the multinomial-based classifier), although the general order in terms of survival risk seemed to be preserved.

Over this set of subjects the concordance indexes of newly fitted CPH models (Table 3.6) showed to be less than the what observed over subjects classified by WHO.

Classification system	**Concordance w.r.t. WHO un-classified subjects (totally 1036)**
Multinomial-based	0.609
MFNCHD-based	0.626

Table 3.6: Performances, in terms of concordance index, of the Cox Proportional Harzards (CPH) models that were fitted for each classification systems. The classes were considered as predictors and age plus gender were added it in as covariates.

3.4 Discussion

Much effort was dedicated to the study of the molecular characterization of AML [62]. The WHO provided guidelines on how to take care of subjects suffering from AML with particular mutated genes or alterated karyotypes, but many cases still remain ambiguous or unknown. In this work, following the WHO classification, only 2012 out 3048 subjects could be actually classified, leaving out almost more than one third of the entire cohort. To deeply characterize the molecular landscape of AML, then, an alternative non-official classification was employed after modelling the AML as a HDMM of underlying multinomials [62]. Such classification, which we refer to as Pappaemmanuil et al., established new clinical criteria for the molecular classification of AML by observing how the multinomials fitted by the HDMM were structured. In this way a new classification based on class-defining molecular alterations was proposed. Nevetheless, ambiguities persisted and, in this work, such new classification was not able to classify 740 subjects, since 335 subjects had more than one class-defining alteration and 405 subjects had none.

In this work the underlying multinomial representation of the HDMM for AML was questioned. To this end, thanks to the HARMONY Alliance, a HDMM of multinomial distributions was fitted from scratch on a huge cohort of 3048 subjects. To handle missing data, only components with a strong representative molecular alterations signature were considered, whereas the ones spurious were eliminated. Twelve components with their respective multinomial distributions were ultimately determined for the HDMM. As expected, ten of them agreed with Pappaemmanuil et al. components (Supplementary Table B.15), while two emerged as novel. The first new component involved IDH2$^{\text{p.140}}$ mutations and the second involved the mono-allelic mutations of CEBPA. Interestingly, these two molecular alterations shared a parent gene with another already known component. Namely, a IDH2$^{\text{p.172}}$-driven component and a CEBPA$^{\text{mono-allelic}}$-driven component were previously reported. Therefore, the molecular landscape of both CEBPA mutations and IDH2 mutations was found to be further characterized in this work, potentially revealing new important co-occurent biological events. In fact, under a thorough observation of the new components and their associated ones, it could be noticed that when both IDH2 and NPM1 were mutated then the mutation on the IDH2 seemed to occur at codon p.140; whereas, when IDH2 was mutated but NPM1 was not, the IDH2 mutation emerged at codon p.172. A similar principle applied for CEBPA mutations. In the absence of a mutated GATA2 or mutated WT1, a mutation of CEBPA was more likely to be mono-allelic. Instead, if one of the two genes mutated, the mutation of CEBPA emerged as bi-allelic. Such new components could then assist to deeply stratify subjects with CEBPA and IDH2 subjects and look for more personalized pharmaceutical options.

To question the efficacy of the multinomial representation of the AML components, an automatic classifier based on the probability mass functions of the multinomial distribution was employed. Afterwards, the comparisons with both the WHO reference system and Pappaemmanuil et al. did not show a highly accurate correspondence between such systems and the multinomial-based classifier. In fact, only 71% subjects were directly associated with the WHO classes, and this fraction was even less (66%) with respect to the Pappaemmanuil et al. reference system. These two reference systems were considered as benchmarks of a clinical input. The goal of the comparison was to control whether a completely clinically un-biased classifier was able to explain a clinical input through the simple combination of molecular alterations.

The poor results of the multinomial-based classifier was the main motivation to define a refinement approach. To explain why the automatic classifier failed to optimally reproduce the reference systems a closer evaluation on the principles of the multinomial distribution was carried out. The multinomial distribution models extractions with replacement. That is, given an AML multinomial component with M parameters, any m-th molecular could be drawn more than one time across several extractions. This principle does not apply for the AML data since each subject is represented by an array of zero or one values. To overcome this principle the hypergeometric distribution could help, since each molecular alteration could be set to be extracted at most one time. In contrast, though, all AML hypergeometric components would be equal because each molecular alteration would have the same probability to be extracted. In such sense, the multinomial distribution helped because their parameters, which indicated the probabilities of the molecular alterations, where the elements that made the components distinct. The idea was to exploit a distribution which had both advantages: a distribution that limited the molecular extractions to be at most one and a distribution with parameters that could help prioritizing alterations. The direct solution was the Multivariate Fisher's Non-Central Hypergeometric distribution (MFNCHD). The further intuitive benefit of the MFNCHD was also the following. Suppose to have a rare molecular alteration j_r that occurs T_{j_r} in the whole cohort and a popular alteration j_p that emerged T_{j_p} times, i.e. $T_{j_p} > T_{j_r}$. Also, assume that all T_{j_r} occurrences are assigned to an AML component and only t_{j_p} are assigned to the same component. Now, if $t_{j_p} > T_{j_r}$, a multinomial distribution will fit a greater probability for the alteration j_p despite all the occurrences for alteration j_r were assigned to the same component. In other words, the multinomial distribution does not assume that there are rare and popular molecular alterations, which biases its fitting procedure to typically give great importance to popular molecular alterations. The MFNCHD on the other hand was assumed to sort this issue and to reward enriched molecular alterations in a component. That is, if a rare alteration is enriched in a component then it will have a high weight.

Given the properties of the MFNCHD, an automatic classifier was created and compared to the multinomial-based one along with the reference classification systems. As expected, rare molecular alterations became drivers of their own component as shown by IDH2 codon mutations, inv(3), t(15;17) and the mutations of CEBPA. In this way, the AML components became much more similar (qualitatively) to the classes defined by Pappaemmanuil et al. and, accordingly, they became more accurate in capturing the clinical effectiveness of both the reference classification systems. The enhanced correspondences encouraged to study the potential survival prediction that the MFNCHD-based classifier could provide. Notably, the concordance of the CPH model fitted over the MFNCHD-based components was higher than concordance obtained when using the multinomial-based components. This was true for both on the subjects that could be classified by WHO and the subject who could not. In particular, a powerful advantage of both automatic classifiers was that they did not yield ambiguous classifications. They could both handle quite smoothly the presence of multiple alterations, which results in all subjects of the cohort being assigned. Interestingly, the classifiers could also be run on new AML subjects that did not take part in the HDMM fitting. In this way they can provide a tool to perform real-time classification. Intuitively, an HDMM with underlying MFNCHD could be the next target for the modelling of AML. Although theoretically possible, the computational limitations due to the exact calculation of the partition functions of the MFNCHDs makes the implementation of an HDMM with underlying MFNCHD unfeasible. Nonetheless, there is still room for improvement on the molecular characterization of AML. For example, HDMM could not separate two clinically relevant alterations: t(15;17) and t(6;9). Their strong co-occurence with the mutations of FLT3 joined them in a unique component that was driven by them when using the MFNCHD and was driven by FLT3$^{\mathrm{ITD}}$ when using the multinomial distribution. It is likely that an increased representation of both cases in the cohort would have forced the single component to split in two.

This work, thanks to the MFNCHD approach, provided new insights to design more advanced classification system for AML and new tools to perform automatic unbiased classification, which helps to shed light on known ambiguous cases.

Chapter 4

Empowering In-depth definition of Simvastatin and Ezetimibe Effects in Humans by Intelligible Heterogeneous Networks

4.1 Introduction

Recent market growth of advanced high-throughtput technologies for molecular biology has fed interest in modeling the complex nature of the human diseases for novel preventive and therapeutical opportunities.

Atherosclerotic cardiovascular disease (ASCVD) is one of the major global health threats. Currently, lipid-lowering therapy remains the cornerstone treatment for ASCVD with statins as the therapeutic mainstay in hypercholesterolemia for both primary and secondary prevention [68]. Despite the established benefit of cholesterol lowering, a significant part of the population fails to reach the guideline-directed low-density lipoprotein (LDL) total cholesterol (LDL-TC) therapeutic goals with statin monotherapy. Hence, combined therapies are recommended, which include the intestinal and biliary cholesterol absorption inhibitor ezetimibe as the first complementary drug-of-choice. Ideally, the preventive and therapeutical approaches should target all the known treatable risks and multiple factors for ASCVD. These therapeutic choices should be based on comprehensive knowledge of the complexity and heterogeneity of their molecular effects on the underlying biological mechanisms which may drive ASCVD and/or other (patho)physiologic processes. A comprehensive understanding of drug-induced perturbations is feasible when the intricate system of molecular interactions is considered as represented by molecular interaction networks [15]. Network medicine [16, 32, 69] serves natural architectures to unravel the complexity of biological processes. This network-based discipline of system's biology exploits graph theory to model bottom constituents of living cells (i.e. genes, proteins, metabolites) into high order organizations [17,30,33]. Such organizations are traditionally isolated in single-omics, which limit the identification of complexity facets of the biologically complex system [29, 31, 70]; however, the increasing availability of multi-omics databases promise to enhance the description of large heterogeneous biological organizational structures and how these are perturbed by drug treatments.

Research on cardiovascular disorders is rapidly evolving within the framework of Network Medicine [71–73], where molecular interaction networks, or interactomes, play pivotal roles. These objects are unbiased maps of cellular interactions across biological constituents (e.g., genes, proteins, and small molecules) represented by graphs or networks [74]. Human protein-protein interactomes, derived from physical protein-protein interaction (PPI) ascertainment, are biological networks that encode the physical associations between proteins. Although their heterogeneity still indicates a much needed effort to solve discrepancies [75] many PPI-based interactomes have been proposed [76,77]. They provide a solid platform to explore higher order, biological systems-wide behaviours rather than simply multiple independent analysis on genes or proteins (i.e. genes products).

In this work, to facilitate interpretation and clinical implementation, heterogeneous networks were created to characterize the complex interactions between the molecular effects in the liver (endophenotype), plasma, and bile (peripheral phenotype) exerted by two of the most prescribed lipid-lowering drugs for prevention and management of ASCVD in humans, i.e., simvastatin and ezetimibe. Such networks revealed to be intelligible heterogeneous networks. The Stockholm Study, in which subjects eligible for cholecystectomy were randomized to simvastatin, ezetimibe, combined treatment (simvastatin and ezetimibe), or placebo for 4 weeks prior to surgery, generated different types of data: liver transcriptomics and methylomics, and biochemical parameters such as biliary lipids, lipoprotein lipid composition, and atherogenic characteristics.

Starting from the human protein-protein interactome and integrating all generated data, heterogeneous networks representing the significant biological perturbations induced by the different treatments in subjects were generated. Their comparative analysis facilitated dimensionality reduction and highlighted the discernible modules within which the unique effects of the intervention are manifest. Thereafter, in order to perform a first experimental validation of some of the findings from the newly identified heterogeneous modules and networks, a genetically modified human hepatocyte-like cells (SOAT2-only-HepG2) was utilized. This cellular model aims at mimicking the human hepatic lipid metabolism [78]. By exploiting Network Medicine, this work defined an innovative approach to create heterogeneous networks able to increase the interpretability of the analytical results and to facilitate their clinical implementation. By exploiting Network Medicine, this work showed how intelligible heterogeneous networks are able to increase the interpretability of the analytical results and to facilitate their clinical implementation.

4.2 Methods

4.2.1 Dataset pre-processing

The Stockholm Study [79] is a single-blind, randomized trial showing that addition of ezetimibe to simvastatin treatment caused a significant reduction of plasma cholesterol, cholesteryl esters, and TG in remnant particles. Forty non-obese, normolipidemic subjects with uncomplicated cholesterol gallstone disease, eligible for elective cholecystectomy, were enrolled in the randomized 4-week, single-blind, placebo-controlled treatment that included: simvastatin 80mg daily (s), ezetimibe 10mg daily (E), simvastatin 80mg and ezetimibe 10mg daily (S+E), or placebo (P). The cohort included 13 fertile females, 12 post-menopausal females and 14 males between 25 and 80 years old. Fasting blood samples were collected at the first and at the end-of-study visits. Biliary BA and liver biopsies were collected after overnight fasting during the surgical procedure. After filtering and pre-processing the cohort studied included 33 subjects, whose data were utilized to ultimately build the intelligible networks.

4.2.2 Human protein-protein interaction network

In this study, the Human Interactome (HI) [80] contained human physical, macromolecular interaction data from different sources, including protein-protein interactions, protein complexes, kinase-substrate interactions, and signalling pathways. High-quality protein-protein interactions were from several high-throughput yeast-two-hybrid studies, mass spectrometry, as well as the curated literature. The latest large-scale binary PPI were retrieved from HuRI [81]. In addition, experimental signalling interactions and kinase-substrate interactions, as well as high-quality literature-based signalling interactions, were also incorporated [82–85]. The HI exploited in this study had 16,470 proteins and 233,957 interactions.

4.2.3 Gene Expression Profiling

RNA-seq data was normalized with the *rlog* method of DESeq2 [86] and Principal Component Analysis (PCA) was performed on the normalized gene expression data to check the presence of strong outliers[1]. Recursively, the strongest outlier on either one of the two first principal components was filtered out until no more strong outliers appeared. Only one sample was removed. Next, the DESeq2 model was used to analyse whether the expression of any gene depended on the treatment protocol or not, independently

[1] Details on RNA-seq normalization and DESeq2 are respectively reported in sections 6.3.1 and 6.3.2

from gender, age and BMI. The Gene Expression Profiling (GEP) analysis was performed for each contrast of treatment groups. In detail, the significance of the regressed coefficients was estimated with the Wald test and the $\log_2$-fold changes (LFC) were reduced for each pair of treatment protocols with the shrinkage estimator ashr as proposed by Stephens [87]. At last, p-values were adjusted for multiple-testing with the Benjamini-Hochberg method [88]. Genes with adjusted p-value below 0.05 were considered as significant.

4.2.4 Genes – biochemical parameters association

Every biochemical parameter underwent a two-step filtering process to keep outlier values out of the further analyses. First, values related to wrong data acquisitions were removed. Next, values out of the range of three times the standard deviation with respect to the average were considered as outliers and therefore removed. Outliers were recursively eliminated until none was found. The statistical association between biochemical parameters and genes expression was still estimated with the DESeq2 model. Here, though, a simple model within each treatment group was employed without controlling for gender, age and BMI. This choice was strongly influenced by the extremely small subcohorts sizes. The Wald test was utilized to evaluate regressed coefficients relevance; yet, the shrinkage per unit of changes was this time performed with the Bayesian shrinkage estimator apeglm proposed by Zhu, et al. [89]. To account for likely under-powered tests, a significant threshold equal to 0.001 was set on Benjamini-Hochberg adjusted p-values. In addition, changes between coefficients were considered significant when greater than 0.1.

4.2.5 DNA methylation analysis

Methylation data was available for 28 subjects (out of 33) with none of them resulting as an outlier on the first two components of a standard PCA. Built-in explicit SNP probes (‘rs’ probes) were removed as well as probes with missing values or belonging to chromosome X or Y. Next, based on the work of Zhoue et al. [90], probes with internal SNPs close to the 3’ end of the probe were taken out, along with probes with non-unique mapping to the bisulfite-converted genome and probes with off-target hybridization due to partial overlap with non-unique elements. At last, methylation β-values were turned into M-values to work on objects with better statistical properties [91]. The identification of CpGs whose methylation values differ according to the treatment was obtained by using a linear regression model [92] for each pair of treatments. An empirical Bayes smoothing of the standard errors was then employed to borrow information across the CpGs and to obtain a more stable inference and improved power [93]. Possible bias and inflation of the test statistic due, for instance, to the presence of unknown confounding variables or violations of the test assumptions, were then adjusted using bacon [94].

Eventually, the p-values were adjusted for multiple testing using the Benjamini- Hochberg method. To identify Differentially Methylated Regions (DMRs) the comb-p method [95] was utilized over the CpGs p-values (see section 6.3.3). P-values were also adjusted for multiple testing using the Benjamini-Hochberg method and the threshold for significance was set to 0.01.

4.2.6 Building treatment-specific heterogeneous networks

The significant differentially expressed (DE) genes were all collected from the GEP analyses. Each treatment resulted to have multiple DE genes when put in contrast with all other three treatments. The set of DE genes across all contrasts of a single treatment was referred to as seeds list. The seeds for each treatment were mapped onto the HI, together with 18 genes previously recognized as significantly affected by the treatments (Supplementary Table C.1). Thereafter, the Network Propagation algorithm [96] (details in section 6.3.5) was run until its convergence, in order to spread the information given by DE genes through the HI. Later, the connected components to the seed genes were identified, and the top 50 genes, if available, were predicted from their neighbourhood, ranked by the diffusion score. Subsequently, it was determined whether genes with statistical associations to biochemical parameters or DMRs were neighbours either of seeds or predicted genes. Lastly, isolated genes and genes that could not be mapped onto the HI were filtered out. Particularly noteworthy, several significant results for all reductionist analyses were not located within any gene region, or were not mappable to the HI, and, thus, were removed.

4.2.7 Gene prioritization and gene module analysis

The network propagation algorithm inherently provided a way to prioritize genes. To consider all data sources, the properties of the individual networks were emphasized. The betweenness centrality [97], which represents how frequently a node occurs in all of the shortest paths of a graph (section 6.3.4), was then utilized to define a gene prioritization rule. To highlight only the genes with very large betweenness values, only genes above the third quartile were considered, and the seed genes were excluded. Gene seeds exclusion was chosen to explore exclusively those genes either not emerging directly from the GEP or not manually added. Furthermore, gene modules were targeted and searched as maximal cliques. In graph theory, cliques are completely connected groups of nodes. Moreover, maximal cliques are the cliques that do not belong to any other clique. A modified version of the Bron–Kerbosch algorithm [98] was run and maximal cliques with less than three genes were neglected. The identified maximal cliques were called modules.

4.2.8 Disease classes exploration and GO enrichment analysis

To link genes to specific diseases, DisGeNet [76] was exploited. To determine general disease classes, the CUI identifier of the specific diseases was linked to the MeSH identifier. Genes modules were thoroughly examined to control whether they enriched GO terms or not. The standard hypergeometric test was performed to establish GO enriched terms. The criteria for significance included p-values lower than 0.01 and number of genes within the interval $[10, 300]$.

4.3 Results

4.3.1 Treatment-specific heterogeneous networks

Gene Expression Profiling (GEP) from the pairwise comparisons between all the treatments generated lists of differentially expressed (DE) genes. These lists included all the DE genes found for each individual treatment in at least one comparison. Totally 19 DE genes for simvastatin, 60 for ezetimibe, 130 for combined treatment (simvastatin and ezetimibe), and 99 for placebo treated subjects were identified. All the lists of DE genes, referred to as seeds as previously mentioned, were next enriched with 18 genes already well characterized to participate in lipid metabolism or to be affected by lipid-lowering treatments (Supplementary Table C.1). For every treatment, the seeds were projected onto the HI and were used to predict via the Network Propagation algorithm the most informative genes in their neighbourhoods. Thereafter, the initial GEP-based networks were extended to include the biochemical parameters as another layer of information. The biochemical parameters included plasma biochemical laboratory analyses, bile lipid analyses, lipoprotein lipid composition, and their atherogenic properties defined as binding to human arterial proteoglycans (PG-Binding). Using generalized linear models and exploring the topology of the HI, the biochemical parameters were incorporated into the networks using the genes statistically associated with at least one of them. Ezetimibe-treated subjects reported the highest number of associations (67 genes associated with 29 biochemical parameters), followed by subjects receiving combined treatment (52 genes associated with six parameters), while placebo-treated subjects reported 24 genes associated with six biochemical parameters. Simvastatin-treated subjects had the least number of associations, ten genes associated with 14 biochemical parameters. Finally, the networks were further extended by including information from DNA methylation profiling and discovered several genes with differentially methylated regions (DMRs). Nevertheless, more than 80% of these regions were either unmappable or isolated in the interactome module for each of the different treatments. Subjects given the combined treatment showed the highest number (18) of differentially methylated (DM) genes, in which at least three consecutive and DM CpG sites were identified. This group was followed by ezetimibe-treated subjects (15), simvastatin-treated subjects (three), and placebo-treated subjects (one). Based on the final heterogeneous networks (Figures 4.1-4.4), diffusion scores indicated that seed genes retained most of the information with respect to all other genes (Figure 4.5a). The genes predicted by the Network Propagation algorithm were next in the ability to retain information. The genes associated with biochemical parameters and DMRs retained a level of information that did not differ from the remaining genes. The heterogeneous networks that characterize the different treatments had different topologies. The heterogeneous network of placebo-treated subjects contained almost entirely (99%) genes gathered by GEP (with limited effects of Network Propagation and biochemical associations) and has the largest number of disconnected regions (15). The heterogeneous network of ezetimibe-treated subjects had the most heterogeneous composition. The genes associated with biochemical and lipoprotein parameters covered 32.5% of the network and were located centrally. Plasma C-peptide (C-peptide), LDL triglycerides (LDL-TG), and insulin showed the largest betweenness values. The heterogeneous network of subjects treated with the combined treatment showed the largest fraction of DM genes (8.7%), which were also poorly connected (median degree was one). The heterogeneous network from simvastatin-treated subjects was mainly dominated by genes from transcriptomics (87%), but it also showed a highly connected constellation of genes (especially SORT1, CXCL8, VSIG4, and NR4A2) and biochemical parameters including PG-Binding and its ratio with plasma apolipoprotein B and AI (ApoB and ApoA-I), respectively levels, and biliary bile acid (BA). Approximately 43% of the overall constituents (genes and biochemical parameters) of the networks emerged in at least two of them, with 20 genes shared by all treatment groups. The placebo-treated subjects reported the heterogeneous network with the lowest number of unique constituents (30), followed by simvastatin-treated subjects (45), subjects treated with the combined treatment (62), and ezetimibe-treated subjects (88).

4.3.2 Prioritizing genes by different types

To understand the importance of genes predicted to be affected by the treatments, the betweenness centrality was chosen. Several non-seed genes were above the third quartile of the betweenness distribution in all heterogeneous networks (Figure 4.5b). Twenty-one genes were highly central for more than one treatment. Only four genes emerged in three treatments, APOA1 and APOA2 (for ezetimibe, combined treatment, and placebo), and LPL and PPARA (for simvastatin, ezetimibe, and combined treatment). SORT1 and 16 other genes were shared by two treatment groups. Ranking non-seed genes according to the data source highlighted more detailed information. In the heterogeneous network of simvastatin-treated subjects, six genes (SORT1, CXCL8, JUN, FCER1G, NR4A2, and CSF1R) were central and statistically associated with at least one of the following parameters: ApoA-I, biliary BA, biliary total cholesterol (Biliary-TC), biliary phospholipids (Biliary-PL), plasma gamma-glutamyl transferase (GGT), plasma high density lipoprotein total cholesterol (HDL-TC), HDL cholesteryl esters (HDL-CE), HDL phospholipids

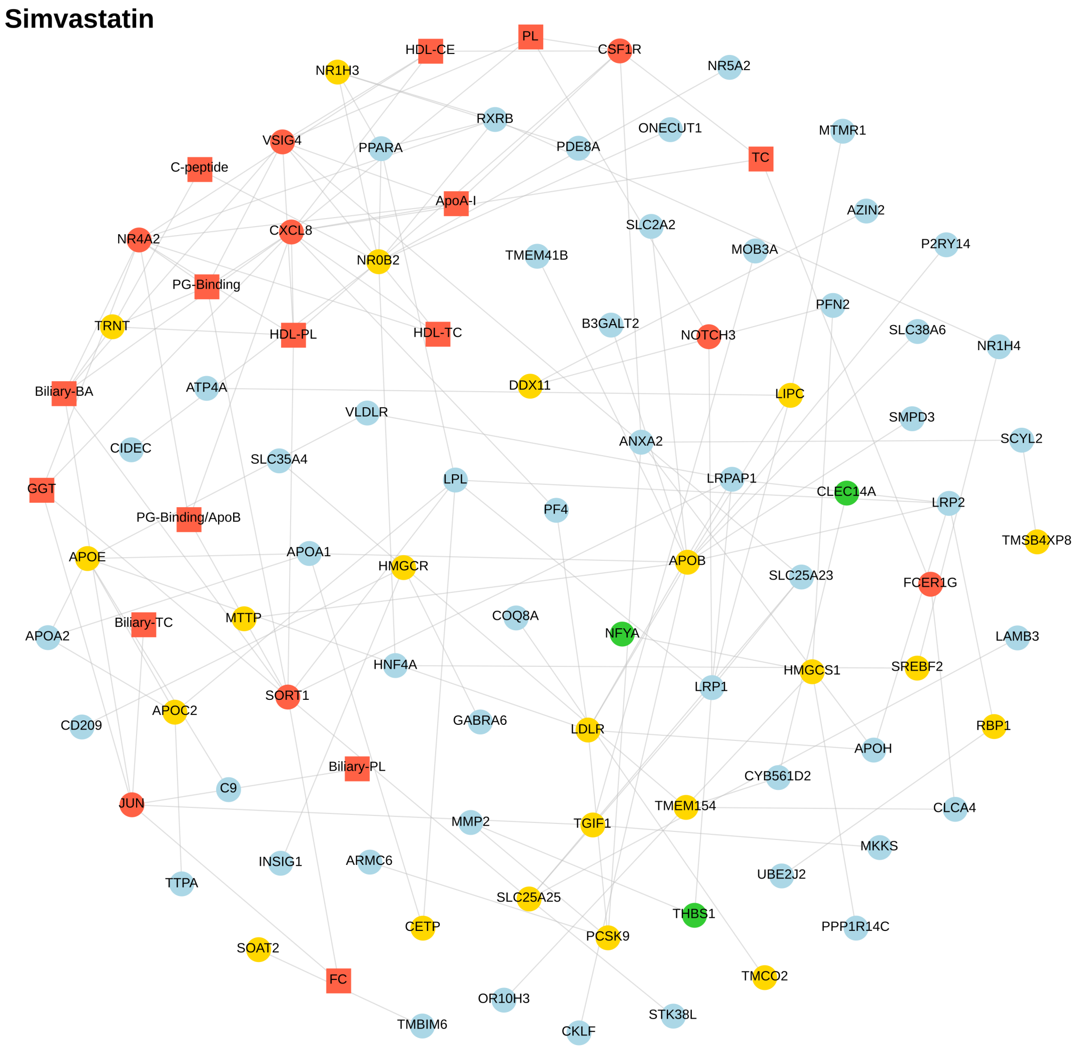


Figure 4.1: Heterogeneous network for simvastatin-treated subjects. Genes are represented by circles and biochemical parameters by squares. The colour of the constituents indicates their data source: genomic seeds (DE genes from GEP, plus genes previously recognized as significantly affected by the treatments) are shown in yellow, predicted DE genes (obtained by Network Propagation algorithm) in light blue, biochemical parameters with their associated genes in red, and DM gene in green. Potentially, DM genes with at least one biochemical parameter association can be found (represented by a green circle with a red border), but none was detected for simvastatin-treated subjects.

(HDL-PL), PG-Binding, PG-Binding/ApoB, plasma total cholesterol (Plasma-TC), total free cholesterol (Plasma-FC), and phospholipids (Plasma-PL). In the heterogeneous network of ezetimibe-treated subjects, THBS1 was included among the top central genes and

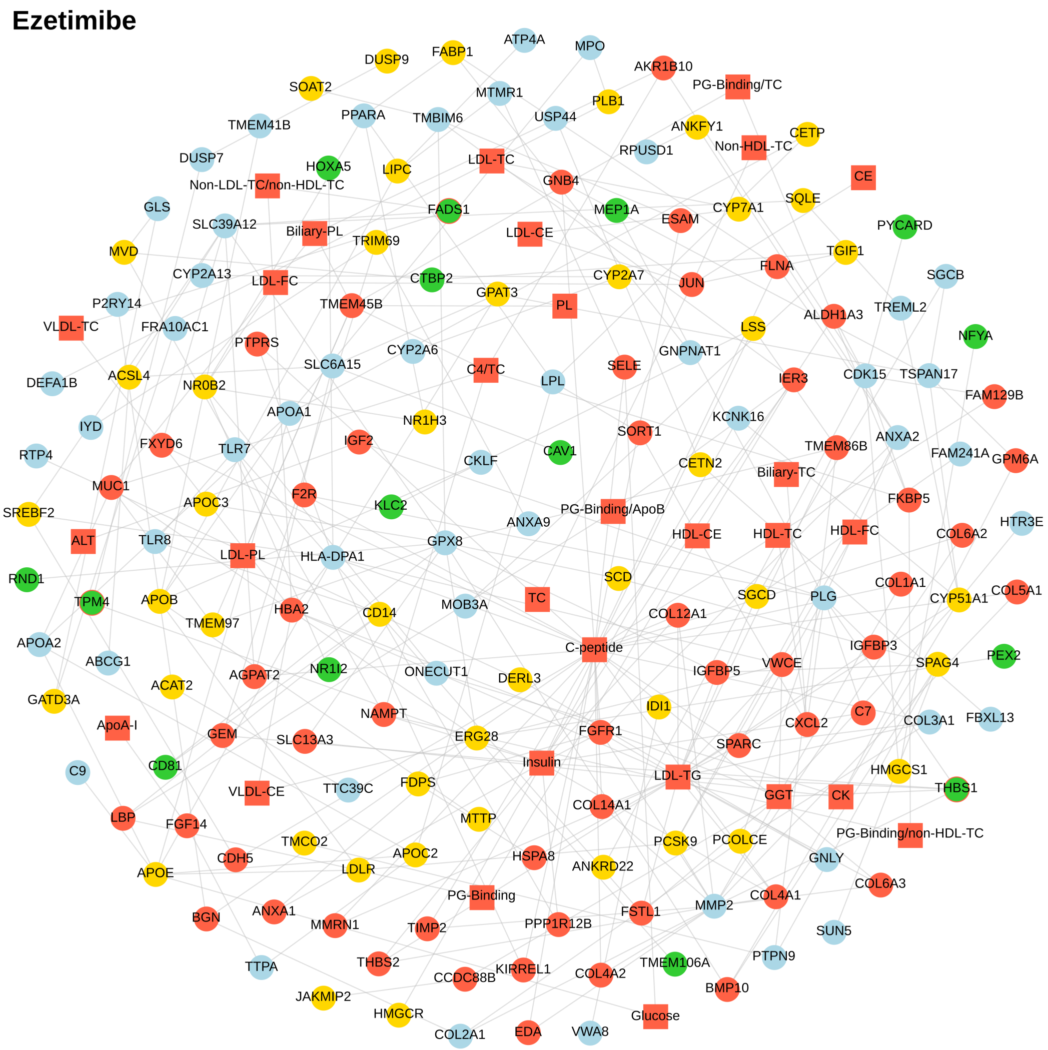


Figure 4.2: Heterogeneous network for ezetimibe-treated subjects. Genes are represented by circles and biochemical parameters by squares. The colour of the constituents indicates their data source: genomic seeds (DE genes from GEP, plus clinically known genes) are shown in yellow, predicted DE genes (obtained by the Network Propagation algorithm) in light blue, biochemical parameters with their associated genes in red, and DM genes in green. Additionally, DM genes with at least one biochemical parameter association are represented by green circles with a red border.

shown to be both DM and associated with three plasma biochemical parameters (insulin, LDL-TG, and GGT). In addition, NAMPT and THBS1 had the largest betweenness, respectively, for the genes with biochemical associations and genes with DMRs. In both the combined treatment and heterogeneous network of placebo-treated subjects, AQP6

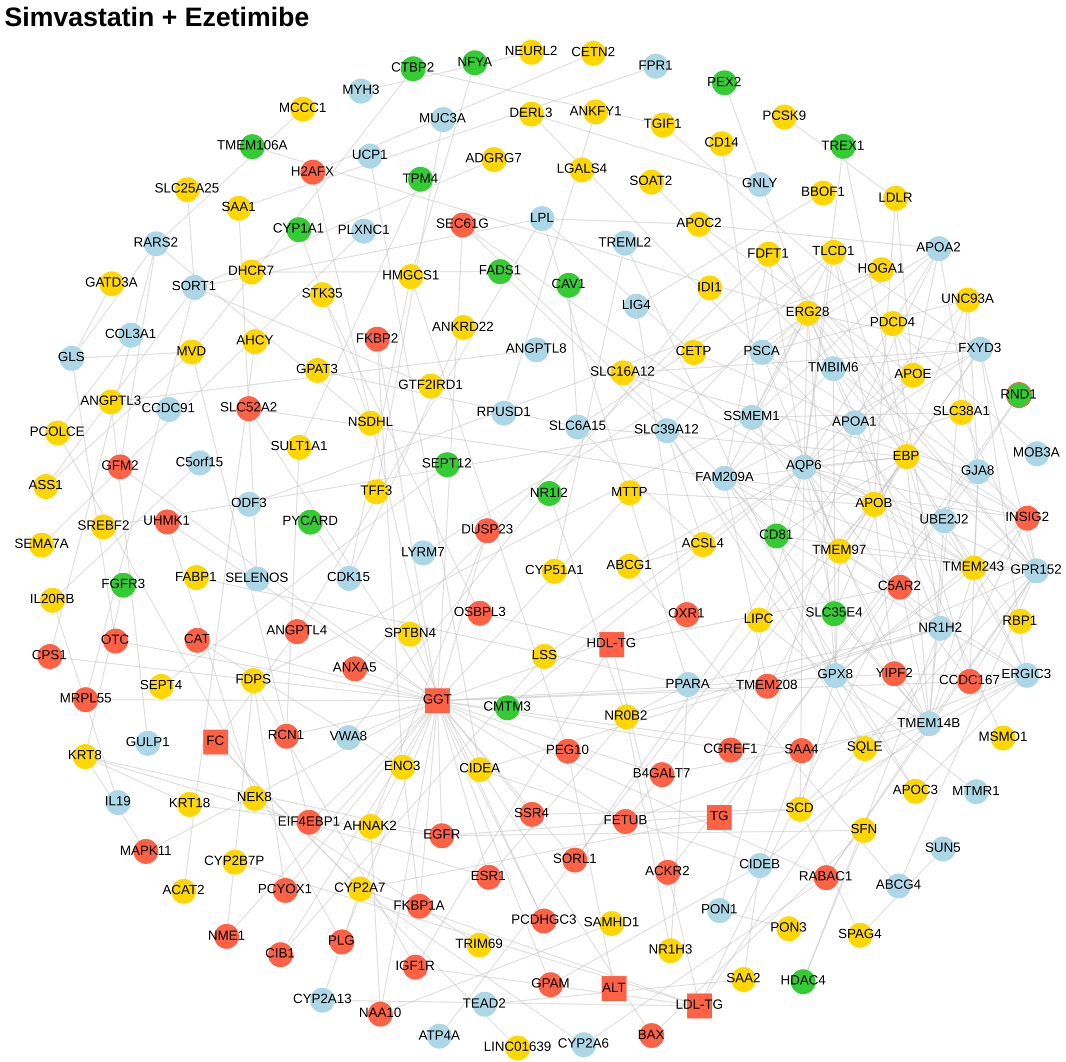


Figure 4.3: Heterogeneous network for subjects treated with combined therapy. Genes are represented by circles and biochemical parameters by squares. The colour of the constituents indicates their data source: genomic seeds (DE genes from GEP, plus genes previously recognized as significantly affected by the treatments) are shown in yellow, predicted DE genes (obtained by the Network Propagation algorithm) in light blue, biochemical parameters with their associated genes in red, and DM genes in green. Potentially, DM genes with at least one biochemical parameter association can be found (represented by a green circle with a red border).

is the most central gene. Notably, several genes in the combined treatment network are central due to their connectivity with plasma GGT, which resulted as the most central constituent. Mapping the top central non-seed genes to DisGeNet revealed associations

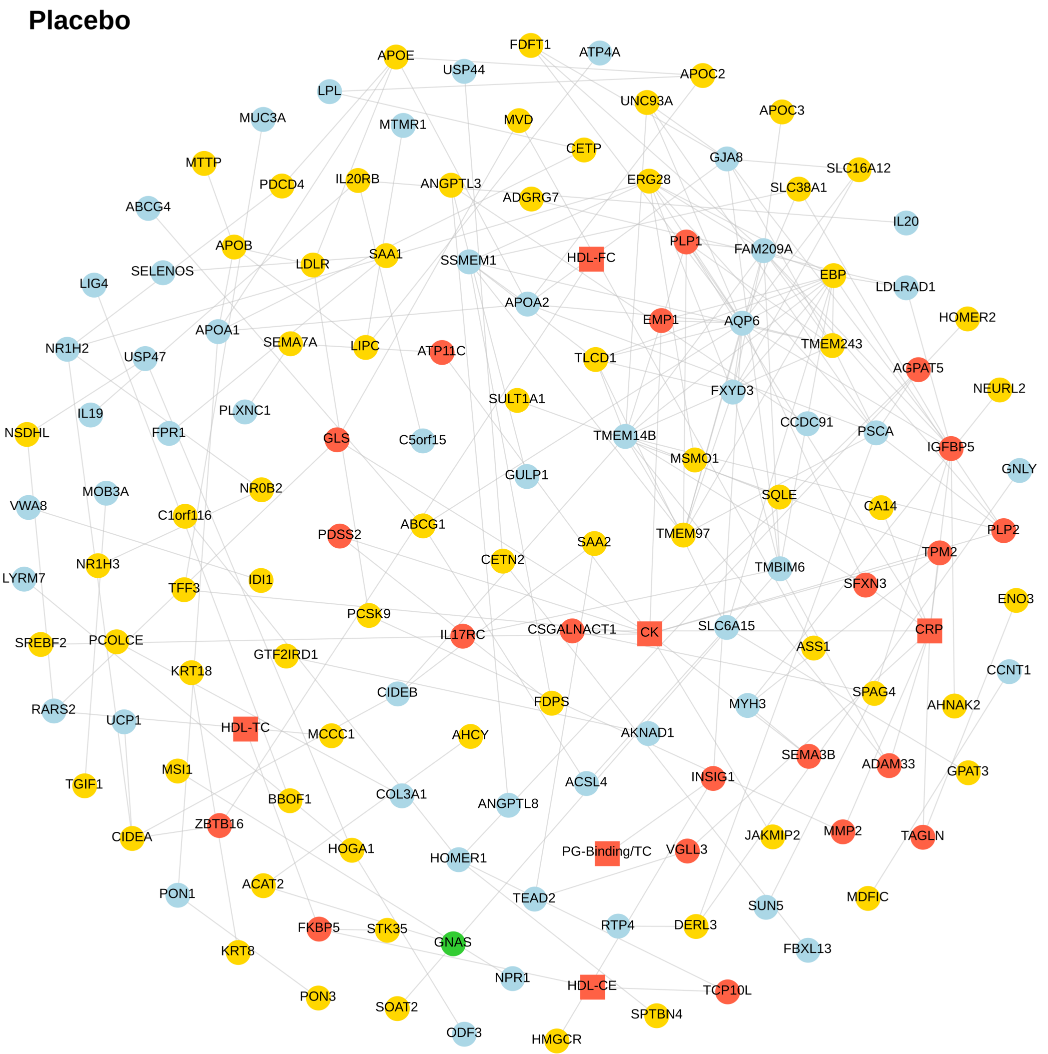
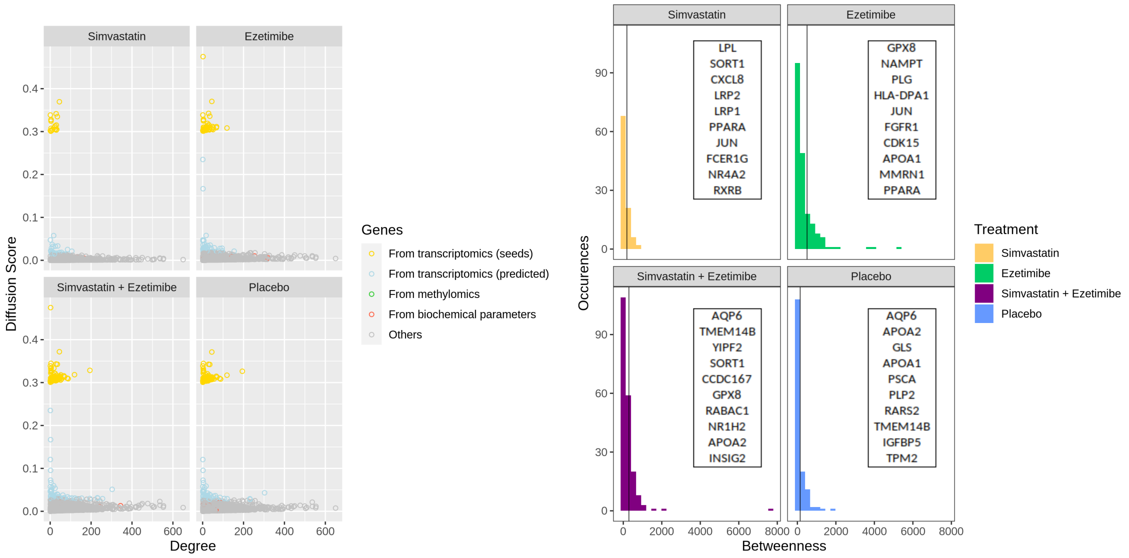


Figure 4.4: Heterogeneous network for placebo-treated subjects. Genes are represented by circles and biochemical parameters by squares. The colour of the constituents indicates their data source: genomic seeds (DE genes from GEP, plus genes previously recognized as significantly affected by the treatments) are shown in yellow, predicted DE genes (obtained by the Network Propagation algorithm) in light blue, biochemical parameters with their associated genes in red, and DM gene in green. Potentially, DM genes with at least one biochemical parameter association can be found (represented by a green circle with a red border), but none was detected for placebo-treated subjects.

to similar diseases classes (under the MeSH identifier) across the treatments (Figure 4.6), and oncologic and digestive system diseases were the most represented ones. Of interest, liver malignant neoplasms were associated to the previously mentioned PPARA and APOA1, which were among the non-seed genes with the highest betweenness in the heterogeneous networks of the treatments. The same genes were also related to digestive system diseases mainly involving alcoholic and non-alcoholic steatohepatitis (ASH and NASH, respectively). Liver cirrhosis was highlighted in all treatments and related to different sets of genes in each multi-source network, including ANXA2, LPL, GPX8, GLS, and MMP2.

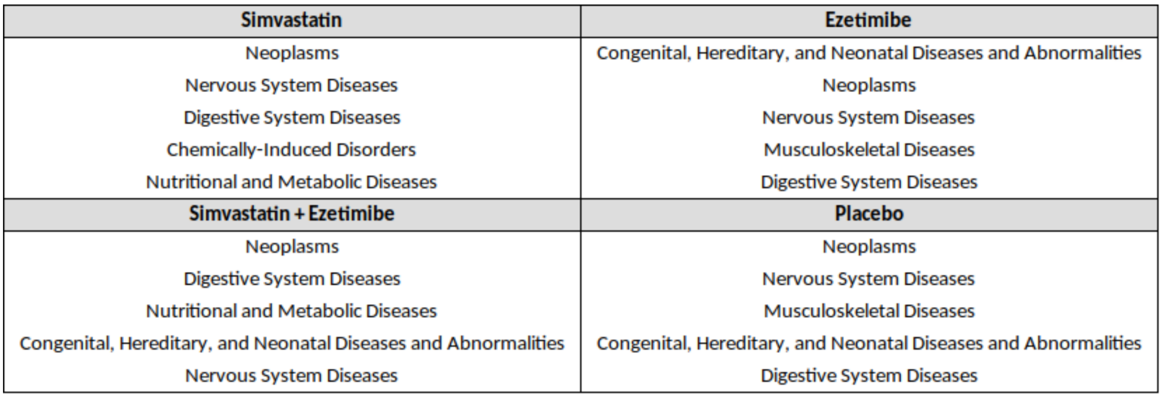


(a) (b)

Figure 4.5: The genes diffusion score distribution as a function of degree is illustrated on the left. The diffusion score quantifies the amount of information, driven by the seed genes, that each gene retains. The degree is the number of associations of a specific node, determining its centrality in the module. On the right, the distribution of betweenness centrality is reported for every treatment, with the vertical axis highlighting the third quartile. Tables within plots are comprised of the ten most central non-seed genes.

4.3.3 Identification of treatment-unique modules

To address possible unique effects of the different treatments, the heterogeneous networks were analysed and the maximal cliques, herein referred to as modules, were searched using a modified version of the Bron–Kerbosch algorithm. There were 47 modules in the network for subjects treated with the combined treatment, 21 in placebo-treated subjects, eight in ezetimibe-treated subjects, and four in simvastatin-treated subjects. The heterogeneous module for subjects given the combined treatment formed a unique



Simvastatin	Ezetimibe
Neoplasms	Congenital, Hereditary, and Neonatal Diseases and Abnormalities
Nervous System Diseases	Neoplasms
Digestive System Diseases	Nervous System Diseases
Chemically-Induced Disorders	Musculoskeletal Diseases
Nutritional and Metabolic Diseases	Digestive System Diseases
Simvastatin + Ezetimibe	**Placebo**
Neoplasms	Neoplasms
Digestive System Diseases	Nervous System Diseases
Nutritional and Metabolic Diseases	Musculoskeletal Diseases
Congenital, Hereditary, and Neonatal Diseases and Abnormalities	Congenital, Hereditary, and Neonatal Diseases and Abnormalities
Nervous System Diseases	Digestive System Diseases

Figure 4.6: The top five general disease classes, extracted from the MeSH identifiers in DisGeNet, that were connected to the most central non-seed genes for every treatment. Neoplasms generally refer to different types of carcinomas, often involving hepatic tissue (i.e., malignant neoplasm of liver). Digestive system diseases included mainly hepatic disorders, such as alcoholic and non-alcoholic steatohepatitis (ASH and NASH, respectively).

densely connected component, similarly to that observed for the placebo-treated subjects. Conversely, simvastatin- and ezetimibe-treated subjects respectively reported two and three connected components. In total, 64 unique modules emerged, 52 uniquely linked to a single treatment, and 12 linked to multiple treatments. Interestingly, two modules, APOE, APOC2, plus APOA2, and APOE, APOA1, plus APOA2, were shared by all treatments (Figure 4.7) and were connected. The heterogeneous networks from subjects given the combined treatment showed the largest overlap of modules (nine) with the heterogeneous network of placebo-treated subjects, and only one overlap of modules with the heterogeneous network of subjects treated with ezetimibe (NR0B2, NR1H3 and PPARA). All genes within modules showed diffusion scores above the third quartile of the score distribution (Figure 4.8a) with no clear pattern with respect to their degree. Additionally, there were 40 genes belonging to a single treatment module. The heterogeneous network for placebo-treated subjects had four uniquely associated genes within its module (EMP1, IGFBP5, PLP1, and PLP2), and simvastatin-treated subjects had four as well (LRP1, RXRB, SLC25A23, and SLC25A25). Ezetimibe-treated subjects, instead, reported eight (CD14, COL3A1, HLA-DPA1, KIRREL1, PTPRS, SPARC, TLR7, and TLR8), and the heterogeneous network for subjects treated with the combined treatment had 24 (AHNAK2, CCDC167, CD81, CYP2A13, CYP2A7, DUSP23, EGFR, ERGIC3, FDFT1, FETUB, GPR152, GPX8, INSIG2, KRT18, KRT8, NAA10, NR1H2, RABAC1, SCD, SLC35E4, TMEM208, TREX1, UNC93A, and YIPF2). Only genes from transcriptomics, both seed and predicted genes, comprised the modules for simvastatin-treated subjects. Even the modules for subjects receiving placebo included only transcriptomic genes, but, in addition, genes having associations with biochemical parameters including lipoprotein features, emerged. Conversely, biochemical parameters were present in both the modules identified in the heterogeneous network from subjects receiving either ezetimibe or the combined treatment. In the module from ezetimibe-treated subjects, gene SPARC was observed to be linked to LDL-TG (positive correlation) and to C-peptide (negative correlation). Furthermore, plasma C-peptide was connected to PTPRS and KIRREL1 (both having negative correlations). In the module from subjects receiving the combined treatment, plasma GT showed eight connections (CCDC167, DUSP23, EGFR, FETUB, NAA10, RABAC1, TMEM208, and YIPF2), while there was only one edge between LDL-TG and INSIG2 (positive correlation). Moreover, the combined treatment module was the only module that included genes with DMRs. CD81 and SLC35E4 had a 3-CpG-long DMR and TREX1 had a 4-CpG long DMR spanning through another gene (ATRIP). Performing Gene Ontology (GO) enrichment analysis on all treatment modules separately, without removing the shared modules, resulted in biological functions dominated by the APOE, APOC2, APOA1, and APOA2 cluster. Chylomicron assembly (GO:0034378), very low-density lipoprotein (VLDL) particle remodelling (GO:0034372), chylomicrons (GO:0042627), and PL efflux (GO:0033700) were at the top of the enriched biological functions list. In contrast, when the shared modules were filtered, enriched biological functions depended on treatment except for placebo, where none was found, as shown in Figure 4.8b. Transmembrane transporter activity-related pathways and the regulation of cholesterol storage-related pathways were also observed for simvastatin-treated subjects, while the regulation of interferon production-related pathways were also observed for ezetimibe-treated subjects. For subjects treated with the combined treatment, hepatocyte apoptotic process (GO:0097284) was at the top of the enriched biological functions list, followed by sterol and cholesterol biosynthetic processes-related pathways.

4.3.4 Experimental validation of putative target genes

By analysing the heterogeneous networks based on different types of data from the Stockholm Study, putative target genes highlighted in treatment-specific disease modules were identified. The solute carrier family 25 member 25 (SLC25A25) and the transmembrane BAX inhibitor motif containing 6 (TMBIM6) genes appeared in different networks as either a DE seed or as a predicted DE gene, respectively. As these two genes have not previously been shown to be affected by lipid-lowering drugs they were experimentally explored in genetically modified cells. SOAT2-only-HepG2 cells were used since, like human hepatocytes in vivo, they could only express sterol-O-acyltransferase 2 (SOAT2) and not SOAT1. Because simvastatin could not be functionally activated by hepatocytes alone, the cells were treated with the active compound atorvastatin instead. The SLC25A25 gene encodes a transmembrane carrier that facilitates transport of solutes across the inner mitochondrial membrane. Moreover, it plays an important role in main-

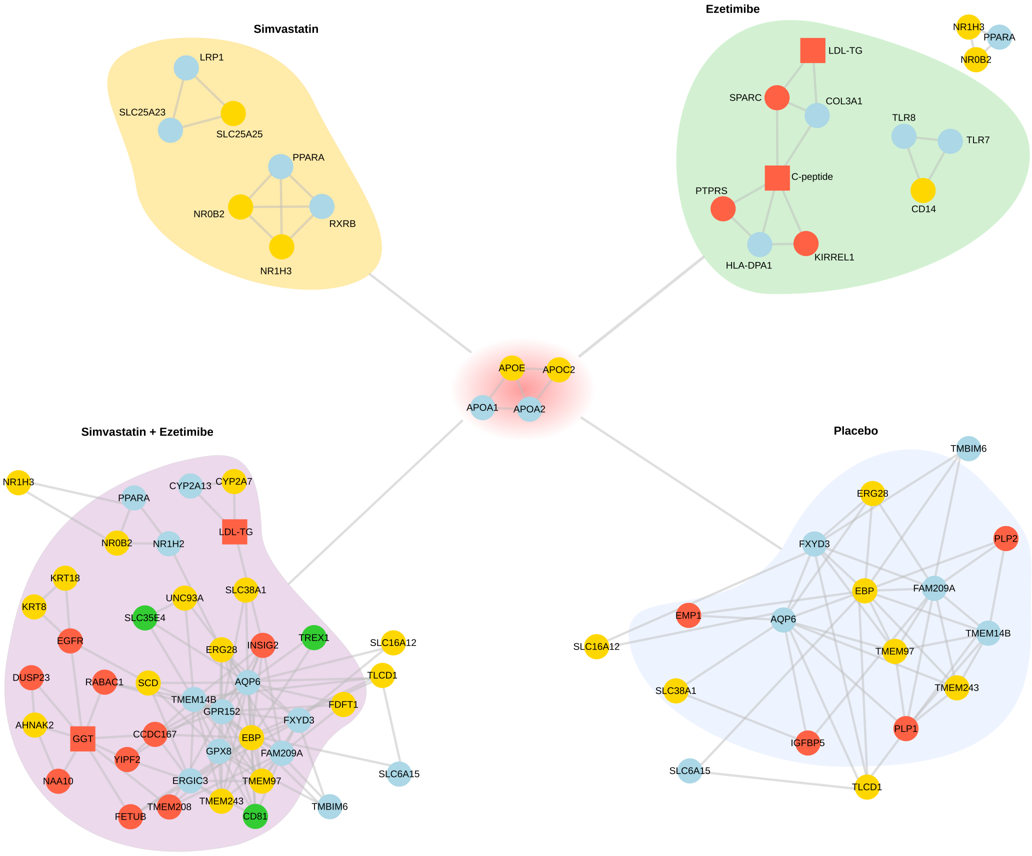
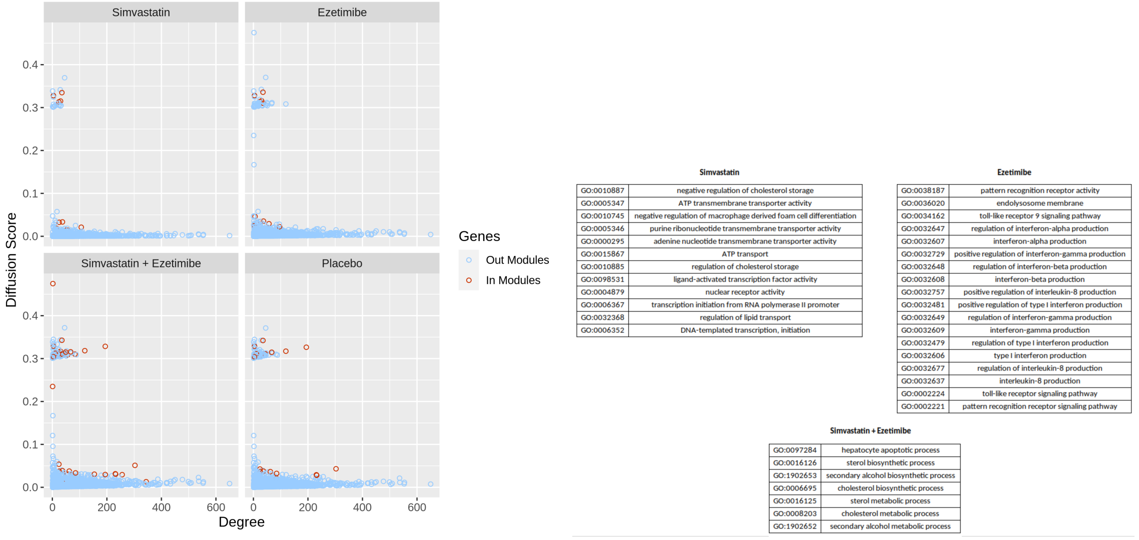


Figure 4.7: Gene modules for the heterogeneous networks of treatments, i.e., simvastatin, ezetimibe, combined therapy, and placebo. The modules, i.e., maximal cliques, formed a unique densely connected component for both the combined therapy and placebo, whereas for simvastatin and ezetimibe, two and three connected components were found, respectively. The central group of genes, comprising two connected modules (APOE-APOC2-APOA2 and APOE-APOA1-APOA2), is common to all treatments. Coloured regions highlight genes belonging to modules uniquely found for the specific treatment.

taining mitochondrial metabolism and ATP production [99]. As shown in Figure 4.9, treatments with atorvastatin alone or in combination with ezetimibe did not increase SLC25A25 expression in the pre-clinical model. TMBIM6, also known as BAX-inhibitor 1 (BI-1), is a transmembrane protein involved in the control of apoptosis through different activities in different subcellular compartments; it is also involved in the progression of several types of cancers in humans. As shown in Figure 4.10, treatment with ezetimibe alone or in combination with atorvastatin increased TMBIM6 expression compared to vehicle, i.e. placebo, and combined treatment exerted a similar effect with respect to statin alone. To be noted, hepatic data from subjects did not show any significant effect



(a)

Simvastatin

GO:0010887	negative regulation of cholesterol storage
GO:0005347	ATP transmembrane transporter activity
GO:0010745	negative regulation of macrophage derived foam cell differentiation
GO:0005346	purine ribonucleotide transmembrane transporter activity
GO:0000295	adenine nucleotide transmembrane transporter activity
GO:0015867	ATP transport
GO:0010885	regulation of cholesterol storage
GO:0098531	ligand-activated transcription factor activity
GO:0004879	nuclear receptor activity
GO:0006367	transcription initiation from RNA polymerase II promoter
GO:0032368	regulation of lipid transport
GO:0006352	DNA-templated transcription, initiation

Ezetimibe

GO:0038187	pattern recognition receptor activity
GO:0036020	endolysosome membrane
GO:0034162	toll-like receptor 9 signaling pathway
GO:0032647	regulation of interferon-alpha production
GO:0032607	interferon-alpha production
GO:0032729	positive regulation of interferon-gamma production
GO:0032648	regulation of interferon-beta production
GO:0032608	interferon-beta production
GO:0032757	positive regulation of interleukin-8 production
GO:0032481	positive regulation of type I interferon production
GO:0032649	regulation of interferon-gamma production
GO:0032609	interferon-gamma production
GO:0032479	regulation of type I interferon production
GO:0032606	type I interferon production
GO:0032677	regulation of interleukin-8 production
GO:0032637	interleukin-8 production
GO:0002224	toll-like receptor signaling pathway
GO:0002221	pattern recognition receptor signaling pathway

Simvastatin + Ezetimibe

GO:0097284	hepatocyte apoptotic process
GO:0016126	sterol biosynthetic process
GO:1902653	secondary alcohol biosynthetic process
GO:0006695	cholesterol biosynthetic process
GO:0016125	sterol metabolic process
GO:0008203	cholesterol metabolic process
GO:1902652	secondary alcohol metabolic process

(b)

Figure 4.8: The genes diffusion score distribution as a function of degree showing where genes inside modules localize (on the left). The diffusion score quantifies the amount of information, driven by the seed genes, that each gene retains. On the right the Gene Ontology (GO) terms enriched by the genes of unique modules are presented.

of treatment on TMBIM6 expression.

4.4 Discussion

Using multidimensional data of different types from the Stockholm Study cohort, intelligible heterogeneous networks were created as tools to investigate the different endophenotypes and peripheral phenotypes resulting from treatment of subjects with two common lipid-lowering drugs, i.e., simvastatin and ezetimibe, alone or in combination. The heterogeneous networks gathered coherent signals from different human biological levels, i.e., gene expression and methylation, as well as plasma and bile biochemical parameters and lipoprotein functionality. By creating heterogeneous networks of these data, this work showed a way to address some of the challenges with big data in the characterization of the complexity of biological effects. The identification of unique and shared modules of heterogenous information increased the interpretability of the integrated information, by reducing their dimensionality, and represented a concrete step toward precision medicine. What this works wants to stress out is that solely through

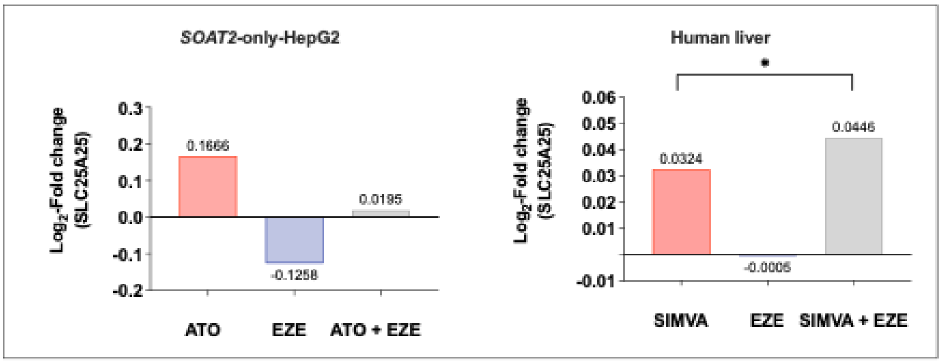


Figure 4.9: On the left, SLC25A25 mRNA expression levels in SOAT2-only-HepG2; on the right, hepatic RNAseq data obtained from liver tissue from subjects in the Stockholm Study. SLC25A25, solute carrier family 25 member 25. Data are expressed as $\log_2$-fold change compared to vehicle (left) or placebo (right). The cells were treated with vehicle (DMSO), atorvastatin $5\mu M$ (ATO), ezetimibe $25\mu M$ (EZE), and atorvastatin $5\mu M$ + ezetimibe $25\mu M$ (ATO + EZE). From the Stockholm Study, human liver samples from subjects treated with simvastatin $80mg/d$ (SIMVA), ezetimibe $10mg/d$ (EZE), combined therapy (simvastatin $80mg/d$ + ezetimibe $10mg/d$; SIMVA + EZE). Statistical analysis on SOAT2-only-HepG2 cells data was performed using multi-variable ANOVA followed by Least Significant Difference test. Human hepatic gene expression data were analysed using DESeq2 according to the presented methodology.

the usage of intelligible networks many information can actually become a resource. Huge sparse networks risk to become a little resource if no clear interpretable and actionable insight can be captured. Here, the intelligible networks revealed to be beneficial for two reasons. The first reason was that the dimension of the networks was little. The second reason was that the heterogeneous constituents of the networks created a context of all biological layers. The ability to provide a context to several reductionist outcomes showed to facilitate the comprehension of treatments effects and to accelerate the decision on what should be investigated. A limitation of this work was the relatively small number of subjects available, determined by the obvious difficulty in recruiting subjects willing to undergo a liver biopsy. Nevertheless, this limitation did not affect the main aim to define an approach for the creation of intelligible heterogenous networks. However, the depicted interactions should be viewed as suggestive of putative, hitherto unknown effects of the lipid-lowering drugs used in the Stockholm Study.

As proof-of-concept, the putative effects of these drugs were preliminary validated on two identified genes in an experimental system. Furthermore, there was also a need to sort out whether the putative effects on genes occur in hepatocytes and/or in other cell types that are present in biopsies from whole liver (e.g., Kupffer cells, stellate cells, liver sinusoidal endothelial cells, and circulating blood cells). Hence, a unique in vitro human

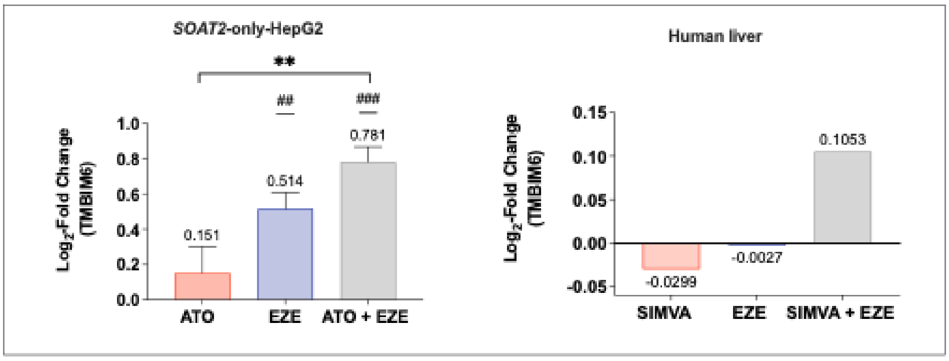


Figure 4.10: On the left, TMBIM6 mRNA expression levels in SOAT2-only-HepG2; on the right, hepatic RNAseq data obtained from liver tissue from subjects in the Stockholm Study. TMBIM6, trans-membrane BAX-inhibiting motif containing 6. Data are expressed as $\log_2$-fold change compared to vehicle (left) or placebo (right). Cells were treated with vehicle (DMSO), atorvastatin $5\mu M$ (ATO), ezetimibe $25\mu M$ (EZE), and atorvastatin $5\mu M$ + ezetimibe $25\mu M$ (ATO + EZE). From the Stockholm Study, human liver samples from subjects treated with simvastatin $80mg/d$ (SIMVA), ezetimibe $10mg/d$ (EZE), combined therapy (simvastatin $80mg/d$ + ezetimibe $10mg/d$; SIMVA + EZE). Multi-variable ANOVA followed by Least Significant Difference test. ## vs. vehicle < 0.005; ### vs vehicle $p < 0.001$; **vs statin therapy alone $p < 0.005$. Human hepatic gene expression data were analysed using DESeq2 according to the presented methodology.

hepatocyte-like cellular model was utilized to better simulate human hepatic lipoprotein and lipid metabolism. The strategy used to construct the heterogeneous networks consisted of the sequential introduction of several sources of biological data to the human protein-protein interactome. Liver transcriptomic data was the backbone of the networks, owing to the rapid transcriptional response to the different treatments, and since the transcriptomic data reflect mechanistically important, underlying molecular effects (i.e., endophenotypes). Only afterwards, plasma and biliary biochemical parameters and the ability of lipoproteins to bind to human arterial PG were added as a second type of information to expand the initial networks of interactions, and to create more heterogeneous and thereby more informative networks. This strategy was chosen because, despite the rapid response to treatments, the biochemical parameters and the lipoprotein functionality characteristics have a much lower representation as important endophenotypes, representing the treatment effects on the peripheral phenotypes most closely related to ASCVD. The epigenetic information was added last because background gene methylation may affect the transcriptomes, the biochemical parameters, and the lipoprotein functionality characteristics. Though, methylation was expected to change at most mildly after only 4-week of treatment. Prior to the addition of the data relative to the biochemical parameters and the lipoprotein functionality characteristics, each individual liver transcriptomic network was enriched with several genes previously recognized as significantly affected by the treatments as seed genes. This addition compensated for the loss of evidence-based information and helped to overcome the difficulty of identifying an optimal network due to the limited cohort size. The heterogeneous networks herein contextualized the putative effects of the lipid-lowering treatments, i.e., simvastatin, ezetimibe, and combined treatment, in a more complex milieu of biological interactions than homogeneous networks from a single type of data could permit. To reduce the dimensionality of the networks and thereby enhance their interpretability, intersection-based steps were carried out in order to add only consistent information from the biochemical parameters and the epigenome. As the putative affected genes differed between the different treatments and between the different sources of information, the gene ranking process was flexible enough to be organized either by treatment or by source of information for a later validation step. CTBP2 was an interesting example, being at the top of the ranked gene list in ezetimibe and combined treatment due to its interactions with several genes in the transcriptome-based network infrastructure but also due to its DM status. Since the CTBP2 methylation status regulates its expression in humans [100], it may be worth investigating whether the methylation of this gene may affect some of the molecular effects of the combined treatment or contribute to some of the variability in lipid-lowering response to the combined treatment [101]. Furthermore, the information from the biochemical parameters and the lipoprotein functionality characteristics could yield nodes having centrality or bridging different part of the transcriptome-based networks. Striking was the centrality of plasma GT and how it conveyed connections across different parts of the network from subjects treated with the combined treatment. To enhance contextuality, the heterogeneous networks were apt to include further layers of information, such as public databases, annotations, and literature-mined data. As matter of fact, the connection of the non-seed genes at the top of the rankings to DisGeNet highlighted classes of diseases that might contribute to a deeper analysis and understanding of the heterogeneity and complexity of the response to the lipid-lowering treatments. The digestive system disease class was common among all treatments and included hepatitis; NASH appeared related to genes that were central in networks for subjects treated with simvastatin and alone or in combination with ezetimibe. These outcomes are not surprising, knowing how hepatitis virus may use machinery involved in hepatic lipid metabolism [102, 103] and recognizing that NASH and ASCVD are considered to comprise the cardiometabolic syndrome [104]. In addition, this work was based on data obtained from subjects suffering from gallstone disease, which was also included in the same disease class. The consistency of these disease classes was additional evidence in support of further investigations into the most central genes of the heterogeneous networks presented here.

Moreover, the intelligibility of the heterogeneous networks indicated distinctive treatment effects based on their modules of gene and biochemical parameters. In fact the distinctive endophenotypes resulting from treatment with simvastatin, and ezetimibe, alone or in combination, physically localized in different areas of the human protein-protein interactome; areas coupled to distinct biological processes. Modules suggested that simvastatin might affect cholesterol metabolism by also altering PPARA and NR1H3, and ATP transport by affecting SLC25A25 and SLC25A23 gene expression. In ezetimibe-treated subjects three modules were found instead. The first one was uniquely composed of predicted genes along with genes statistically associated with biochemical parameters, namely LDL-TG and C-peptide. Predicted genes in this unique module were COL3A1 and HLA-DPA1, directly interacting with TPTRS, SPARC, and KIRREL1. Although this larger module highlighted physical interactions of gene products related to biochemical parameters, GO annotations did not give information about their involvement in biological processes. Conversely, the small unique module formed by the interaction of TLR7, TLR8, and CD14 was characterized by biological processes such as Toll-Like Receptor (TLR) signalling pathways and regulation of interferon (IFN) production, both important in early immune response activation. The last unique module in ezetimibe-treated subjects also contained PPARA and NR1H3. Among all treatments, the combined treatment reported the largest and most heterogeneous unique module. This result could partly reflect the unique effects of the two lipid-lowering drugs on the molecular regulation of cholesterol metabolism when given as combined treatment. As expected, cholesterol and sterol biosynthetic and metabolic processes were especially enriched in this sub-network. The emergence of three DM genes, coupled with the INSIG2 – LDL-TG association, namely TREX1, CD81, and SLC35E4, suggested an instrumental interplay across DNA methylation and peripheral plasma lipoprotein phenotype. What was clearly seen after complementing the network analyses with high-level bioinformatic data, such as DisGeNet and GO, was that contextuality was achievable and could improve the ability to explain processes perturbed by the combined treatment. The heterogeneous network of the placebo-treated subjects reflected the sum of all differentially expressed genes compared to the individual lipid-lowering treatments. Hence, this heterogeneous network might define in part a true placebo-effect and in part the effects of the individual drugs since the differential expression of the genes was partly inclusive of the specificity of the different treatments as well as other unrelated genes. Of interest, the observation that the network of placebo-treated subjects included the largest number of disconnected regions and was almost entirely composed of genes from GEP, due to the limited effects of Network Propagation and paucity of biochemical associations, possibly reflected the less cohesive molecular effects that placebo should have.

Given the limitations discussed above, the herein heterogeneous networks gave molecular insights that suggested, rather than proved, the impact of the lipid-lowering drugs studied on molecular pathways. To make further progress, as proof-of-concept, two highly relevant genes proposed by the different treatment modules were investigated with a genetically modified human hepatocyte-like model able to resemble the human hepatic lipid metabolism. Gene SLC25A25, encoding a mitochondrial solute carrier responsible for energy homeostasis and regulation, with involvement in ATP transport across the membrane was first investigated. This gene was found to be differentially expressed by the GEP for simvastatin-treated subjects and for subjects receiving the combined treatment. In addition, SLC25A25 belonged to one of the modules of simvastatin and together with SLC25A23 was involved in ATP transport. When studying the pre-clinical model, no significant effects for atorvastatin were seen. The discordant results might be secondary to the heterogeneous cell composition of liver tissue, whereas SOAT2-only HepG2 cells originated from hepatocytes. In addition, we could not exclude that the usage of atorvastatin, instead of simvastatin, might also be the reason for this lack of effect. The second gene investigated was TMBIM6, which was predicted to be affected in the unique module for combined treatment and was present in all the other heterogeneous networks. Interestingly, this gene has a direct interaction with SOAT2, suggesting a role in cholesterol metabolism, insulin signalling, and lipid oxidation [105]. TMBIM6 encodes a protein involved in the prevention of apoptosis, and its downregulation was found to be linked to the progression of hepatocellular carcinoma in chronic cirrhotic and hepatitis C-infected subjects [106]. Ezetimibe alone or in combination with atorvastatin significantly upregulated TMBIM6 gene expression when compared to vehicle; this was also true when the combined treatment in subjects was compared to the use of atorvastatin alone. Hence, the intelligible heterogeneous networks facilitated to identify effects by lipid-lowering drugs on TMBIM6 that were previously unknown, providing insight into this novel target for possible therapeutic advances in different disease scenarios.

In summary, the intelligible heterogeneous networks herein reported incorporated what makes the integration of multiple biological sources beneficial: contextuality and actionability. On the Stockholm Study the networks helped to explore the unique effects of two of the most common lipid-lowering drugs, i.e., simvastatin and ezetimibe. In fact, the networks showed the feasibility of integrating diverse types of big data into a context that could be exploited in practice, to guide further investigations, and, theoretically, to provide the basis for mechanistic insights. This work once again stressed that to make the networks manageable in practice they must be intelligible, which is a characteristic strongly dependent on their dimension. Reducing the size of the networks and increasing the consistency of their content were, indeed, crucial, as they facilitated the identification of unique treatment modules with practical implications. The results of this work highlighted how an holistic approach oriented to yield networks with lower dimensionality can drive the characterization of complex drug effects and simplify the discovery of new targets/biomarkers. An important next step in the journey of bringing big data to clinical practice is the demonstration that use of intelligible heterogenous networks, such as those described herein, has a meaningful impact on therapeutic decision-making.

Chapter 5

Multi-modal integration via infomax-trained Neural Networks

5.1 Introduction

Personalized Medicine is the great goal of the current biomedical and clinical medicine research community [107]. Since the early 2000's [108] the pursuit of therapeutical opportunities tailoring their impact and implications on specific molecular characteristics and clinical conditions began. The future scenario of medical practice is commonly foreseen to provide person-centered pharmaceutical solutions. To this end, a wide and transversal knowledge of the inner interplay between the human health status and its underlying biological system must be achieved. Understanding the consequences of genomic alterations and their role in favoring the growth and persistence of diseases is therefore crucial. Though this does not suffice, because the genome interacts with other biological constituents like the epigenome and the transcriptome. In principle all biological constituents may play a role in the development of a disease. Besides, the genome-level constituents are not the only layer of the biological complex that is the human body. Genes, proteins and metabolites instrumentally work to form up-scaling functional pathways [15] that progressively composes every aspect of each cell type of the human body, together with their ability to cooperate. Hunting how such higher-level cooperation manifests is also crucial to shift towards Personalized Medicine.

The human body could then be studied as an integrative biological complex system where all its possible statuses result from biological processes harmonically working together. It follows that nowadays many efforts are devoted to integrative methods and approaches. Thanks to the rapid commercialization of modern high-throughput biotechnologies large collections of heterogeneous molecular data can be collected simultaneously. This, along with the growing availability of digital clinical records, favors holistic studies of human diseases. Many different data types, which are referred to as modalities, are expected to harbour different biological processes and, ideally, their integration should reproduce the comprehensive harmonic cooperation of such processes. Though, since such underlying cooperation is unknown, the concept of integration is hard to define. Assumptions are then necessary and many techniques for integrating data [109–112] suppose that the modalities reinforce a common biological signature, i.e., a biomarker. Although being intuitive, such supposition does not consider that modalities may not only contain the same signature but also may be complementary to each other. In this work, multi-modal integration was tackled by using Artificial Neural Networks, which have a long history of been appelated with the term *black-box*. This term entails that unknown and hard to interpretate mechanisms are used by these frameworks to model data. Here, the black-box nature was leveraged to free the multi-modal integration from strong assumptions. This choice was also motivated by the real aim of this work, which is to demonstrate that it is not always necessary to integrate all available modalities to improve the prediction a human health outcome. In fact, even before questioning how modalities interact with each other, the real effectiveness of integrating multiple modalities should be at least observed. This was also the additional reason to employ black-box frameworks, that is to take advantage of their unmatched flexibility to see whether multi-modal integration is really more advantegeous than using a reductionist approach, i.e., the single modalities are analysed separately.

To pursue such aim a large publicly available dataset of breast cancer was selected and multi-modal data were downloaded. Namely, imaging data and molecular data were used. The integration of images and genomics drew a lot of interest in recent years and contributions in this direction are quickly increasing. In particular, the imaging data analysed in this work derived from digital histopathology. A digital image in histopathology is the outcome of a laboratory procedure that starts from a biopsy and ends with an image generation using a scanner [113]. A slide of thin tissue usually undergoes a hematoxylin staining, which colors elements in the slide, especially the nuclei, in purple shades. After other possible cleaning and rinsing steps, eosin is added as a counterstain to highlight the difference between cells nuclei and the surrounding cytoplasm elements through pink shades. Eventually, the slide is digitized by a scanner. Several technical aspects can affect the properties and quality of the final image, from the type of the reagents, the paraffine embedding, to the time spent for rinsing. Therefore different outcomes across laboraties and also within the same laboratory can be expected. The characteristics of histopathological images are then challenging and what makes them even more challenging is their dimension. Typically an image includes several layers of magnification and, when analyses are conducted at high-resolution scales, multiple patches are generated. The patching procedure complicates the analysis because the images become collections of many patches. Thus, although the main focus of this work was multi-modal integration several insights on how to handle, process and use histopathological images are provided.

Here, an agnostic-method to perform multi-modal integration was proposed. The core idea of this method succeeded to preserve important characterizing signature after simulated multi-modal integration tests were carried out. Additionally, the application on a breast-cancer cohort preliminarly confirmed that it was not fruitful to integrate modalities regardless of what they harboured. In fact, it was observed that only a specific subset of modalities characterized the data with a strong unknown biological signature.

5.2 Methods

5.2.1 Data filtering

The results shown here were in whole based upon data generated by The Cancer Genome Atlas (TCGA) Research Network. Breast cancer multi-modal data were collected for a total of 1098 subjects. Namely, the modalities were: Hematoxylin-Eosin (H&E) stained slides from digital pathology, gene expression from RNA-seq experiments and somatic mutations from Whole Genome Sequencing (WGS). Demographics (age and gender), tumour type and staging information were also used.

Initially, solely subjects with tumour samples extracted from fresh or frozen tissue were selected. Next, for each subject reporting multiple tissue portions, only the portion where the highest number of modalities were derived from was considered. Besides, tissue portions were kept only if they were located at the top of the block used for molecular analysis. To futher uniform the selection of the H&E slides, exclusively slides with a 40X magnification were preserved. Male subjects (12) were excluded from the analyses. At this point the available subjects were 1022.

5.2.2 Data selection

To ease the premises for the integration only the Invasive, or Infiltrating, Ductal Carcinoma (IDC) type was considered in this work[1]. This type covered 714 subjects in herein cohort. Two classes were then defined and eventually used as target of a classification task: early-stage and late-stage subjects [114]. All subjects up to second tumour stage were considered at an early stage while the remainings were considered at a late stage (Table 5.1). The staging followed the AJCC system [115]. Totally, only 440 subjects were provided with all three modalities. Hence, the integration method was employed exclusively on this smaller cohort, which included 329 early-stage subjects and 111 late-stage subjects. Ten-fold cross validation (CV) was adopted to assess the performance of the integration. All folds were stratified based on the proportion of the classes in the whole cohort. To create a validation set, 10% of the subjects in each training folds was randomly picked in a stratified way.

[1] The IDC is the major breast cancer type in the world (approximately 80% of all cases).

Early-stage	**Late-stage**
Stage I	
Stage IA	Stage IIIA
Stage IB	Stage IIIB
Stage II	Stage IV
Stage IIA	Stage X
Stage IIB	

Table 5.1: Definition of the early and late stage classes. These two classes were later used as target for a classification task.

5.2.3 Modalities pre-processing

Patching Whole Slides Images

The pre-processing of H&E slides from digital pathology, which fall under the appelative Whole Slides Images (WSIs), was carried out as following [116]. First, 1024×1024 patches with a 100 pixels overlap were generated for each WSI. Second, the patches were resized to 512 × 512 while using an anti-aliasing filter. Third, blurred (or poorly informative) patches were roughly determined via the magnitude of the gradients and later removed. That is, Sobel operators on both directions were run on each gray-converted patch and their absolute values were summed up. If the magnitudes of gradients were found to be large, over more than 60% of the patch, then such patch was discarded.

Normalizing RNA-seq gene counts

The rlog method from DESeq2 [86] was exploited to normalize gene counts. To make the normalization more robust, the rlog method was fed by all available filtered subjects. Namely, 1013 of the total 1022 filtered subjects were provided with RNA-seq data and underwent normalization. By doing so, rlog could borrow gene variance information from a much higher number of subjects. Once normalized, only protein coding genes were kept and scaled to avoid high variable genes to bias downstream analyses. Besides, genes with very low standard deviation, i.e., less than 10^{-2}, were removed. At the end the total number of genes provided with counts were 19516.

Mutations

Annotated variants from WGS experiments were transformed to a binary matrix, where genes with at least one variant where considered as mutated (1) and unmutated (0) otherwise. Next, genes with no Entrez identifier [117] were removed as well as genes with standard deviation lower than 10^{-2}. Eventually there were 13447 genes with mutation status for each subject.

5.2.4 Tumour stage classifier

In this work two steps were held independent: integration and classification. In particular, the classifier was chosen to be a single layer as wide as the dimension of the multi-modal embedding supplied by the multi-modal integration method. The classifier and the integration method were trained independently one from another. Therefore the multi-modal integration method was agnostic, in the sense that it was not designed to solve a unique and specific classification problem. Herein, the classifier was trained to distinguish early-stage tumours and late-stage tumours (two classes).

5.2.5 Multi-modal integration method

To perform multi-modal integration over the TCGA breast cancer cohort an innovative neural network was customly designed and implemented. Though, the true novelty does not lie in the network but in the learning rule driving the training.

Infomax core learning rule

Say M modalities are collected for N subjects. Each m-th modality is represented by an object X_n^m for every n-th subject. Then the core idea herein proposed to induce integration is to maximize the Mutual Information (MI) between each modality and an object yielded by the combination of all modalities. Details on MI can be found in section 6.4.5. Similarly to what proposed by the Deep InfoMax [118] (see details in section 6.4.6), given a global encoder $E_\psi(\cdot)$ and a local encoder $C_\psi^m(\cdot)$, the learning rule can be defined as:

$$\underset{\omega,\psi}{\operatorname{argmax}} \frac{1}{M} \sum_{m=1}^{M} \mathrm{MI}_\omega(C_\psi^m(X_n^m); E_\psi(\{X_n^m\})) \ , \tag{5.1}$$

for each n-th subject. This objective was built similar to the local objective of Deep InfoMax. Explicitly, the global encoder $E_\psi(\cdot)$ is trained so that the global embedding of all modalities maximizes (on average) the MI with each of them. In this way, the global embeddings of the subjects are expected to provide a signature that shares characterizing information with all modalities without any privileges towards only a few of them. In other words, the characterizing modal-signatures are jointly embedded into a unique multi-modal signature. Inspired by the previous work on f-divergences [119], MI was not estimated by the KL-divergence but by the Jensen-Shannon divergence (JSD). Following the f-divergences formulation (details in section 6.4.5) the JSD between two distributions P and Q is defined by:

$$F_\omega^{JSD}(P,Q) = \mathbb{E}_{x\sim P}[\log(2) - \log(1+e^{-V_\omega(x)})] - \mathbb{E}_{x\sim Q}[\log(1+e^{V_\omega(x)}) - \log(2)] \ . \tag{5.2}$$

Thus, equation 5.1 becomes:

$$\underset{\omega,\psi}{\operatorname{argmax}} \frac{1}{M} \sum_{m=1}^{M} F_{\omega}^{JSD}(C_{\psi}^{m}(X_{n}^{m}); E_{\psi}(\{X_{n}^{m}\})) \ . \tag{5.3}$$

Three functions need a definition: the global encoder $E_{\psi}(\cdot)$, the local encoder $C_{\psi}^{m}(\cdot)$ and the discriminator $V_{\omega}(\cdot)$. Also, an approach to generate both the positive cases $x \sim P$ and the negative cases $x \sim Q$ from equation 5.2 is required.

In this work, in which the modalities were three, the global and local encoder were Neural Networks (NNs), as well as the discriminator, which consisted of a multiple single layer NNs. The architectures of the encoders are reported in the next sections, while the one of the discriminator is showed in Table 5.2. For each modal-embedding a 128-wide single layer NN is employed and then its outcome is multiplied by the outcome of another 128-wide single layer NN fed with the global embedding. Furthemore, to yield the positive and negative cases, during the training, subjects in the batch are used. Positive cases correspond to the right pairs $(C_{\psi}^{m}(X_{n}^{m}); E_{\psi}(\{X_{n}^{m}\})$ while negative cases took all random pairs between local and global embeddings. The larger the batch, the greater will be the number of positive and especially negative cases.

Operation	**Size**	**Output**
Modal embedding $C_{\psi}^{m}(X_{n}^{m}) \rightarrow$ Linear layer	128	Output 1
Multi-modal embedding $E_{\psi}(\{X_{n}^{m}\}) \rightarrow$ Linear layer	128	Output 2
Output 1 $\odot$ Output 2		

Table 5.2: Architecture of the discriminator $V_{\omega}(\cdot)$ necessary to estimate the Mutual Information between each modality and the final multi-modal embedding.

Global Encoder

The global encoder $E_{\psi}(\cdot)$ was designed to first encode all modalities separately and then to produce a final multi-modal embedding. The modality-encoders are the actual local encoders, which took in the m-th pre-processed modality data X_{n}^{m} and mapped it to a low dimensional latent space. After all m-th modality-embeddings $C_{\psi}^{m}(X_{n}^{m})$ are generated they get pushed forward through another encoder $f_{\psi}(\cdot)$ that performs integration. Therefore the global encoder can be formulated as $E_{\psi}(\cdot) = (f_{\psi} \circ C_{\psi})(\cdot)$. Figure 5.1 illustrates such architecture. As mentioned, the multi-modal integration occurs explicitly through the encoder $f_{\psi}(\cdot)$, defined as a simple four-layer feed-forward NN. The concatenation of all modal-embeddings $C_{\psi}^{m}(\cdot)$, which supposedly harbour modal-signatures, goes through the $f_{\psi}(\cdot)$ to generate a unique multi-modal signature that globally embeds them. The layers were designed to scale down the dimension of the input by an half (Figure 5.2). Starting from a 2048-wide layer the dimension drops to 256 at the end,

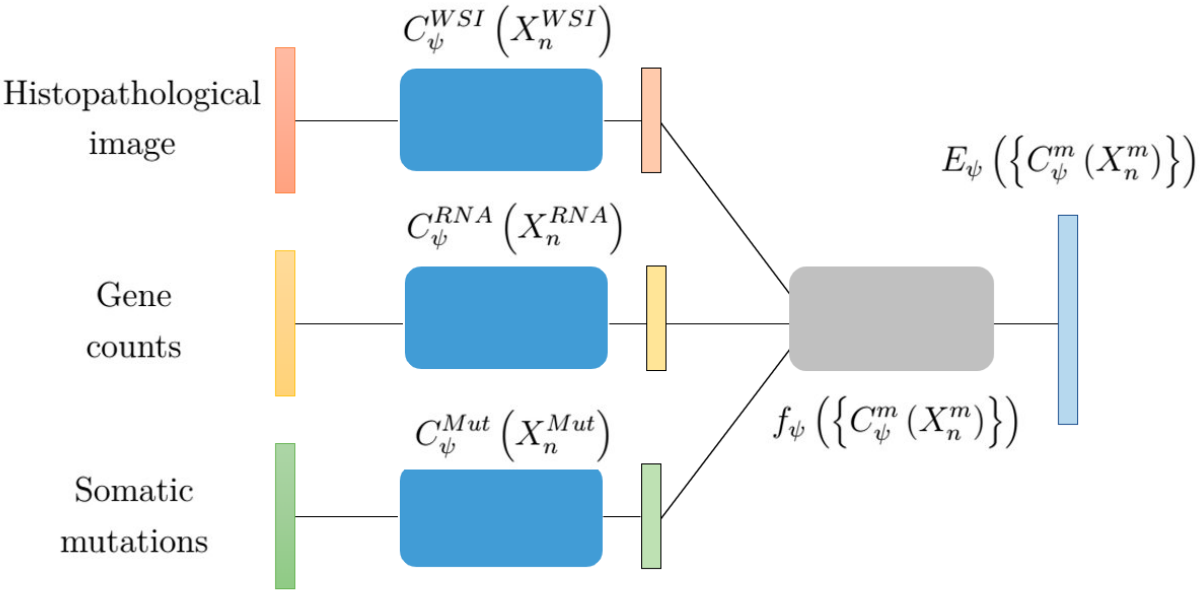

Histopathological image
$C_\psi^{WSI}\left(X_n^{WSI}\right)$
Gene counts
$C_\psi^{RNA}\left(X_n^{RNA}\right)$
Somatic mutations
$C_\psi^{Mut}\left(X_n^{Mut}\right)$
$E_\psi\left(\left\{C_\psi^m\left(X_n^m\right)\right\}\right)$
$f_\psi\left(\left\{C_\psi^m\left(X_n^m\right)\right\}\right)$

Figure 5.1: Overview of the global encoder $E_\psi(\cdot)$ architecture used by the multi-modal integration framework on the TCGA breast cancer cohort. The local encoder $C_\psi^m(\cdot)$ embeds each input modality and passes it on to the encoder $f_\psi(\cdot)$ which yields the final multi-modal embedding.

which determines the dimension of the multi-modal embedding. Except for the last layer, the ReLU function was utilized to activate the neurons. In this work the three modalities (WSI, gene counts and somatic mutations) were endowed with three diverse local encoders to handle the different properties and dimension of the modal-data. Though, gene counts and somatic mutations encoders were variants of the same framework, an adjusted version of a feature sparsity-oriented network known as LassoNet [120] (see details in section 6.4.4).

WSI encoder

Mapping a WSI to a single embedding is a challenging task recently tackled by Multiple Instance Learning (MIL) frameworks [121]. As mentioned, a WSI is a collection of patches and each of them might portrait a region carrying crucial histopathological information. Both theoretically and practically the design of an efficient WSI encoder is therefore not-trivial. Here, inspired by previous work [122], an intuitive approach was undertaken. First, given a WSI, all its patches were independently mapped by a DenseNet-121 (section 6.4.2) onto a vector latent space, i.e., patch-embeddings were generated. Second, a unique WSI embedding was obtained by using a weighted-average of the patch-embeddings. The weights, or scores, of the weighted-average were obtained by

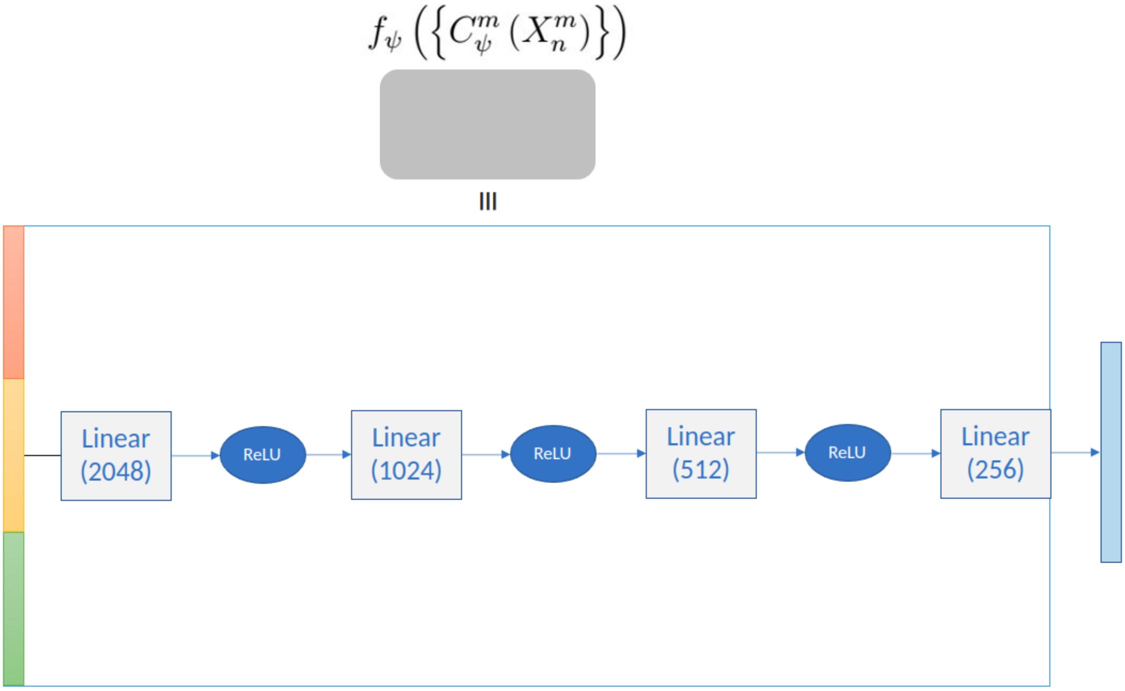


Figure 5.2: Details on the on-top encoder $f_\psi(\cdot)$ fed by the embedding of the three modalities. A four-layer feed-forward NN was designed with intermediate of ReLU activation functions. Such NN performs integration over the concatenation of the multi-modal embeddings. The newly introduced learning rule (equation 5.1) drives the encoder to preserve modal-signatures and combine them into a unique multi-modal signature.

another feed-forward NN independently fed by each patch-embedding. In short, given the n-th subject WSI X_n^{WSI} and its P_n patches,

$$C_\psi^{WSI}\left(X_n^{WSI}\right) = \sum_{p=1}^{P_n} A\left(\vec{h}_{n,p}\right)\vec{h}_{n,p} \ , \tag{5.4}$$

where

$$\vec{h}_{n,p} = \text{Dense}\left(X_{n,p}^{WSI}\right) \ . \tag{5.5}$$

The two functions Dense$(\cdot)$ and $A(\cdot)$ respectively defines the DenseNet-121 network used to generate patch-embeddings and the network used to score them. The weighted-average plus the network $A(\cdot)$ provide the WSI with an attention mechanism (refer to section 6.4.3). The architecture of the WSI encoder is showed in Figure 5.2 while the structure of the attention module is reported in Supplementary Figure D.1.

The attention module was defined as shown in Table 5.3. To be noted, DenseNet-121 was only selected after multiple standard deep architectures were trained and tested (details in D.2).

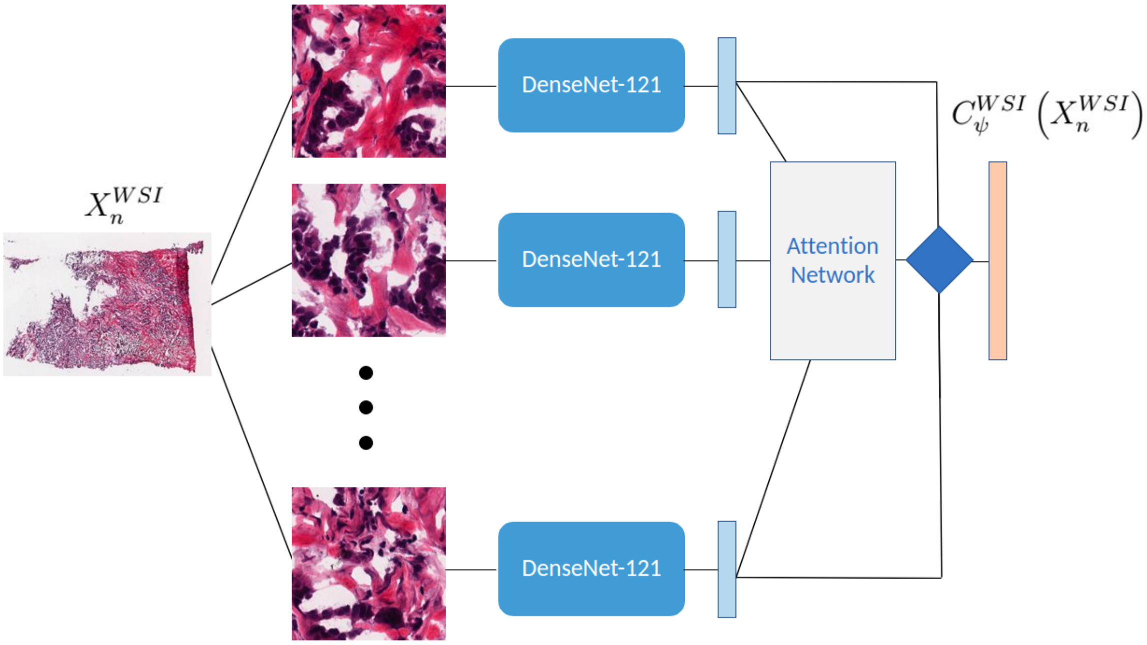


Figure 5.3: Illustration of the Neural Network encoding the WSIs. A DenseNet-121 turns each patch into an embedding vector, which is then scored by an attention module and combined with all the others (through a weighted average) to form a final unique embedding for the WSI.

Operation	**Size**	**Activation**	**Output**
Input → Linear layer	256	tanh	Output 1
Input → Linear layer	256	sigmoid	Output 2
Output 1 ⊙ Output 2			
Linear	1		

Table 5.3: Layers of the attention module on top of the DenseNet-121.

Gene counts and somatic mutations encoders

As mentioned, the LassoNet was used to perform the encoding of both normalized gene counts and somatic mutations. The caveat of LassoNet is that it combines the residual network output with the feed-forward output to directly yield a prediction. This means no layer is fed in the LassoNet by both linear and non-linear effects. Since ideally an embedding for the input would include both effects, the LassoNet architecture was adjusted to address this caveat. Intuitively, adding a layer where the residual layer and the feed-forward NN connect sorts any caveat. By doing so, prior to perform any type of prediction, the effects from linear and non-linear effects are integrated. The

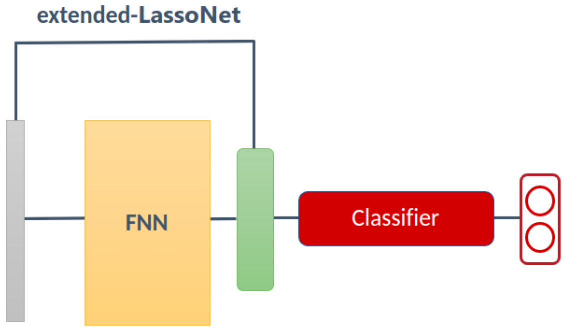


Figure 5.4: Illustration of the extended-LassoNet, where an additional layer is added before the classifier in order to absorb both the linear effects, i.e., residual layer, and non-linear effects, i.e., feed-forward Neural Network (FNN).

architecture of the feed-forward NNs, as well as the penalty parameter of LassoNet, was chosen based on the modality after a hyper-parameter optimization procedure (details in Supplementary section D.3). For gene counts the optimal feed-forward NN had two-layers of 128 and 64 neurons, whereas for somatic mutations the optimal architecture used four-layers with decreasing width ($64 \to 128 \to 256 \to 512$). Further, the penalty parameter for the two modalities were respectively 81.59 and 114.24. It follows that the embeddings of gene counts and somatic mutations were of different sizes (64 and 512).

5.2.6 Training

Ideally, the entire integration network illustrated in Figure 5.1 would be trained end-to-end. That is, all the modal-encoders $C_\psi^m(\cdot)$ along with the on-top encoder $f_\psi(\cdot)$ would be iteratively updated to maximize the MI according to equation 5.3. Unfortunately, the end-to-end training is hard to run due to the high memory demands of the WSI encoder. In alternative the training was divided in two steps. Modal-encoders were first trained separately on the pre-processed data to correctly classify the tumour stages. Their final layers then were taken to represent the modal-embeddings. In detail, the encoder for the WSIs (plus the classifier) was trained in 20 epochs and the epoch with the highest Matthews correlation coefficient (MCC) over the validation set was taken as the optimal encoder. As explained in section D.3, a similar approach was followed to deduce the optimal LassoNet for gene counts and somatic mutations (together with the optimal hyper-parameters). Next, the on-top encoder $f_\psi(\cdot)$ was trained on the modal-embeddings according the MI maximization learning rule. Afterwards, the training of the classifier was performed. All these steps were executed for each CV fold.

As for WSI, separate Adam optimizers with a learning rate of 10^{-4} and a weight decay of 10^{-4} were used to update the DenseNet-121, the attention module and the local classifier. On the same note, the extended-LassoNets for gene counts and somatic mutations were driven by Adam optimizers with 10^{-3} learning rates and no weights decay. Even the on-top encoder and the discriminators (needed for the MI estimation) used Adam optimizers with 10^{-4} learning rates and 10^{-4} weights decay. The final classifier, instead, was updated by an Adam optimizer with learning rate equal to 10^{-4} and no weights decay, while the weighted cross-entropy loss was used again as the objective to minimize. The weights for the loss for each tumour stage class were calculated as:

$$w^{c}_{loss} = \frac{N}{n_{classes} \cdot N_c} \tag{5.6}$$

where N is the total number of subjects and N_c is the number of subjects per class c. Clearly in this work: $n_{classes} = 2$.

5.2.7 Testing on simulated data

Due to the agnostic nature of the infomax core learning rule, the integration method can potentially work on general multi-modal data. Two simulated tests were then performed to ascertain the validity of the core idea behind integration. In both tests, following the scenario of the TCGA breast cancer data, 440 simulated subjects were generated in two unbalanced classes.

The goal of the first test was to establish whether the classification of two classes with strong signatures from a single modality were influenced by the integration with a second noisy modality. The two classes were generated as sine and cosine signals and composed one modality, while the other noisy modality was simulated as additional white gaussian noise (AWGN).

Differently, the second test simulated two totally complementary modalities that only together allowed to predict the right class. In this case, the goal was to control whether the integration of the two modalities was able to successfully embed both signatures. One class was defined as the combination of either two sine or cosine functions, while the other was defined for the alternative cases (i.e., sine plus cosine, cosine plus sine).

For each test the approach to training and testing was the same. The simulated data was split in training/validation/test sets and no local encoders were used. In other terms, the local encoders were identity functions. The on-top encoder $f_{\psi}(\cdot)$, as well as the discriminator, was the one used for TCGA data (portraited in Figure 5.1). Training was set to run for at most 5000 epochs and an early stopping criterion was added upon the validation set. Afterwards, the global embeddings were used to train the classifier (built as for the TCGA case) until the validation loss converged. Training and testing was run 10 times to reproduce the ten-fold CV scenario of the TCGA analysis.

5.3 Results

5.3.1 TCGA breast cancer data

The integration of the multi-modal data for breast cancer was performed for all possible combination of modalities. Totally, then, integration was performed seven times: for the three separate modalities (gene counts, somatic mutations and WSI), for the three modalities couples (gene counts plus somatic mutations, gene counts plus WSI, and somatic mutations plus WSI) and once for the modalities triplet. The tumour stage classification performances of all seven are showed in Figure 5.5, where accuracy, MCC and area under the roc curve (AUC) are the metrics. All modalities struggled to achieve optimal classification power. Median accuracies hovered around 0.65 and 0.70, which was a close value to the proportion of the early-stage subjects (approximately 0.75). Across multi-modal integrations, Table 5.4, gene counts plus WSI stood out especially in terms of MCC and variability. This two modalities yielded the only positive non-null MCC (0.14) while preserving accuracy (0.70). Also, they produced the least oscillating performances in terms of accuracy and MCC.

Multi-modal embedding	**Accuracy (Median)**	**Accuracy (IQR)**	**MCC (Median)**	**MCC (IQR)**	**AUC (Median)**	**AUC (IQR)**
Gene counts + Mutations	0.70	0.14	0.00	0.23	0.52	0.14
Gene counts + WSI	0.70	0.07	0.14	0.11	0.60	0.12
Mutations + WSI	0.61	0.20	−0.03	0.11	0.49	0.12
Gene counts + Mutations + WSI	0.69	0.11	0.00	0.12	0.53	0.08

Table 5.4: Overview of all classification performance metrics obtained by the multi-modal integration approach applied to each combination of modalities.

Since the local encoders were trained with a topstream classifier separately from the integration framework, performance baselines for the single modalties were established. The single local-embeddings were, indeed, already fine-tuned to solve the classication performances. It was natural then to compare the performances directly obtained by the local-embeddings with their performances after being further embedded by the on-top encoder. Table 5.5 highlights a slight drop in performances after the further embedding.

5.3.2 Simulated data

No groundtruth was provided for the TCGA data that could be used to fully evaluate the effectiveness of the infomax based integration. To this end simulated data were used instead. The classification of sine and cosine signals (coming from one modality) integrated with a modality solely filled by AWGN was tested for several levels of noise. Figure 5.6

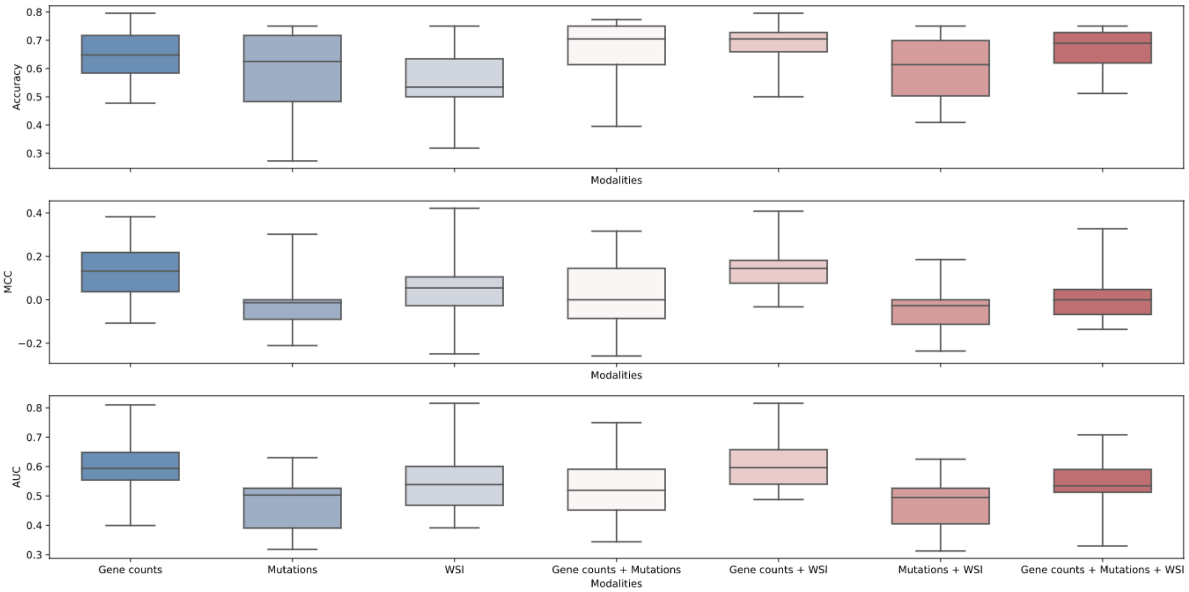


Figure 5.5: Performance metrics of the infomax based integration method run over all combination of modalities for the TCGA breast-cancer dataset. Accuracy at the top shows that multi-modal embeddings formed by gene counts and either WSI or somatic mutations yields better performances than the single gene counts embedding. Matthews correlation coefficient (MCC) in the middle highlights that it was the multi-modal embedding of gene counts and WSI to outperform all the others. At the bottom, the area under the roc curve (AUC) confirms even further that the embedding of gene counts and WSI has the best performance.

shows that more noise hinders classification. Nevetheless, even when the sine/cosine signals are weak compared to the amount of noise, the classification performances does not abruptly drops. Notably even when signal and noise are added the integration showed robustness (Supplementary section D.4).

The integration framework also managed to retrieve complementary signatures based on the results of the second simulated test (Figure 5.7). Subjects were given two sine and/or cosine signals. Those with two sine signals or cosine signals defined one class, all remaining subjects formed the other class. The two signals were split in two modalities (M1 and M2) so that only the contemporary knowledge of the signals was needed to assign subjects. To benchmark the integration method, modalities were trained both individually and concatenated to directly perform classification. The results showed that training the multi-modal embeddings via the infomax principle definitely outperformed all other approaches.

Single-modal input	Accuracy (Median)	Accuracy (IQR)	MCC (Median)	MCC (IQR)	AUC (Median)	AUC (IQR)
Gene counts direct	0.69	0.11	0.16	0.20	0.64	0.10
Gene counts embedding	0.65	0.13	0.13	0.18	0.59	0.09
Mutations direct	0.65	0.21	0.00	0.08	0.50	0.13
Mutations embedding	0.62	0.23	−0.01	0.09	0.50	0.14
WSI direct	0.56	0.09	0.04	0.20	0.54	0.18
WSI embedding	0.53	0.13	0.05	0.13	0.54	0.13

Table 5.5: Comparison between the performances of the single-modalities trained to directly classify tumour stages and the single modalities first embedded (by the same integration framework) and then trained. Performance did not change considerably but a recurrent small drop by the single-modal embeddings can be noticed.

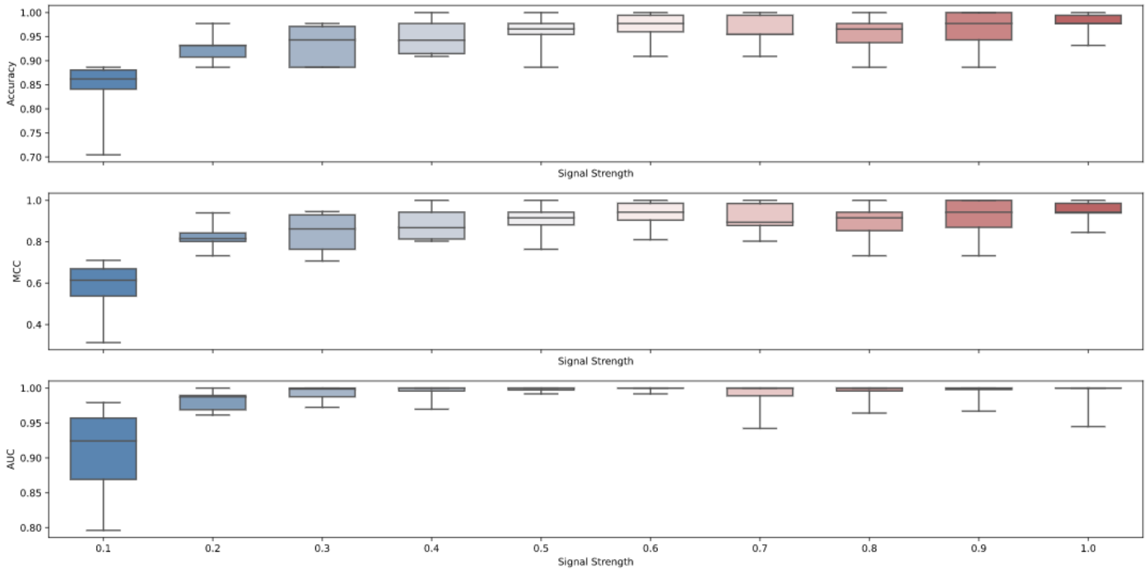


Figure 5.6: Impact of an additional white gaussian noise (AWGN) modality integrated with a modality harbouring a strong signature for two classes: sine and cosine. The signal strength indicates the fraction of the signal if it was simply added to the noise.

5.4 Discussion

The main assumption of multi-modal data integration is that using many multiple modalities together is better than using them separately [123]. Therefore, when several modalities are available in a study, every modality is supposed to carry relevant information

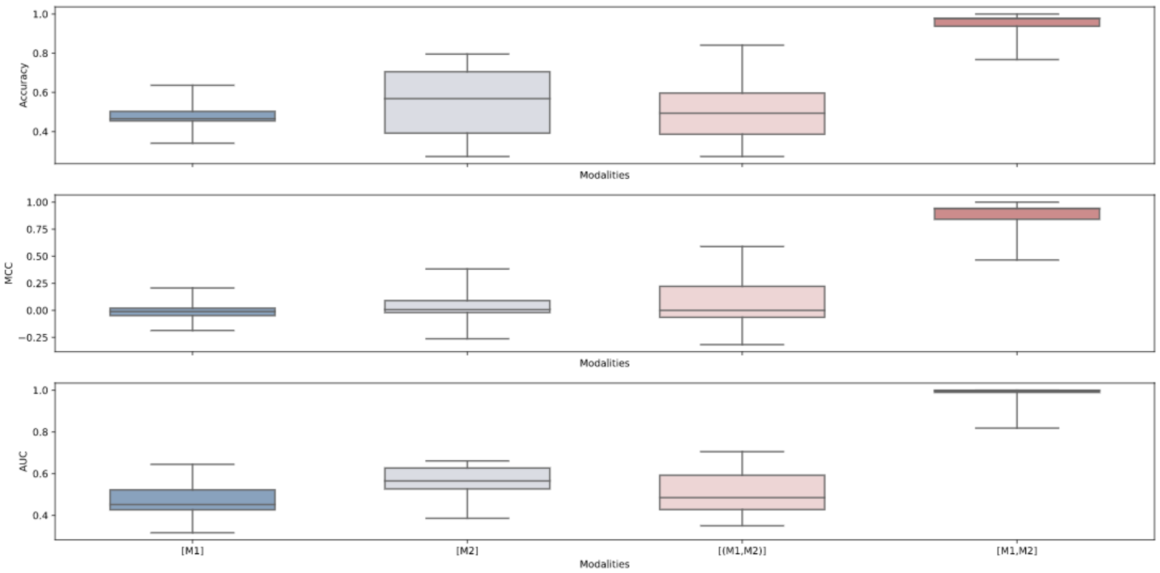


Figure 5.7: Comparison of the classification performances when two simulated modalities, M1 and M2, are totally complementary. Classes were simulated so that they could be predicted exclusively if both modalities are provided. The cases where only one modality is provided ([M1] or [M2]), where they are simply concatenated ([M1, M2]) and where they are integrated ([M1,M2]) are showed in the plot.

to answer the research question. Previous studies showed a glimpse of this benefit for multi-modal integration [124–126]. Here, a TCGA breast cancer dataset was analysed to study whether the stage of tumours could be identified by the available modalities, i.e., gene counts, somatic mutations and WSIs from digital pathology. The initial assumption on this study, then, presumes that integrating gene counts, somatic mutations and WSIs improves the tumour stage prediction compared to using them separately. However, it is intuitive to imagine that modalities are not equally important. In other words, some modalities may be highly informative while others may result superfluous. If a multi-modal integration model does not have control over what information from which modality partecipates to the integration process, then integration may even not occur. In fact, the model can just pick some minimal information in order to optimize learning, which does not necessarily imply that all modalities are used.

To make sure integration occurs, this work proposed a self-supervised training based upon the infomax principle. A multi-modal embedding was defined as a representation of the modalities that maximizes the MI with each of them. Practically, information from modalities are preserved if they contribute to generate a multi-modal embedding that is not statistically independent from any modality. In other words, modal-information inducing the final embedding to be statistically independent from at least one modality are filtered out. Such constraint restricts what information passes through but forces integration.

In the TCGA analysis multi-modal integration was performed for all modalities together, for the three multi-modal pairs and also for the single modalities. The infomax principle was set as the learning rule for a NN encoding the input modalities. Before integration, other NNs were built to encode gene counts, somatic mutations and WSI separately. Training the final embeddings to classify tumour stages clearly showed that multi-modal integration can easily underperform. The integration of the three modalities (gene counts, somatic mutations and WSIs) did not yield the best performances (median MCC=0) and the multi-modal pairs had comparable but diverse performances. Somatic mutations plus WSIs had the worst performances (median accuracy=0.61 and median MCC=-0.03) while gene counts plus WSIs had the best (median accuracy=0.70 and median MCC=0.14). Gene counts identified the single-modality with the highest accuracy (0.69) and MCC (0.16) when they were not further embedded, i.e., simple modal-embeddings. This indicated that the final embeddings given by the integration models lost bits of information. The reason behind this drop could be due to the agnostic nature of the integration, whereas the modal-embeddings were obtained after training with the classifier. Although the best performance in terms of MCC was provided by the modal-embeddings of gene counts, they had a high variability (MCC-IQR=0.20). More stable and similar performances resulted from the multi-modal embeddings of gene counts plus WSIs. Despite the little improvement, this result underlined how critical it was to ascertain the occurence of integration before exploring their meaning. Noteworthy, the classification of tumour stages might not be strictly related to the available modalities. This point further discredits the assumption that multi-modal integration should always be beneficial, and suggests that the effectiveness of integration is strongly related to the research question. There might be research questions where multi-modal integration highly performs because the modalities provide essential insights, and other research tasks where multi-modal integration results poor because modalities do not carry information. Hence, the evaluation of multi-modal integration should be both based on the research question and focused on identifying the right modalities. Moreover, knowing if only a few modalities carry information has several benefits. First it facilitates data selection and second it can result more affordable. In fact, if a specific target is found to be related to only a couple of modalities (or potentially even a single of them) instead of a many, lots of savings on data acquisition and collection are possible.

To be noted the pre-processing on input modalities might be crucial to prepare data for integration. In fact, the poor performances of somatic mutations seen for the TCGA analysis could also be due to the non-optimal preparation of the data. For example, mutations could be transformed from binary data to mutational signatures [127] before taking part to the integration. Besides, the approach to create modal-embeddings could be refined. As mentioned, due to the computational demands of the WSIs, gene counts, somatic mutations and WSIs were encoded separately before passing on to the integration encoder. Their modal-encoders were trained using the classifier on top. Ideally, modal-encoders and the integration encoder would be trained altogether to feed the classifier with totally task-unbiased multi-modal embeddings. This could lead modal-encoders to agnostically represent the modalities while indirectly influencing each other.

To make sure the TCGA breast cancer data results were not providing misleading feedbacks on the integration method, simulated tests were run. As expected their results underlined the ability of the infomax based integration to capture complementarity and resist to white noise. It follows, that the next step for this work is to examine the biological implications related to the integration method results. Thanks to what observed on the TCGA breast cancer dataset, in fact, it could be interesting to see whether patch scores, provided by the attention network, highlight true informative tissue regions from a clinical perspective. Similarly this could be theoretically explored also for gene counts and mutations owing to the LassoNet feature-sparsity prowess. The herein integration method, then, could both make sure that multiple modalities effectively improve the prediction of a clinical target and reveal which specific inputs drove the improvement. However, further testing is needed to confirm the goodness of the method in terms of providing clinical and biological insights. The classification of tumour stages showed here was one of many possible tasks to test the integration method on. Interestingly, though, the global multi-modal embedding are totally unbiased from any specific task and classification might not be the only application to pursue. For instance, unsupervised clustering [128] could be performed on the multi-modal embeddings to observe whether multi-modal groups of subjects emerge. This could lead to determine new clinical and biological groups that are stratified not by a single modal signature (e.g., somatic mutations) but by multiple ones. Such multi-modal stratification could help to identify rare heterogeneous groups, which would ultimately contribute to Personalized Medicine.

In ultimate analysis, the method presented here provided the technical groundwork to build artificial intelligent frameworks that effectively integrate multiple modalities. Tests on simulated data demonstrated the robustness of the method to noise and the ability to integrate complementary modalities. Further, an application on TCGA breast cancer preliminarly confirmed that only a subset of modalities impact on tumour stage classification and that a pair of them seemed to slightly outstand (gene counts and WSIs). Future work will be dedicated to prove a widespread efficacy of the integration method on several real multi-modal data.

Chapter 6

Methodologies

6.1 Methodology behind Transcription Factors Bi-Clustering

6.1.1 RMA

The Robust Multiarray Average [129], simply known as RMA, is the most popular framework to pre-process microarray data. Microarray intensity of any i-th probe and j-th subject, or more generally array, is usually represented by

$$I_{i,j} = K_j \times \phi_i \times \theta_{i,j} \times \epsilon_{i,j} + O_{i,j} \ . \tag{6.1}$$

Clearly any intensity is then the result of multiple effects, one (K_j) depending strictly on the subject, one (ϕ_i) depending on the probe and the remainings ($\theta_{i,j}$, $\epsilon_{i,j}$, and $O_{i,j}$) depending on either of them. The subject-specific effect K_j entails that every subject carries a unique additional signal, which therefore needs to be accounted for when comparing multiple subjects. Similarly, the probe-specific effect ϕ_i contributes to the final intensity differently across probes. Lastly, there are three terms that relate to both subject and probe: a quantity proportional to the true RNA abundance ($\theta_{i,j}$), its associated uncertainty ($\epsilon_{i,j}$) and local artifacts ($O_{i,j}$). Affymetrix GeneChip arrays have two types of probes: perfect match (PM) and mismatch (MM). The former are oligonucleotide of 25 base-pairs, or 25-mer, complementary to a reference sequence of interest, like a transcript. The latter are similar but the middle base-pair, i.e., the thirteenth, is purposedly changed to potentially help identifying non-specific binding and other artifacts.

Background correction

To render model 6.1 into a linear framework the first step of RMA is background correction, i.e., removal of $O_{i,j}$. As mentioned, background correction removes local artifacts and for Affymetrix GeneChip arrays it does so by using a normexp model. Let represent equation 6.1 simply as follows [130],

$$I = S + O \tag{6.2}$$

upon

$$\begin{aligned} S = s &\sim \exp\left(\frac{1}{\alpha}\right) \\ O = o &\sim \mathrm{N}(\mu, \sigma) \ , \end{aligned} \tag{6.3}$$

and assuming the actual signal to be positive, i.e., $s > 0$. In order to estimate S we target the expectation value of the conditional distribution:

$$f_{S|I}(s|i; \alpha, \mu, \sigma) = \frac{f_{I,S}(i, s; \alpha, \mu, \sigma)}{f_I(i; \alpha, \mu, \sigma)} \ . \tag{6.4}$$

The joint distribution $f_{I,S}$ can be calculated from the product of the two density functions in 6.3,

$$f_{S,O}(s, o; \alpha, \mu, \sigma) = \frac{1}{\alpha}\exp\left(-\frac{s}{\alpha}\right) \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{o-\mu}{\sigma}\right)^2\right] \tag{6.5}$$

and using the substitution $o = i - s$. In fact, replacing o and transforming the exponentials arguments, the joint distribution $f_{I,S}$ is:

$$f_{I,S}(i, s; \alpha, \mu, \sigma) = \frac{1}{\alpha}\exp\left(\frac{\sigma^2}{2\alpha^2} - \frac{i-\mu}{\alpha}\right) \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{s-\mu_{S \cdot I}}{\sigma}\right)^2\right] \tag{6.6}$$

with $\mu_{S \cdot I} = i - \mu - \frac{\sigma^2}{\alpha}$. Given the joint distribution, the marginal function for I can be obtained by integrating over s,

$$\begin{aligned} f_I(i; \alpha, \mu, \sigma) &= \frac{1}{\alpha}\exp\left(\frac{\sigma^2}{2\alpha^2} - \frac{i-\mu}{\alpha}\right) \frac{1}{\sigma\sqrt{2\pi}} \int_0^{+\infty} \exp\left[-\frac{1}{2}\left(\frac{s-\mu_{S \cdot I}}{\sigma}\right)^2\right] ds \\ &= \frac{1}{\alpha}\exp\left(\frac{\sigma^2}{2\alpha^2} - \frac{i-\mu}{\alpha}\right) (1 - F(0; \mu_{S \cdot I}, \sigma)) \ , \end{aligned} \tag{6.7}$$

where $F(\cdot)$ stands for the Normal distribution function. Replacing the joint and marginal distribution in equation 6.4 yields

$$f_{S|I}(s|i; \alpha, \mu, \sigma) = \frac{N(s; \mu, \sigma)}{1 - F(0; \mu_{S \cdot I}, \sigma)} \tag{6.8}$$

whose expectation value can be calculated by

$$E\left[f_{S|I}(s|i; \alpha, \mu, \sigma)\right] = \mu_{S \cdot I} + \sigma^2 \frac{N(0; \mu, \sigma)}{1 - F(0; \mu_{S \cdot I}, \sigma)} \ . \tag{6.9}$$

Thus, the background-corrected signal can be estimated by equation 6.9, upon parameters setting. For Affymetrix GeneChip arrays this procedure is performed only over PM intensity values; therefore without using MM probes.

Normalization

Following background correction, model 6.1 can be $\log_2$-transformed to become

$$Y_{i,j} \equiv \log_2(S_{i,j}) = \log_2(K_j) + \log_2(\phi_i) + \log_2(\theta_{i,j}) + \log_2(\epsilon_{i,j}) \ . \tag{6.10}$$

Now, without normalization it is hard to conclude that $Y_{i,j_1} < Y_{i,j_2}$ actually means $\theta_{i,j_1} < \theta_{i,j_2}$. Hence, to overcome the diverse subjects effects normalization is necessary. The most adopted method to perform normalization with RMA is quantile normalization [131]. This normalization makes the statistical distribution of all subjects the same. Given N subjects and the k-th quantile over all subjects, i.e., $\vec{q_k} = (q_{k1}, .., q_{kN})$, the goal is to project all $\vec{q_k}$ onto the N-dimension diagonal $\vec{d} = (d_1, .., d_N)$, which is in unit vector form. The projections $\vec{q}_{k\vec{d}} = (\vec{q_k} \cdot \vec{d})\vec{d}$ turn out to be equal to $\vec{q}_{k\vec{d}} = (\overline{q_k}, .., \overline{q_k})$, where every coordinate is the average k-th quantile across subjects, i.e., $\overline{q_k} = \frac{1}{N}\sum_{n=1}^{N} q_{kn}$. Setting the number of quantiles as the number of probes for each subjects (which is the same), the quantiles of a single subject are intuitively obtained by sorting probes intensities in ascending order. Hence, after sorting, averaging across subjects is performed for each quantile. Eventually, the k averages are sorted back for every subject into the original probes order. By doing so, subjects effects are bypassed and inconsistencies such $Y_{i,j_1} < Y_{i,j_2} \not\rightarrow \theta_{i,j_1} < \theta_{i,j_2}$ are resolved.

Noteworthy, the same inconsistencies holds for the probe affinity effects ϕ_i. As a matter of fact, $Y_{i_1,j} < Y_{i_2,j}$ does not necessarily imply $\theta_{i_1,j} < \theta_{i_2,j}$ because of distinct ϕ_i. Though, this issue is intrinsically sorted out by relative operations, i.e. $Y_{i,j_1} - Y_{i,j_2} \propto \log_2\left(\frac{\theta_{i,j_2}}{\theta_{i,j_2}}\right)$, which are commonly used in gene expression profiling (e.g. log-fold changes).

Summarization

Ultimately, since multiple probes are designed to express a probeset (typically a set of transcripts), their intensity levels are summarized to generate one single estimate of RNA expression. It follows, after normalization, that model 6.10 becomes

$$Y_{i,j,k} = \log_2(\phi_{i,k}) + \log_2(\theta_{j,k}) + \log_2(\epsilon_{i,j,k}) \tag{6.11}$$

where now the $Y_{i,j,k}$ indicates the $\log_2$-transformed intensity value for the i-th probes in the k-th probeset. Summarization within RMA is commonly carried out by median polishing [132]. To make model 6.11 even more suitable for the median polish fitting procedure, let assume $\log_2(\theta_{j,k}) = \log_2(\theta_k) + \log_2(\delta_{j,k})$, which yields,

$$Y_{i,j,k} = \log_2(\theta_k) + \log_2(\delta_{j,k}) + \log_2(\phi_{i,k}) + \log_2(\epsilon_{i,j,k}) \ . \tag{6.12}$$

For the sake of simplicity this equation can be written as: $Y_{i,j,k} = \Theta_k + \Delta_{j,k} + \Phi_{i,k} + \mathcal{E}_{i,j,k}$. In such setting one variable goes along subjects ($\Delta_{j,k}$), another one moves through probes

($\Phi_{i,k}$) and one is independent from both subjects and probes (Θ_k). This setting can be portrayed as a two-way table with subjects on the rows and probes on the columns. The fitting procedure of median polish is then executed for each k-th probeset as shown in Algorithm 4[1]. At the end of the procedure a single intensity value is provided for each subject's probeset.

Algorithm 4 Median polish fitting procedure for a certain k_{th} probeset. Notably, residuals convergence also means that subject medians and probe medians move progressively towards 0.

```
Compute overall median across subjects $\hat{\Theta}_{k_{th}} = \text{median}_{i,j}\left\{Y_{i,j,k=k_{th}}\right\}$
Calculate residuals $\hat{\mathcal{E}}_{i,j,k=k_{th}} = Y_{i,j,k=k_{th}} - \hat{\Theta}_{k_{th}}$
Set $\hat{\Delta}_{j,k=k_{th}} = 0 \; \forall j$
Set $\hat{\Phi}_{i,k=k_{th}} = 0 \; \forall i$
procedure Repeat(until residuals converge)
    Obtain subject medians $d_{j,k=k_{th}} = \text{median}_i\left\{Y_{i,j,k=k_{th}}\right\}$ and probe contributions median $dp_{k=k_{th}} = \text{median}_i\left\{\hat{\Phi}_{i,k=k_{th}}\right\}$
    Update subject contributions $\hat{\Delta}_{j,k=k_{th}} = \hat{\Delta}_{j,k=k_{th}} + d_{j,k=k_{th}}$
    Update probe contributions $\hat{\Phi}_{i,k=k_{th}} = \hat{\Phi}_{i,k=k_{th}} - dp_{k=k_{th}} \; \forall i$
    Update overall term $\hat{\Theta}_{k_{th}} = \hat{\Theta}_{k_{th}} + dp_{k=k_{th}}$
    Update residuals $\hat{\mathcal{E}}_{i,j,k=k_{th}} = \hat{\mathcal{E}}_{i,j,k=k_{th}} - d_{j,k=k_{th}} \; \forall j$
    Obtain probe medians $p_{i,k=k_{th}} = \text{median}_j\left\{Y_{i,j,k=k_{th}}\right\}$ and subject contributions median $pd_{k=k_{th}} = \text{median}_j\left\{\hat{\Delta}_{j,k=k_{th}}\right\}$
    Update probe contributions $\hat{\Phi}_{i,k=k_{th}} = \hat{\Phi}_{i,k=k_{th}} + p_{i,k=k_{th}}$
    Update subject contributions $\hat{\Delta}_{j,k=k_{th}} = \hat{\Delta}_{j,k=k_{th}} - pd_{k=k_{th}} \; \forall j$
    Update overall term $\hat{\Theta}_{k_{th}} = \hat{\Theta}_{k_{th}} + pd_{k=k_{th}}$
    Update residuals $\hat{\mathcal{E}}_{i,j,k=k_{th}} = \hat{\mathcal{E}}_{i,j,k=k_{th}} - p_{i,k=k_{th}} \; \forall i$
end procedure
```

6.1.2 fRMA

Although RMA is extremely popular and well-performing, it has strong limitations when it comes to compare independently pre-processed dataset or when a small number of subjects (or arrays) need to be analyzed. The latter scenario is especially true in a clinical setting when even one array could need pre-processing.

The shortcomings of RMA have been addressed by the proposal of the frozen Robust Multiarray Analysis, i.e. fRMA [44]. The idea behind the method is to gather several

[1] The median polish can also be imagined as a smoothing technique.

microarray datasets acquired by the same platform technology and determine global parameters to utilize when processing a new dataset. Fundamentally, fRMA changes the quantile normalization and the summarization steps, since background-correction is an array specific manipulation. First, given the chosen datasets, fRMA has already estimated a reference distribution that new arrays will be projected on to perform normalization. Second, model 6.12 is rewritten as:

$$Y_{i,j,k,b} = \Phi_{i,k} + \Theta_{j,k} + \gamma_{i,k,b} + \mathcal{E}_{i,j,k,b} \tag{6.13}$$

where the new index b introduces the batch effect harboured by every new dataset and $\gamma_{i,k,b}$ defines a random effect due to the i-th probe intensity variability over different batches, with $Var(\gamma_{i,k,b}) = \tau^2_{ik}$. Further, fRMA supposes $Var(\mathcal{E}_{i,j,k,b}) = \sigma^2_{ik}$. Parameters τ^2_{ik}, σ^2_{ik}, $\Phi_{i,k}$ are estimated by fRMA using the collected datasets and are then considered fixed (frozen) parameters to be used on others. Therefore, if a certain b_{th} dataset needs pre-processing; first it is background-corrected, then it is quantile-normalized with respect to the reference distribution and next it goes through the following summarization procedure. Given the frozen probe affinity effects $\hat{\Phi}_{i,k}$, the intensity are corrected, i.e., $Y^*_{i,j,k,b=b_{th}} = Y_{i,j,k,b=b_{th}} - \hat{\Phi}_{i,k}$, which leaves $Y^*_{i,j,k,b=b_{th}} = \Theta_{j,k} + \gamma_{i,k,b=b_{th}} + \mathcal{E}_{i,j,k,b=b_{th}}$. At this point, $\Theta_{j,k}$ needs to be estimated and for the sake of clarity the extreme case with one subject (or array) is described. In such setting j and b indeces are dropped and the model becomes,

$$Y^*_{i,k} = \Theta_k + \gamma_{i,k} + \mathcal{E}_{i,k} \ , \tag{6.14}$$

with

$$\begin{aligned} Var(Y^*_{i,k}) &= Var(\Theta_k + \gamma_{i,k} + \mathcal{E}_{i,k}) = \\ &= Var(\Theta_k) + Var(\gamma_{i,k}) + Var(\mathcal{E}_{i,k}) = \\ &= Var(\gamma_{i,k}) + Var(\mathcal{E}_{i,k}) = \\ &= \tau^2_{i,k} + \sigma^2_{i,k} \ . \end{aligned} \tag{6.15}$$

If $Y^*_{i,k}$ is divided by its variance and every probe is given a weight w_{in}, the RNA abundance for probeset k can be estimated by

$$\hat{\Theta}_k = \frac{\sum_i \frac{w_{in} Y^*_{i,k}}{\tau^2_{i,k} + \sigma^2_{i,k}}}{\sum_i \frac{w_{in}}{\tau^2_{i,k} + \sigma^2_{i,k}}} \tag{6.16}$$

using an M-estimator routine. In typical M-estimator fashion, probes with a low weight can be considered outliers and have a low impact on the final intensity. This is also true for probes with very high variability. This robust estimation is consistently extended by fRMA to entire batches to determine the RNA abundances acquired by every probe for all subjects, i.e., $\Theta_{j,k}$.

6.1.3 Microarray quality control

A crucial initial stage for all Affymetrix microarray studies is quality control. There are two complementary approaches to tackle quality control: image qualitative diagnostics and expression quality assessment. The first approach entails a visual inspection of each subject (or array) to determine the presence of anomalies, such as scratches, spots, hazes and other unusual shapes [133]. Visual inspection can be performed on both raw data and also on pre-processed data. It is very common for the latter case to use outcomes from probe-level fitting models. These models, after RMA background-correction and normalization, are meant to fit probes intensity values. That is, given a probeset, its associated probes values are used to fit,

$$Y_{i,j} = \Phi_i + \Theta_j + \mathcal{E}_{i,j} \tag{6.17}$$

through an M-estimator[2]. The M-estimator generates a weight for probe intensities $Y_{i,j}$, which can all be visualized for the same subject to observe whether specific areas are abnormally highlighted. Besides, residuals and residuals-related quantities are also quite utilized for visual inspections.

The second approach focuses on expression-related measurements. Affymetrix provides several metrics [42] both given for each single array and for multi-arrays. Several graphical approaches are also being utilized to portrait arrays quality, such as RLE and NUSE. The relative log expression, referred to as RLE, is a metric formulated as

$$\text{RLE}(\Theta_{j,k}) = \Theta_{j,k} - \text{median}_j \left\{\Theta_{j,k}\right\} \tag{6.18}$$

and is meant to establish whether the distribution of positive expressed and negative expressed probesets are approximately balanced. As a matter of fact, assuming the probesets to be genes, it is widely accepted that only a few biological changes occur in a disease, which results in equal popularity of upregulated and downregulated genes, as well as absolute values around 0, i.e. small variance. Differently, the normalized unscaled standard error, or NUSE, overlooks how each expression value $\Theta_{j,k}$ variates for subject j with respect to the whole cohort. That is,

$$\text{NUSE}(\Theta_{j,k}) = \frac{\text{SE}(\Theta_{j,k})}{\text{median}_j \left\{\text{SE}(\Theta_{j,k})\right\}} \ , \tag{6.19}$$

with SE standing for the standard error. A global NUSE was also proposed as GNUSE and directly follows the fRMA approach. With NUSE the median standard error at the denominator of the formula 6.19 does not operate on the subjects in the actual dataset but over the subjects in the reference cohort of fRMA, which makes even these standard errors frozen parameters. As for fRMA, GNUSE works trasversally across several microarray datasets acquired by the same platform technology, which makes it a global quality metric for microarrays.

[2] The reported model is the most popular but many others exist.

6.1.4 Removal of Unwanted Variation

Batch harmonization, known also as batch correction, is the procedure to remove systematic noise due to both known and unknown factors. Systematic noise in the biomedical field is typically related to known factors, such as raw data provider and biological sources, but can be also detected regardless. The removal of such noise is pivotal for the accuracy of final outcomes, since it can both arise false results and prevent the discovery of true ones.

One of the first and most used method to correct systematic noise is ComBat [134]. This method supposes that the source of systematic noise is known and it adds it to a linear model. Therefore, ComBat works only when a variable representing the systematic noise can be defined. When this is not possible, i.e. unknown systematic source, the most intuitive approach to correct systematic noise is provided by the Surrogate Variable Analysis (SVA) [135]. The goal of SVA is to capture patterns from the signal and build surrogate variables representative of them, so that they can be used to improve correction of systematic noise.

Along with SVA, methods known for the removal of unwanted variation, or RUV [136], were developed to deal with unwanted factors. Despite of the similarity with SVA, the surrogate variables are built by RUV methods exclusively using a subset of the entire signal, i.e., negative controls, expected to play no role with the factor of interest. By definition then, negative controls are entities, such as genes or probes, totally independent from the factors of interest, which makes them ideal candidates to showcase systematic noise. The RUV underlying model for m samples and n genes or probes is:

$$Y_{m\times n} = X_{m\times p}\beta_{p\times n} + Z_{m\times q}\gamma_{q\times n} + W_{m\times k}\alpha_{k\times n} + \epsilon_{m\times n} \tag{6.20}$$

which also supposes conditionally independent ϵ_{ij} and $\epsilon_{ij} \sim \mathrm{N}(0, \sigma_j^2)\ \forall i, j$. The p features of interest and the q covariates are represented respectively by X and Z, whereas the additional W stands for the k unknown factors. To ease the notation, the sizes subscripts and the matrix of covariates Z are mostly neglected in this section. Therefore the RUV model mainly under study is:

$$Y = X\beta + W\alpha + \epsilon \tag{6.21}$$

Since unknown factors are unobserved per definition, the indeterminate value of k poses a further challenge and forces to make assumptions.

The main assumption of RUV lies in the estimation of W using negative controls. That is, assuming the independency of these ones from X, the model becomes:

$$Y_{ctrl} = W\alpha_{ctrl} + \epsilon_{ctrl} \tag{6.22}$$

which can be tackled by factor analysis techniques like Singular Value Decomposition, i.e. SVD. This setting allows to estimate the unwanted factors $\hat{W}$ and then to regress $\hat{\beta}$ with the simple Ordinary Least Square (OLS):

$$\hat{\beta} = (X^T R_{\hat{W}} X)^{-1} X^T R_{\hat{W}} Y\ , \tag{6.23}$$

where the matrix operator $R_{\hat{W}} = [I - W(W^TW)^{-1}W^T]$ projects onto the orthogonal space of W. This procedure is also referred to as RUV-2 (RUV in two steps). Noteworthy, the RUV procedure has been extended into four steps, known as RUV-4, in order to better control the amount of biological signal removed by the factor analysis estimation of $\hat{W}$.

When the target of the analysis is the identification of differentially expressed genes or probes across groups, RUV-2 and RUV-4 are excellent tools to filter out the systematic batch effect. Though, when the objective of a research is to perform other advanced methods, such as network analysis and clustering, which require a fully adjusted matrix of Y, the RUV procedure must change. Such context can be further generalized to cases where the X matrix of observed features of interest is unobserved. As a matter of fact, even when X is known but is not surely uncorrelated to systematic noise, one can assume it to be unobserved. The model 6.20 when X is considered as unknown becomes:

$$Y = W\alpha + \epsilon \ . \tag{6.24}$$

Then one can make use of negative controls and equation 6.22 to calculate $\hat{W}$ and $\hat{\alpha}$:

$$\hat{\alpha} = (\hat{W}^T\hat{W})^{-1}\hat{W}^TY \ . \tag{6.25}$$

Such estimation of α is strongly biased towards the removal of biological signal if X and W are correlated. Thus, the RUV-random [45] method proposes to Ridge penalize the coefficient in order to control the amount of variability, systematic noise plus biological signal, to be removed. Therefore, the estimation of α becomes:

$$\hat{\alpha} = (\hat{W}^T\hat{W} + \nu)^{-1}\hat{W}^TY \ , \tag{6.26}$$

where ν is the penalty term. Eventually the fully adjusted matrix can be retrieved by $Y_{adj} = Y - \hat{W}\hat{\alpha}$. To be noted, centering Y before performing the adjustment is strongly recommended to avoid the addition of spurious correlations.

Further, the RUV-random approach can be integrated when one desires to remove systematic noise as well as known covariates (Z). Assuming Z to be negligible for negative controls, the adjusted version of RUV-random follows the previous line, which estimates W from 6.22 and then calculates $\hat{\alpha}$ with the closed form solution:

$$\hat{\alpha} = (\hat{W}^TR_Z\hat{W} + \nu)^{-1}\hat{W}^TR_ZY \ . \tag{6.27}$$

This also leads to the estimation of Z coefficients by:

$$\hat{\gamma} = (Z^TZ)^{-1}Z^T(Y - \hat{W}\hat{\alpha}) \ . \tag{6.28}$$

In the end, the adjusted matrix is obtained from the removal of both $\hat{W}\hat{\alpha}$ and $\hat{Z}\hat{\gamma}$ from Y.

Notably, all RUV procedures are not capable of addressing the case where X and W are too correlated, which implies an almost complete removal of interesting signal. Unfortunately, at the state of art, no method to tackle this problem exists. Therefore, RUV-random seems to fill the need of finding an overall good tradeoff between the removal of systematic noise and biological signal.

6.1.5 Estimating negative controls

As previously stated, RUV procedure depends on a set of negative controls. The choice of this set can be either knowledge-driven or empirical. In the empirical case a RUV-affiliated approach to estimate negative controls exists. Let suppose that all the true biological signal is carried by term $X_{m\times p}\beta_{p\times n}$ in model 6.20. Intuitively, the least variable entities (genes) across a dataset could be considered negative controls. In order to determine them, the proposed procedure first calculates the average expression value for each gene and then bins these averages to stratify the different levels of expression. Afterwards, it selects from every bin the nc genes with the lowest IQR and collect them together. Lastly, it uniformly samples from this collect nc genes, which are considered negative controls.

Yet, this approach may not be optimal and alternatives exist. A more sophisticated method to define negative controls explicitly targets the Least Variable Set (LVS) genes [137]. Let consider the probeset-based model 6.17 and the subject (array) effects Θ_j. The fitting of this model pivots on the residual variance σ^2, similarly to what shown in 6.15, which implies an association between the regressed $\hat{\Theta}_j$ and σ^2. Consequently, the variability χ^2 of $\hat{\Theta}_j$, i.e., $\chi^2 = \vec{\hat{\Theta}}^t \operatorname{cov}(\vec{\Theta})\,\vec{\hat{\Theta}}$, depends on σ^2. Now, the main assumption of LVS is that both the true signal and the systematic noise are additive in $\hat{\Theta}_j$ and for negative controls only the latter holds. That is, given a probeset,

$$\begin{aligned} \hat{\Theta}_j &= \text{signal}_j + \text{noise}_j \\ \text{signal}_j &= 0 \;\text{ if probeset is a negative control .} \end{aligned} \tag{6.29}$$

Therefore, if both signal and noise are random effects, probesets with a true signal are expected to have a variability χ^2 greater than negative controls (since they only wavers according to the noise). Hence, the least variable set of probesets can be estimated as negative controls. Though, probesets are directly comparable when their models have common residual variance. To account for residual variances, LVS models the dependency non-parametrically with $\chi^2_k \sim f(\log_2(\sigma^2_k))$, where k indexes probesets, and utilizes quantile regression instead of the standard regression. Doing so, the non-parametric model, usually a B-spline, enables to capture the possible non-linear dependency, whereas quantile regression is flexible enough to target any quantile of the output distribution. In fact, if standard regression models the average output, quantile regression intuitively models a selected quantile of the output distribution. That is, if negative probesets are expected to be below the q_{thres}-quantile of the variability distribution, LVS fits the q_{thres}-quantile of χ^2 and considers negative controls the probesets under the fitted curve.

6.1.6 PANDA

PANDA [46], standing for Passing Attributes between Networks for Data Assimilation, is a method to integrate different data sources and it is commonly used to create regulation networks. Regulation networks represent connections between two types of entities: genes and transcription factors (TFs). Transcription factors are proteins that bind to the promoter region of a gene and trigger the transcription process of its DNA sequence.

PANDA assumes two general kind of nodes: effector node and affected nodes. For simplicity in the following the gene-TFs setting is considered, where TFs are the effector nodes and genes are the affected ones. Given these two types, PANDA takes into account all possible edges: TF to TF, TF to gene and gene to gene. Such edges inherently define three networks interacting with each other. The first one is the cooperative network (P), composed solely by TF to TF edges, and represents how the TFs work together. The second one is the regulatory network (W) that estimates how TFs regulate the genes (or alternatively how the genes are regulated by the TFs). The third one is the co-regulatory network (C), which contains all the gene to gene connections, and expresses how genes are similarly regulated.

Further, for the regulatory network, PANDA defines two quantities for each edge: the responsability (R_{ij}) and the availability (A_{ij}). The former, R_{ij}, determines the impact of the i-th TF on the j-th gene, given that the j-th gene is regulated by other TFs. The latter, A_{ij}, indicates the impact of the i-th TF on the j-th gene, knowing that the i-th TF influences other genes. These two quantities measure the same edge strength in the regulatory network W from two complementary angles.

Therefore, PANDA takes in three networks and searches for an agreement across them. First, the networks need to be normalized because the types of edges are inevitably on different scales. The normalization is performed by converting the values of each network adjacent matrix to Z-scores. In detail, assuming M_{pq} is the value at the p-th row and q-th column of a general adjacent matrix M, the corresponding Z-score can be calculated by:

$$Z(M_{pq}) = \frac{1}{\sqrt{2}} Z_p(M_{pq}) + \frac{1}{\sqrt{2}} Z_r(M_{pq}) \ , \tag{6.30}$$

where Z_p and Z_r respectively compute the Z-score across the p-th row and the q-th column[3].

The integration performed by PANDA requires a similarity metric that measures the level of agreement between the networks and is able both to penalize edges unsupported by the data and to strengthen edges supported by data. PANDA chosen metric is strongly influenced by the Tanimoto similarity score:

$$T(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{\|x\|^2 + \|y\|^2 - \vec{x} \cdot \vec{y}} \ . \tag{6.31}$$

Though, since the range of Tanimoto score is bounded between 0 (total disagreement) and 1 (total agreement) and the adjancent matrices are normalized to Z-scores, the

[3] Explicitly, the Z-score formula is: $Z(x) = \frac{x-\mu}{\sigma}$, where μ and σ are the mean and standard deviation across the values across either a row or a column.

authors of PANDA proposed a similar metric that generates approximated Z-scores:

$$T_Z(\vec{x},\vec{y}) = \frac{\vec{x}\cdot\vec{y}}{\sqrt{\|x\|^2 + \|y\|^2 - |\vec{x}\cdot\vec{y}|}} \ . \tag{6.32}$$

After the normalization and the choice of a similarity metric, the PANDA underlying algorithm can perform data integration. First, the responsability R_{ij} and the availability A_{ij} are calculated by:

$$\begin{aligned} R_{ij}^{(t)} &= T_Z(P_{i\cdot}^{(t)}, W_{\cdot j}^{(t)}) \\ A_{ij}^{(t)} &= T_Z(C_{j\cdot}^{(t)}, W_{i\cdot}^{(t)}) \ , \end{aligned} \tag{6.33}$$

where the responsability $R_{ij}^{(t)}$ is estimated based on how similar the i-th TF and the j-th gene interact with all the TFs and the availability $A_{ij}^{(t)}$ is estimated based on how similar the j-th gene and the i-th TF interact with all the genes. Second, the regulatory network W is updated by: $W_{ij}^{(t)} = \frac{1}{2}A_{ij}^{(t)} + \frac{1}{2}R_{ij}^{(t)}$. Third, the cooperative network P and the co-regulatory network C are updated according to:

$$\begin{aligned} P_{iv}^{(t)} &= T_Z(W_{i\cdot}^{(t)}, W_{v\cdot}^{(t)}) \\ C_{wj}^{(t)} &= T_Z(W_{\cdot w}^{(t)}, W_{\cdot j}^{(t)}) \ . \end{aligned} \tag{6.34}$$

In the case of the cooperative network, the edge weight between two TFs is computed based on how similar such TFs influence all the genes. In analogy, for the co-regulatory network, the edge weight between two genes is estimated based on how similar such genes are affected by all TFs. These three update steps are repeated until a convergence criterion is met. In addition, to regularize the update procedure PANDA applies the standard rule $X^{(t+1)} = (1-\alpha)X^{(t)} + \alpha X^{(t)}$ for all the networks, given the learning rate α. Eventually, the three networks reach a consensus and can be further analysed to unveil patterns across all kind of edges and nodes.

6.1.7 Bi-clustering and tri-clustering with δ-max

Clustering is the basic operation of determining groups of objects that share a similar pattern. Several techniques exist and are heavily used by all the scientific community and even beyond that. The underlying concept of clustering is intuitive: objects with similar features or with similar relationships between features are considered closed. That is, given n objects and m features in $n \times m$ matrix form X, clustering identify row-groups, which are m-long vectors. Besides, this concept extends to higher dimensions. For example, suppose to add the time dimension to X, so that l timepoints are available in the data, i.e., X has dimension $n \times m \times l$. Clustering still finds row-groups but with matrix size $m \times l$.

Now, when measuring the similarity between objects over m features (or $m \times l$) only a subset of them may be actually similar, which favors clustering, and only a subset may make them look diverse. This more-detailed landscape of similarity [138] is tackled by bi-clustering (or tri-clustering) approaches. In a two-dimensional setting bi-clustering identifies sub-matrices of X, while in a three-dimensional setting tri-clustering finds sub-blocks of X. For microarray expression data the CC bi-clustering algorithm [52] is the one of the most popular. The main focus of the algorithm is to determine sub-matrices A of different sizes $I \times J$ with a mean squared residue (MSR) value lower than a predefined δ. Assume each element of A is formulated as

$$a_{ij} = m + r_i + c_j \ , \tag{6.35}$$

with overall average $a_{I,J}$ rows and columns averages a_{iJ} and a_{Ij}. Upon considering $m = a_{I,J}$, $r_i = a_{iJ} - a_{IJ}$ and $c_j = a_{Ij} - a_{IJ}$, each element can be estimated as

$$\hat{a}_{ij} = a_{iJ} + a_{Ij} - a_{IJ} \ . \tag{6.36}$$

The squared residue between is then:

$$\begin{aligned} \mathrm{SR}(a_{ij}) &= (a_{ij} - \hat{a}_{ij})^2 \\ &= (a_{ij} - a_{iJ} - a_{Ij} + a_{IJ})^2 \ , \end{aligned} \tag{6.37}$$

and the MSR is simply the average of the SRs, i.e., $\mathrm{MSR}(A) = \frac{1}{IJ} \sum_i \sum_j \mathrm{SR}(a_{ij})$. Along with MSR, also the row squared residue $d_i(A) = \frac{1}{J} \sum_j \mathrm{SR}(a_{ij})$ is defined, as well as the column squared residue $d_j(A) = \frac{1}{I} \sum_i \mathrm{SR}(a_{ij})$.

Defined the fundamental quantities, the CC algorithm loops four steps (Algorithm 5): multiple nodes deletion, nodes deletion, node addition, masking. Multiple node deletion first removes at once all rows with large residue and then removes all columns

Algorithm 5 Overview of the CC bi-clustering algorithm.

```
Set threshold $\delta$
Set multiple node deletion positive parameter $\alpha$
Set clustering stopping criterion
Set $A = X$, with $I$ and $J$ being the rows and columns respectively
procedure REPEAT(until stopping criterion is met)
    Multiple node deletion: removal of $i$s and $j$s with $d > \alpha\mathrm{MSR}(A)$
    Single node deletion: iterative removal of a single $i$ or $j$ with the largest $d$ until $\mathrm{MSR}(A) < \delta$
    Node addition: addition of $\{i | i \notin I\}$ and $\{j | j \notin J\}$ with $d < \mathrm{MSR}(A)$
    Masking for next bi-cluster discovery: $A = \mathrm{mask}(A)$
end procedure
```

Algorithm 6 Multiple node deletion.

```
Given parameter $\alpha$, with $\alpha > 0$
if MSR($A$) $< \delta$ then
    return $A$
end if
procedure Repeat(until no changes occur)
    Compute MSR($A$), plus $\{d_i|i \in I\}$ and $\{d_j|j \in J\}$
    Remove each $i$ with $d_i > \alpha$MSR($A$)
    Compute MSR($A$), plus $\{d_i|i \in I\}$ and $\{d_j|j \in J\}$
    Remove each $j$ with $d_j > \alpha$MSR($A$)
end procedure
```

accordingly (Algorithm 6). This step quickly subsets the input sub-matrix, especially when it is high-dimensional. Single node deletion follows the same idea of multiple node deletion but it does so at finest level (Algorithm 7). Rows and columns with the largest average residue are erased one by one sequentially and between every two deletions MSR and average residues are updated. This step terminates once MSR drops below δ. Next, since the whole procedure thus far does not guarantee to find maximal bi-cluster, i.e., its solution are sub-optimal, extra rows and columns can be added to the bi-cluster (Algorithm 8). It can be proved that adding either a row or a column that does not belong to A but has average residue lower than the current MSR(A), does not increase the MSR. Therefore, all rows out of A holding such property are first added, and then, upon updating the residues, all analogous columns out of A are added as well. The discovery of a bi-cluster equals the discovery of a pattern. Therefore, to avoid the rediscovery of every found pattern and also to avoid mixing patterns between early and late bi-clusters, the algorithm masks every element in X belonging to A with random values. Found a bi-cluster, the algorithm restarts the loop and it does so until the stopping criterion is met, which is usually after a certain number of bi-clusters are collected.

Algorithm 7 Single node deletion.

```
Compute MSR($A$), plus $\{d_i|i \in I\}$ and $\{d_j|j \in J\}$
if MSR($A$) $< \delta$ then
    return $A$
end if
procedure Repeat(until MSR($A$) $< \delta$)
    Remove row or column with $\max_{i,j}(\{d_i\}, \{d_j\})$
    Compute MSR($A$), plus $\{d_i|i \in I\}$ and $\{d_j|j \in J\}$
end procedure
```

The simplicity of the CC algorithm makes it extendable to higher dimensions. A

Algorithm 8 Node addition.

```
procedure Repeat(until no changes occur)
    Compute MSR($A$), plus $\{d_i | i \notin I\}$ and $\{d_j | j \notin J\}$
    Add each $i$ with $d_i <$ MSR($A$)
    Compute MSR($A$), plus $\{d_i | i \notin I\}$ and $\{d_j | j \notin J\}$
    Add each $j$ with $d_j <$ MSR($A$)
end procedure
```

tri-clustering version of such sort has been proposed with the appellative δ-trimax [51]. Initial model is adjusted to include a third-dimension effect and, consequently, the MSR formulation becomes:

$$\begin{aligned} \mathrm{MSR}(A) &= \frac{1}{IJK}\sum_{ijk}(a_{ijk} - \hat{a}_{ijk})^2 \\ &= \frac{1}{IJK}\sum_{ijk}(a_{ijk} - a_{iJK} - a_{IjK} - a_{IJk} + 2a_{IJK})^2 \ . \end{aligned} \tag{6.38}$$

Accordingly, all steps in the algorithm implement additional operations to address the third dimension. Nevertheless, the backbone idea of the CC bi-clustering algorithm remains the same for the δ-trimax algorithm.

6.2 Methodology behind Automatic Molecular Driver Identification

6.2.1 Multiple Imputation by Chained Equations

Missing data are a critical problem in scientific research. Lack of potential important information can lead to misleading results, or, in parallel, limit studies to draw partial conclusions. Further, missing data can encourage researches to take out samples with the idea of preserving only the complete cases. This choice could dramatically reduce the size of a dataset and it could neglect either unknown sub-populations or vital predictors.

The field of data imputation deals with the substitution of missing data. Several approaches are used to achieve completeness in a dataset without the need to filter out cases. The most basic approach consist in filling out missing data with constant or expected values (e.g., average and median). Nevetheless, it is crucial to evaluate the type of missing data a research has to deal with. There are three classes of missing data [139]: MCAR, MAR and MNAR. The first class, MCAR, indicates cases where observation are missing completely at random. In other words, all cases share the same probability of missing data. With the second class, MAR, such probability is shared only within groups of cases. Therefore, data are simply missing at random. Lastly, the third class, MNAR, represents studies where there is a pattern driving the missingness of data, i.e. data are missing not at random. This class covers all situations where MCAR and MAR do not hold and methods to address such situations vary from case to case.

Although MCAR is the ideal class to perform imputation on, it is hard to assume. Conversely, it is widely popular to assume that data are MAR, imaging that there are different layers of missingness within the observations. That is, the missingness does not depend on unknown patterns and the available observations contain all the information to fill it out. Multiple Imputation by Chained Equations, simply MICE [64], are a family of imputation algorithms that grew popular in the last two decades . The common ground for these methods is that missing data are replaced more than once (i.e., multiple imputation technique), and are MAR [140]. Without specifying the underlying model performing imputation, the core of the MICE procedure can be summarized by Algorithm 9.

Algorithm 9 Core of a general MICE algorithm.

```
Dataset $X$ of size $N \times M$, with $N$ cases and $M$ variables with missingness (assumed to be MAR), plus a general model function $f(\cdot)$
Fill out missingness singularly (e.g. with average values)
repeat
    for each $m \in \mathcal{M} = \left\{m \in M | \exists X_{n,m}^{miss} \text{ for at least one } n\right\}$ do
        Train $X_{n,m}^{obs} = f(\{X_{n,m'}\}_{m'!=m})$, where $X_{n,m}^{obs}$ indicates the observed cases for the m-th variable
        Predict missing cases for the m-th variable using: $X_{n,m}^{miss} = f(\{X_{n,m'}\}_{m'!=m})$
    end for
until order of variables does not affect the prediction
```

The logic behind MICE is to mutually predict missing cases for a variable based on the other variables (or a subset of them) that have been previously imputed. Since the internal loop follows an order of variables, it is common to burn the first imputations to avoid dependance over such order. After this state of convergence is achieved, multiple dataset with different imputation values can be generated and the downstream analysis can be run over each of the imputed dataset. By doing so, it is eventually possible to control whether the imputed values impact on the final outcomes and how they do so. Ultimately, the choice of function $f(\cdot)$ can range over many types of models, from simple linear regression models to random forests.

6.2.2 Dirichlet Processes

Mixture models enable to model observations as a combination of multiple distributions [141]. In the well-known single distribution case, an observation x_n for a sample X_n

is drawn from a probability density function $f(X_n = x_n)$. Mixture models evolve this case by assuming an observation can derive from G components (or groups) with prior probabilities $\{\pi_g\}_{g=1..G}$ and with conditional densities given by $f_g(X_n = x_n|\pi_g)$. By definition, mixture models do not constraint an observation to be sampled from a unique g-th component since it can be drawn from multiple ones. Consequently, the density function for sample X_n is the marginal density:

$$f(X_n = x_n) = \sum_{g}^{G} f_g(X_n = x_n|Z_n = g)\pi(Z_n = g) \ , \tag{6.39}$$

where the proxy variable Z_n indicates the component an observation is drawn from. Though, upon drawing from the mixture model, the proxy Z_n becomes a binary random variable equals to one for a single g-th and zero for all the remainings. It follows that Z_n can be replaced by a random vector $\vec{Z}_n$ whose observations are single draws from a multinomial distribution with parameters equal to the prior component probabilities. That is,

$$(Z_{n1} = z_{n1}, .., Z_{nG} = z_{nG}) \sim \text{Multi}(\pi_1, .., \pi_g) \ . \tag{6.40}$$

Hence, the sampling procedure from a mixture model first requires to draw a g-th component according to 6.40 and, secondly, it requires to sample from the corresponding density $f_g(\cdot)$. Although mixture models enable a heterogeneous derivation of the observations, the number of components is crucial but rarely known a priori .

Chinese Restaurant Process Mixture Model

The Chinese Restaurant Process (CRP) is a common intuitive mechanism that can handle a flexible number of components. Suppose a restaurant has several tables, which are equivalent to the components, and some of them have clients sit at while others are empty. Clients correspond to observations, and when a new client comes in, it can choose whether to sit at a table with others or sit on its own at an empty table. In statistical terms the CRP models the probability of a new n observation (client) being drawn from a g component (table) as:

$$P(Z_n = g; \pi_{[n-1]}, \alpha) = \left\{ \begin{matrix} \frac{|g|}{\alpha+(n-1)} \text{ if } g \in \pi_{[n-1]} \\ \frac{\alpha}{\alpha+(n-1)} \text{ otherwise.} \end{matrix} \right\} \tag{6.41}$$

In words, $\pi_{[n-1]}$ is a partition of the $n-1$ clients already inside the restaurant before n entered and the probability of sitting to a non-empty table is proportional to the number of clients already at it. Contemporarily, the new n client has always the possibility to sit at an empty table and the probability of this event is proportional to the α parameter.

So far, CRP plays the role of the prior distribution over components, like the Multinomial in 6.40. To render the actual mixture model 6.39, a density function $f_g(\cdot)$ is associated with each component. By doing so, the probability for the n observation of being drawn from the g-th component depends on both the number of observations already derived from such component and the probability density of x_n being from g, given by $f_g(x_n|Z_n = g)$. Chosen a class of functions $f(\vec{\theta})$, parameterized by the parameters vector $\vec{\theta}$, the G component densities differ due to the diverse $\vec{\theta}$. Therefore, even the parameters vector $\vec{\theta}$ of a component needs to be drawn as well from a random probability measure $G_0(\cdot)$. In summary, the CRP mixture model for N clients (observations) first draws a partition $\pi_{[N]}$ from a CRP (equation 6.42), then generates the parameters vectors from the chosen measure G_0 (equation 6.43), and, at last, samples from the density $f(\vec{\theta})$ (equation 6.44).

$$\pi_{[N]} \sim \text{CRP}(\alpha, N) \tag{6.42}$$

$$\vec{\theta}_g|\pi_{[N]} \overset{iid}{\sim} G_0 \text{ for } g \in \pi_{[N]} \tag{6.43}$$

$$x_n|\vec{\theta}_g, \pi_{[N]} \overset{ind}{\sim} f(\vec{\theta}) \text{ for } g \in \pi_{[N]}, n \in g \tag{6.44}$$

Dirichlet Mixture Model

The CRP mixture model [142] is exemplary because it shows how the problem of a finite number of components can be worked around. Yet, the CPR prior of the mixture model focuses on partitions of the observations, i.e. how clients are distributed over the tables, while it is more common to work directly on random variables, such as the parameters vectors $\vec{\theta}$. To fully comprehend the importance of passing from partitions to variables, suppose to change the process from seating clients to drawing balls from an urn. Namely, once a ball is randomly extracted, another two balls with the same parameters of the former (e.g. color) are added to the urn. This process, known as the Pólya Urn, is analogous to the CRP but the first draws parameters (of the balls), whereas the second draws a component (i.e., a table for the clients). Intuitively, the two different processes are related since balls with the same parameters form a component, as well as CRP components can share parameters (equation 6.43).

To make the two processes isomorphic, the Pólya Urn is extended to the Blackwell-MacQueen (BM) model urn, which is defined as:

$$\vec{\phi}_n| \left\{\vec{\phi}_i\right\}_{i'=[1,n-1]} \sim \frac{\alpha}{\alpha + (n-1)} G_0 + \frac{1}{\alpha + (n-1)} \sum_{i=1}^{n-1} \delta_{\vec{\phi}_i} \ . \tag{6.45}$$

The analogy between the BM model urn 6.45 and CRP 6.41 is clear. The BM places a random base measure G_0 to draw any new parameters vectors $\vec{\phi}_n$ and it uses the atom measures δ_{ϕ_i} for the already drawn vectors. It follows that a sampling from the BM mixture model employs:

$$\left\{\vec{\phi}_n\right\}_{n=[1,N]} \sim \text{BM}(N, G_0, \alpha) \tag{6.46}$$

$$x_n|\vec{\phi}_n \overset{ind}{\sim} f(\vec{\phi}_n) \text{ for } n = [1, N]. \tag{6.47}$$

Both the CPR and BM model urn provide a flexible prior for the mixture model that relieves from the need of choosing the number of components. The CRP prior generates a partition of the observations, i.e., a sequence of components ($\pi_{[N]}$), whereas the BM prior generates a sequence of random variables for the observations, meaning $\left\{\vec{\phi}_n\right\}_{n=[1,N]}$. Thanks to the de Finetti theorem it is viable to generalize the BM model to a process that allows to independently and identically draw the random variables. This process, whose associated mixture model is reported below, is referred to as Dirichlet Process (DP) [143].

$$G \sim \text{DP}(\alpha, G_0) \tag{6.48}$$

$$\vec{\phi}_n|G \overset{iid}{\sim} G \text{ for } n = [1, N] \tag{6.49}$$

$$x_n|\vec{\phi}_n \overset{ind}{\sim} f(\vec{\phi}_n) \text{ for } n = [1, N] \tag{6.50}$$

The DP yields a random measure G over the parameters vectors (equation 6.48) and, after this measure is instantiated, such random variables can be drawn both independently and identically. Hence, the DP does not need to sample a whole sequence of either components (CRP) or parameters (BM) and any observation can be generated independently upon fixed G.

Hierarchical Dirichlet Mixture Model

The DP mixture model can be also extended when the observations are believed to be organized in levels [144]. In other words, when the observations derive from a hieararchy structure the mixture model can be integrated with further prior conditions. Therefore, if the observations are supposed to derive from M DPs, the Hierarchical Dirichlet Mixture Model (HDMM) becomes:

$$G_0 \sim \text{DP}(\gamma, H) \tag{6.51}$$

$$G_m \sim \text{DP}(\alpha, G_0) \tag{6.52}$$

$$\vec{\phi}_{n,m}|G_m \overset{iid}{\sim} G_m \text{ for } n = [1, N_m] \tag{6.53}$$

$$x_{n,m}|\vec{\phi}_{n,m} \overset{ind}{\sim} f(\vec{\phi}_{n,m}) \text{ for } n = [1, N_m] \ . \tag{6.54}$$

The HDDMs can take in any number of levels. Greater the number of levels, greater will be the modelling of deep shared statistical properties.

6.2.3 Multivariate non-central hypergeometric distributions

Suppose an urn contains N balls of K different colors. Drawing n balls randomly without replacement in this scenario can be modelled using the multivariate hypergeometric distribution. Yet, if the balls are endowed with a weight or more generally a particular characteristic that may bias the drawing towards one color rather than another, the random response variable follows a so-called non-central hypergeometric distribution [145]. In detail, there are two non-central hypergeometric distribution: the Fisher's non-central hypergeometric distribution and Wallenius' non-central hypergeometric distribution. Both distributions are non-central in the sense that the drawing is biased by some characteristic proportional to the probability of extraction. The two non-central hypergeometrics distributions are different from a theoretical standpoint and their application change based on the assumption of the targeted observations. If the n balls are drawn one at a time, then the Wallenius' non-central distribution should be used. If the n balls can be assumed to be independently drawn, or alternatively if the number of n is not known prior to the drawings, the process can be modelled by the Fisher's non-central distribution. The difficulty in the usage of these two distribution is the partition function. Since the balls have different weights, they should be accounted for in the partition function. Yet, since two different non-central hypergeometric distributions have different weights, the comparison between the two is only viable given their partition functions. If N is high and n is sufficiently large the computation of the partition function becomes unfeasible and the non-central hypergeometric distribution not applicable.

6.2.4 Survival Analysis

The analyses of survival observations establish whether a group of predictors is able to model the occurence of failure events, such as death and relapse [146]. Survival analysis represents today one crucial stage to ultimately test a biomedical research question. Many methods addressing survival analysis currently exist, but the Kaplan-Meier model approach and the Proportional Hazards model emerged as the most popular among the scientific community.

Hazards

A well-recognized concept in survival statistics is hazard [147]. In particular, an hazard function is an abstract function that represents the infinitesimal probability of a failure ($R = 1$) occuring over an infinitesimal time interval, given that it has not occured yet. Formally,

$$\lambda(t) = lim_{\Delta t \to 0} \frac{P(R = 1, T \in [t, t + \Delta t[\, | T > t)}{\Delta t} \ , \tag{6.55}$$

where T stands for the time of the event. Hence an hazard is a proxy for the actual probability of an event. The relationship with the density probability $f_T(t)$ of the failure event is provided by the so-called survival function:

$$\begin{aligned} S_T(t) &= 1 - F_t(t) \\ &= 1 - \int_0^t f_T(t')dt' \ , \end{aligned} \tag{6.56}$$

where $F_t(t)$ is the cumulative distribution function. In fact, the hazard function can be also formulated as:

$$\lambda(t) = \frac{f_T(t)}{S_T(t)} \ . \tag{6.57}$$

This relationship entails that greater hazards mean greater risks.

Kaplan-Meier approach

The survival function 6.56 is at the core of the Kaplan-Meier approach. This function describes the probability of survival up to a certain time. Given a set of N samples with failures occuring at $\{t_i\}_{i=1..N}$, the Kaplan-Meier (KM) approach creates a curve that estimates the survival function behaviour along time, starting before any failure and ending at the last one. That is,

$$\hat{S}_T(t) = \prod_{i|t_i<t}^{N} \left(1 - \frac{x_i}{y_i}\right) \ , \tag{6.58}$$

which takes in the number of failures x_i at the i-th time and the number of survived samples y_i at the same time. By doing so, the KM survival curve delineates the characterists of the survival probability descend. The KM approach is especially effective when multiple groups of samples are compared through the survival curves. Qualitatively the direct comparison of the curves within a plot assists to explore and describe the diverse survival chances across groups. Quantitatively, on the other hand, different statistical tests can be employed to test significant differences between the KM survival curves.

Cox Proportional Hazards Model

When comparing the risks of two subjects the hazard ratios are typically exploited. The ratio determines the proportion between the hazard of two different subjects. So far, no specific assumption on the hazards was stated. If the hazards are assumed to depend on time t and a set of predictors $\vec{b}$, but the effects of time and predictors are mutually independent, the hazards ratios become time-independent. Such assumption is called the PH assumption.

The PH assumption is the cornerstone of the wide-spread Cox Proportional Hazards model [148], whose hazards are defined as:

$$\lambda(t, \vec{\beta}|\vec{x}_i) = \lambda_0(t) \exp\left[\vec{x}_i \cdot \vec{\beta}\right] \ , \tag{6.59}$$

in which $\vec{x}_i$ are the predictors (covariates) for a general i-th sample and the time-dependent $\lambda_0(t)$ is referred to as baseline hazard function. It is obvious then that the baseline hazards cancels out in the hazards ratio formulation, which makes the ratio depends solely on the predictors.

Although hazards are not probabilities, they can help estimate a partial likelihood function for a set of failure events[4]. Suppose to collect N statistically independent samples with different and sorted failure events $\{T_i\}_{i=1}^{N} = \{t_1 = T_1, .., t_N = T_N\}$. Then, the partial likelihood can be defined as:

$$L(\vec{\beta}) = \prod_{i=1}^{N} P(j = i, R_j = 1 | T_j \in [t_i, t_i + \Delta t[\,, T_j > t_i, \vec{x}_j) \ , \tag{6.60}$$

where the argument indicates the probability that the i-th sample is the one to fail, knowing that a failure occurs (infinitesimely) around t_i for a sample at risk at t_i and with covariates $\vec{x}_j$. By definition 6.55, the hazards for the samples at $\{t_1, .., t_N\}$ become:

$$\lambda(t_i, \vec{\beta} | \vec{x}_j) = \lim_{\Delta t \to 0} \frac{P(j = i, R_j = 1, T_j \in [t_i, t_i + \Delta t[\, | T_j > t_i, \vec{x}_j)}{\Delta t} \tag{6.61}$$

from which the infinitesimal probabilities can be approximately determined by:

$$P(j = i, R_j = 1, T_j \in [t_i, t_i + \Delta t[\, | T_j > t, \vec{x}_j) \approx \lambda(t_i, \vec{\beta} | \vec{x}_j) \Delta t. \tag{6.62}$$

Using Bayes theorem, the likelihood becomes:

$$\begin{aligned} L(\vec{\beta}) &= \prod_{i=1}^{N} \frac{P(j = i, R_j = 1, T_j \in [t_i, t_i + \Delta t[\, | T_j > t_i, \vec{x}_j)}{P(T_j \in [t_i, t_i + \Delta t[\, | T_j > t_i, \vec{x}_j)} \\ &= \prod_{i=1}^{N} \frac{P(j = i, R_j = 1, T_j \in [t_i, t_i + \Delta t[\, | T_j > t_i, \vec{x}_j)}{\sum_k P(j = k, R_j = 1, T_j \in [t_i, t_i + \Delta t[\, | T_j > t_i, \vec{x}_j)} \ . \end{aligned} \tag{6.63}$$

Notably $P(j = k, R_j = 1, T_j \in [t_i, t_i + \Delta t[\, | T_j > t_i, \vec{x}_j) > 0$ only for samples at risk after t_i. Thus, differently from other likelihood estimations, in survival models, which lack a response function, the order of events plays the role of a response. Lastly, using 6.62 in 6.63, along with the Cox hazard function 6.59, the partial likelihood reveals to be:

$$L(\vec{\beta}) \approx \prod_{i=1}^{N} \frac{\exp(\vec{x}_i \cdot \vec{\beta})}{\sum_{k | T_k > T_i} \exp(\vec{x}_k \cdot \vec{\beta})} \ . \tag{6.64}$$

It follows that, with no need of baseline hazards, the partial likelihood can be maximized to obtain the best predictors that best yield the order of the failures. To be noted, in the presence of censored subjects, the partial likelihood does not change. Conversely, in case of tied-failures, adjustments are required and several approaches offer solutions [149].

[4] Only a partial likelihood can be targeted since the Cox hazards are non-parametric, i.e., no functional form for the baseline hazards is set.

6.3 Methodology behind Intelligible Heterogeneous Networks

6.3.1 RNA-seq quantification

Thanks to the rapid technological advancements of the biotechnology field over the last three decades, nowadays high-throughput of molecular biology data, such as DNA and RNA is viable [150]. In particular, next-generation sequencing to analyse the transcriptome, usually referred to as RNA-seq [151], contribute to delineate how genes, or more generally transcripts, express in different conditions.

The laboratory process of RNA-seq is affected by random processes hard to handle. First, the available biological sample harbours a biological variability, which can be noticed when the several samples are indipendently sequenced. This is due both the wavery nature of RNA-seq, where transcripts abundances may vary even significantly. Secondly, the technical bias caused by the laboratory stages that prepare the sequencing experiment (e.g., library construction) is another source of variability. Thus, each transcript might theoretically be characterized by a random variable that estimates the fragments it yielded. Nevetheless, fragments overlap between transcripts, which makes the usage of such random variable unfeasible.

To make further progress, the number of times a transcript is found in a fragments can instead be considered as a random variable. In this way the analysis of RNA-seq data enters the field of count modelling. The raw read counts (RC) are the read counts directly returned by an RNA-seq experiment, but they need to be efficiently quantified before passing on to downstream tasks [152]. The intuitive total read counts (TC), which are obtained by dividing the RCs for the total number of counts and multiplying for the average total read counts across subjects, does not suffice since it ignores the case of possible different targeted RNA. Plus it is easily biased by large counts. The most common quantification are then: the transcripts per kilobase million (TPM), the reads per kilobase per million mapped reads (RPKM), and the fragments per kilobase per million mapped reads (FPKM). As for the TPM, transcripts read counts are divided by the length of a transcripts (expressed in kilobases) to generate the RPKs, which are then further divided by their total sum over one million (10^6). TPMs are comparable across subjects since the sum of them is constant. This is not true for RPKM and FPKM, whose calculation is reported as follows. First, the read (fragment) counts of a transcript are divided by the number of total reads of its subject over one million, which yields the RPM. Secondly, the RPMs are divided by the length of their respective transcript. Noteworthy, RPKM and FPKM differ only notionally since fragments are simply the mate paired-end reads that count as one. As mentioned, FPKM (RPKM) are difficult quantification units to base a subjects comparison on. This is because their sum is not equal from subject to subject.

6.3.2 Modelling RNA-seq with DESeq

Several other scaling factors exist that provide units to quantify the read (fragment) counts. One that gauged interest in the recent years is the scaling factor of the DESeq model [153], given by:

$$\hat{s}_j = \text{median}_i \frac{k_{ij}}{(\prod_{l=1}^{m} k_{ij}\ .)^{\frac{1}{m}}} \tag{6.65}$$

Such formulation defines the scaling factor for the j-th subject as the median ratio, across transcripts, between transcripts counts, k_{ij}, and the geometric mean of such transcripts counts over all the subjects. Once read counts are scaled by the factors, subjects comparison is viable but transcripts comparison within subject is not. As matter of fact, the scaling factors accounts for the differences of total counts but does not consider that transcripts can have diverse lengths.

The definition of the median-of-ratios scaling factors is one of key characteristics of the DESeq model. Formally, DESeq assumes counts to be distributed according to a Negative Binomial distribution. That is,

$$K_{ij} \sim \text{NB}(\mu_{ij}, \alpha_i)\ , \tag{6.66}$$

in which the counts for the i-th of the j-th subjects has average counts μ_{ij} and dispersion α_i. The reason behind the negative binomial is the following. Imagine that the real concentration of the fragments (paired-end reads) of a transcript i in a subject j is proportional to a random variable R_{ij}, such that the rate of the fragments equals $s_j r_{ij}$. Then, the real expression of the transcript, embodied by the random variable K_{ij}, follows the conditioned Poisson distribution:

$$P(K_{ij} = k_{ij} | r_{ij}, s_j) = \frac{e^{r_{ij}s_j}(-r_{ij}s_j)^{k_{ij}}}{k_{ij}!}\ . \tag{6.67}$$

Because r_{ij} is unknown and represents the observation of the random variable proportional to the real concentration of the fragments, it is necessary to marginalize it out in order to estimate the probability of K_{ij}. Assuming that the j-th subject belongs to a group ρ_j and $R_{ij} \overset{iid}{\sim} P(r_{ij}|q_{i\rho(j)}, \nu_{i\rho(j)})$, the marginal probability becomes:

$$P(K_{ij} = k_{ij} | \mu_{ij}, \sigma^2_{ij}) = \int P(K_{ij} = k_{ij} | r_{ij}, s_j) P(r_{ij}|q_{i\rho(j)}, \nu_{i\rho(j)}) dr_{ij} \tag{6.68}$$

with $\mu_{ij} = s_j q_{i\rho(j)}$ and $\sigma^2_{ij} = \mu_{ij} + s^2_j \nu_{i\rho(j)}$. Now, if the prior probability of R_{ij} is supposed to be a Gamma distribution, the real expression are distributed according to the Negative Binomial 6.66, with $\alpha_i = \sigma^2_{ij} = \mu_{ij} + s^2_j \nu_{i\rho(j)}$.

In short, DESeq model is conditioned on the scaling factors s_j, whose estimates are calculated by 6.65, on the variable $q_{i\rho(j)}$, which is the mean value from R_{ij} (i.e., it is proportional to the real concentration of the fragments) and on the corresponding variance $\nu_{i\rho(j)}$ for R_{ij}. The presence of this latter variance is the reason why the counts k_{ij} are over-dispersed, i.e., $\alpha_i > \mu_{ij}$.

Although an advancement of DESeq was introduced, DeSeq2 [154], the main model remained intact, while choosing how to model the parameters $q_{i\rho(j)}$ and α_i changed. Since the technical considerations on the parameters modelling fall outside the scope of this work, only a few relevant details are reported in the following. First, with DESeq2 the the $q_{i\rho(j)}$ are related to potential counts predictors, i.e., $logq_{ij} = \sum_r x_{jr}\beta_{ir}$. By doing so, DeSeq2 makes further progress in modelling the counts with respect to DESeq, where only the groups determined by $\rho(\cdot)$ were compared. Second, DESeq2 defines a prior for the predictors $\vec{\beta}$ to regularize their values. Third, the overdispersion α_i are estimated by models that borrow information across genes DESeq2. Fourth, DESeq2 exploits a Generalized Linear Model (GLM) for the Negative Binomial to regress the coefficients and then propose to use either the Wald test or the Likelihood Ratio Test (LRT) to test their significance. Lastly, the DESeq2 defines a new transformation, the rlog, to quantify and normalize counts. The basic idea is to fit the NB model without predictors, except for an indicator variable representing the samples, and then to use the regressed coefficient β_{ij} plus the intercept β_{i0} to estimate the normalized count. That is,

$$\text{rlog}(K_i j) = \log_2 q_{ij} = \beta_{i0} + \beta_{ij} \ . \tag{6.69}$$

For genes with high counts, the rlog-transformation provides similar values to a simple log-transformed normalized counts. Though, for genes with low counts, the rlog-transformed counts are pushed around the average of genes across subjects. This property makes the r-log transformation particularly helpful to visualize and compare subjects, since the transformed values tend to be homoscedastic.

6.3.3 Combining p-values

Multiple single analysis are typically performed within a study to test multiple hypothesis simultaneously. Afterwards, it is common to adjust the p-values of such tests to account for the number of independent tests that may potentially yield false significant outcomes. Eventually, only those outcomes with p-values above a certain threshold are filtered out.

When the number of independent tests is high, like for genome-wide data, many of the outcomes turn out to be non-significant [155], although their proximity to significant outcomes may suggest otherwise. This is also true when similar hypotheses are tested in multiple datasets, where the same outcome may reveal to be non-significant in most datasets but not in all.

The methods to combine p-values are apt to channel multiple hypothesis tests into a compressed statistic [156]. Assume N independent hypothesis tests are performed and yields N p-values. If the null distribution of the corresponding N statistics is continuous, then the p-values are independent and identically distributed according to a uniform distribution between zero and one. Thus, given a statistic defined as

$$C^P = \sum_{i=1}^{N} F^{-1}(p_i) \ , \tag{6.70}$$

in which $F(\cdot)$ is a cumulative function, such statistic has a null distribution provided by the additive property of the independent and identically distributed p-values.

Stouffer-Lipták test-statistic and dependence adjustments

The compressed test-statistic C^P follows, under null hypothesis, a standard normal distribution if $F(\cdot)$ is the cumulative distribution function of a standard normal distribution, i.e. $\mathrm{N}(0, 1)$. The corresponding test is defined as the Stouffer-Lipták test for combining p-values [157].

So far, tests have been considered independent. Though, forms or dependency can can be quite common in several fields, especially when the tests are run on quantities with a defined distance metric. Spatial correlation is therefore the most intuitive form of dependency that needs to be addressed. To this end, suppose to have a non-degenerate positive definite correlation matrix Σ that completely embeds the dependencies between the p-values $\vec{p}$. Then, using the Cholesky decomposition, $\Sigma = LL^T$ and transforming the quantiles $\vec{q^*} = L^{-1}F^{-1}(\vec{p})$, the Stouffer-Lipták test holds since the transformed quantiles reveal to be independent [158]. Clearly Cholesky matrix L depends on the correlation matrix Σ, whose estimation needs to be addressed carefully in order to respect the expected properties.

6.3.4 Network Centralities

Networks, or graphs, are mathematical frameworks to represent systems of interconnected objects [159]. The objects are referred to as nodes and the connections between nodes are called either edges or links. Edges can be direct, when they show a direction, or indirect. Network theory is the field covering all facets of networks, from their type of topology to the algorithms that can be run on them.

The most curious action when observing a network is to look for the most important nodes or edges. The concept of "importance" in network theory is association with the concept of *centrality*. More central is a node, more its role to form the network topology is. The commonest centrality metrics for nodes are the degree and the betweenness.

The degree of a node is equivalent to its total number of edges, i.e. both incoming and outgoing. Therefore, nodes with a high degree are connected to many other nodes.

Unlike the degree, the betwenness centrality has a more complex meaning and its linked to the concept of *path*. A path between two nodes is a set of edges, when it exists, that consent to leave from a node and arrive to the other. In particular, the shortest path between two nodes is the path between them with the minimum number of edges. The betweenness for a node x is then formulated as:

$$c_B(x) = \sum_{a!=b!=x} \frac{p_{ab}(x)}{p_{ab}} \tag{6.71}$$

whose argument shows the ratio between the number of shortest paths between nodes a and b going through x ($p_{ab}(x)$), and the simple total number of shortest paths between a and b. By definition, the betweenness captures the bridging role of nodes. A node with high betweenness has the crucial role in a network to assist the exchange of information between nodes. Notably degree centrality and betweenness centrality are not strictly proportional, meaning that a node with a high degree may not have a high betweenness and vice versa.

6.3.5 Network Propagation

In network analysis it is possible to study how a certain information diffuses across the topology [160]. The idea behind the diffusion is clear: if a certain a amount of information is given to each node and is passed on, different areas of the network will gather different amount of such information. Therefore, the structure of a network can be studied in detail using network diffusion based methods.

The Network Propagation algorithm [161] is a diffusion-based method [162] with a well-established physical interpretation. It employs a random walk with restarts over all the network simultaneously according to the equation:

$$\vec{y}_{t+1} = \alpha W \vec{y}_t + (1-\alpha)\vec{y}_0 \tag{6.72}$$

where $\vec{y}(t) = (y_1(t), y_2(t), .., y_N(t))$ represents the probability of a random walker to occur at node i at time t for all N nodes. The initial state y_0 is the vector encoding input binary data like patient-specific genomic alterations (e.g., i-th entry is 1 if i-th gene is mutated, 0 otherwise). Further, matrix W is the adjacency matrix of the network and α is a parameter, with interval $[0, 1]$, controlling the probability for the walker to be retained at the nodes with respect to the probability of the walker moving away from the actual nodes (to adjacent nodes). As t increases, the algorithm converges to a stationary distribution $\vec{y^*}$ that can be seen as a network smoothing of the initial state $\vec{y}_0$. To help imaging the phenomen captured by the Network Propgation algorithm, the Euler-forward numerical implementation of a Laplacian source (and sink) hydrodynamical model can be exploited. When an ideal fluid flows in the network at constant input rate from source nodes (the non-zeros in the initial vector state), it spreads across the network by moving along the edges, to flow out eventually from each node at a constant sink rate. This observation highlights how the stationary distribution $\vec{y^*}$ can be interpreted as a steady-flow distribution. Therefore, upon convergence, the areas where the fluid gathers the most express important structures with respect to the initial state.

6.4 Methodology behind infomax-based multi-modal integration

6.4.1 Convolutional Neural Networks

A Convolutional Neural Network (CNN) is an evolution of a Multi-Layer Perceptron (MLP) [163] that leverages on convolutional operations to favor the extraction of local informative patterns [164]. Nowadays, numerous CNNs exist and they are well recognized to successfully work on images. The concept behind CNNs addresses two problems of the straight implementation of MLPs: the rising number of parameters and the invariantless nature to shifting, rotating and other transformations. In fact, a MLP entails all fully connected layers and, when used on a image, each layer needs to learn one weight per input unit. Then, the high number of parameters tilts towards overfitting during training. Though, most importantly, MLPs are not natively capable of handling spatial information since by flatting images, pixels are not initially considered mutually related. That is, a modification of the image, that exhibits the same true information but portraits it distinctly, is hardly captured by MLPs. Yet, this not the case for CNNs.

The building blocks of CNNs to enhance local learning are the receptive fields [165]. Namely, a receptive field is the region around a single input unit (i.e., the first layer unit are the image pixels) weighting in on the unit signal to produce a feature. To practically yield a feature for each input unit, a kernel multiplies the receptive field element-wisely and the products are summed. The kernel contains a weight for each position of the receptive field and when the kernel slides along all input units it performs convolution on the whole input. This convolution mechanism highlights the second pivotal property of the CNNs: weight sharing. In fact, since the same kernel slides thrugh the whole input, only the weights of the kernel actually can learn (plus biases). Therefore, regardless of the input size the weights are the same, which enormously reduces the trainable parameters with respect to MLPs. A convolutional layer, then, generates one feature for each input unit, usually referred to as feature map. Commonly, convolutional layers produces several feature maps, typically called channels, to model multiple key local patterns. Though, since receptive fields comprise limited regions around an input unit, the discovery of long-distance patterns during learning may resent it. To tackle this issue CNNs utilize pooling layers on the feature maps to combine together the neighbor features, which enables the reception of less-local patterns. Although, pooling layers reduce dimensionality, which is handy for implementation purposes, they can also blur the flowing local information. Thus, the typical trade-off of CNNs is to use enough wide convolutional layers, i.e., many kernels, to cover the potential plethora of local and distant patterns. The sequential pass formed by convolutional layer plus pooling layer represents the fundamental mechanism of modern CNNs, whose architecture consists typically of many of these layers piled up. Ultimately, CNNs are endowed on top with a feed-forward neural network that is fed by the latest feature maps, which are globally pooled and flatten before passing on.

6.4.2 DenseNet

A simple architecture of many sequential convolutional layers may provide good results but, nowadays, the modern architectures do not directly work this way. The reason behind it is that the input signal of interest degrades after passing through a high number of layers. To tackle this defect, the common idea is to create shortcuts for the signal so that it may skip portion of the deep architecture.

Dense Convolutional Networks [166], or DenseNets, incorporate this idea by connecting each layer output to the input of all the following layers. That is, a convolutional l-th layer receives all $(l-1)$ earlier feature maps $\{\mathcal{I}_{l'}\}_{l'>0}$ along with the input $\mathcal{I}_0$ and yields its feature maps:

$$\mathcal{I}_l = H_l\left(\left[\mathcal{I}_0, \mathcal{I}_1, ..., \mathcal{I}_{(l-1)}\right]\right) \ . \tag{6.73}$$

Yet, the joint input for a convolutional layer is only feasible when all feature maps have common shape. Equal shape preserves, indeed, the consistency of the convolution operations. In detail, when the input of a layer is a three dimensional object (e.g., width × height × number of channels) each kernel slides independently per channel and then adds up all cross-channel results. If the feature maps were down-sampled by pooling layers, cross-channels summation would not be possible anymore. To take advantage of both dense connections and pooling implications, DenseNets main body consistent of two alternating constituents: Dense Blocks and transition layers. A Dense Block is an area of the network where a stack of convolutional layers works according to 6.73 and no pooling operation occurs, i.e. feature maps shape is unaffected. DenseNets uses a common number K of output channels, known as growth factor, for all its Dense Blocks. On the other hand, transition layers are the layers employed between Dense Blocks and they consist of a bottleneck layer [167] followed by a pooling layer. So-called bottleneck layers are 1×1 convolutional layers used to lower the number of channels. As a matter of fact, the number of input channels pass on to the l-th layer of a Dense Block goes as $K(l-1) + K_0$, with K_0 being the number of channels of the input, which can be very high. To control this number, at the end of a Dense Block the bottleneck layer compresses the number of channels (to $4K$ according to the original implementation). Only afterwards, a pooling layer is placed to down-sample the feature maps before supply them to the next Dense Block. Eventually, the final Dense Block feeds a fully connected neural network following a global average pooling over the channels (plus flattening). DenseNets show that carrying the past feature maps along relieves the network from re-learning information from past feature maps. In real testing, this collective mechanism eases the need of many parameters since DenseNets performs optimally with small growth rates (e.g. $K = 12$).

6.4.3 Attention Modules

Today, a popular concept in Machine Learning is *attention*, as demonstrated by the growing interest towards recent models like Transformers [168]. Attention is the ability to evaluate the relevance of several inputs differently, which allows to neglect part of the received information and to focus on what information actually matters. Thus, an attention mechanism eventually provides context [169] to a collection of information. The formulation of attention is tied with the definition of expectation since the result of attention can be formulated as a weighted average between all information with normalized weights. That is,

$$\begin{gathered} y = \sum_i a_i x_i \\ \text{with} \\ \sum_i a_i = 1 \ . \end{gathered} \tag{6.74}$$

An attention score a_i can be imagined as the probability of the output y to be x_i, which makes the actual sum over all possible x_i the expectation value of y. The novel application of this intuitive mechanism in ML is to model the attention score for a single information as function of the information itself. Namely,

$$a_i = \frac{f(x_i)}{\sum_i f(x_i)} \ , \tag{6.75}$$

where $f(\cdot)$ is typically a feed-forward neural network. Hence, attention modules consist of trainable blocks used to provide context to the flow of information through a neural network. The implicit advantage of attention modules, upon training, is the production of the attention scores, which can support to explain what patterns a network learnt and what information is really valuable to achieve optimal performances.

6.4.4 LassoNet

Regularization techniques are popular methods to reduce the complexity of a model to yield better generalization properties. Two techniques are regularly used in the Machine Learning world: $L2$ and $L1$ regularizations. Given the linear model $y_i = \vec{x}_i \vec{\beta} + \epsilon_i$ with $\epsilon_i \sim \mathrm{N}(0, \sigma)$ $\forall i$, the Maximum Likelihood (ML) regression is known to perform the following optimization:

$$\begin{gathered} \text{argmax}_{\vec{\beta}} \prod_i \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{y_i - \vec{x}_i \vec{\beta}}{\sigma}\right)^2} \\ \text{and} \\ \text{argmin}_{\vec{\beta}} \sum_i \left(y_i - \vec{x}_i \vec{\beta}\right)^2 \ . \end{gathered} \tag{6.76}$$

What regularization $L2$ and $L1$ techniques do is equivalent to add a prior for the parameters. As a matter of fact, the maximization task in 6.76 derives from the probability of observing each i, i.e., $\text{argmax}_{\vec{\beta}} \prod_i \text{N}(y_i - \vec{x}_i\vec{\beta}, \sigma)$, and in Bayesian terms this is the product of the conditional probabilities $P(y_i|\vec{\beta})$. A prior to the parameters $P(\vec{\beta})$ turns the optimization into $\text{argmax}_{\vec{\beta}} \prod_i P(\vec{y}|\vec{\beta})P(\vec{\beta})$. The $L2$ regularization, known as Ridge Regression, implements a Gaussian prior:

$$P(\beta_j;\delta) \sim \text{N}(0, \frac{1}{\delta}) \ , \tag{6.77}$$

and performs minimization according to:

$$\text{argmin}_{\vec{\beta}} \left[\sum_i \left(y_i - \vec{x}_i\vec{\beta} \right)^2 + \delta^2 \sum_j \beta_j^2 \right] \ . \tag{6.78}$$

Differently, the $L1$ regularization, called Lasso (regression), adopts a Laplace prior:

$$P(\beta_j;\delta) \sim \text{Laplace}(0, \delta) \ . \tag{6.79}$$

to perform:

$$\text{argmin}_{\vec{\beta}} \left[\sum_i \left(y_i - \vec{x}_i\vec{\beta} \right)^2 + \frac{1}{\delta} \sum_j |\beta_j| \right] \ . \tag{6.80}$$

From tasks 6.78 and 6.80 it is clear that Ridge Regression penalizes the L2-norm of the parameters, whereas Lasso does penalize their L1-norm. It follows that both regularizations force the parameters to be small; yet, only the Lasso can cancel coefficients out, i.e., they regress to 0. This property makes features sparse, which is extremely beneficial when high-dimensional data are analysed and dimensionality reduction is demanded. In fact, differently from Ridge Regression, Lasso can be used also as features selection technique.

To extend the benefit of Lasso to higher degree models than linear regression, the LassoNet procedure has recently been proposed [120]. Given a $N \times M$ feature matrix X, with N samples and M features, alongside a chosen outcome y, the ultimate goal of the LassoNet is to perform:

$$\min_{f \in \mathcal{F}, s \subseteq \mathcal{S}} \hat{E} \left[L(f(X_s), y) \right] \ , \tag{6.81}$$

where $\mathcal{S} = \{1, 2, ..M\}$ is the group of all features and $\mathcal{F}$ is an arbitrary class of functions. Thus, the LassoNet seeks to determine a combination of a function f and a subset s of the features such that the error (estimated by the loss function L) between the expected outcome and $f(X_S)$ is minimized.

To tackle the problem, which is unfeasible to solve by a combinatorial grid search (i.e., NP-hard), the LassoNet employs a residual feed-forward neural network (Figure

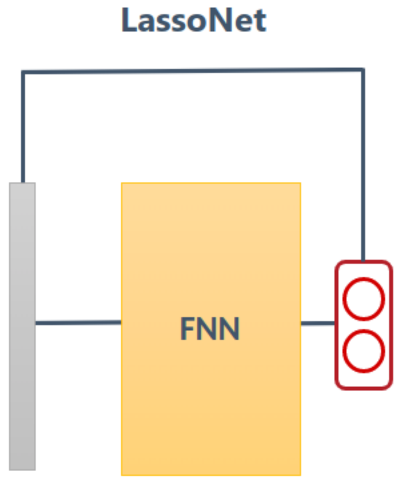


Figure 6.1: LassoNet architecture, which consists of a residual neural network. The residual layer directly connects the input to the output, while a feed-forward neural network (FNN) is employed in parallel.

6.1) and introduces a novel designed objective function. The class of functions $\mathcal{F}$ is then defined as

$$\{f \equiv f_{\theta,W} : X \mapsto \vec{\theta}^T X + g_W(X)\} \ , \tag{6.82}$$

in which $\vec{\theta}$ are the weights of the residual layer and $g_W(\cdot)$ is the feed-forward neural network with weight matrix W. As expected, the main body of the objective function is the standard objective function of Lasso, equation 6.80, where the L1-norm regularization is set only over the weights of the residual layer. In addition to this, LassoNet restricts the weights of the first layer of $g_W(\cdot)$ to be smaller than the residual layer weights. Therefore LassoNet performs training according to:

$$\begin{gathered} \min_{\theta,W} L_{\theta,W}(f(X), y) + \lambda ||\theta||_1 \\ \text{subject to } ||W_j^{(1)}||_\infty \leq M|\theta_j| \ , \ \forall j \in [1, M] \ , \end{gathered} \tag{6.83}$$

where M is a parameter referred to as hierarchy multiplier. The condition on $W^{(1)}$ controls the role of non-linearity for each j-th input feature, which may definitely increase if M is significantly above 0. Under $M = 0$, the non-linear effects totally disappear and only the linear component, provided by the residual layer, holds, which exactly reproduces the standard linear Lasso. Contrarily, if $M \to \infty$, $g_W(\cdot)$ is completely unaffected by the residual layer. Hence, the LassoNet prioritizes the linear-order effects of the residual layer over the non-linear effects of $g_W(\cdot)$, which is then forced to push forward sparse input features.

Changes on the training update rule are necessary to account for the condition on the weights of the first hidden layer. Sequentially, all parameters (θ, W) are first updated with the gradient descend algorithm and then a novel numerically high-performing routine, called hierarchical proximal operator or HIER-PROX, is employed on the constrained parameters, i.e., $(\theta, W^{(1)})$[5]. The procedure of LassoNet is reported by Algorithm 10, where it is clear that the main parameter to set is the hierarchy multiplier M. In addition, a condition for the end of training must be set being either a fixed number of epochs or a convergence criterion. As the well-known Lasso, the LassoNet reveals to be a tool to seek sparse patterns in the input features that can help both to simplify the modelling of y given X and to guide the selection of a relevant subset of input features for arbitrary downstream tasks.

Algorithm 10 Description of the LassoNet procedure.

```
Set the hierarchy parameter $M$
Set the learning rate $\alpha$ for gradient descend and the current stopping criterion
procedure REPEAT(until stopping criterion is met)
    Compute loss and all the gradients for $\theta$ and $W$
    Perform gradient descent on $\theta$ and $W$ with learning rate $\alpha$
    Update $\theta$ and $W$ with the HIER-PROX routine
end procedure
```

6.4.5 Mutual Information

Mutual Information (MI) [170] is a well-recognized construct in information and probability theories. Intuitively, the mutual information between two random variables X and Y measures how much information on X is carried by Y and viceversa. That is, given an observation y of Y, X can be partially estimated. The definition of MI is tied with the Kullback-Leibler (KL) divergence [171]. Assuming $P_{(X,Y)}$ as the joint probability of X and Y, with P_X and P_Y being the respective marginal probabilities, the MI between such two random variables equals:

$$\mathrm{MI}(X,Y) = D_{KL}(P_{(X,Y)}||P_X \otimes P_Y) \ , \tag{6.84}$$

where:

$$P_{(X,Y)}(x,y) \log\left(\frac{P_{(X,Y)}(x,y)}{P_X(x)P_Y(y)}\right) \ . \tag{6.85}$$

[5] The computational details of the HIER-PROX are not provided here because they are beyond the scope of this work. As a matter of fact, the numerical routine underlying the algorithm was customly designed to reproduce the simultaneous mathematical constraints imposed by LassoNet.

In both real ($\mathbb{R}$) and discrete spaces, the KL integrates (i.e., sums in the discrete case) (x, y) over $X \times Y$. Clearly, if two random variables are completely independent the MI equals to 0.

Infomax principle and estimation

In the field of ML the infomax principle holds when the response of a function maximizes the MI with its inputs [172,173]. By definition 6.84, the infomax principle is tied to the estimation of the KL divergence, which is one among a wide range of divergence measures known as *f-divergences* [174]. Namely, any f-divergence has the following general form:

$$D_f(P||Q) = \int_{\mathcal{X}} q(x) f\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x \tag{6.86}$$

where

$$f : (0, \infty) \to \mathbb{R} \text{ , } f \text{ is convex and lower semi-continuous with } f(1) = 0 \text{ .} \tag{6.87}$$

Equation 6.86 becomes the KL divergence once the so-called *generator* is: $f(u) = u\log(u)$.

To estimate f-divergences between two distributions, variational methods can be exploited [119, 175]. These methods target the lower bound of the divergence measure and, to this end, they typically use the convex conjugate of $f(\cdot)$, i.e., $f^*(\cdot)$. The convex conjugate of a function $f(\cdot)$, defined as in 6.87, is given by,

$$f^*(t) = \sup_{u\in(0,\infty)} [tu - f(u)] \tag{6.88}$$

and since $(f(\cdot), f^*(\cdot))$ are mutually dual, plus $f^*(\cdot)$ inherits convexity and lower-continuity, $f(\cdot)$ turns out to be the convex conjugate of $f^*(\cdot)$. Therefore, equation 6.87 is equivalent to:

$$D_f(P||Q) = \int_X q(x) \sup_{t\in dom_{f^*}} [tu(x) - f^*(t)] \, \mathrm{d}x \tag{6.89}$$

in which $u(x) = \frac{p(x)}{q(x)}$. To yield a lower bound, Jensen inequality suffices, while t can be represented as a functional $T : X \to \mathbb{R}$. It follows that:

$$\begin{aligned} D_f(P||Q) &= \int_{\mathcal{X}} q(x) \sup_{t\in dom_{f^*}} \left[t\frac{p(x)}{q(x)} - f^*(t)\right] \mathrm{d}x \\ &\geq \sup_{T\in\mathcal{T}} \int_X q(x) \left[T(x)\frac{p(x)}{q(x)} - f^*(T(x))\right] \mathrm{d}x \\ &= \sup_{T\in\mathcal{T}} \left[\int_X T(x)p(x)\mathrm{d}x - \int_X q(x) f^*(T(x))\mathrm{d}x\right] \text{ .} \end{aligned} \tag{6.90}$$

Lastly, the lower bound can be further represented in terms of expectations:

$$D_f(P||Q) \geq \sup_{T\in\mathcal{T}} \{\mathbb{E}_{x\sim P}[T(x)] - \mathbb{E}_{x\sim Q}[f^*(T(x))]\} \text{ .} \tag{6.91}$$

Hence, an estimation of any f-divergence can be obtained by maximizing the objective

$$F_\omega = \mathbb{E}_{x\sim P}[T_\omega(x)] - \mathbb{E}_{x\sim Q}[f^*(T_\omega(x))] \tag{6.92}$$

over the parameters ω of function $T(\cdot)$ that belongs to a class of functions $\mathcal{T}$. Equation 6.92 can be further generalized to address the constraint on $T_\omega(x)$ co-domain, which coincides with dom_{f^*}. In fact, chosen a proper function $g_f(\cdot)$ that respects such co-domain, $T_\omega(x)$ can be represented as the composition $(g_f \circ V_\omega)(x)$, where $V_\omega : X \to \mathbb{R}$ is unconstrained. Consequently, the final objective is:

$$F_\omega = \mathbb{E}_{x\sim P}[g_f\left(V_\omega(x)\right)] - \mathbb{E}_{x\sim Q}[f^*(g_f\left(V_\omega(x)\right))] \ . \tag{6.93}$$

The variational approach for the KL divergence first implies $f(u) = \text{ulog}(u)$, that makes $f^*(t) = e^{(t-1)}$ with $\text{dom}_{f^*} \equiv \mathbb{R}$ and, secondly, it sets $g_f(\cdot) = I$ since no constraints need to be held. Thus, the objective for the KL divergence results in:

$$F_\omega = \mathbb{E}_{x\sim P}[V_\omega(x)] - \mathbb{E}_{x\sim Q}[e^{V_\omega(x)-1}] \ . \tag{6.94}$$

To respect the infomax principle, then, the $\text{MI}(X, Y)$ can be both estimated and maximized by maximizing the $D_{KL}(P_{(X,Y)}||P_X \otimes P_Y)$ through objective 6.94, whose complete form depends on the crucial choice of V_ω, called the discriminator. The use of encoder plus discriminator, i.e. a feed-forward neural network, is becoming popular thanks especially to the Generative Adversarial Networks (GAN) [176] that are heavility built upon the KL divergence.

6.4.6 Deep InfoMax

Deep InfoMax [118], or simply DIM, is a recently-developed method in Computer Science that extends the infomax principle to Deep Learning architectures. The method was designed for images optimal representation, i.e., embedding, that can be exploited efficiently for downstream tasks. The main idea resumes the goal of Contrastive Learning [177], which seeks to learn representation through several comparisons with similar instances (positive cases) and different instances (negative cases). Ideally, a general framework should learn for each object a representation that is closer to objects with similar properties and farther to those different. In short, the desired representations are task-agnostic and extremely representative of the objects they encode when compared to all the others. Contrastive Learning techniques gauge a lot of interest nowadays because they can boost self-supervised models [178].

Aiming at the infomax principle, DIM seeks an optimal image encoder $E_\psi(\cdot)$, a neural network with parameters ψ, that is able to maximize the MI between the input X and the final global embedding $E_\psi(X)$. DIM defines three objectives, each with its own purpose: a global objective, a local objective and a prior matching objective. The global objective is built to maximize the MI between the global embedding $E_\psi(X)$ and intermediate feature maps. Ideally, the global objective would not use such feature maps but input X; yet, the input may be too uninformative at pixel level. This is why DIM utilizes the feature maps of a chosen convolutional layer, which are expected to embed local significant features to pass on to the next layer. It follows that the encoder is a composition of the function that yields the feature maps, $C_\psi(\cdot)$, and a function that transforms these maps into the global embeddings, i.e., $E_\psi(\cdot) = (f_\psi \circ C_\psi)(\cdot)$. Accordingly, the global objective seeks:

$$\underset{(\psi,\omega_G)}{\text{argmax}}\, \text{MI}_{\omega_G}(C_\psi(X), E_\psi(X)) \tag{6.95}$$

where ω_G stands for the parameters of a fully connected discriminator $V_{\omega_G}(\cdot)$.

When computing the global objective, the feature maps are flattened and concatenated to the global embedding before feeding the discriminator. By doing so, there is no mechanism to prevent the most variable features in the maps to dominate the MI estimation and maximization, washing out the contribution of all the other features. Therefore, DIM formulates a local objective as well, whose goal is to maximize the MI between the global embedding and the local features in the maps. In particular, the feature maps $C_\psi(X)$ correspond to a volume of down-sampled images, with depth equal to the number of channels, and the values of a feature along channels represent all its local embeddings. Then, to drive the global embedding to be equally similar with each feature, DIM defines the local objective:

$$\underset{(\psi,\omega_L)}{\text{argmax}}\, \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \text{MI}_{\omega_L}(C_\psi^{(i,j)}(X), E_\psi(X)) \ . \tag{6.96}$$

Additionally, DIM implements a statistical constraint on the global embedding. Similar to GANs, DIM defines the prior matching objective:

$$\text{argmin}_\psi\, \text{argmax}_\theta\ \text{MI}_\theta(\mathcal{P}_{E_\psi}, \mathcal{Q}) \tag{6.97}$$

so that the distribution of the global embeddings $\mathcal{P}_{E_\psi}$ matches the desired distribution $\mathcal{Q}$. In this way, the statistical properties of the global embeddings reflect those of $\mathcal{Q}$.

At the end, DIM linearly combines the three objectives and trains the encoder (along with the discriminators) to perform:

$$\text{argmax}_{(\psi,\omega_G,\omega_L)}\ \left[\alpha \text{MI}_{\omega_G}(C_\psi(X), E_\psi(X)) + \tfrac{\beta}{IJ} \textstyle\sum_i \sum_j \text{MI}_{\omega_L}(C_\psi^{(i,j)}(X), E_\psi(X))\right] + argmin_\psi\, argmax_\theta\ \gamma \text{MI}_\theta(\mathcal{P}_{E_\psi}, \mathcal{Q}) \ . \tag{6.98}$$

To be noted, the weights α, β and γ reveal to be crucial for the encoder. Ultimately, although the KL divergence should be used to estimate MI, DIM proposes to use other f-divergences to perform the optimization task. In fact, since the main goal is the maximation of MI rather than its actual estimation, using different divergence estimators could help to achieve better performances.

Chapter 7

Conclusion

Four studies were reported in this thesis and they all contributed methodologically to the analysis of biological and medical data. Each study pointed out the benefit of data integration approaches in hunting specific heterogenous signatures, i.e., biomarkers. The efforts made for this thesis were encouraged by the growth of available heterogenous data and were motivated to bring Personalized Medicine closer. The first study showed how horizontal integration created new opportunities to characterize the inner workings of onco-hematological diseases. Clustering fine-level regulatory connections across 13 diseases and more than 30 data providers revealed to be a productive strategy to discover shared gene regulation mechanisms. Vertical integration characterized the other three studies. In the second study the unsupervised stratitification of Acute Myeloid Leukemia was tackled by an early integration over genomic data. Mutations and karyotypes aberrations were concatenated together if they were one single data source and then they were modelled altogether without emphasizing their different biological layer. The refined results, though, were apt to interpretation and clearly identified how the two layers together defined groups of people that corresponded to diverse clinical prognosis. In turn, intelligible heterogeneous networks were built in the third study to provide a context to the effects of conventional lowering-lipids drugs. The conducted late integration joined together a reduce number of interacting constituents that were found to be potentially significant in their respective single-source analysis. Intelligibility of the heterogeneous networks propelled hypothesis generation on the underlying interplay between transcriptome, methylome and (partial) lipidome. A pre-clinical validation supported the exploitation of the networks to drive hypothesis, which led to find evidence on the involvement of a gene in artherosclerosis that was so far unknown to play any role. The fourth study counted on deep Convolutation Neural Networks (CNNs) to perform intermediate integration thanks to an innovative learning rule. Mixing the flexible black-box nature of CNNs with a learning rule forcing data types (gene counts, somatic mutations and histopathological images) to maximize their Mutual Information with a final heterogeneous representation turned out to be a potential optimal strategy to ascertain the goodness of integrating multiple biological layers. Preliminary results on publicly available breast cancer data suggested that it is the combination of gene counts with imaging data to best classify tumour stages in terms of accuracy and variability. The evidence reported in this thesis to support data integration in biomedical research emerged in all four studies. This evidence particularly stressed how multiple single-type data inserted into a certain integrative context allowed to achieve better modelling and to lay optimal groundwork for interpretation. Heterogeneous biomarkers can then reveal to be more effective than the traditional ones because they account for a broader complex biological background of diseases. This augment in specificity comes down to the design of progressively better and precised medical counter-measurements, i.e., for prevention and treatment, and pharmaceutical solutions (e.g., new drugs), which will constitute the principles of Personalized Medicine.

Appendix A

Supplementary for Transcription Factors Bi-Clustering

A.1 List of GEO datasets

Totally 34 datasets from the Gene Expression Omnibus (GEO) database were utilized. The list of all datasets is reported in Table A.1.

A.2 Systematic effects after RMA

The fRMA procedure was the one utilized for the main study, but also RMA was tested. If every single dataset (i.e., GSE) was pre-processed by the standard RMA individually then the systematic noise would have increased. Figure A.1 clearly illustrates, over the first two principal components, how the summarized expression values distributed and how they strongly depended on the batches.

A.3 Bi-clusters overlap

The detected bi-clusters do not overlap, which means that all disease−TF couples are assigned only once. Though, bi-clusters may share a subset of diseases. For example, a bi-cluster can contain ALL, AML plus CLL and another one can contain only ALL plus AML. Therefore, a subset of diseases can be observed over all TFs it has been found on. Table A.2 reports all subsets of diseases with their respective number of shared TFs.

GEO Dataset
GSE14671
GSE39133
GSE12195
GSE118238
GSE19069
GSE66006
GSE93291
GSE36000
GSE25550
GSE24080
GSE53786
GSE15434
GSE19429
GSE35426
GSE127462
GSE79196
GSE58445
GSE21261
GSE6338
GSE17920
GSE11318
GSE132929
GSE35348
GSE39577
GSE13314
GSE69034
GSE39671
GSE12417
GSE19784
GSE34171
GSE6891
GSE93261
GSE50006
GSE13159

Table A.1: Collected datasets from the Gene Expression Omnibus (GEO).

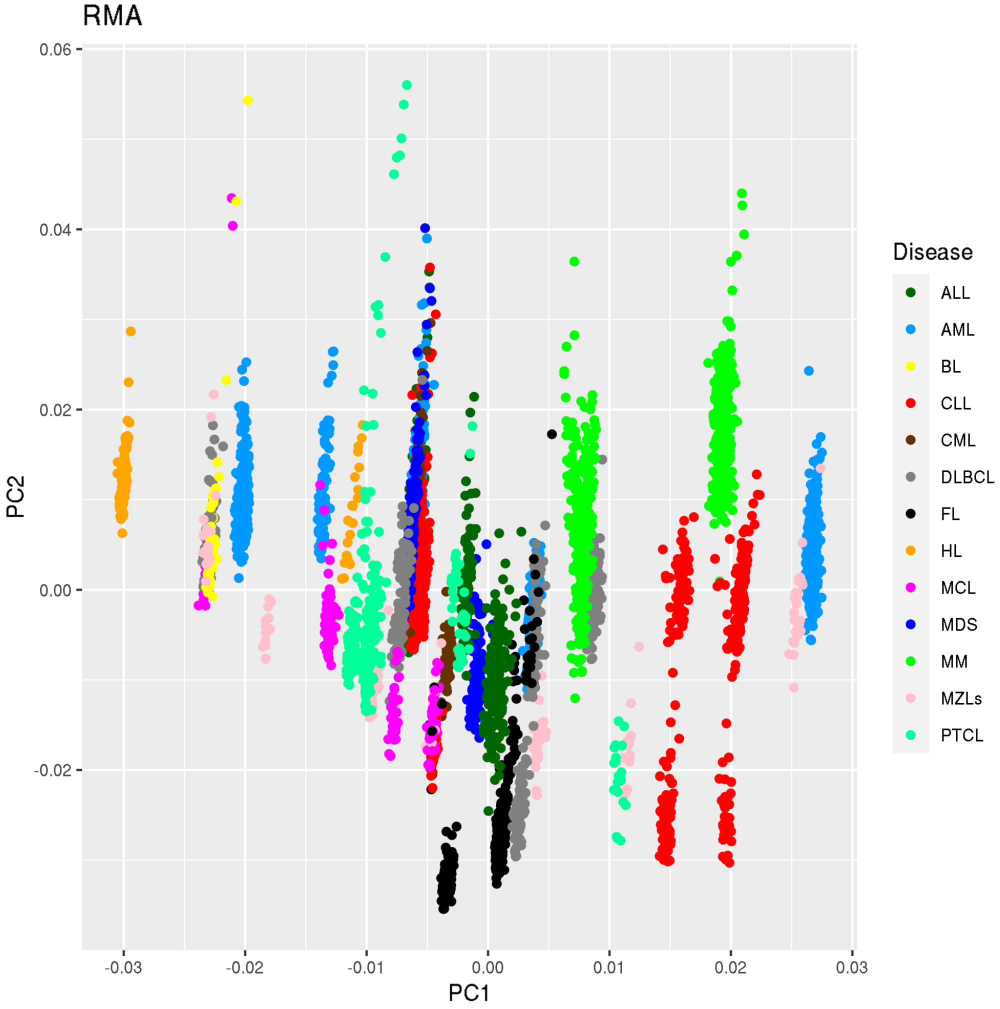


Figure A.1: Organization of diseases over the two principal components when data are normalized according to the standard RMA approach.

A.4 Bi-clusters according to partitions

For the study the bi-clusters were used also to observe how diseases split over diverse set of TFs. Knowing how diseases partition across the TFs might bring further insights in explaining genes co-regulation. A total of 16 partitions were found, which are following reported in Figures A.2, A.3, A.4 and A.5.

Cluster	**Number of TFs**
AML	36
ALL	37
CLL	98
MDS	114
MCL	142
MZLs	146
MM	147
CML	152
PTCL	185
HL	188
DLBCL	188
BL	224
ALL, AML, BL, CLL, CML, DLBCL, HL, MCL, MDS, MM, MZLs, PTCL	786
MCL, PTCL	789
DLBCL, HL, PTCL	822
ALL, AML, CLL, CML, MCL, MDS, MM, MZLs	858
MCL, MM, MZLs	861
AML, CLL, MCL	861
MCL, MZLs	862
MM, MZLs	863
ALL, AML, CLL, MDS	894
ALL ,MDS	896
ALL, AML, CLL	909
ALL, AML	971
FL	1010

Table A.2: Subsets of diseases along with their number of similar transcription factors (TFs).

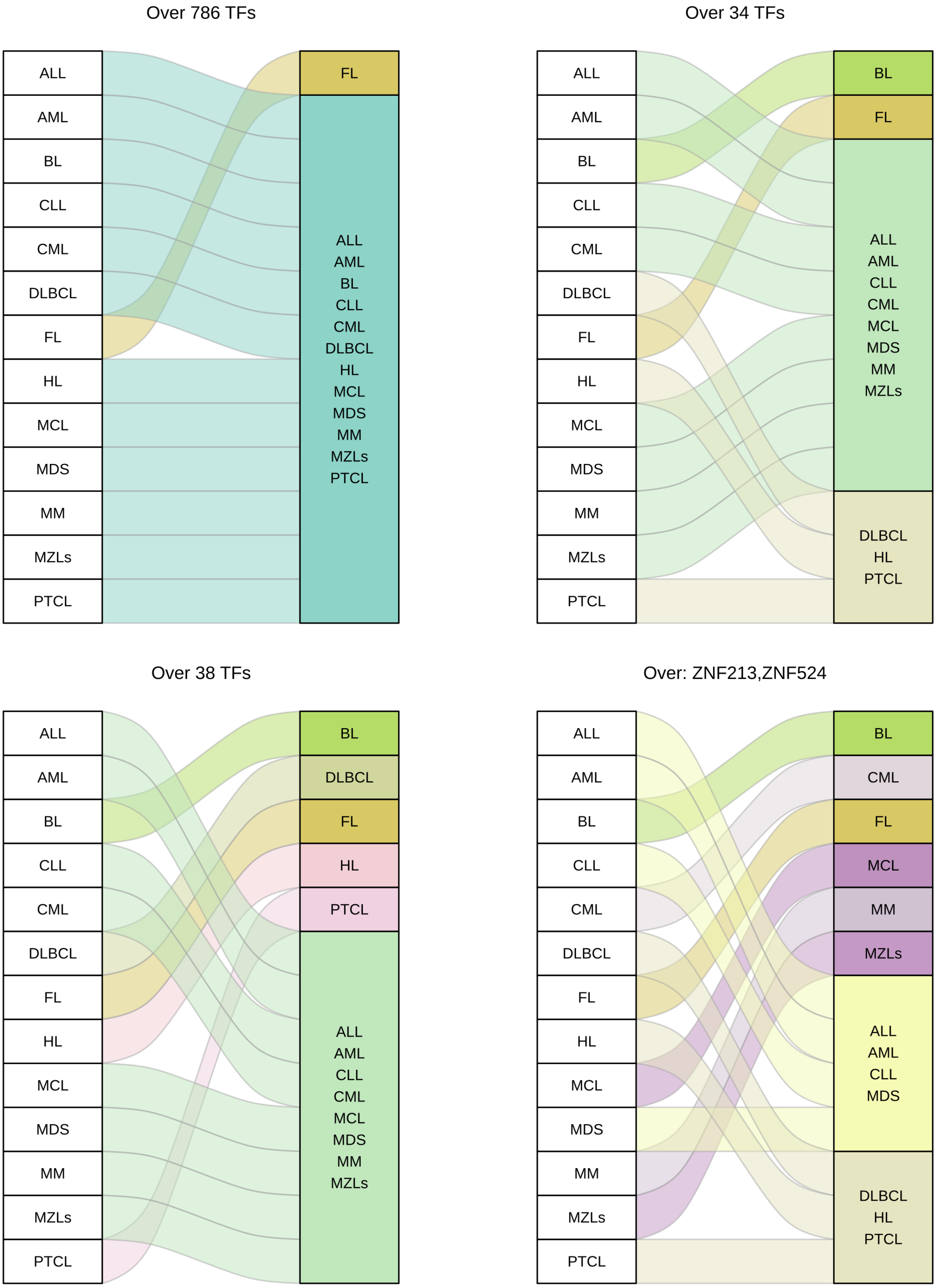


Figure A.2: First four-diseases partitions.

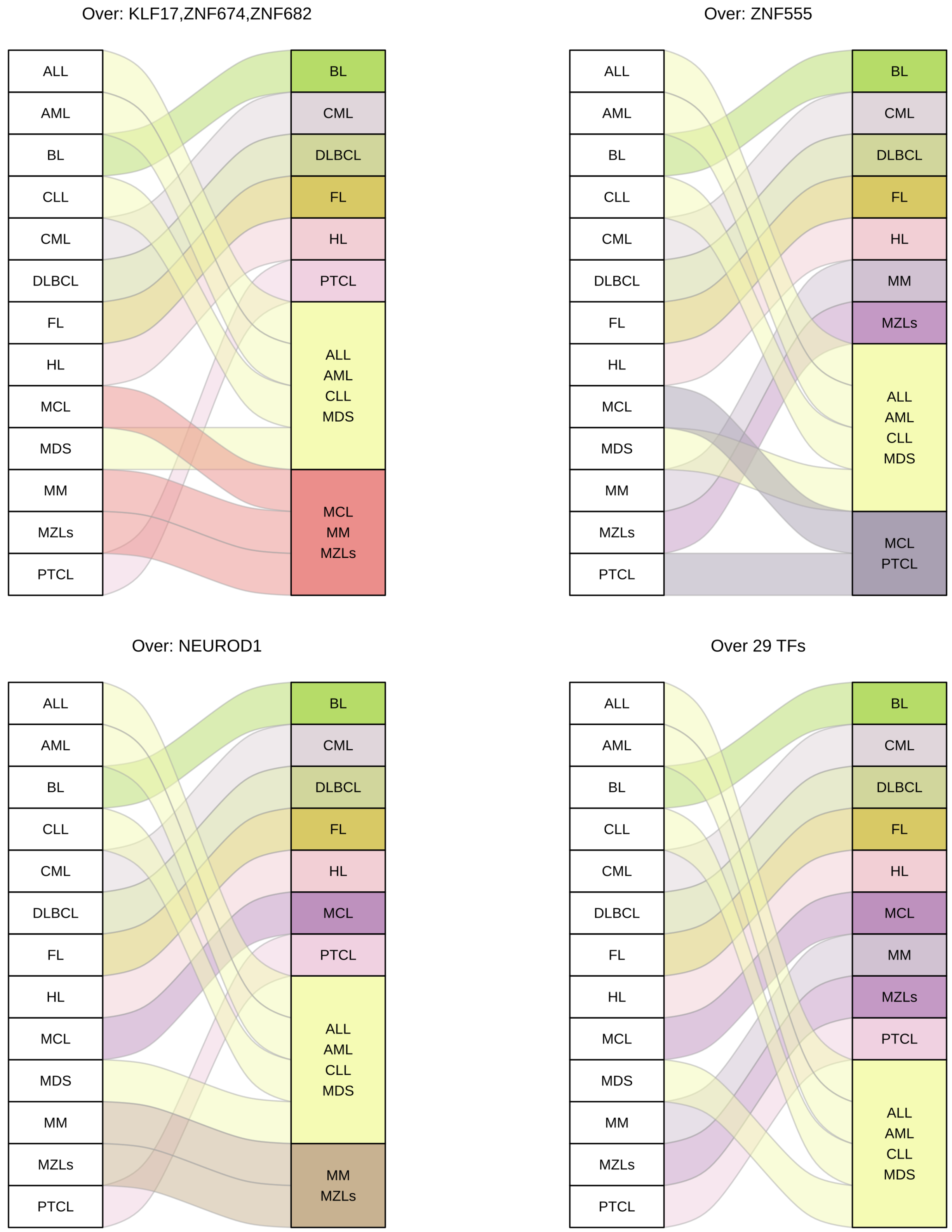


Figure A.3: Second four-diseases partitions.

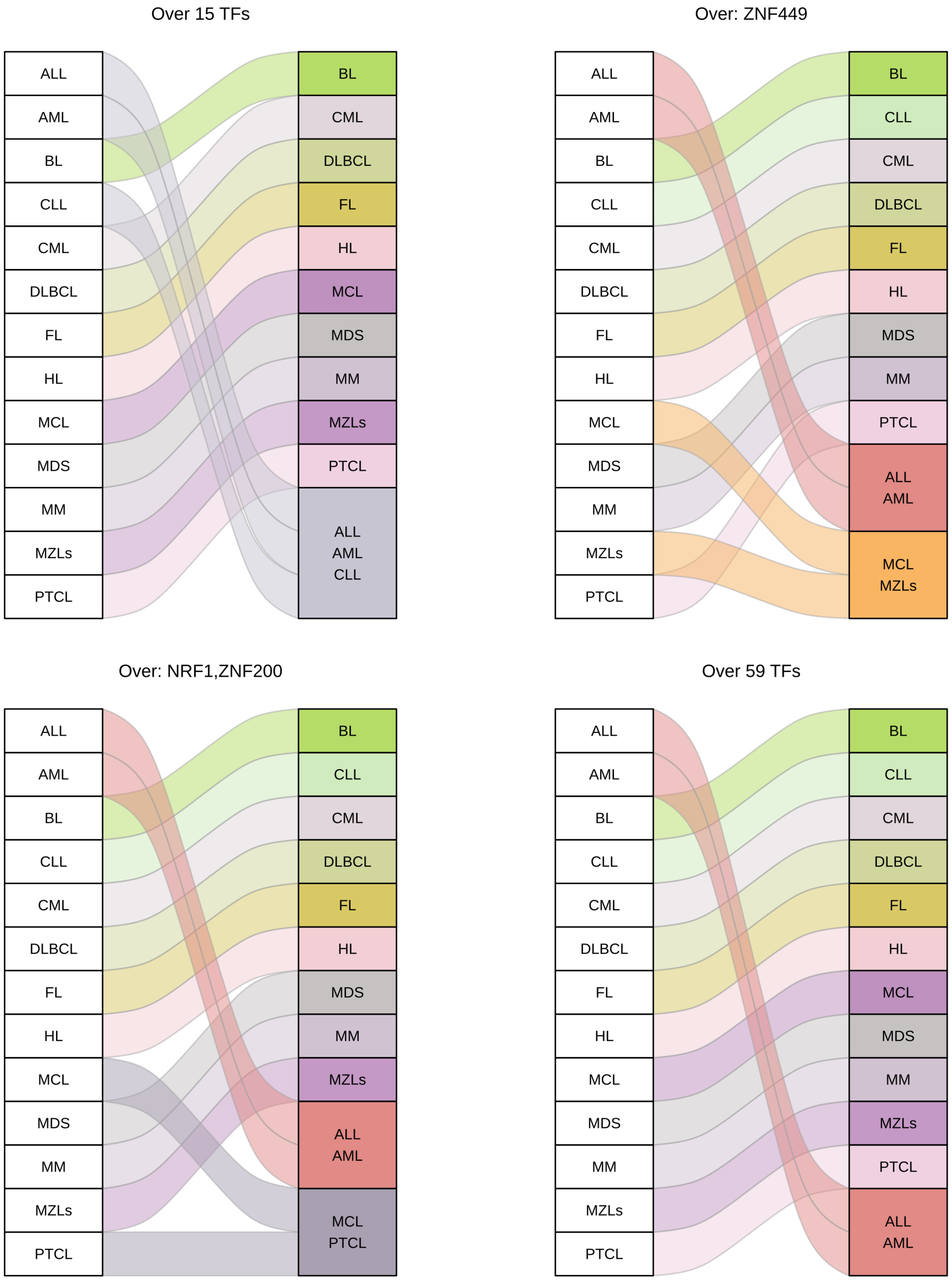
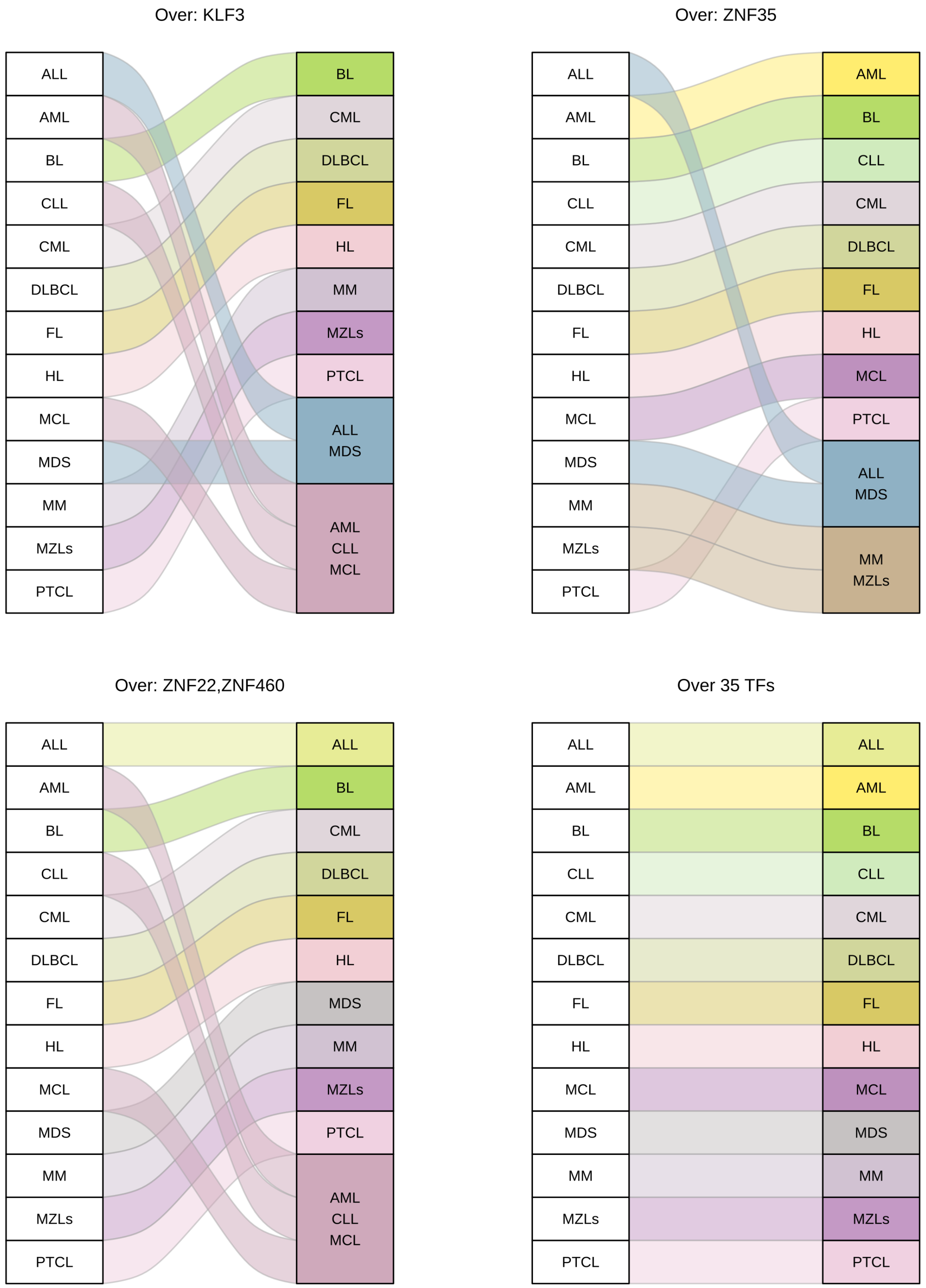


Figure A.4: Third four-diseases partitions.


Over: KLF3

Over: ZNF35

Over: ZNF22,ZNF460

Over 35 TFs

Figure A.5: Fourth four-diseases partitions.

Appendix B

Supplementary for Automatic Molecular Driver Identification

B.1 Potential available molecular variables

Here, the entire list of the monitored genomic alterations is reported.

Karyotypic aberration	**Gene mutation**
abn(3q)	ABCB1
abn(7p)	ABCG2
Complex Karyotype	ABL1
del(7q)	ASXL1
del(9q)	ASXL2
inv(16)	ATRX
inv(3)	BCOR
-12/abn(12p)	BCORL1
-17/abn(17p)	BRAF
-18/del(18p)	BRINP3
-20/del(20q)	CALR
-4/abn(4q)	CBL
-5/del(5q)	CBLB
-7	CBLC
-X	CDKN2A
-Y	CEBPA
+11/dup(11q)	CEBPA$^{\text{bi-allelic}}$
+13	CEBPA$^{\text{mono-allelic}}$
+21	CREBBP
+22	CSF3R

+8/dup(8q)	CUX1
11q23 rearragements	DCK
t(10;11)	DCLK1
t(11;19)	DIS3
t(15;17)	DNMT3A
t(3;5)	EP300
t(6;11)	EPOR
t(6;9)	ETV6
t(8;21)	EZH2
t(9;11)	FBXW7
t(9;22)	FLT3$^{\text{ITD}}$
	FLT3$^{\text{other}}$
	FLT3$^{\text{p.Asp835}}$
	FLT3$^{\text{TKD}}$
	GATA1
	GATA2
	GNAS
	HNRNPK
	HRAS
	IDH1
	IDH2
	IDH2$^{\text{p.140}}$
	IDH2$^{\text{p.172}}$
	IKZF1
	JAK1
	JAK2
	JAK3
	KDM5A
	KDM6A
	KIT
	KIT$^{\text{exon 17}}$
	KIT$^{\text{exon 8}}$
	KRAS
	MIR142
	MLL
	MLL2
	MLL3
	MLL5

	MLL$^{\text{PTD}}$
	MPL
	MYC
	NF1
	NOTCH1
	NPM1
	NRAS
	PHF6
	PRPF40B
	PTEN
	PTPN11
	PTPRT
	RAD21
	RB1
	RUNX1
	SETBP1
	SF1
	SF3A1
	SF3B1
	SH2B3
	SMC1A
	SMC3
	SRSF2
	STAG2
	TERC
	TET2
	TP53
	U2AF1
	U2AF2
	VHL
	WAC
	WT1
	ZRSR2

Table B.1: All genomic and karyotypic alterations selected from the HARMONY Alliance database to characterize Acute Myeloid Leukemia (AML).

B.2 AML components details

After the HDMM chains were fitted, a total of 12 components were eventually extracted. These components were found to be robust because they emerged in most chains and their signature was representative across them. Even when one of these components emerged in a particular chain as multiple components, its signature was found to be extremely recognizable in all of them. The AML components are completely reported in Tables B.2-B.13. The order of the anomalies in each component is based on importance, where importance is quantified by the weight of the Multivariate Fisher's Non-Central Hypergeometric distribution. The higher an anomaly is in the rank, the greater weight it has.

1^{st} Component (Multinomial-based)	1^{st} Component (MFNCHD-based)
NPM1	NPM1
DNMT3A	DNMT3A
FLT3^{ITD}	FLT3^{ITD}
$\text{FLT3}^{\text{other}}$	$\text{FLT3}^{\text{other}}$
NRAS	NRAS
TET2	TET2
IDH1	IDH1
FLT3^{TKD}	FLT3^{TKD}
PTPN11	PTPN11
RAD21	RAD21
WT1	WT1
$\text{FLT3}^{\text{p.Asp835}}$	$\text{FLT3}^{\text{p.Asp835}}$
KRAS	KRAS
STAG2	STAG2
+8/dup(8q)	+8/dup(8q)
CBL	CBL
GATA2	GATA2
del(9q)	del(9q)
KDM6A	KDM6A
MLL	MLL
BRAF	BRAF

Table B.2: Complete signatures of the first AML component.

2^{nd} Component (Multinomial-based)	2^{nd} Component (MFNCHD-based)
Complex Karyotype	Complex Karyotype
TP53	TP53
-5/del(5q)	-5/del(5q)
-17/abn(17p)	-17/abn(17p)
-7	-7
-12/abn(12p)	-12/abn(12p)
+8/dup(8q)	+8/dup(8q)
-18/del(18p)	-18/del(18p)
abn(3q)	-4/abn(4q)
-4/abn(4q)	abn(3q)
-20/del(20q)	-20/del(20q)
del(7q)	del(7q)
+21	+21
+11/dup(11q)	+11/dup(11q)
+22	+22
DNMT3A	PTPN11
NRAS	-Y
PTPN11	abn(7p)
-Y	t(x;11q23)
abn(7p)	del(9q)
del(9q)	+13
t(x;11q23)	-X
+13	NRAS
-X	DNMT3A
KDM6A	KDM6A

Table B.3: Complete signatures of the second AML component.

B.3 Known classifications of AML

The World Health Organization (WHO) characterized Acute Myeloid Leukemia (AML) with subtypes mostly related to clinical prognosis. Ten of them could be used for the setting of the study and are reported in B.14.

Another popular unofficial AML classification [62] was used during the study and since it inspired the work with the refinement technique, its component are reported in Table B.15.

3^{rd} Component (Multinomial-based)	3^{rd} Component (MFNCHD-based)
RUNX1	RUNX1
ASXL1	ASXL1
SRSF2	SRSF2
TET2	STAG2
STAG2	BCOR
NRAS	TET2
+8/dup(8q)	EZH2
BCOR	U2AF1
U2AF1	+8/dup(8q)
EZH2	PHF6
IDH1	IDH1
PHF6	SF3B1
DNMT3A	NRAS
$\mathrm{FLT3}^{\mathrm{ITD}}$	CBL
SF3B1	PTPN11
$\mathrm{FLT3}^{\mathrm{other}}$	ETV6
PTPN11	+13
CBL	ZRSR2
ETV6	DNMT3A
ZRSR2	JAK2
+13	$\mathrm{FLT3}^{\mathrm{other}}$
JAK2	$\mathrm{FLT3}^{\mathrm{ITD}}$
+11/dup(11q)	+11/dup(11q)
KDM6A	NOTCH1
NOTCH1	KDM6A
-20/del(20q)	-20/del(20q)

Table B.4: Complete signatures of the third AML component.

B.4 Automatic classifiers survival curves

In the main text the Kaplan-Meier (KM) survival curves were reported separately for subjects that could be classified by WHO and those who could not. Here the KM survival curves are reported for all subjects, where none was left un-assigned.

Accordingly, herein the computed indexes of the CPH models fitted over the whole cohort are reported (Table 3.6).

4^{th} Component (Multinomial-based)	4^{th} Component (MFNCHD-based)
IDH2	$\text{IDH2}^{\text{p.140}}$
$\text{IDH2}^{\text{p.140}}$	IDH2
NPM1	NPM1
DNMT3A	SRSF2
FLT3^{ITD}	DNMT3A
$\text{FLT3}^{\text{other}}$	FLT3^{ITD}
NRAS	$\text{FLT3}^{\text{other}}$
SRSF2	PTPN11
PTPN11	NRAS

Table B.5: Complete signatures of the fourth AML component.

5^{th} Component (Multinomial-based)	5^{th} Component (MFNCHD-based)
IDH2	$\text{IDH2}^{\text{p.172}}$
$\text{IDH2}^{\text{p.172}}$	IDH2
$\text{IDH2}^{\text{p.140}}$	$\text{IDH2}^{\text{p.140}}$
DNMT3A	SRSF2
SRSF2	STAG2
RUNX1	ASXL1
ASXL1	BCOR
STAG2	RUNX1
BCOR	DNMT3A
+8/dup(8q)	+8/dup(8q)
PHF6	PHF6
+11/dup(11q)	+11/dup(11q)
del(7q)	del(7q)

Table B.6: Complete signatures of the fifth AML component.

6^{th} Component (Multinomial-based)	6^{th} Component (MFNCHD-based)
t(8;21)	t(8;21)
-Y	-Y
KIT	KIT
NRAS	-X
-X	EZH2
EZH2	RAD21
RAD21	del(9q)
del(9q)	$\text{KIT}^{\text{exon 17}}$
TET2	ETV6
ASXL1	KDM6A
$\text{KIT}^{\text{exon 17}}$	$\text{FLT3}^{\text{p.Asp835}}$
KDM6A	JAK2
ETV6	CBL
$\text{FLT3}^{\text{p.Asp835}}$	ASXL1
JAK2	$\text{KIT}^{\text{exon 8}}$
CBL	NRAS
$\text{KIT}^{\text{exon 8}}$	TET2

Table B.7: Complete signatures of the sixth AML component.

7^{th} Component (Multinomial-based)	7^{th} Component (MFNCHD-based)
CEBPA	$\text{CEBPA}^{\text{bi-allelic}}$
$\text{CEBPA}^{\text{bi-allelic}}$	CEBPA
GATA2	GATA2
WT1	WT1
NRAS	KIT
KIT	$\text{CEBPA}^{\text{mono-allelic}}$
$\text{CEBPA}^{\text{mono-allelic}}$	del(9q)
del(9q)	+21
+21	NRAS
FBXW7	FBXW7

Table B.8: Complete signatures of the seventh AML component.

8^{th} Component (Multinomial-based)	8^{th} Component (MFNCHD-based)
-7	inv(3)
NRAS	-7
inv(3)	abn(3q)
KRAS	SF3B1
RUNX1	ETV6
PTPN11	KRAS
abn(3q)	U2AF1
SF3B1	GATA2
WT1	BCOR
U2AF1	PHF6
GATA2	JAK2
BCOR	PTPN11
ETV6	WT1
ASXL1	RUNX1
PHF6	NRAS
JAK2	ASXL1

Table B.9: Complete signatures of the eigth AML component.

9^{th} Component (Multinomial-based)	9^{th} Component (MFNCHD-based)
inv(16)	inv(16)
NRAS	KIT
KIT	KRAS
$\mathrm{FLT3}^{\mathrm{other}}$	NRAS
KRAS	+22
$\mathrm{FLT3}^{\mathrm{TKD}}$	$\mathrm{FLT3}^{\mathrm{TKD}}$
+22	$\mathrm{KIT}^{\mathrm{exon\ 8}}$
$\mathrm{KIT}^{\mathrm{exon\ 8}}$	del(7q)
del(7q)	$\mathrm{FLT3}^{\mathrm{other}}$
$\mathrm{KIT}^{\mathrm{exon\ 17}}$	$\mathrm{KIT}^{\mathrm{exon\ 17}}$

Table B.10: Complete signatures of the ninth AML component.

10^{th} Component (Multinomial-based)	10^{th} Component (MFNCHD-based)
FLT3^{ITD}	t(15;17)
$\text{FLT3}^{\text{other}}$	t(6;9)
t(15;17)	WT1
WT1	FLT3^{ITD}
FLT3^{TKD}	FLT3^{TKD}
t(6;9)	$\text{FLT3}^{\text{other}}$
+8/dup(8q)	+8/dup(8q)
del(7q)	del(7q)

Table B.11: Complete signatures of the tenth AML component.

11^{th} Component (Multinomial-based)	11^{th} Component (MFNCHD-based)
t(x;11q23)	t(x;11q23)
t(9;11)	t(9;11)
NRAS	t(6;11)
+8/dup(8q)	KRAS
$\text{FLT3}^{\text{other}}$	t(11;19)
KRAS	+8/dup(8q)
t(6;11)	ZRSR2
Complex Karyotype	ASXL1
ASXL1	BRAF
t(11;19)	NRAS
ZRSR2	Complex Karyotype
BRAF	$\text{FLT3}^{\text{other}}$
+21	+21

Table B.12: Complete signatures of the eleventh AML component.

12^{th} Component (Multinomial-based)	12^{th} Component (MFNCHD-based)
CEBPA	$\text{CEBPA}^{\text{mono-allelic}}$
$\text{CEBPA}^{\text{mono-allelic}}$	CEBPA
KIT	KIT
FLT3^{ITD}	TET2
NPM1	FLT3^{ITD}
DNMT3A	DNMT3A
TET2	NPM1

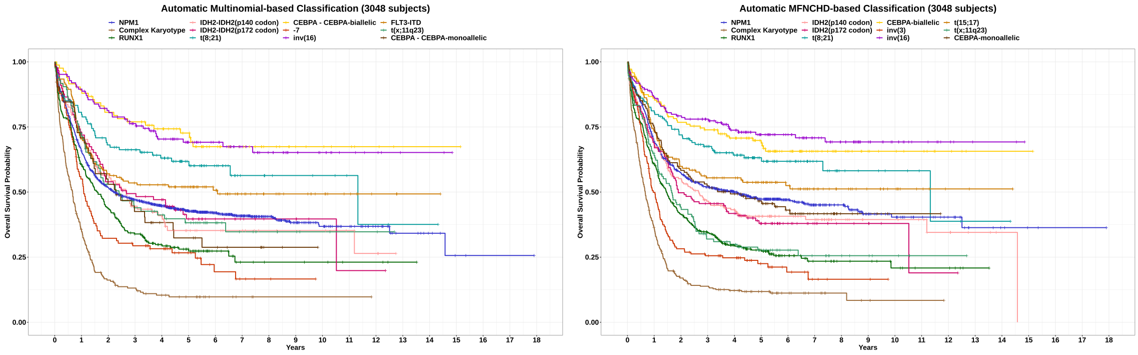
Table B.13: Complete signatures of the twelveth AML component.

WHO subtypes for AML
inv(16)
inv(3)
t(8;21)
t(9;11)
t(6;9)
t(15;17)
mutated RUNX1
mutated TP53
mutated NPM1
biallelic mutation of CEBPA

Table B.14: Overview of the WHO molecular subtypes used in the study for comparison with the AML components identified by the HDMM refined approach.

Genomic classification of AML
mutated NPM1
mutated chromatin, RNA-splicing genes, or both (RUNX1, SRSF2, ASXL1, STAG2)
mutated TP53, chromosomal aneuploidy, or both (Complex Karyotype, -5/5q, -7/7q, TP53, -17/17p, -12/12p)
inv(16)
CEBPA$^{\text{bi-allelic}}$
t(15;17)
t(8;21)
t(x;11q23)
inv(3)
IDH2$^{\text{p172}}$
t(6;9)

Table B.15: Overview of the Pappaemmanuil et al. [62] molecular subtypes used in the study for comparison with the AML components identified by the HDMM refined approach.



(a) (b)

Figure B.1: Kaplan Meier survival curves for all subjects in the cohort according to the automatic multinomial-based classifier and the automatic MFNCHD-based classifier.

Classification system	**Concordance w.r.t. the whole cohort (3048 subjects)**
Multinomial-based	0.653
MFNCHD-based	0.672

Table B.16: Performances, in terms of concordance index, of the Cox Proportional Harzards (CPH) models that were fitted for each classification systems over the whole cohort. The classes were considered as predictors and age plus gender were added it in as covariates.

Appendix C

Supplementary for Intelligible Heterogeneous Networks

C.1 Biologically known seeds

Although seeds were estimated by the GEP analyses, a group of 18 genes were added to the seeds of each group. The rationale was to encourage the Network Propagation algorithm to focus on areas of the Human Interactome (HI) known to be related to lipid metabolism.

C.2 Identification of treatment-related differentially expressed genes

Comparisons of patients treated with the combined treatment reported the largest numbers of DE genes: 83 (compared with placebo-treated patients), 49 (compared with ezetimibe-treated patients) and 13 (compared with simvastatin-treated patients). Of the remaining comparisons, simvastatin-treated patients showed slightly higher numbers of DE genes compared with both placebo-treated patients and ezetimibe-treated patients, 13 and 7, respectively, than the comparison between these latter two groups where only 4 were detected. Simvastatin-treated patients showed several strongly down-regulated genes ($<$ 2-fold) and a few up-regulated ones in every comparison (Figure C.1a). Conversely, subjects treated both with placebo and with ezetimibe had similar up-regulated and down-regulated distributions when compared with subjects treated with the combined therapy and reported only one strongly modulated DE gene when compared with each other. For each treatment group, the DE genes within all comparisons were binded to determine a unique set of DE genes, which we will refer to as seeds. Subjects given the combined treatment had the highest number of seeds, 130. Differently, simvastatin-

Entrez	Gene Symbol	Gene Description
344	APOC2	Apolipoprotein C2
345	APOC3	Apolipoprotein C3
338	APOB	Apolipoprotein B
348	APOE	Apolipoprotein E
1581	CYP7A1	Cytochrome P450 Family 7 Subfamily A Member 1
8431	NR0B2	Nuclear Receptor Subfamily 0 Group B Member 2
8435	SOAT2	Sterol O-Acyltransferase 2
7050	TGIF1	TGFB Induced Factor Homeobox 1
6721	SREBF2	Sterol Regulatory Element Binding Transcription Factor 2
3156	HMGCR	3-Hydroxy-3-Methylglutaryl-CoA Reductase
3157	HMGCS1	3-Hydroxy-3-Methylglutaryl-CoA Synthase 1
3949	LDLR	Low Density Lipoprotein Receptor
255738	PCSK9	Proprotein Convertase Subtilisin/Kexin Type 9
4547	MTTP	Microsomal Triglyceride Transfer Protein
29881	NPC1L1	NPC1 Like Intracellular Cholesterol Transporter 1
10062	NR1H3	Nuclear Receptor Subfamily 1 Group H Member 3
1071	CETP	Cholesteryl Ester Transfer Protein
3990	LIPC	Lipase C, Hepatic Type

Table C.1: List of the biologically known genes that were manually added as seeds in the network analysis

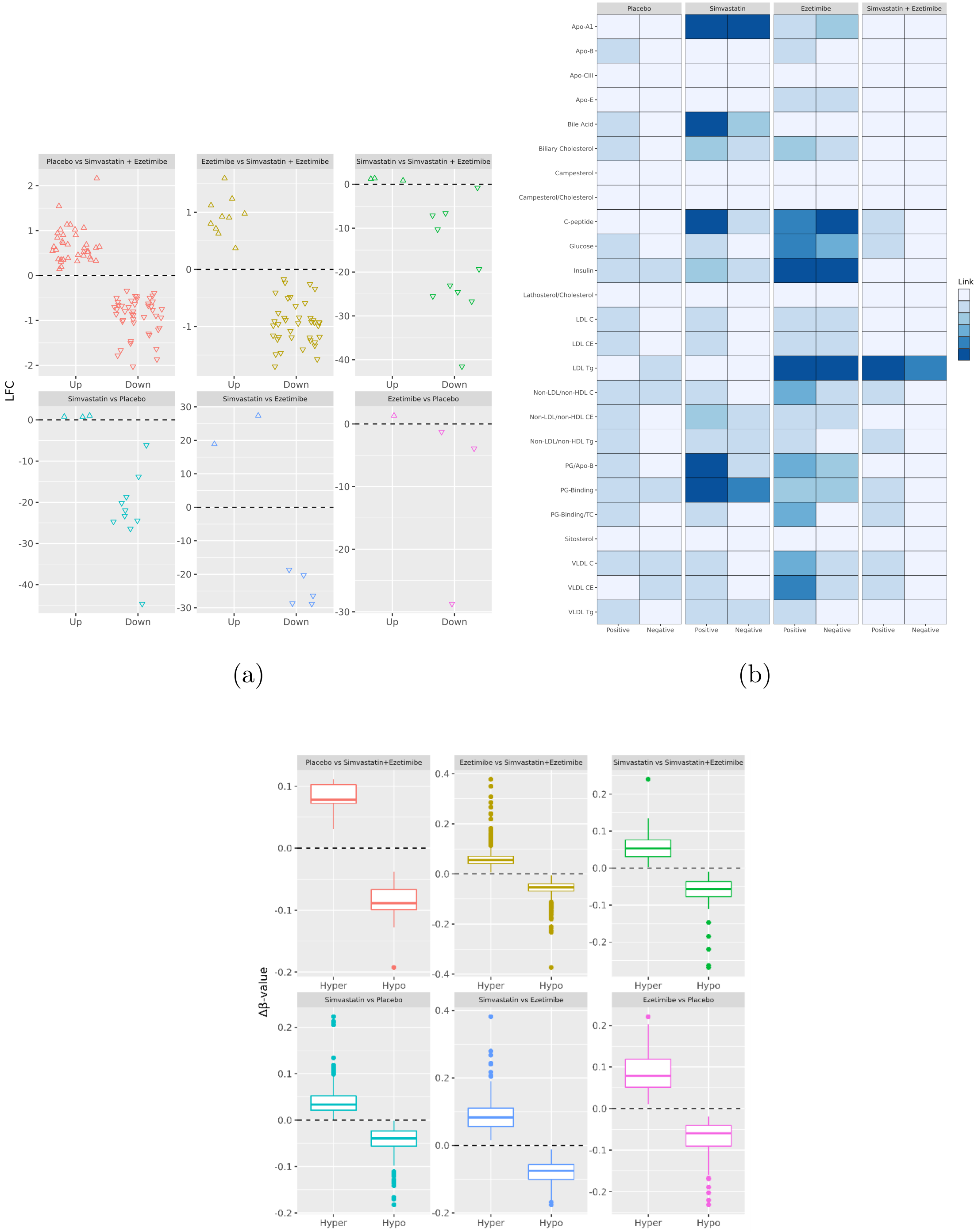
treated patients showed the lowest number of seeds, 19, while subjects receiving ezetimibe reported 60. Lastly, subjects with placebo identified 99 seeds.

C.3 Association of clinical parameters to the treatments

To establish links between RNA-seq data and common laboratory measurements acquired for clinical purposes, the association of gene expression was analysed separately with each of the 50 used biochemical parameters. Such analyses were also performed using generalized linear models and were run individually within each treatment group to highlight the unique characterization of the therapies. Subjects treated with placebo had the smallest number of links between genes and biochemical parameters, 262, both positively (146) and negatively associated (116). Similarly, subjects receiving the combined treatment revealed 351 associations, 206 positively and 145 negatively associated. In contrast, both simvastatin-treated and ezetimibe-treated subjects were found to have strong links to several biochemical parameters, with 681 and 900 associations, respectively. A total of 519 positive and 162 negative links were found for subjects given simvastatin and 480 positive and 420 negative links were for subjects given ezetimibe.

C.4 Detection of differentially methylated regions depending on treatment

The methylation profile of each treatment was analysed by fitting a linear model for each contrast and by using an empirical Bayesian approach to improve the statistical accuracy of the outcomes (like explained in the Methodolody section). Totally 3302 differentially methylated (DM) CpG sites were estimated between subjects receiving the combined treatment and ezetimibe-treated patients. Among such sites, 998 were found to be in intergenic regions. The combined treatment also reported 129 DM sites compared to simvastatin treatment, with almost balanced hyper- and hypo-methylated sites, and 24 DM sites compared with placebo-tested subjects, with only five hyper-methylated sites for placebo. When drug treatments are compared with placebo, 130 DM sites and 676 DM sites were detected, respectively, for ezetimibe and simvastatin treatment. For the latter, six DM sites were discovered within IL32 region. Lastly, 362 DM sites were identified between simvastatin-treated and ezetimibe-treated subjects, with 191 hyper-methylated and 171 hypo-methylated sites with respect to simvastatin. Noteworthy, among the hyper-methylated sites, the atherosclerosis-linked gene SGK1 was found. Generally, DM sites differed in methylation values mostly below 0.1, in terms of β-values, with few exceptions in all comparisons (Figure C.1c). To improve further the treatment profiling and assess the small single-site differences, the results on contiguous single CpG sites were combined in order to determine differentially methylated regions, or DMRs. The comparison with the largest number of DMRs is that of combined therapy vs. ezetimibe treatment, which was expected since it also showed the highest number of DM sites. In fact, 143 DMRs were identified, 79 of which covered more than three CpG sites. Additionally, eight DMRs were identified for other genes. Conversely, subjects given the combined treatment showed only one and three DMRs when compared with placebo-treated subjects (within gene ANKRD2) and simvastatin-treated subjects (within MUC4, KIF26B and a region overlapping NFYA and LOC221442). Placebo-treated subjects yielded four DMRs when compared with ezetimibe-treated patients, three of which were intragenic. Here, the longest DMR was associated with the atherosclerosis-related gene OXT. Instead, the comparison of placebo-treated subjects and simvastatin-treated subjects highlighted 17 DMRs, six of which were intergenic and six others were intragenic and covering more than three CpG sites. In particular, a 5 sites-long region inside IL32 was retrieved. Lastly, 13 DMRs between simvastatin-treated and ezetimibe-treated patients were observed, three of which were located within atherosclerosis-associated regions: TMEM232, SGK1, and a BUD31-PTCD1 overlap region.



(a)

(b)

(c)

Figure C.1: Reductionist analyses outcomes. Figure on the left shows the point distributions of $\log_2$-fold changes of DE genes for every constrast of treatments. Figure on the right represents the level of association (positive and negative) between a biochemical parameter and a treatment. Figure at the bottom reports the distributions of the average difference in β values of DM CPGs for every contrast of treatments.

Appendix D

Supplementary for infomax-based multi-modal integration

D.1 Attention Module

Here, the architecture of the attention network used to score the patches of a Whole Slide Image (WSI) is showed (Figure D.1).

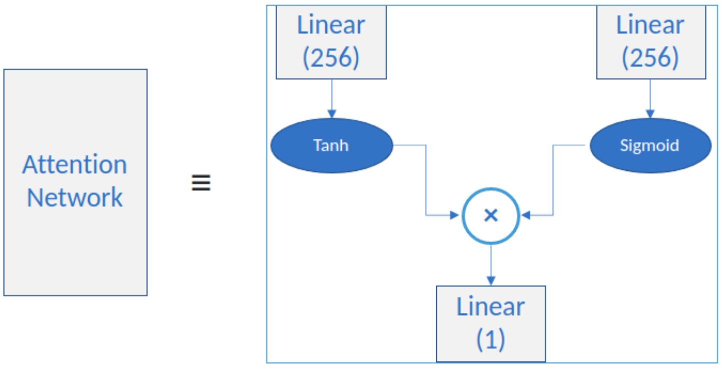


Figure D.1: Illustration of the attention module. Two 256-wide linear layer are applied to the input and then two different activation functions are employed (tanh and sigmoid). The product of their outcome is then computed and passed it on to a single perceptron.

D.2 Optimal Network for WSI

The DenseNet-121 was picked as the deep Convolutional Neural Network (CNN) to encode WSI patches after comparison with ResNet-34 [167], Inception-V4 [179] and

DenseNet-161. All four CNNs were trained to perform patch classification, in which the class of a patch coincided with the class of WSI it was extracted from. With a random stratified train/test split over subjects (396/43) the CNNs were trained on 260653 patches and tested on 33504 patches. Mini-batch training of 32 patches was conducted using the cross-entropy loss and Adam optimizer (learning rate=10^{-4} and weight decay=10^{-4}) for five epochs. No augmentation techniques were utilized but all CNNs were pre-trained on the ImageNet dataset [180]. Both DenseNets showed to be the least CNNs prone to overfitting and with increasing accuracy and Matthews correlation coefficients (MCC) in the first epochs. Evenutally DenseNet-121 was preferred because it has many less parameters than DenseNet-161, i.e., 7 mln against 26 mln.

D.3 Extended-LassoNet fine tuning

To determine optimal depth and width for the feed-forward NN of the extended-LassoNet the following hyper-parameter optimization procedure was run for gene counts and somatic mutations separately. For the sake of clarity the full procedure will be explained for the gene counts.

Each subjects fold (from the ten-fold CV) was further cross validated with ten inner folds. For every inner fold the extended-LassoNet was trained on a regularization path to classify early and late stage tumours based on all genes counts. The training was completed over six configurations of the feed-forward neural networks (Supplementary D.1). Besides, the hierarchy multiplier M was fixed to ten. Training was conducted using a weighted cross-entropy loss and gradients were updated according to the Adam optimizer (learning rate=10^{-3} and momentum=0.9). Also, early stopping was utilized; if the loss did not decrease of, at least, 1% over ten consecutive epochs, the training was interrupted. Number of epochs was set to 10^4. The Matthews correlation coefficient (MCC) was the accuracy metric to base the optimal selection upon. That is, the architecture

Widths of consecutive linear layers
512 → ReLU → 256 → ReLU → 128 → ReLU → 64
256 → ReLU → 128 → ReLU → 64
128 → ReLU → 64
64 → ReLU → 128
64 → ReLU → 128 → ReLU → 256
64 → ReLU → 128 → ReLU → 256 → ReLU → 512

Table D.1: Overview of the depths and widths of the architectures which the fine tuning procedure for the extended-LassoNet was run on. Intermediate ReLU functions were added between the linear layers.

and penalty term with the highest validation median MCC over the 100 inner folds (ten inner folds per ten outer folds) were considered as the optimal hyper-parameters.

D.4 Integrated simulated signature with noise

In addition to the integration of a noisy modality with a modality carrying a clear strong signature, it was observed also the case where noise is simply added to the signal, and then integration is performed (Figure D.2).

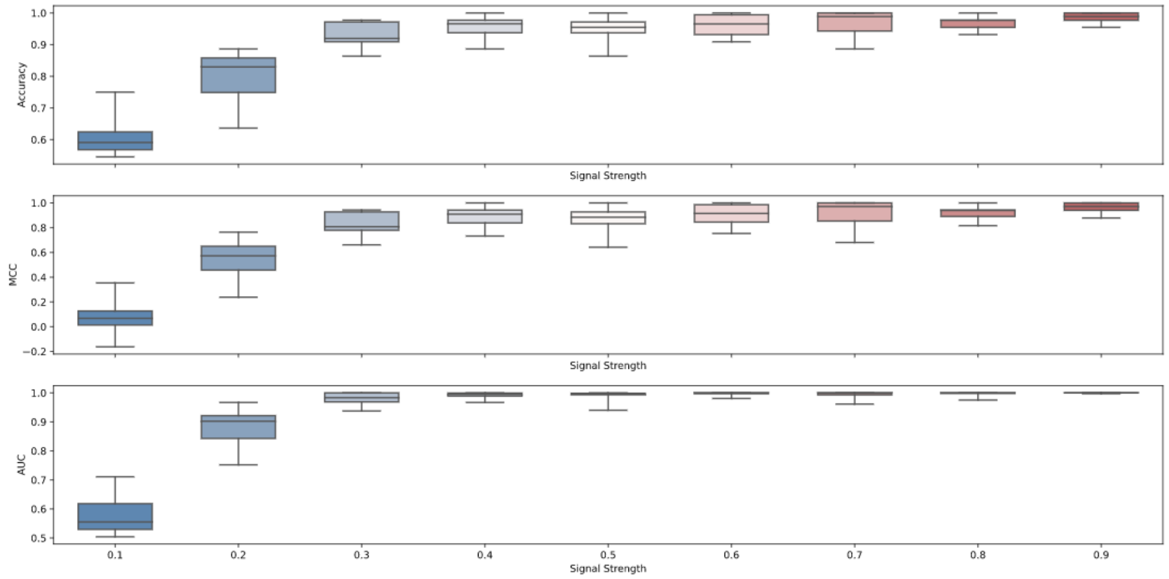


Figure D.2: Effect of additional white gaussian noise (AWGN) on the classification of sine and cosine signals. After being generated they are mixed with AWGN in different proportion.

Interestingly, noise started to interfere with classification when it was extremely dominating over the signal. As a matter of fact, the main drop in performance occurred between 0.2 and 0.1 signal strengths, which mean that the noise contributes approximately $80 - 90\%$ to simulated data passed to the infomax based integration framework.

Bibliography

[1] Simon Tripp and Martin Grueber. Economic impact of the human genome project. *Battelle Memorial Institute*, 58:1–58, 2011.

[2] Sergey Nurk, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V. Bzikadze, Alla Mikheenko, Mitchell R. Vollger, Nicolas Altemose, Lev Uralsky, Ariel Gershman, Sergey Aganezov, Savannah J. Hoyt, Mark Diekhans, Glennis A. Logsdon, Michael Alonge, Stylianos E. Antonarakis, Matthew Borchers, Gerard G. Bouffard, Shelise Y. Brooks, Gina V. Caldas, Nae Chyun Chen, Haoyu Cheng, Chen Shan Chin, William Chow, Leonardo G. de Lima, Philip C. Dishuck, Richard Durbin, Tatiana Dvorkina, Ian T. Fiddes, Giulio Formenti, Robert S. Fulton, Arkarachai Fungtammasan, Erik Garrison, Patrick G.S. Grady, Tina A. Graves-Lindsay, Ira M. Hall, Nancy F. Hansen, Gabrielle A. Hartley, Marina Haukness, Kerstin Howe, Michael W. Hunkapiller, Chirag Jain, Miten Jain, Erich D. Jarvis, Peter Kerpedjiev, Melanie Kirsche, Mikhail Kolmogorov, Jonas Korlach, Milinn Kremitzki, Heng Li, Valerie V. Maduro, Tobias Marschall, Ann M. McCartney, Jennifer McDaniel, Danny E. Miller, James C. Mullikin, Eugene W. Myers, Nathan D. Olson, Benedict Paten, Paul Peluso, Pavel A. Pevzner, David Porubsky, Tamara Potapova, Evgeny I. Rogaev, Jeffrey A. Rosenfeld, Steven L. Salzberg, Valerie A. Schneider, Fritz J. Sedlazeck, Kishwar Shafin, Colin J. Shew, Alaina Shumate, Ying Sims, Arian F.A. Smit, Daniela C. Soto, Ivan Sovi, Jessica M. Storer, Aaron Streets, Beth A. Sullivan, Françoise Thibaud-Nissen, James Torrance, Justin Wagner, Brian P. Walenz, Aaron Wenger, Jonathan M.D. Wood, Chunlin Xiao, Stephanie M. Yan, Alice C. Young, Samantha Zarate, Urvashi Surti, Rajiv C. McCoy, Megan Y. Dennis, Ivan A. Alexandrov, Jennifer L. Gerton, Rachel J. O'Neill, Winston Timp, Justin M. Zook, Michael C. Schatz, Evan E. Eichler, Karen H. Miga, and Adam M. Phillippy. The complete sequence of a human genome. *Science*, 376(6588):44–53, apr 2022.

[3] Richard A Gibbs. The human genome project changed everything. *Nature Reviews Genetics*, 21(10):575–576, 2020.

[4] Sam Behjati and Patrick S. Tarpey. What is next generation sequencing? *Archives of Disease in Childhood - Education and Practice*, 98(6):236–238, dec 2013.

[5] Jeantine E Lunshof, Jason Bobe, John Aach, Misha Angrist, Joseph V Thakuria, Daniel B Vorhaus, Margret R Hoehe, and George M Church. Personal genomes in progress: from the human genome project to the personal genome project. *Dialogues in clinical neuroscience*, 2022.

[6] Li Yan, Neal Rosen, and Carlos Arteaga. Targeted cancer therapies. *Chinese Journal of Cancer*, 30(1):1, 2011.

[7] Indhupriya Subramanian, Srikant Verma, Shiva Kumar, Abhay Jere, and Krishanpal Anamika. Multi-omics data integration, interpretation, and its application. *Bioinformatics and biology insights*, 14:1177932219899051, 2020.

[8] Satya P. Yadav. The Wholeness in Suffix -omics, -omes, and the Word Om. *Journal of Biomolecular Techniques : JBT*, 18(5):277, dec 2007.

[9] Bruce Alberts, Alexander Johnson, Julian Lewis, David Morgan, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell.* 2017.

[10] C. David Allis and Thomas Jenuwein. The molecular hallmarks of epigenetic control. *Nature Reviews Genetics 2016 17:8*, 17:487–500, 6 2016.

[11] Mark Larance and Angus I. Lamond. Multidimensional proteomics for cell biology. *Nature Reviews Molecular Cell Biology 2015 16:5*, 16:269–280, 4 2015.

[12] Yang Woo Kwon, Han Seul Jo, Sungwon Bae, Youngsuk Seo, Parkyong Song, Minseok Song, and Jong Hyuk Yoon. Application of proteomics in cancer: Recent trends and approaches for biomarkers discovery. *Frontiers in Medicine*, 8:1644, 9 2021.

[13] Caroline H. Johnson, Julijana Ivanisevic, and Gary Siuzdak. Metabolomics: beyond biomarkers and towards mechanisms. *Nature Reviews Molecular Cell Biology 2016 17:7*, 17:451–459, 3 2016.

[14] Qiang Yang, Ai Hua Zhang, Jian Hua Miao, Hui Sun, Ying Han, Guang Li Yan, Fang Fang Wu, and Xi Jun Wang. Metabolomics biotechnology, applications, and future trends: a systematic review. *RSC Advances*, 9:37245–37257, 11 2019.

[15] Zoltán N Oltvai and Albert-László Barabási. Life's complexity pyramid. *science*, 298(5594):763–764, 2002.

[16] Joseph Loscalzo, Albert-Lȁszlø' Barabȁsi, and Edwin K. Silverman. *Network medicine : complex systems in human disease and therapeutics.* 2017.

[17] Baruch Barzel, Amitabh Sharma, and Albert-László Barabási. Graph theory properties of cellular networks, 2013.

[18] Yan V. Sun and Yi Juan Hu. Integrative analysis of multi-omics data for discovery and functional studies of complex human diseases. *Advances in genetics*, 93:147–190, 2016.

[19] Yehudit Hasin, Marcus Seldin, and Aldons Lusis. Multi-omics approaches to disease. *Genome Biology 2017 18:1*, 18:1–15, 5 2017.

[20] Konrad J Karczewski and Michael P Snyder. Integrative omics for health and disease. *Nature Publishing Group*, 19, 2018.

[21] Johannes Griss, Guilherme Viteri, Konstantinos Sidiropoulos, Vy Nguyen, Antonio Fabregat, and Henning Hermjakob. Reactomegsa - efficient multi-omics comparative pathway analysis. *Molecular & Cellular Proteomics: MCP*, 19:2115, 12 2020.

[22] Pallavi Tiwari, Satish Viswanath, George Lee, and Anant Madabhushi. Multimodal data fusion schemes for integrated classification of imaging and non-imaging biomedical data. In *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 165–168. IEEE, 2011.

[23] Yehudit Hasin, Marcus Seldin, and Aldons Lusis. Multi-omics approaches to disease. *Genome biology*, 18(1):1–15, 2017.

[24] Paul Pavlidis, Jason Weston, Jinsong Cai, and William Stafford Noble. Learning gene functional classifications from multiple data types. *Journal of computational biology*, 9(2):401–411, 2002.

[25] Iliyan Mihaylov, Maciej Kańduła, Milko Krachunov, and Dimitar Vassilev. A novel framework for horizontal and vertical data integration in cancer studies with application to survival time prediction models. *Biology direct*, 14(1):1–17, 2019.

[26] Shuaichao Wang, Xingjie Shi, Mengyun Wu, and Shuangge Ma. Horizontal and vertical integrative analysis methods for mental disorders omics data. *Scientific reports*, 9(1):1–12, 2019.

[27] Yifeng Li, Fang-Xiang Wu, and Alioune Ngom. A review on machine learning principles for multi-view biological data integration. *Briefings in bioinformatics*, 19(2):325–340, 2018.

[28] Vladimir Gligorijević and Nataša Pržulj. Methods for biological data integration: perspectives and challenges. *Journal of the Royal Society Interface*, 12(112):20150571, 2015.

[29] Albert László Barabási and Zoltán N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics 2004 5:2*, 5:101–113, 2 2004.

[30] Georgios A. Pavlopoulos, Maria Secrier, Charalampos N. Moschopoulos, Theodoros G. Soldatos, Sophia Kossida, Jan Aerts, Reinhard Schneider, and Pantelis G. Bagos. Using graph theory to analyze biological networks. *BioData Mining*, 4:1–27, 4 2011.

[31] Jingwen Yan, Shannon L. Risacher, Li Shen, and Andrew J. Saykin. Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. *Briefings in Bioinformatics*, 19:1370–1381, 11 2018.

[32] Network medicine in pathobiology. *The American Journal of Pathology*, 189:1311, 7 2019.

[33] Mikaela Koutrouli, Evangelos Karatzas, David Paez-Espino, and Georgios A. Pavlopoulos. A guide to conquer the biological network era using graph theory. *Frontiers in Bioengineering and Biotechnology*, 8:34, 1 2020.

[34] Minoru Kanehisa and Susumu Goto. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28:27, 1 2000.

[35] Minoru Kanehisa. Toward understanding the origin and evolution of cellular organisms. *Protein Science : A Publication of the Protein Society*, 28:1947, 11 2019.

[36] Guanming Wu and Robin Haw. Functional interaction network construction and analysis for disease discovery. *Methods in Molecular Biology*, 1558:235–253, 2017.

[37] Frank Emmert-Streib, Matthias Dehmer, and Benjamin Haibe-Kains. Gene regulatory networks and their applications: Understanding biological and medical problems in terms of networks. *Frontiers in Cell and Developmental Biology*, 2:38, 8 2014.

[38] Marouen Ben Guebila, Camila M Lopes-Ramos, Deborah Weighill, Abhijeet Rajendra Sonawane, Rebekka Burkholz, Behrouz Shamsaei, John Platig, Kimberly Glass, Marieke L Kuijjer, and John Quackenbush. Grand: a database of gene regulatory network models across human conditions. *Nucleic Acids Research*, 9 2021.

[39] Todd Andrew Stephenson. An introduction to bayesian network theory and usage. Technical report, Idiap, 2000.

[40] Olivier Gevaert, Frank De Smet, Dirk Timmerman, Yves Moreau, and Bart De Moor. Predicting the prognosis of breast cancer by integrating clinical and microarray data with bayesian networks. *Bioinformatics*, 22(14):e184–e190, 2006.

[41] Luca Falzone, Salvatore Salomone, and Massimo Libra. Evolution of cancer pharmacological treatments at the turn of the third millennium. *Frontiers in Pharmacology*, 9(NOV):1300, nov 2018.

[42] Affymetrix: GeneChip Expression Analysis: Data Analysis Fundamentals. *Santa Clara, CA*, 2002.

[43] Matthew N. McCall, Peter N. Murakami, Margus Lukk, Wolfgang Huber, and Rafael A. Irizarry. Assessing affymetrix GeneChip microarray quality. *BMC Bioinformatics*, 12(1):1–10, may 2011.

[44] Matthew N. McCall, Benjamin M. Bolstad, and Rafael A. Irizarry. Frozen robust multiarray analysis (fRMA). *Biostatistics (Oxford, England)*, 11(2):242, apr 2010.

[45] Systematic noise degrades gene co-expression signals but can be corrected. *BMC Bioinformatics 2015 16:1*, 16(1):1–17, sep 2015.

[46] Kimberly Glass, Curtis Huttenhower, John Quackenbush, and Guo Cheng Yuan. Passing Messages between Biological Networks to Refine Predicted Interactions. *PLOS ONE*, 8(5):e64832, may 2013.

[47] Matthew T Weirauch, Ally Yang, Mihai Albu, Atina G Cote, Alejandro Montenegro-Montero, Philipp Drewe, Hamed S Najafabadi, Samuel A Lambert, Ishminder Mann, Kate Cook, et al. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, 158(6):1431–1443, 2014.

[48] W James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. The human genome browser at ucsc. *Genome research*, 12(6):996–1006, 2002.

[49] Timothy L Bailey, James Johnson, Charles E Grant, and William S Noble. The meme suite. *Nucleic acids research*, 43(W1):W39–W49, 2015.

[50] Damian Szklarczyk, Annika L Gable, Katerina C Nastou, David Lyon, Rebecca Kirsch, Sampo Pyysalo, Nadezhda T Doncheva, Marc Legeay, Tao Fang, Peer Bork, et al. The string database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic acids research*, 49(D1):D605–D612, 2021.

[51] Titin Siswantining, Noval Saputra, Devvi Sarwinda, and Herley Shaori Al-Ash. Triclustering Discovery Using the δ-Trimax Method on Microarray Gene Expression Data. *Symmetry 2021, Vol. 13, Page 437*, 13(3):437, mar 2021.

[52] Yizong Cheng and George M Church. Biclustering of expression data. In *Ismb*, volume 8, pages 93–103, 2000.

[53] Manzhi Wang, Yan Guo, Lin Dong, and Kehong Bi. Coexistence of nodal marginal zone b-cell lymphoma and multiple myeloma: A case report and literature review. *Medicine*, 101(17):e29219, 2022.

[54] Hong-Wei Wang, Wen Yang, Lin Wang, Yun-Long Lu, and Jiang-Yang Lu. Composite diffuse large b-cell lymphoma and classical hodgkin's lymphoma of the stomach: case report and literature review. *World Journal of Gastroenterology: WJG*, 19(37):6304, 2013.

[55] Ida Münster Ikonomou, Anne Tierens, Gunhild Troen, Hege Vangstein Aamot, Sverre Heim, Grete F Lauritzsen, Helen Vålerhaugen, and Jan Delabie. Peripheral t-cell lymphoma with involvement of the expanded mantle zone. *Virchows Archiv*, 449(1):78–87, 2006.

[56] Qinglian Zhai, Maaike van der Lee, Teun van Gelder, and Jesse J Swen. Why we need to take a closer look at genetic contributions to cyp3a activity. *Frontiers in Pharmacology*, 13:912618, 2022.

[57] Maria Solovey, Ying Wang, Christian Michel, Klaus H Metzeler, Tobias Herold, Joachim R Göthert, Volker Ellenrieder, Elisabeth Hessmann, Stefan Gattenlöhner, Andreas Neubauer, et al. Nuclear factor of activated t-cells, nfatc1, governs flt3 itd-driven hematopoietic stem cell transformation and a poor prognosis in aml. *Journal of hematology & oncology*, 12(1):1–12, 2019.

[58] David S Wishart, Craig Knox, An Chi Guo, Savita Shrivastava, Murtaza Hassanali, Paul Stothard, Zhan Chang, and Jennifer Woolsey. Drugbank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research*, 34(suppl_1):D668–D672, 2006.

[59] Leonidas Benetatos and Georgios Vartholomatos. On the potential role of dnmt1 in acute myeloid leukemia and myelodysplastic syndromes: not another mutated epigenetic driver. *Annals of hematology*, 95(10):1571–1582, 2016.

[60] Zhang Zhaojun, Xiong Qian, Wang Shaobin, Wang Hai, Zhang Qian, Qi Heyuan, Li Yanming, Sun Hongying, Chang Kai-Hsin, George Stamatoyannopoulos, et al. Comprehensive investigation of the molecular mechanism of primitive hematopoiesis regulating by klf3. *Blood*, 120(21):4729, 2012.

[61] S.H. Swerdlow, E. Campo, N.L. Harris, E.S. Jaffe, A.S. Pileri, H. Stein, and J. Thiele. *WHO classification of tumours of haematopoietic and lymphoid tissues.* Lyon : International Agency for Research on Cancer, 2017., 2017.

[62] Elli Papaemmanuil, Moritz Gerstung, Lars Bullinger, Verena I. Gaidzik, Peter Paschka, Nicola D. Roberts, Nicola E. Potter, Michael Heuser, Felicitas Thol,

Niccolo Bolli, Gunes Gundem, Peter Van Loo, Inigo Martincorena, Peter Ganly, Laura Mudie, Stuart McLaren, Sarah O'Meara, Keiran Raine, David R. Jones, Jon W. Teague, Adam P. Butler, Mel F. Greaves, Arnold Ganser, Konstanze Döhner, Richard F. Schlenk, Hartmut Döhner, and Peter J. Campbell. Genomic Classification and Prognosis in Acute Myeloid Leukemia. *New England Journal of Medicine*, 374(23):2209–2221, jun 2016.

[63] Paul Madley-Dowd, Rachael Hughes, Kate Tilling, and Jon Heron. The proportion of missing data should not be used to guide decisions on multiple imputation. *Journal of Clinical Epidemiology*, 110:63–73, jun 2019.

[64] Joseph L Schafer and John W Graham. Missing data: our view of the state of the art. *Psychological methods*, 7(2):147, 2002.

[65] Pierre E. Jacob, John O'Leary, and Yves F. Atchadé. Unbiased Markov chain Monte Carlo with couplings. *J. R. Stat. Soc. B*, 82(2):1–32, aug 2017.

[66] Stephen P. Brooks and Andrew Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455, 1998.

[67] Viet-An Nguyen, Jordan L. Boyd-Graber, and Philip Resnik. Sometimes Average is Best: The Importance of Averaging for Prediction using MCMC Inference in Topic Modeling, jan 2014.

[68] Francois Mach, Colin Baigent, Alberico L Catapano, Konstantinos C Koskinas, Manuela Casula, Lina Badimon, M John Chapman, Guy G De Backer, Victoria Delgado, Brian A Ference, et al. 2019 esc/eas guidelines for the management of dyslipidaemias: lipid modification to reduce cardiovascular risk. *Atherosclerosis*, 290:140–205, 2019.

[69] Kivilcim Ozturk, Michelle Dow, Daniel E. Carlin, Rafael Bejar, and Hannah Carter. The emerging potential for network analysis to inform precision cancer medicine. *Journal of molecular biology*, 430:2875, 9 2018.

[70] Joseph Loscalzo, Isaac Kohane, and Albert Laszlo Barabasi. Human disease classification in the postgenomic era: A complex systems approach to human pathobiology. *Molecular Systems Biology*, 3:124, 1 2007.

[71] Giuditta Benincasa, Raffaele Marfella, Nunzia Della Mura, Concetta Schiano, and Claudio Napoli. Strengths and opportunities of network medicine in cardiovascular diseases. *Circulation journal : official journal of the Japanese Circulation Society*, 84:144–152, 1 2020.

[72] Federica Sarno, Giuditta Benincasa, Markus List, Albert Lazlo Barabasi, Jan Baumbach, Fortunato Ciardiello, Sebastiano Filetti, Kimberly Glass, Joseph Loscalzo, Cinzia Marchese, Bradley A. Maron, Paola Paci, Paolo Parini, Enrico Petrillo, Edwin K. Silverman, Antonella Verrienti, Lucia Altucci, Claudio Napoli, Albert Lazlo Barabasi, and Lucia Altucci. Clinical epigenetics settings for cancer and cardiovascular diseases: real-life applications of network medicine at the bedside. *Clinical Epigenetics 2021 13:1*, 13:1–38, 3 2021.

[73] Bradley A. Maron, Rui Sheng Wang, Sergei Shevtsov, Stavros G. Drakos, Elena Arons, Omar Wever-Pinzon, Gordon S. Huggins, Andriy O. Samokhin, William M. Oldham, Yasmine Aguib, Magdi H. Yacoub, Ethan J. Rowin, Barry J. Maron, Martin S. Maron, and Joseph Loscalzo. Individualized interactomes for network-based precision medicine in hypertrophic cardiomyopathy with implications for other clinical pathophenotypes. *Nature Communications 2021 12:1*, 12:1–11, 2 2021.

[74] Marc Vidal, Michael E. Cusick, and Albert László Barabási. Interactome networks and human disease. *Cell*, 144:986, 3 2011.

[75] Ettore Mosca, Matteo Bersanelli, Tommaso Matteuzzi, Noemi Di Nanni, Gastone Castellani, Luciano Milanesi, and Daniel Remondini. Characterization and comparison of gene-centered human interactomes. *Briefings in Bioinformatics*, 22:1–16, 11 2021.

[76] Janet Piñero, Juan Manuel Ramírez-Anguita, Josep Saüch-Pitarch, Francesco Ronzano, Emilio Centeno, Ferran Sanz, and Laura I. Furlong. The disgenet knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research*, 48:D845, 1 2020.

[77] Damian Szklarczyk, Annika L. Gable, Katerina C. Nastou, David Lyon, Rebecca Kirsch, Sampo Pyysalo, Nadezhda T. Doncheva, Marc Legeay, Tao Fang, Peer Bork, Lars J. Jensen, and Christian von Mering. The string database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research*, 49:D605, 1 2021.

[78] Camilla Pramfalk, Tomas Jakobsson, Cristy RC Verzijl, Mirko E Minniti, Clara Obensa, Federico Ripamonti, Maria Olin, Matteo Pedrelli, Mats Eriksson, and Paolo Parini. Generation of new hepatocyte-like in vitro models better resembling human lipid metabolism. *Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids*, 1865(6):158659, 2020.

[79] Osman Ahmed, Karin Littmann, Ulf Gustafsson, Camilla Pramfalk, Katariina Öörni, Lilian Larsson, Mirko E. Minniti, Staffan Sahlin, German Camejo, Paolo

Parini, and Mats Eriksson. Ezetimibe in combination with simvastatin reduces remnant cholesterol without affecting biliary lipid concentrations in gallstone patients. *Journal of the American Heart Association: Cardiovascular and Cerebrovascular Disease*, 7, 12 2018.

[80] Rui-Sheng Wang and Joseph Loscalzo. Network module-based drug repositioning for pulmonary arterial hypertension. *CPT: pharmacometrics & systems pharmacology*, 10(9):994–1005, 2021.

[81] Katja Luck, Dae-Kyum Kim, Luke Lambourne, Kerstin Spirohn, Bridget E Begg, Wenting Bian, Ruth Brignall, Tiziana Cafarelli, Francisco J Campos-Laborie, Benoit Charloteaux, et al. A reference map of the human binary protein interactome. *Nature*, 580(7803):402–408, 2020.

[82] Arunachalam Vinayagam, Ulrich Stelzl, Raphaele Foulle, Stephanie Plassmann, Martina Zenkner, Jan Timm, Heike E Assmus, Miguel A Andrade-Navarro, and Erich E Wanker. A directed protein interaction network for investigating intracellular signal transduction. *Science signaling*, 4(189):rs8–rs8, 2011.

[83] Dénes Türei, Tamás Korcsmáros, and Julio Saez-Rodriguez. Omnipath: guidelines and gateway for literature-curated signaling pathway resources. *Nature methods*, 13(12):966–967, 2016.

[84] Peter V Hornbeck, Bin Zhang, Beth Murray, Jon M Kornhauser, Vaughan Latham, and Elzbieta Skrzypek. Phosphositeplus, 2014: mutations, ptms and recalibrations. *Nucleic acids research*, 43(D1):D512–D520, 2015.

[85] Robert H Newman, Jianfei Hu, Hee-Sool Rho, Zhi Xie, Crystal Woodard, John Neiswinger, Christopher Cooper, Matthew Shirley, Hillary M Clark, Shaohui Hu, et al. Construction of human activity-based phosphorylation networks. *Molecular systems biology*, 9(1):655, 2013.

[86] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 2014.

[87] Matthew Stephens. False discovery rates: A new deal. *Biostatistics*, 2017.

[88] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 1995.

[89] Anqi Zhu, Joseph G. Ibrahim, and Michael I. Love. Heavy-tailed prior distributions for sequence count data: Removing the noise and preserving large differences. *Bioinformatics*, 2019.

[90] Wanding Zhou, Peter W. Laird, and Hui Shen. Comprehensive characterization, annotation and innovative use of infinium dna methylation beadchip probes. *Nucleic acids research*, 2017.

[91] Pan Du, Xiao Zhang, Chiang Ching Huang, Nadereh Jafari, Warren A. Kibbe, Lifang Hou, and Simon M. Lin. Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, 2010.

[92] Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. Limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Research*, 2015.

[93] Gordon K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 2004.

[94] Maarten van Iterson, Erik W. van Zwet, Bastiaan T. Heijmans, Peter A.C. 't Hoen, Joyce van Meurs, Rick Jansen, Lude Franke, Dorret I. Boomsma, René Pool, Jenny van Dongen, Jouke J. Hottenga, Marleen M.J. van Greevenbroek, Coen D.A. Stehouwer, Carla J.H. van der Kallen, Casper G. Schalkwijk, Cisca Wijmenga, Sasha Zhernakova, Ettje F. Tigchelaar, P. Eline Slagboom, Marian Beekman, Joris Deelen, Diana van Heemst, Jan H. Veldink, Leonard H. van den Berg, Cornelia M. van Duijn, Bert A. Hofman, Aaron Isaacs, André G. Uitterlinden, P. Mila Jhamai, Michael Verbiest, H. Eka D. Suchiman, Marijn Verkerk, Ruud van der Breggen, Jeroen van Rooij, Nico Lakenberg, Hailiang Mei, Michiel van Galen, Jan Bot, Dasha V. Zhernakova, Peter van 't Hof, Patrick Deelen, Irene Nooren, Matthijs Moed, Martijn Vermaat, René Luijk, Marc Jan Bonder, Freerk van Dijk, Wibowo Arindrarto, Szymon M. Kielbasa, Morris A. Swertz, and Peter Bram 't Hoen. Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution. *Genome Biology*, 2017.

[95] Brent S. Pedersen, David A. Schwartz, Ivana V. Yang, and Katerina J. Kechris. Comb-p: Software for combining, analyzing, grouping and correcting spatially correlated p-values. *Bioinformatics*, 2012.

[96] Matteo Bersanelli, Ettore Mosca, Daniel Remondini, Gastone Castellani, and Luciano Milanesi. Network diffusion-based analysis of high-throughput data for the detection of differentially enriched modules. *Scientific Reports*, 2016.

[97] Linton C Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977.

[98] David Eppstein and Darren Strash. Listing all maximal cliques in large sparse real-world graphs. In *International Symposium on Experimental Algorithms*, pages 364–375. Springer, 2011.

[99] Denuja Karunakaran, A Brianne Thrush, My-Anh Nguyen, Laura Richards, Michele Geoffrion, Ragunath Singaravelu, Eleni Ramphos, Prakriti Shangari, Mireille Ouimet, John P Pezacki, et al. Macrophage mitochondrial energy status regulates cholesterol efflux and is enhanced by anti-mir33 in atherosclerosis. *Circulation research*, 117(3):266–278, 2015.

[100] Charles E Mordaunt, Dorothy A Kieffer, Noreene M Shibata, Anna Członkowska, Tomasz Litwin, Karl-Heinz Weiss, Yihui Zhu, Christopher L Bowlus, Souvik Sarkar, Stewart Cooper, et al. Epigenomic signatures in liver and blood of wilson disease patients include hypermethylation of liver-specific enhancers. *Epigenetics & chromatin*, 12(1):1–16, 2019.

[101] Elisenda Climent, Ana M Bea, David Benaiges, Ángel Brea-Hernando, Xavier Pintó, Manuel Suárez-Tembra, Verónica Perea, Núria Plana, Francisco Blanco-Vaca, and Juan Pedro-Botet. Ldl cholesterol reduction variability with different types and doses of statins in monotherapy or combined with ezetimibe. results from the spanish arteriosclerosis society dyslipidaemia registry. *Cardiovascular Drugs and Therapy*, 36(2):301–308, 2022.

[102] Jairo Aldana-Bitar, Jeff Moore, and Matthew J Budoff. Ldl receptor and pathogen processes: Functions beyond normal lipids. *Journal of Clinical Lipidology*, 2021.

[103] Karina González-Aldaco, Luis A Torres-Reyes, Claudia Ojeda-Granados, Alexis José-Ábrego, Nora A Fierro, and Sonia Román. Immunometabolic effect of cholesterol in hepatitis c infection: implications in clinical management and antiviral therapy. *Annals of Hepatology*, 17(6):908–919, 2018.

[104] Daniela Stols-Gonçalves, G Kees Hovingh, Max Nieuwdorp, and Adriaan G Holleboom. Nafld and atherosclerosis: two sides of the same dysmetabolic coin? *Trends in Endocrinology & Metabolism*, 30(12):891–902, 2019.

[105] Camilla Pramfalk, Osman Ahmed, Matteo Pedrelli, Mirko E Minniti, Serge Luquet, Raphaël GP Denis, Maria Olin, Jennifer Härdfeldt, Lise-Lotte Vedin, Knut R Steffensen, et al. Soat2 ties cholesterol metabolism to β-oxidation and glucose tolerance in male mice. *Journal of Internal Medicine*, 2022.

[106] Andromachi Kotsafti, Fabio Farinati, Romilda Cardin, Patrizia Burra, and Marina Bortolami. Bax inhibitor-1 down-regulation in the progression of chronic liver diseases. *BMC gastroenterology*, 10(1):1–8, 2010.

[107] Margaret A Hamburg and Francis S Collins. The path to personalized medicine. *New England Journal of Medicine*, 363(4):301–304, 2010.

[108] Kewal K Jain. Personalized medicine. *Current opinion in molecular therapeutics*, 4(6):548–558, 2002.

[109] Patrick Danaher, Pei Wang, and Daniela M Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):373–397, 2014.

[110] Max Bylesjö, Daniel Eriksson, Miyako Kusano, Thomas Moritz, and Johan Trygg. Data integration in plant biology: the o2pls method for combined modeling of transcript and metabolite data. *The Plant Journal*, 52(6):1181–1191, 2007.

[111] Nimrod Rappoport and Ron Shamir. Nemo: cancer subtyping by integration of partial multi-omic data. *Bioinformatics*, 35(18):3348–3356, 2019.

[112] Charles J Vaske, Stephen C Benz, J Zachary Sanborn, Dent Earl, Christopher Szeto, Jingchun Zhu, David Haussler, and Joshua M Stuart. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using paradigm. *Bioinformatics*, 26(12):i237–i245, 2010.

[113] Elizabeth Chlipala, Christine M. Bendzinski, Kevin Chu, Joshua I. Johnson, Miles Brous, Karen Copeland, and Brad Bolon. Optical density-based image analysis method for the evaluation of hematoxylin and eosin staining precision. *https://doi.org/10.1080/01478885.2019.1708611*, 43(1):29–37, jan 2020.

[114] Shikha Roy, Rakesh Kumar, Vaibhav Mittal, and Dinesh Gupta. Classification models for invasive ductal carcinoma progression, based on gene expression data-trained supervised machine learning. *Scientific reports*, 10(1):1–15, 2020.

[115] Mahul B Amin, Frederick L Greene, Stephen B Edge, Carolyn C Compton, Jeffrey E Gershenwald, Robert K Brookland, Laura Meyer, Donna M Gress, David R Byrd, and David P Winchester. The eighth edition ajcc cancer staging manual: continuing to build a bridge from a population-based to a more “personalized” approach to cancer staging. *CA: a cancer journal for clinicians*, 67(2):93–99, 2017.

[116] Yu Fu, Alexander W Jung, Ramon Viñas Torne, Santiago Gonzalez, Harald Vöhringer, Artem Shmatko, Lucy R Yates, Mercedes Jimenez-Linan, Luiza Moore, and Moritz Gerstung. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nature Cancer*, 1(8):800–810, 2020.

[117] Donna Maglott, Jim Ostell, Kim D Pruitt, and Tatiana Tatusova. Entrez gene: gene-centered information at ncbi. *Nucleic acids research*, 33(suppl_1):D54–D58, 2005.

[118] R. Devon Hjelm, Karan Grewal, Phil Bachman, Alex Fedorov, Adam Trischler, Samuel Lavoie-Marchildon, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[119] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization. *Advances in Neural Information Processing Systems*, pages 271–279, jun 2016.

[120] Ismael Lemhadri, Feng Ruan, and Rob Tibshirani. Lassonet: Neural networks with feature sparsity. In *International Conference on Artificial Intelligence and Statistics*, pages 10–18. PMLR, 2021.

[121] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019.

[122] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021.

[123] Sijia Huang, Kumardeep Chaudhary, and Lana X Garmire. More is better: recent progress in multi-omics data integration methods. *Frontiers in genetics*, 8:84, 2017.

[124] Zhaoxiang Cai, Rebecca C Poulos, Jia Liu, and Qing Zhong. Machine learning for multi-omics data integration in cancer. *Iscience*, page 103798, 2022.

[125] Nikola Simidjievski, Cristian Bodnar, Ifrah Tariq, Paul Scherer, Helena Andres Terre, Zohreh Shams, Mateja Jamnik, and Pietro Liò. Variational autoencoders for cancer data integration: design principles and computational practice. *Frontiers in genetics*, 10:1205, 2019.

[126] Ioana Bica, Petar Velickovic, Hui Xiao, and P Liò. Multi-omics data integration using cross-modal neural networks. In *ESANN*, 2018.

[127] Damiano Fantini, Vania Vidimar, Yanni Yu, Salvatore Condello, and Joshua J Meeks. Mutsignatures: an r package for extraction and analysis of cancer mutational signatures. *Scientific reports*, 10(1):1–12, 2020.

[128] Giulia Tini, Luca Marchetti, Corrado Priami, and Marie-Pier Scott-Boyer. Multi-omics integration—a comparison of unsupervised clustering methodologies. *Briefings in bioinformatics*, 20(4):1269–1279, 2019.

[129] Rafael A. Irizarry, Bridget Hobbs, Francois Collin, Yasmin D. Beazer-Barclay, Kristen J. Antonellis, Uwe Scherf, and Terence P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, apr 2003.

[130] Jeremy D. Silver, Matthew E. Ritchie, and Gordon K. Smyth. Microarray background correction: maximum likelihood estimation for the normal–exponential convolution. *Biostatistics (Oxford, England)*, 10(2):352, apr 2009.

[131] B. M. Bolstad, R. A. Irizarry, M. Åstrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, jan 2003.

[132] Frederick Mosteller and John W. (John Wilder) Tukey. Data analysis and regression : a second course in statistics. page 588, 1977.

[133] B. M. Bolstad, F. Collin, J. Brettschneider, K. Simpson, L. Cope, R. A. Irizarry, and T.P. Speed. Quality Assessment of Affymetrix GeneChip Data. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, pages 33–47, dec 2005.

[134] Johnson WE, Li C, and Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics (Oxford, England)*, 8(1):118–127, jan 2007.

[135] Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLOS Genetics*, 3(9):1–12, 2007.

[136] Using control genes to correct for unwanted variation in microarray data. *Biostatistics (Oxford, England)*, 13(3):539, jul 2012.

[137] Stefano Calza, Davide Valentini, and Yudi Pawitan. Normalization of oligonucleotide arrays based on the least-variant set of genes. *BMC Bioinformatics*, 9(1):1–11, mar 2008.

[138] Iven Van Mechelen, Hans-Hermann Bock, and Paul De Boeck. Two-mode clustering methods: astructuredoverview. *Statistical Methods in Medical Research*, 13(5):363–394, 2004. PMID: 15516031.

[139] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[140] Melissa J Azur, Elizabeth A Stuart, Constantine Frangakis, and Philip J Leaf. Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research*, 20(1):40–49, 2011.

[141] Geoffrey J McLachlan, Sharon X Lee, and Suren I Rathnayake. Finite mixture models. *Annual review of statistics and its application*, 6:355–378, 2019.

[142] David M Blei, Thomas L Griffiths, and Michael I Jordan. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM (JACM)*, 57(2):1–30, 2010.

[143] Yee Whye Teh et al. Dirichlet process. *Encyclopedia of machine learning*, 1063:280–287, 2010.

[144] Yee Teh, Michael Jordan, Matthew Beal, and David Blei. Sharing clusters among related groups: Hierarchical dirichlet processes. *Advances in neural information processing systems*, 17, 2004.

[145] Agner Fog. Calculation methods for wallenius' noncentral hypergeometric distribution. *Communications in Statistics—Simulation and Computation®*, 37(2):258–273, 2008.

[146] David G Kleinbaum, Mitchel Klein, et al. *Survival analysis: a self-learning text*, volume 3. Springer, 2012.

[147] Dhananjay Kumar and Bengt Klefsjö. Proportional hazards model: a review. *Reliability Engineering & System Safety*, 44(2):177–188, 1994.

[148] Ross L. Prentice. *Introduction to Cox (1972) Regression Models and Life-Tables*, pages 519–526. Springer New York, New York, NY, 1992.

[149] Haiqun Lin and Daniel Zelterman. Modeling survival data: extending the cox model, 2002.

[150] Jason A Reuter, Damek V Spacek, and Michael P Snyder. High-throughput sequencing technologies. *Molecular cell*, 58(4):586–597, 2015.

[151] Samuel Marguerat and Jürg Bähler. Rna-seq: from technology to biology. *Cellular and molecular life sciences*, 67(4):569–579, 2010.

[152] Yingdong Zhao, Ming-Chung Li, Mariam M Konaté, Li Chen, Biswajit Das, Chris Karlovich, P Mickey Williams, Yvonne A Evrard, James H Doroshow, and Lisa M McShane. Tpm, fpkm, or normalized counts? a comparative study of quantification measures for the analysis of rna-seq data from the nci patient-derived models repository. *Journal of translational medicine*, 19(1):1–15, 2021.

[153] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Nature Precedings*, pages 1–1, 2010.

[154] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):1–21, 2014.

[155] Brent S Pedersen, David A Schwartz, Ivana V Yang, and Katerina J Kechris. Comb-p: software for combining, analyzing, grouping and correcting spatially correlated p-values. *Bioinformatics*, 28(22):2986–2988, 2012.

[156] Nicholas A Heard and Patrick Rubin-Delanchy. Choosing between methods of combining-values. *Biometrika*, 105(1):239–246, 2018.

[157] Katerina J Kechris, Brian Biehs, and Thomas B Kornberg. Generalizing moving averages for tiling arrays using combined p-value statistics. *Statistical applications in genetics and molecular biology*, 9(1), 2010.

[158] Dmitri V Zaykin, Lev A Zhivotovsky, Peter H Westfall, and Bruce S Weir. Truncated product method for combining p-values. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 22(2):170–185, 2002.

[159] Mark Newman. *Networks*. Oxford university press, 2018.

[160] Naoki Masuda, Mason A Porter, and Renaud Lambiotte. Random walks and diffusion on networks. *Physics reports*, 716:1–58, 2017.

[161] Matteo Bersanelli, Ettore Mosca, Daniel Remondini, Gastone Castellani, and Luciano Milanesi. Network diffusion-based analysis of high-throughput data for the detection of differentially enriched modules. *Scientific Reports*, 6(1):1–12, 2016.

[162] Matteo Bersanelli, Ettore Mosca, Daniel Remondini, Enrico Giampieri, Claudia Sala, Gastone Castellani, and Luciano Milanesi. Methods for the integration of multi-omics data: mathematical aspects. *BMC bioinformatics*, 17(2):167–177, 2016.

[163] Simon Haykin. *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1994.

[164] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2323, 1998.

[165] Christopher M Bishop. *Pattern Recognition and Machine Learning*, volume 16. 2007.

[166] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[167] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[168] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[169] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[170] C. E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423, jul 1948.

[171] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, mar 1951.

[172] Ralph Linsker. Self-Organization in a Perceptual Network. *Computer*, 21(03):105–117, mar 1988.

[173] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159, 1995.

[174] S. M. Ali and S. D. Silvey. A General Class of Coefficients of Divergence of One Distribution from Another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28(1):131–142, jan 1966.

[175] XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, nov 2010.

[176] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[177] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

[178] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2020.

[179] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

[180] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.