THE UNIVERSITY
of EDINBURGH

Stratification of patient subgroups using high-dimensional and time-series observations.

Lucile P. A. Neyton

Doctor of Philosophy

The University of Edinburgh

2020

Declaration

I declare that this thesis has been composed solely by myself and that it has not been submitted, in whole or in part, in any previous application for a degree. Except where states otherwise by reference or acknowledgment, the work presented is entirely my own.

Lucile P. A. Neyton

February 2020

Abstract

Precision medicine and patient stratification are expanding as a result of innovations in high-throughput technologies applied to clinical medicine. Stratification can explain differences in disease trajectories and outcomes in heterogeneous cohorts. Thus, approaches employed for patient treatment can be tailored by taking into account individual variabilities and specificities.

This thesis focuses on clustering approaches and how they can be applied to both single time points and time-series high-dimensional data for the identification of disease subtypes defined by distinct mechanisms, also called endotypes, in complex and/or heterogeneous diseases. Multiple carefully selected clustering strategies were compared to highlight which would produce the most relevant stratification in terms of mathematical robustness and biological meaning, both of which quantified using standardised methods.

More specifically, this strategy was applied to time-series multi-omics data from a cohort of patients with acute pancreatitis, an inflammatory disease of the pancreas. Using this high-dimensional multi-omics data as well as routine lab and clinical measurements, the cohort was stratified into four subgroups.

Findings from the analysis of acute pancreatitis data showed that two of the four subgroups could be detected in another syndrome, acute respiratory distress syndrome, suggesting that inflammatory signatures are comparable between diseases.

With the aim of applying these principles to other diseases and using preliminary results from other studies suggesting that relevant subgroups might be highlighted, data from inflammatory bowel disease and Parkinson's disease cohorts was analysed. Results from our analyses confirmed that disease knowledge could be gained using this approach.

Work from this thesis provides novel approaches for the application and evaluation of stratification methods. Furthermore, results may constitute a basis for the development of tailored treatment approaches for acute pancreatitis, acute respiratory distress syndrome, inflammatory bowel disease and Parkinson's disease. Also, the observation of commonalities between distinct inflammatory diseases will broaden the perspectives when analysing disease data and more specifically, in biomarker discovery and drug development processes.

Lay Summary

Precision medicine is the tailoring of medical care to subgroups of patients based on each individual's characteristics. Taking into account variability within cohorts of patients can improve the prevention, understanding, and treatment of diseases. For example, blood samples can be collected and used to identify and quantify molecules of interest which can then be used to separate individuals into different subgroups.

For this project, different methods to highlight patient subgroups using many measurements were selected. Once the results were obtained, we compared them using mathematical and biological metrics to quantify the success of each one of them and select the most relevant one.

Acute pancreatitis is a disease of the pancreas, a vital organ producing and releasing molecules involved in the digestion and the regulation of blood sugar. There are many causes of acute pancreatitis. However, disease mechanisms are still poorly understood. During acute pancreatitis, the pancreas becomes inflamed and this inflammation can spread to other vital organs such as the kidneys or lungs which can pose threats to the life of affected individuals.

Acute pancreatitis patients evolve in many different ways which makes it very hard to predict if a patient will be more at risk of developing organ failure than others. No specific treatment exists to this date.

In this PhD project, we highlighted four subgroups of interest among a cohort of patients with acute pancreatitis, using a range of different measurements. To do so we grouped individuals to obtain subgroups of patients as similar as possible to each other and as different as possible from individuals in other groups. Measurements important to each one of the four identified acute pancreatitis subgroups were compared to those used to tell apart two subgroups of another life-threatening syndrome, acute respiratory distress syndrome, characterised by respiratory failure. We found common features

and concluded that there were similarities between individuals affected with different types of critical illnesses. We performed similar analyses on data collected on inflammatory bowel disease patients, a term encompassing two diseases characterised by gut inflammation, and Parkinson's disease patients, a neurodegenerative disease. We aimed to try to explain why patients with the same disease could be so different and classify them into subgroups which would be relevant for the understanding and treatment of the disease.

We expect this to be important for future care approaches and more specifically for the development of treatments tailored for subgroups of individuals across different diseases.

Acknowledgements

First and foremost, I would like to express my gratitude to my mother who always supported and advised me. Thank you for being the role model anyone could have hoped for.

Damian and Kenny, you have been absolutely perfect, and I feel incredibly lucky to have had such a great supervision team most PhD students would have wished for.

I would like to thank the members of my thesis committee, Christ Ponting and Dave Robertson who provided great guidance.

To the administrative team, Susan and Kate, you have been of great help too and have always been available.

To the Baillie lab team at the Roslin Institute, thank you for your daily support. More specifically, thank you Sara for being always available and supportive. Thank you also to Tim and Evangelos for your great conversations.

To Iain, Philippe, Alex, Nikolina and the computational biology team at GSK, thank you for making a placement possible, for welcoming me for a few weeks and for your kindness and advice.

Also, I would like to thank Corentin for his support, for being an amazing proof-reader (and more specifically for making sure the way the thesis was written did not hamper understanding!) and exploring the incredible beauty of Scotland with me.

Julie, you have always been a great and supportive listener. Thank you for your advice.

Contents

Figures

Tables

Abbreviations

AGNES	AGglomerative NESting
ANOVA	ANalysis Of Variance
AP	Acute Pancreatitis
ARDS	Acute Respiratory Distress Syndrome
AUC	Area-Under-the-Curve
BMI	Body Mass Index
BMU	Best Matching Unit
CAGE	Cap Analysis of Gene Expression
CD	Crohn's Disease
cDNA	Complementary DNA
CF	Cystic Fibrosis
CSF	CerebroSpinal Fluid
CT	Computerised Tomography
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DIANA	DIvisive ANAlysis
DIGE	Differential In-Gel Electrophoresis
eGFR	Estimated Glomerular Filtration Rate
EHR	Electronic Health Record
ELBO	Evidence Lower BOund
EM	Expectation-Maximisation
FANTOM5	Functional ANnoTation Of the Mammalian genome 5
FDR	False Discovery Rate
FPKM	Fragments Per Kilobase Million
FPR	False Positive Rate

GEO	Gene Expression Omnibus
GMM	Gaussian Mixture Models
GO	Gene Ontology
GWAS	Genome-Wide Association Study
HDU	High-Dependency Unit
HGNC	HUGO Gene Nomenclature Committee
HUGO	HUman Genome Organisation
IBD	Inflammatory Bowel Disease
ICU	Intensive Care Unit
IGP	In-Group Proportion
IMOFAP	Inflammation, Metabolism and Organ Failure in AP
IQR	InterQuartile Range
JI	Jaccard Index
KAPVAL	Kynurenine pathway in AP VALidation
KEGG	Kyoto Encyclopedia of Genes and Genomes
LD	Linkage Disequilibrium
MDS	Movement Disorder Society
MDS-UPDRS	MDS – Unified Parkinson's Disease Rating Scale
MJFF	Michael J. Fox Foundation
mMODS	Modified Multiple Organ Dysfunction Score
MODS	Multiple Organ Dysfunction Syndrome
MRI	Magnetic Resonance Imaging
mRNA	Messenger RNA
MS	Mass Spectrometry
NCBI	National Centre for Biotechnology Information
NDA	Network Density Analysis

NGS	Next Generation Sequencing
NMI	Normalised Mutual Information
NMR	Nuclear Magnetic Resonance
OPTICS	Ordering Points To Identify the Clustering Structure
PAM	Partitioning Around Medoids
PC	Principal Component
PCA	Principal Component Analysis
PD	Parkinson’s Disease
PET	Positron Emission Tomography
PLS-DA	Partial Least Squares – Discriminant Analysis
PPMI	Parkinson’s Progression Markers Initiative
PRS	Polygenic Risk Score
QTL	Quantitative Trait Loci
RPKM	Reads Per Kilobase Million
RPLC	Reversed-Phase Liquid Chromatography
rRNA	Ribosomal RNA
SD	Standard Deviation
SIRS	Systemic Inflammatory Response Syndrome
SNP	Single Nucleotide Polymorphism
SOM	Self-Organising Map
TPM	Transcripts Per Kilobase Million
UC	Ulcerative Colitis
UPGMA	Unweighted Pair Group Method with Arithmetic mean
UPLC	Ultra Performance Liquid Chromatography
VIP	Variable Importance in Projection
VST	Variance Stabilising Transformation

1. Chapter 1 – Introduction

This introductory chapter is organised around five main sections. The first section will lay out the project's foundation and give some background information relevant to the field of precision medicine and the different projects undertaken as part of this PhD. In the second section, the motivation behind the project and how it currently fits into the field will be discussed. In the same section, a brief summary of this project's contribution to the field will be presented. The general strategy behind the project will be described into the third section of this introduction. Finally, the thesis' structure and contributions to knowledge will be outlined.

1.1 Focus and background

1.1.1 An introduction to precision medicine

Precision medicine is an approach aimed at the prevention, diagnosis and treatment of diseases that uses information and measurements specific to individuals.

Precision medicine exists in contrast to the "one-size-fits-all" strategy. Traditionally, medicine takes a single course for patients affected by a disease without necessarily taking an individual's specific variation into account. Given the analysis results of surgical biopsies, histology and imaging for example, a diagnosis was made and, based on the disease trajectory (the evolution of a patient as the disease progresses) of the average patient, a care strategy was chosen.

The term precision medicine recently gained in popularity and especially from 2015. It has since been widely used (Figure 1.1).

Potential factors which could have driven this gained popularity compared to the personalised medicine term are the publication of the report "Towards Precision Medicine" which was published in 2011[1] and the launch of the Precision Medicine Initiative in 2015. Usually this term is preferred over "personalised medicine" because the later suggests that one treatment per patient is available while in truth treatments are often targeting subgroups of individuals. However, they are sometimes used interchangeably in the literature.

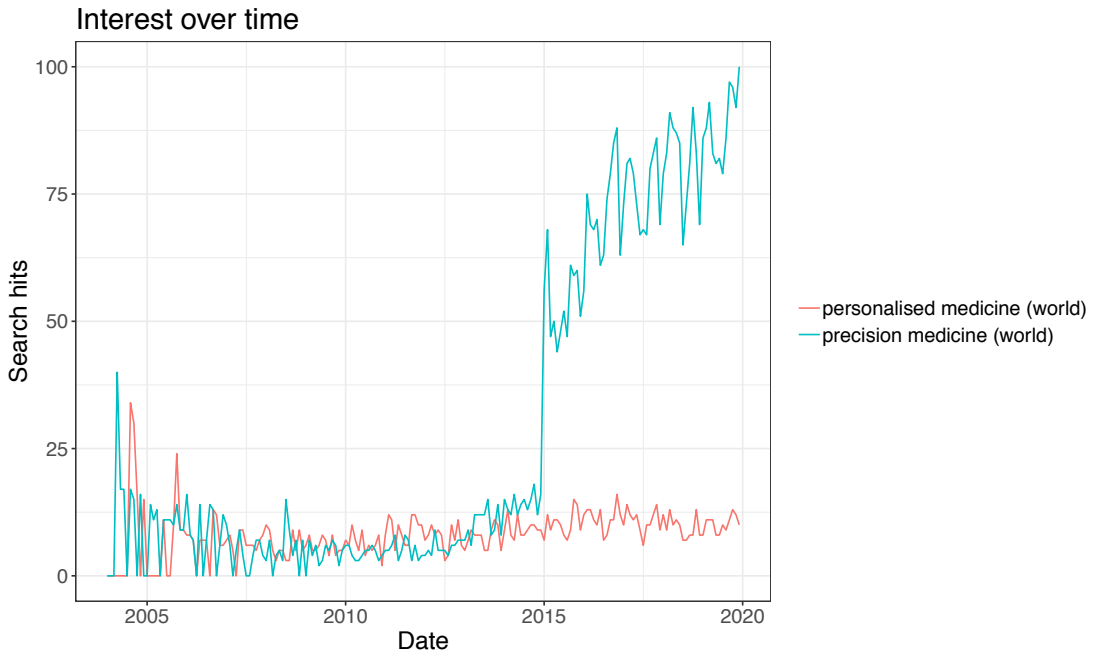


Figure 1.1 – Trend plot for the 'personalised medicine' and 'precision medicine' terms between January 2004 and November 2019 (inclusive). The y axis represents worldwide popularity (relative to the number of hits) and values are scaled between 0 and the maximum popularity for the term. Data was fetched from Google Trends (https://trends.google.com/trends/) using the R package gtrendsR[2].

A first step towards precision medicine is stratification, consisting of defining subgroups of individuals among a cohort of patients affected by a disease. These subgroups can be referred to as endotypes.

For a given disease, endotypes are subtypes which are characterised by a distinct functional or pathobiological mechanism[3].

For each one of these subgroups, specific strategies can be applied to diagnose and/or treat affected individuals more effectively. Furthermore, once defined, such information can be applied to disease prevention.

The term “endotype” can be compared to “phenotype”, the later referring to the set of observable characteristics of an individual. A given phenotype usually presents no direct relationship to underlying mechanisms of a disease.

Specifically, the term “precision medicine” refers to the integration of molecular profiles, such as genetic information, into the prevention, diagnostic and treatment strategies for a disease in order to provide patients with a tailored approach[4–7].

In summary, precision medicine is defined as a strategy taking into account the uniqueness of individuals, as characterised by observations and measurements for each patient, and allowing to choose a suitable treatment course accordingly.

The field has been further enhanced by the development of technologies allowing researchers to measure many variables for a single individual simultaneously, at a given point in his/her disease trajectory. These measurements can be integrated to gain a better understanding of disease processes and identify individuals or subgroups with similar trajectories. Moreover, the cost of generating such measurements has greatly reduced over time and is even reducing faster than Moore’s law[8]. Moore’s law states that the number of transistors per chip doubles every two years, thus halving its price. Costs of DNA sequencing used to be compared to the slope of Moore’s law. However, since around 2007[9], when next generation sequencing became available, the price of sequencing has seen an even steeper decrease. Also, the sharing of such datasets has been facilitated by dedicated online platforms. This has allowed the integration of such data by many

research groups with diverse interests, permitting to maximise scientific discovery.

1.1.2 How can stratification help inform decisions in precision medicine?

As illustrated in the previous section, stratification is a cornerstone of precision medicine. To define subgroups, either presenting distinct disease mechanisms or treatment responses, many patient measurements can be used such as biosensors measurements, data from images, omics data (corresponding to the measurement of different pools of molecules for an organism) and electronic health record (EHR) data.

This is especially true for heterogenous diseases, which can involve distinct underlying mechanisms and consequently have different aetiologies, evolutions and outcomes, and are thus characterised by multiple disease trajectories. Indeed, even though patients present a similar syndrome, or collection of symptoms, they might in reality be better described, and understood, as subtypes, which are yet to be defined. Defining subtypes for these diseases could not only help in understanding the mechanisms behind them but also in identifying measurements of interest which could be considered as candidate drug targets for the design of new treatments. One widespread issue is that some treatments commonly given nowadays only benefit a small proportion of affected individuals[10]. The identification of individuals with a given disease likely to benefit from a given treatment would greatly benefit patients and might be permitted by disease stratification.

However, the choice of measurements to use to define subgroups might not be obvious. This is not a trivial choice and results will depend on this decision. To understand the complex interplay between the different types of data and how it relates to potential subtypes it is advantageous to measure as many things as possible. However, because of cost constraints this might not be possible for most projects. Depending on the question asked, measuring too

many things might not be relevant as well. Once all measurements are generated, the ones relevant to the question will have to be identified. Detecting signal from noise might pose a challenge as a subgrouping could appear solely because of noise rather than a relevant stratification.

Moreover, there are many ways to stratify a cohort of patients and only some of them will be relevant to the disease studied. For example, in acute pancreatitis, a highly heterogenous inflammatory condition affecting the pancreas, patients can be stratified according to the severity for example. However, this does not correlate with outcome[11,12]. For this reason, collecting more measurements, and more specifically omics measurements, might help in stratifying the disease into relevant subgroups which could help in understanding and treating the condition.

1.1.3 Early applications of precision medicine

Even though precision medicine is a relatively recent term and has lately been gaining popularity, its principles are not new and have been commonly applied in medicine.

Indeed, many early examples of stratification exist. Perhaps one of the most known examples is the stratification according to blood type, discovered in 1901 and which is still used today. By identifying the blood types of the donor and receiver and matching them accordingly, successful transfusions can be performed[13].

Another early example includes diabetes mellitus, a disease in which the ability to produce or respond to insulin, a hormone involved in the control of blood sugar levels, is affected. Diabetes was first described as two underlying diseases[14] in the 1930s, based on insulin sensitivity and which were later referred to as type 1 and type 2 diabetes[15]. In this work, building upon observations made previously and noting the varying sensitivity of individuals to insulin, the author describes how individuals with diabetes reacted

differently to the injection of insulin, one subgroup showing an immediate decrease in blood sugar and the other showing little or no effect. The author concludes that one of these subtypes is not due to a lack of insulin but rather to an altered sensitivity to insulin (later referred to as type 2 diabetes). The implications of this distinction were great as it led to more discovery in the field and the design of management therapies specific to each one of the two subtypes.

Another well-known example is the discovery of the HER2 target in metastatic breast cancers, which was validated in the 1990s[16]. This type of metastatic breast cancer is characterised by the overexpression of the HER2 gene (human epidermal growth factor receptor 2) and the resulting protein is used as a marker (associated measurable trait) of efficacy for a drug, trastuzumab. This drug, when given to HER2-positive breast cancer patients, binds to the HER2 receptor, significantly reducing mortality and increasing remission[17].

More recently, in the early 2000s, mutations in the BRAF gene were identified in malignant melanoma[18] providing a new therapeutic target which has, later on, been used for its treatment. For example, vemurafenib, a BRAF inhibitor has been shown to improve survival in affected individuals[19].

1.1.4 Emerging applications of precision medicine

There are many ways to integrate patient data, which can come from many different sources[20,21], in a precision medicine setting. Some of them are described in this section, which is in no way exhaustive, especially in this fast-moving field.

1.1.4.1 Biosensor-based precision medicine applications

Biosensors, whether wearable or implantable, can allow continuous measurements to be recorded. This is especially important in the medical context in which individuals are seen by health professionals on a discrete

basis. These sensors can fill the gaps between visits by continuously monitoring a patient. This kind of data can be used for monitoring, prevention, diagnosis and treatment[22].

In cases where continuous observation and treatment are required, such as in chronic illnesses, biosensors can be coupled with treatment delivery systems to improve compliance and minimise side effects[23] .

A well-known example is for individuals affected with diabetes mellitus. Implantable biosensors allow, through the real-time monitoring of glucose blood level, to predict and prevent hypoglycaemic (low blood sugar) and hyperglycaemic (high blood sugar) episodes. According to the values obtained, insulin can be automatically administered according to the condition of each patient.

1.1.4.2 Imaging data and precision medicine applications

Imaging data and more specifically Computerised Tomography (CT), Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET) images can be used for precision medicine applications as well.

Quantitative measurements such as size, shape and homogeneity can be extracted from these images and can be used to characterise variability between patients. These measurements are referred to as radiomics and can be correlated with other variables. For example, it has been shown that radiomics from MRI images could be used to predict endotype classification for invasive breast cancers[24].

1.1.4.3 Omics data in precision medicine

Blood or biopsy samples can be used to qualify and quantify molecules of interest. For example, DNA, RNA, proteins and metabolites (small molecules) can be identified and measured on these samples. The study of these

molecules is usually referred to as omics. Respectively, genomics, transcriptomics, proteomics and metabolomics.

All these measurements are not currently used at the same level in precision medicine and some still need maturing, such as improved acquisition techniques or specific bioinformatics tools, to allow for them to be fully integrated to clinical practice.

Genomics, or the study of genomes, involves DNA sequencing, and is an example of well-integrated data in the field of precision medicine.

For example, cystic fibrosis (CF) is a disease characterised by a build-up of sticky mucus in the lungs and digestive system. CF is an inherited disease which is the result of a defective gene[25] (cystic fibrosis transmembrane conductance regulator). This gene encodes for a protein and when mutations occur, they can lead to the production of an abnormal protein which will then cause a defect in epithelial ion transport (transport of charged atom or molecule between compartments). Across patients with cystic fibrosis there is not a single possible mutation and stratifying individuals based on these mutations has helped in designing suitable drugs[26,27]. For example, carriers of missense mutations (characterised by an amino acid change), resulting in the ion channel being closed more often, can be treated using compounds called potentiators.

Recently, transcriptomics, or the study of gene expression, has been a great research interest and especially in cancer research.

For example, the expression of eight genes in the Interferon/Stat1 pathway, measured using microarrays, has been linked to outcome in glioblastoma, a type of tumour occurring in the brain and spinal cord[28]. More specifically, upregulation of genes from this pathway predicts a poor outcome in glioblastoma patients.

Another example of the use of gene expression for precision medicine is colorectal cancer. Indeed, gene expression of gene groups measured using

microarrays, has been shown to be predictive of outcome after surgery[29] for patients with this type of cancer.

The study of protein data, or proteomics, has also been well researched in the precision medicine context. For example, drug response/sensitivity was successfully predicted in cancer patients, using two cell line panels, based on a range of different protein abundances[30].

Finally, metabolomics data corresponding to the profiling of metabolites (or small molecules), while perhaps being a least mature field[31], is also promising. For example, the levels of two metabolites, spermine and citrate, were associated with the aggressiveness of prostate cancer[32]. Lower levels of these two metabolites were associated with more aggressive forms of cancer.

1.1.4.4 Electronic health records data in precision medicine

Another resource consists of EHRs containing a breadth of data collected by health professionals and usually covering years. EHRs can contain laboratory tests, clinical measurements and prescription data for example. A study showed that a model could be used to classify drugs into harmful or safe categories for pregnant patients given data from EHRs[33].

Clinical data especially is usually a cornerstone of any precision medicine analysis as it can be linked to molecular data and usually gives indication of a patient outcome.

1.1.5 Data integration for precision medicine

Each one of the biomedical data types reported in the previous section yields a huge amount of measurements on its own but these can be combined to gain additional knowledge of studied diseases and/or biological processes. However, this is not trivial due to the number of variables, which is usually very high, and the number of samples, usually much lower. Moreover, the different

data types are heterogeneous, making this task even more complex. Regardless, this has been shown to be possible and relevant for precision medicine[34]. To deal with this amount of data and take full advantage of it, computational structure and tools are essential, as well as specialist knowledge of data and bioinformatics algorithms.

1.1.6 Precision medicine, potential pitfalls

The term precision medicine, in its broad sense, aims at taking into account variability using a range of different measurements. While stratification, consisting of separating individuals into meaningful subgroups, can help in understanding a disease, it can raise the concern that not all patients will be part of one of the subgroups or will be part of a subgroup that will not benefit from specific care or treatment[35].

Another potential challenge lies in the fact that there might not be enough funding, storage, technologies or bioinformatic tools to take full advantage of the precision medicine approach.

1.2 Motivation

Acute pancreatitis (AP), an inflammatory disease of the pancreas, has an incidence of 34 per 100,000 persons-years[36] and is the most common gastrointestinal cause for admission to hospital emergency services. In 1 in 4 affected individuals, AP will lead to multi-organ dysfunction syndrome (MODS)[37] and among those, 1 in 5 will die[38]. Although acute pancreatitis' aetiologies are known, it remains unclear what will cause a patient to evolve in a certain way and how this will relate to MODS and death. To this date, no specific therapy targeting AP is available to patients. Thus, great benefit could be obtained from the study of AP's heterogeneity.

In this project, the first precision medicine study we know of focusing solely on acute pancreatitis patients, we hypothesised that endotypes (also referred to

as condition subtypes) existed and could be defined using a range of molecular measurements. Moreover, we believe that describing such endotypes would be relevant for the field, disease knowledge and would open new therapeutic avenues.

We compared acute respiratory distress syndrome (ARDS) endotypes[39] with AP endotypes identified in this thesis using clinical and blood measurements and found similarities.

Furthermore, a similar approach could be applied to other heterogeneous diseases. As such, clinical and genetic variants data was used to study heterogeneity in inflammatory bowel disease (IBD). Finally, using imaging-derived, DNA methylation, gene expression and clinical measurements, we studied a cohort of Parkinson's disease (PD) patients and aimed at correlating this data to the carrier status for a PD-linked mutation.

1.3 Strategy

To extract disease endotypes for AP, we performed a clustering analysis, a term encompassing different techniques allowing to divide a cohort of samples into smaller sets. This analysis was done using a cohort of AP-affected patients referred to as IMOFAP[40] (Inflammation, Metabolism and Organ Failure in AP) for which different measurements were available. Samples for this cohort were collected for 79 individuals at the Royal Infirmary of Edinburgh between September and December 2013. We then systematically assessed different clustering solutions and selected the one which was identified as being the most relevant, using a priori criteria, in terms of stability and biological relevance. We then compared the obtained groupings to subgroups of ARDS[39] previously described and showed some similarities between AP and ARDS endotypes.

For other diseases, and more specifically IBD and PD, heterogeneity was also studied using different clustering strategies. Here, we aimed at highlighting

subgroups in disease cohorts which could be of interest for the understanding of the disease using different types of data and methodologies. These analyses were motivated by previous findings suggesting the potential existence of disease subgroups for both IBD and PD.

1.4 Thesis structure

Following this introduction, chapter 2 will consist of a presentation of the different types of data which can be used for the study of diseases and for precision medicine-based projects and will be followed by a literature review presenting clustering analyses as well as some of the challenges relating to the type of data used in this context.

Chapter 3 will consist of a presentation of the main project, looking at data for a cohort of AP-affected patients. AP will be introduced along with details about the available data. Methods will then be presented, starting from the clustering strategy and giving more details about the results' evaluation and validation. Finally, results will be exposed and followed by conclusions.

Chapter 4 will be dedicated to the comparison of endotypes described in chapter 2 with other critical illnesses such as ARDS. The basis for this study and how it could benefit the study of critical illnesses will be laid out. Methods chosen to make the comparisons will be presented along with selected datasets. After summarising the main findings, some conclusions will be reported. Future directions will be discussed too.

In chapter 5, a study of the heterogeneity in IBD will be shown and potential analyses will be outlined.

In chapter 6, I will describe how a Parkinson's disease dataset was used to try to highlight relevant stratification. In a first section, the dataset and goal will be exposed. This will be followed by a presentation of employed methods. A results section will report obtained results. A discussion and conclusions will be presented as well.

The last chapter, chapter 7, will consist of a review of the main conclusions of this thesis and its limitations. Future directions will be evoked as well as general thoughts about disease stratification and precision medicine in the context of this work.

1.5 Contributions to knowledge

This thesis offers new perspectives on diverse illnesses: acute pancreatitis, inflammatory bowel disease, Parkinson's disease. The similarity between different types of critical illnesses was also studied.

More specifically, it produces a new way of stratifying acute pancreatitis individuals, different from the traditional aetiology or severity-based strategy. Indeed, the current classifications do not permit a clear understanding of AP's underlying mechanisms nor the identification of potential biomarkers and/or pathways of interest that could be targeted for new therapy strategies.

Then, different critical illnesses were shown to share the same signal. This consists of a novel way to look at critical illnesses. Moreover, this could be further explored and lead to new discoveries which would be of great interest for the study and treatment of critical illnesses.

In conclusion, this thesis describes new ways to study heterogenous diseases and promising new stratification analyses increasing the current knowledge and providing further study avenues for the analysis of diseases.

A way of studying previously identified single nucleotide polymorphisms located in co-expressed genomic regions for patients affected with IBD is described in this thesis and could be beneficial for the understanding of both Crohn's disease and ulcerative colitis. Ultimately this could lead to new therapeutic strategies for their treatment.

Finally, stratification in Parkinson's disease was looked at, using a novel approach, and as far as the analyses showed, no relationships could be

established between the stratification and a single nucleotide polymorphism related to Parkinson's disease. Ultimately, this suggests that the mutation carriers do not present specific molecular phenotypes when compared to other PD individuals

2. Chapter 2 – Literature review and concepts

This chapter introduces core concepts that are relevant to the research subject presented in this thesis. It is divided into three main parts.

The first part introduces omics data, that is large datasets resulting from the identification and quantification of different pools of molecules in a cell, tissue, or organism.

The second part discusses clustering principles, that is the process of dividing a set into subsets. As the main project is based on human acute pancreatitis and uses time-series multi-omics data, when presenting the context, emphasis is placed on human subjects as well as clustering techniques that are especially relevant to the analysis of omics data.

The third section specifically describes clustering strategies that would be useful when looking at time-series multi-omics data.

2.1 Omics data

For one to understand the underlying mechanisms of a disease, the analysis of biological samples is essential.

Within cells, genetic information is transmitted through the processes of transcription (from DNA to messenger RNA and non-coding RNA, the latter being involved in many processes, notably the regulation of both transcription and translatation[41]) and translation (from messenger RNA to proteins). This transfer is unidirectional and is referred to as the central dogma of biology[42], which will be influenced by many factors.

The main omics layers, DNA, RNA, proteins and metabolites, have complementary roles and must be analysed together to provide an overview of involved mechanisms. Indeed, the characterisation of a single omics layer

does not suffice to provide a complete description of a biological process or a disease[43].

Independent acquisition methods are used to provide omics measurements that can then be combined using different strategies to extract relevant information. Omics can be measured in different media (such as plasma, serum, urine or tissue extract) depending on the hypothesis one wishes to test and the constraints inherent to the experiment.

The four major omics fields, genomics, transcriptomics, proteomics and metabolomics will be briefly presented in the following section.

2.1.1 Genomics

The study of an individual's DNA, also referred to as genome, is called genomics. The genome of an individual can be sequenced and analysed using genomics technologies such as arrays or next generation sequencing (NGS)[44,45]. As whole genome sequencing is still an expensive process[46], other strategies have been designed to only sequence regions of interest, for example whole exome sequencing, targeting protein-coding regions. Many study design variations are possible but will most likely follow the same logic as the following steps:

- **DNA fragmentation:** The DNA sequence is extracted and digested into small DNA fragments.
- **Ligation with adapters:** DNA fragments are ligated to adapters (short synthetized sequences).
- **Amplification:** Sequences are cloned (using polymerase chain reaction) prior to sequencing or hybridisation onto a genome chip.

Not only the DNA sequence itself can be studied, epigenomics (the study of reversible modifications that can potentially affect gene expression) is also an area of great interest[47].

2.1.2 Transcriptomics

Transcriptomics is the discipline studying the transcriptome of an organism, corresponding to the whole set of RNA molecules of an individual. RNA molecules in a cell or tissue can be identified and quantified in order to study gene regulation and function.

As in genomics, the transcriptome is usually sequenced using high-throughput technologies such as micro-array or NGS technologies like RNA-Seq, the former measuring expression using a pre-defined set of probes and the latter sequencing the whole transcriptome (or mRNA landscape, depending on the application).

Whole transcriptome sequencing is used to measure mRNA as well as non-coding RNAs whereas mRNA sequencing only measures mRNA. Library preparation will mostly follow the same steps:

- **Transcript type selection/depletion:** As ribosomal RNA makes up the most of RNA reads, it is usually depleted, if not of interest for the study. Regarding mRNA, Poly-A selection can be used to only keep mRNA reads as they present polyadenylated tails.
- **RNA fragmentation:** Usually, after these first steps, the RNA is fragmented.
- **Complementary DNA synthesis and amplification:** Following ligation of fragments with adapters, cDNA is synthetized then amplified.

This is then followed by sequencing producing as output nucleotide sequences for each RNA fragment. These fragments will need to be assembled and mapped against a reference genome before being able to quantify gene expression[48]. When no reference genome is available, for example when studying non-model organisms, de novo transcriptome assembly can be performed.

2.1.3 Proteomics

The whole set of proteins present in a sample is called the proteome. Its study is referred to as proteomics. To some extent, proteome and transcriptome are correlated but many more factors will reflect the differences observed[49]. As with transcriptomics, high-throughput technologies are used to study the proteome as well. Commonly, mass-spectrometry (MS)-based methods are employed.

When using MS-based technologies, common workflows can consist of the following steps:

- **Abundant protein depletion:** Plasma or serum can be used depending on the analysis but in both cases the samples will be dominated by a small group of proteins, usually housekeeping proteins that are not specific to the studied condition[50]. Special resins can be used to perform the depletion[51].
- **If using a differential in-gel electrophoresis (DIGE)-based protocol:**
    - **Protein separation:** In the case of DIGE, this step is performed on an electrophoresis gel using undigested proteins that can then be selected based on their expression levels.
    - **Protein digestion:** Proteins are digested into peptides (smaller sequences of amino acids, usually under 50 in length) after selection so that they can be analysed.
- **If using a chromatography-based protocol:**
    - **Protein digestion:** Proteins are digested into peptides so that they can be analysed.
    - **Protein separation:** Commonly, chromatography is performed before the MS step to separate the different peptides before performing tandem MS.

MS will then be performed and can consist of tandem MS (one MS step followed immediately by another one) or a simple MS. The MS step will separate the molecules according to their mass to charge ratio. Obtained spectra will be searched against databases of known peptides based on

peptides unique properties (mass-to-charge ratio and retention time) in order to generate reliable protein identification and quantification.

2.1.4 Metabolomics

While changes in the metabolome, consisting of all the small molecules in a biological sample, will partly be driven by changes in the proteome, environment will also account for its variation. Thus, measuring it can provide phenotype-related information that can be used to decipher processes of interest. Two common metabolomics acquisition methods are nuclear magnetic resonance (NMR) and MS[52]. Metabolomics technologies used share similarities with proteomics and preparation usually follow the step below.

- **Separation:** When using MS, analytes from the samples can be separated before measuring mass-to-charge ratios.

Finally, detection and quantification are performed. This can be done using NMR or MS. NMR uses radiofrequency and measures the molecules responses whereas MS isolates and fragments ions (charged molecules) to measure their mass-to-charge ratios and quantify the molecules by looking at the MS spectrum peaks.

Metabolomics studies can be either targeted (when focusing on a certain type of metabolites) or untargeted (when measuring all the molecules).

2.1.5 Other measurements

Aside from the different omics presented above, which can be measured from different biological samples, other characteristics, that might be routinely collected (for example during an hospital stay in the case of human) can be useful to measure. Age or body mass index may be collected as part of such collection. These measurements can be helpful to correlate omics variables with phenotypic properties.

2.2 Clustering, definition and aims

Clustering can be defined as the process of dividing a set of samples into smaller sets, named clusters. Samples in a same cluster should present similar features whereas samples from different sets should present dissimilar features.

Cluster analysis is referred to as an unsupervised learning technique, as the groups are not known a-priori. This is particularly useful when one wants to learn about the structure of a dataset and especially to determine if devising subgroups would help in better describing it. Applications are numerous and go well beyond the sole area of biology.

2.2.1 Cluster analysis steps

Performing a cluster analysis involves different steps which will greatly depend upon the aim to be reached[53,54]. An example of steps taken during the cluster analysis process is described here:

- **Feature selection:** One can choose to use all features or a subset of pre-selected features to perform the analysis.
- **Similarity measure:** Similarity between samples will need to be quantified using a chosen metric
- **Criterion choice:** Any number of clusters (from one cluster containing all samples to one cluster per sample) can be derived from a cluster analysis. By choosing a criterion (usually a cost function), quantifying for example the compactness of the clusters, one can determine an optimal partition for the analysed dataset.
- **Clustering algorithm selection:** This will determine how the clusters are defined. Some require the user to choose the number of clusters to extract (such as partitioning clustering) and some do not (such as hierarchical clustering).

- **Assessment:** Internal measurements can be used to assess the quality a clustering solution and will reflect the compactness of the clusters and/or the separation between different clusters. Stability testing techniques can be used as well, they will determine how much a clustering solution is sensitive to change in the input data. In the case of biological data, known biological annotations can be used to assess the quality of a clustering.
- **Interpretation:** After cluster allocations are determined one must interpret them to determine what the results are based on, to highlight distinctive and shared features between clusters for example.

2.2.2 Feature selection

Reasons for feature selection and a brief overview of employed strategies are presented in the following two paragraphs.

2.2.2.1 Considerations

When performing clustering on a dataset, the number of variables, or features, describing the samples can greatly vary. There can be a few variables or many, with sometimes numbers far exceeding the number of samples (often referred to as high dimensionality).

The development of high-throughput technologies has led to the availability of many high-dimensional datasets. When dealing with a great number of variables, there can be many advantages in selecting only a subset of the initial variables. Indeed, using less variables will decrease the computational burden and can also improve the quality of the clustering by getting rid of noise/irrelevant variation present in a dataset[55,56]. Such considerations will also apply to supervised learning problems like classification as a minimum number of variables will be desired for efficient and simple models. The process of selecting only a subset of variables is referred to as feature selection. The idea behind it is that selected features must not be redundant

and must provide discrimination power that can be used to separate the samples.

2.2.2.2 Strategies

Two main strategies can be applied in this case, filtering and wrapping.

Filtering is applied before running the clustering algorithm. Filtered variables will be selected based on properties of the data. For example, one might drop highly correlated variables as they will provide redundant information. Variance can also be considered as a metric to filter features as low-variance features might provide very little power to discriminate samples.

The wrapping strategy employs a different concept. The clustering algorithm will be run using the complete set of features. The output of the clustering will then be used to perform a selection on the variables. Indeed, contribution measures can be extracted for each variable and used to apply a filter (for example features with the lowest contribution may be dropped as part of the process). Usually, this is repeated several times until an optimum is reached (usually defined by a pre-selected criterium). The choice of criterion might result in different subsets being selected.

One can also choose to extract features of interest by creating new variables (using principal component analysis, for example) that will be used as input to the clustering algorithm to help reduce the number of irrelevant variability included in the model[57].

2.2.3 Proximity measure

Once the input set of samples and variables has been chosen, one must define similarity values between samples, quantified as ‘distances’, prior to applying the clustering algorithm[58]. Distances computed will be a direct measure of how close/different two samples are. This crucial step will be used to define

clusters. Indeed, the aim is to create groups of samples that are close together and apart from others.

2.2.3.1 Considerations

Commonly, dissimilarity is measured as a distance. However, the choice of metric is inherent to the type of each one of the features. Distances metrics used for continuous variables will be different from the ones used for categorical variables. Some metrics might also take into account important characteristics of the variables, such as the distribution of a feature.

2.2.3.2 Strategies

Metrics used in clustering tasks can be classified in two classes, dissimilarity and similarity distances. Some of the most common, with emphasis being placed on the ones which can be applied to continuous variables, will be presented here.

2.2.3.2.1 Dissimilarity metrics

The most popular dissimilarity metric in the case of continuous variables is the Euclidean distance[59].

$$d_2(x_i, x_j) = \sqrt{\sum_{k=1}^{n} (x_{i,k} - x_{j,k})^2}$$

As the Euclidean distance integrates squared differences in values within each one of the spaces (corresponding to features), one must make sure that the data is normalised beforehand so that each variable has a variance of the same order. Indeed, if not, bias might be induced by features with larger variances.

This can be compared to the Minkowski metric[60], of which it is a special case with order equal to 2. Likewise, the Manhattan distance[60] is a special case of the Minkowski distance but with an order of 1.

Another common dissimilarity metric is the Canberra distance[60]. It is a weighted version of the Manhattan distance, the weight being proportional to the sum of absolute values in a given feature space.

$$d(x_i, x_j) = \sum_{k=1}^{n} \frac{|x_{i,k} - x_{j,k}|}{|x_{i,k}| + |x_{j,k}|}$$

Small differences in values will have a different influence on the final distance depending on whether those values are close to 0 or not. For a difference in values of the same order in magnitude, for samples close to 0, the Canberra distance will be much larger compared to values further away.

Hamming distance[59] calculates the number of different elements between two vectors and can be used as well. The magnitude of differences for any given feature will not influence the final distance.

2.2.3.2.2 Similarity metrics

Similarity metrics can be computed and converted back to dissimilarity so that they can be used for the clustering task.

Correlation-based metrics, such as Pearson's or Spearman's correlation coefficients[61] can both be valid measures of similarity when performing clustering.

Pearson correlation coefficient is the correlation between two sample vectors and will be a number between -1 and 1.

$$r(x_i, x_j) = \frac{\sum_{k=1}^{n}(x_{i,k} - \bar{x})(y_{i,k} - \bar{y})}{\sqrt{\sum_{k=1}^{n}(x_{i,k} - \bar{x})^2}\sqrt{\sum_{k=1}^{n}(y_{i,k} - \bar{y})^2}}$$

A correlation value of 1 or -1 indicates that one sample can be predicted from the other using a linear equation. A value of 0 indicates that no linear relationship was identified between the two samples of interest.

Spearman correlation can be described as the comparison of ranks of values between two vectors. It can be computed using the following formula.

$$r(x_i, x_j) = 1 - \frac{6\sum_{k=1}^{n}(r_{x_{i,k}} - r_{y_{i,k}})^2}{n(n^2 - 1)}$$

Spearman's r will be equal to Pearson's r when replacing actual variable values by ranks in Pearson correlation formula. If the changes are not proportional but the variables are still ordered in the same way, then Spearman's r will have an absolute value higher than Pearson's. Both will detect similarities in shape rather than similarities in magnitude.

These similarity metrics can be converted to dissimilarity values using an appropriate transformation, one of the simplest being stated below.

$$d(x_i, x_j) = \frac{1 - r(x_i, x_j)}{2}$$

2.2.4 Criterion

2.2.4.1 Considerations

Depending on the clustering algorithm chosen, there can be many ways to partition a set. One must choose the algorithm which provides the optimal clustering given a specific problem.

In the case of unsupervised clustering, as the truth will not be known, one must use the features of identified clusters, such as cluster centroids, the separation

between clusters or the compactness of clusters to determine the quality of a solution. Depending on the aim of the study and the type of analysed data, other measures can be defined to assess the quality of a clustering solution.

2.2.4.2 Strategies

The most common way to estimate and visualise the homogeneity of clusters, as well as to identify an optimal number of clusters is to compute and plot silhouette scores for all samples[62]. The term silhouette refers to the outline of the score values when represented graphically. For each sample, a silhouette score reflecting how well the observation fits into its cluster can be computed. Its value can range between -1 and 1, a value of 1 meaning that the current cluster is a perfect match for this sample. The following formula is used to compute the silhouette score for a sample.

$$s(\mathrm{x}) = \frac{b(x) - a(\mathrm{x})}{\max\left(a(x), b(x)\right)}$$

$a(\mathrm{x})$ is defined as the average dissimilarity between sample x and all other objects that are from the same cluster. $b(\mathrm{x})$ is the minimum average dissimilarity between sample x and objects from different clusters (computed per cluster). Intuitively this gives a value reflecting how tight samples are within a cluster and how separated this cluster is from the others. This can be averaged over clusters and clustering to have an overview of the partition quality. A strategy to choose a partition is to try and maximise the average silhouette value.

2.2.5 Clustering algorithm

2.2.5.1 Introduction

After having chosen a method to compute pairwise dissimilarities and a criterion to define an optimal cluster solution, the clustering algorithm itself has to be selected. The clustering output can consist of hard or soft partitions, the

former consisting of a sample being assigned to a single cluster whereas the latter consists of a sample belonging to each cluster to a certain degree.

2.2.5.2 Clustering methods

An overview of some commonly used clustering methods, along with their advantages and disadvantages, is described in the following subsections.

2.2.5.2.1 Hierarchical methods

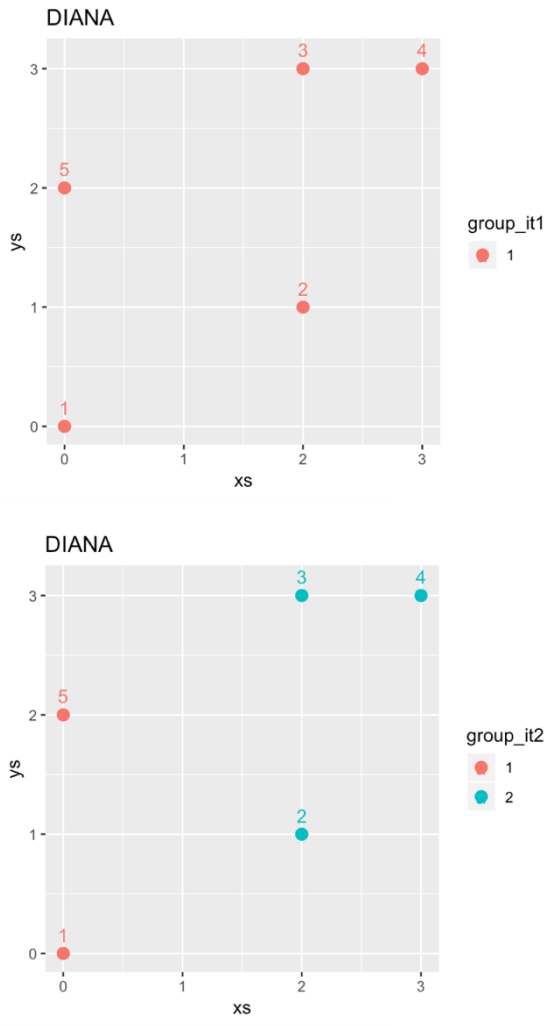
Hierarchical clustering is a clustering method producing nested groups[53]. Depending on the strategy chosen, the starting point will consist of a single group containing all samples (divisive strategy) or a set of groups containing each one a single sample (agglomerative strategy). Divisive strategies are often referred to as DIANA[63] (DIvisive ANAlysis) and agglomerative as AGNES[63] (AGglomerative NESting). This has been successfully applied to cluster gene expression patterns in organisms[64] or a specific disease[65]. Hierarchical methods can be sensitive to outliers and noise and might not work when handling clusters of different sizes.

2.2.5.2.1.1 DIANA

To determine how a group should be divided in DIANA clustering, all possible partitions of the data would have to be considered. In large datasets, this constitutes a substantial computational burden. One can implement the following solution to greatly reduce the number of considered partitions when dividing a set[66]. Briefly, the first step consists of highlighting the sample from a cluster with the greatest average dissimilarity to the other elements of this same cluster. This sample will be used to create a new cluster. Then, all elements of the original cluster will be either kept in the former cluster or moved to the new one. This will depend on the difference between the average dissimilarity between the remaining members of the original cluster and the

average dissimilarity with the samples in the newly created group. If this difference is positive the object is then moved to the new cluster. This is repeated until stability is reached (

Figure 2.1).



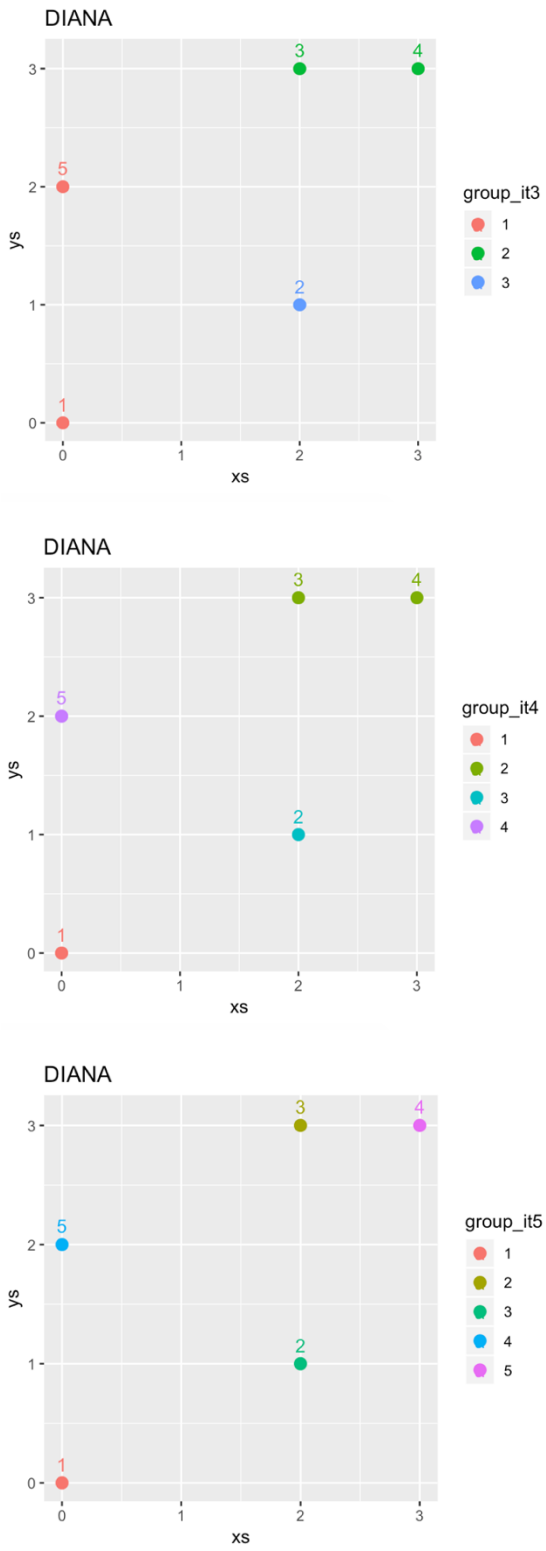
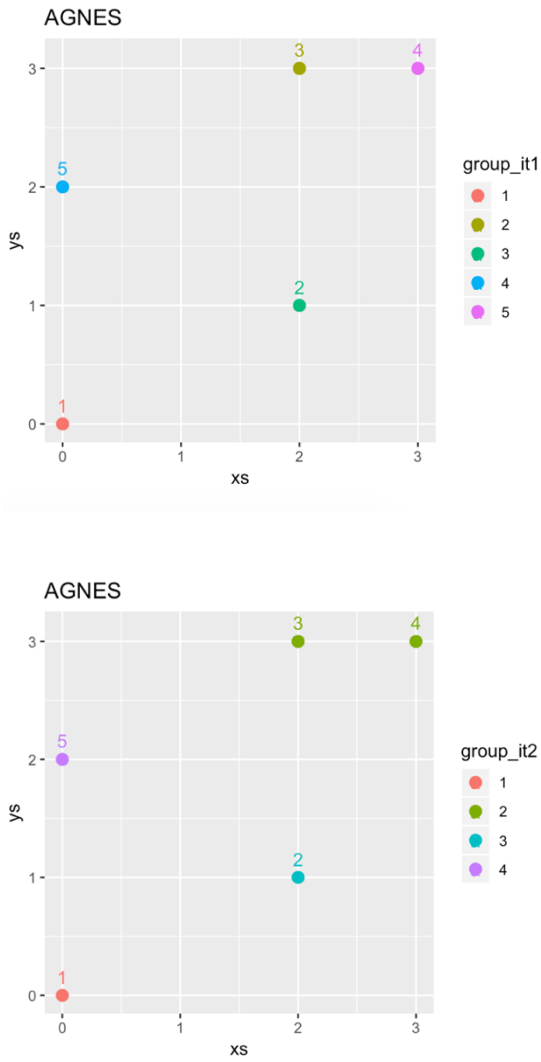


Figure 2.1 - DIANA clustering iterations on dummy data produced using diana function in R with Euclidean distance and default parameters (cluster package).

2.2.5.2.1.2 AGNES

The most commonly used hierarchical clustering strategy is agglomerative clustering, a so-called 'bottom up' approach. There are many criteria to define how clusters should be merged as part of the agglomerative process. These are referred to as linkage methods[63]. Four of them will be presented in this section. Starting from a partition with n 1-element clusters, these will be merged in a pairwise fashion, given the closest clusters as defined by the linkage method, until only one cluster, containing all samples, is generated (Figure 2.2).



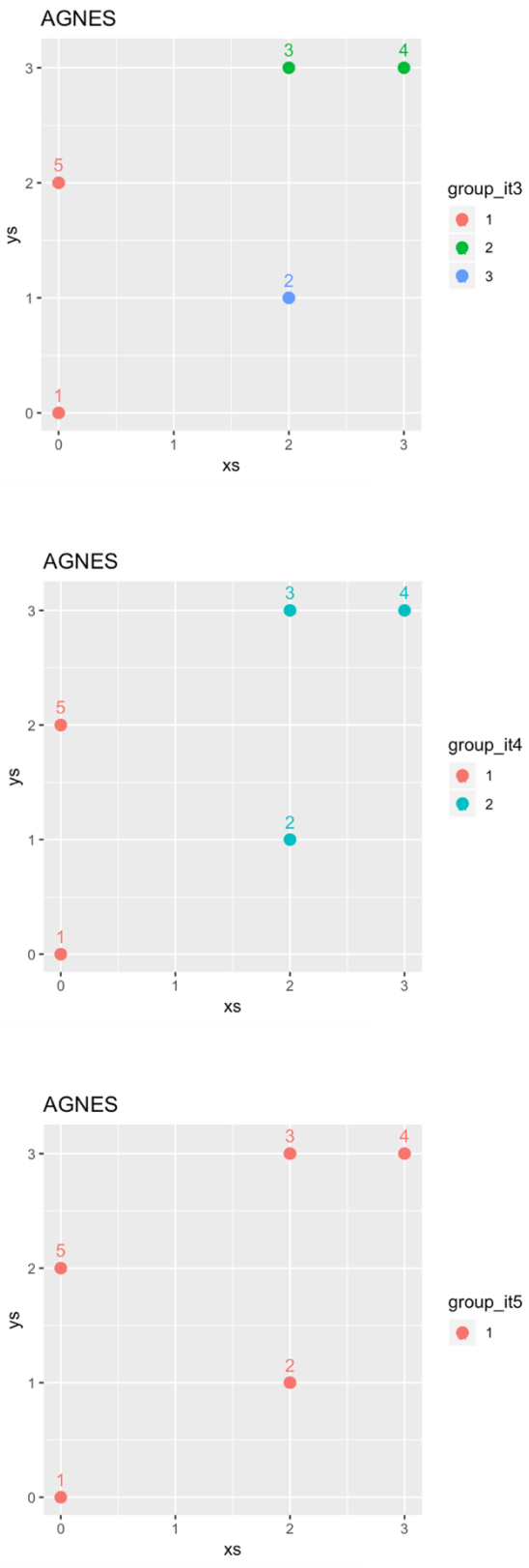


Figure 2.2 - AGNES clustering iterations on dummy data produced using agnes function in R with Euclidean distance, average linkage and default parameters (from the cluster package, plots were generated with ggplot2 library).

Two linkage methods, utilising distances between pairs of samples, can be used. The first one, single linkage, defines dissimilarity between two clusters as the smallest distance between pairs of points, one from each cluster. The second one, complete linkage, uses the largest distance between pairs of points from different clusters in place of the smallest.

Depending on the cluster shapes one of these two methods can be preferred, for example the distance between two elongated clusters will be much smaller when using the single linkage method, whereas the same two clusters would appear much further away in the case of complete linkage.

Another popular linkage method is average linkage, also referred to as UPGMA (Unweighted Pair Group Method with Arithmetic mean), built upon the definition published previously[67]. The decision whether to merge two clusters will be made given the distances between clusters computed as the average distance between elements from the two clusters (performed in a pairwise fashion). This technique avoids the use of extreme dissimilarity values such as when using single and complete linkage and can offer a compromise.

Another commonly used linkage method is Ward's minimal increase of sum-of-squares method[68] and is based on a slightly different principle from the linkage methods previously introduced. Indeed, rather than directly using dissimilarities between samples, it computes sum-of-squares. The merging of samples will be performed based on the minimum total sum of squares increase, in other words, for each candidate pairs of groups, it will compare the sum of squares of their union with the sum of squares for each one of them separately. Ward's method usually performs well when group sizes are of the same order. If group sizes are very different, it might be difficult to highlight them using this algorithm.

2.2.5.2.1.3 Dendrograms

To illustrate the structure of the identified clusters, a dendrogram can be used. It consists of a tree representation for which each leaf is a sample. Dividing branches represent the divisions/fusions performed as part of the chosen clustering algorithm. The vertical value, referred to as ‘height’, for which a split between two branches occurs represent the distance between the two corresponding clusters (Figure 2.3). Here, for example, samples 3 and 4 are the two most similar samples as the branches joining them splits at the smallest height value.

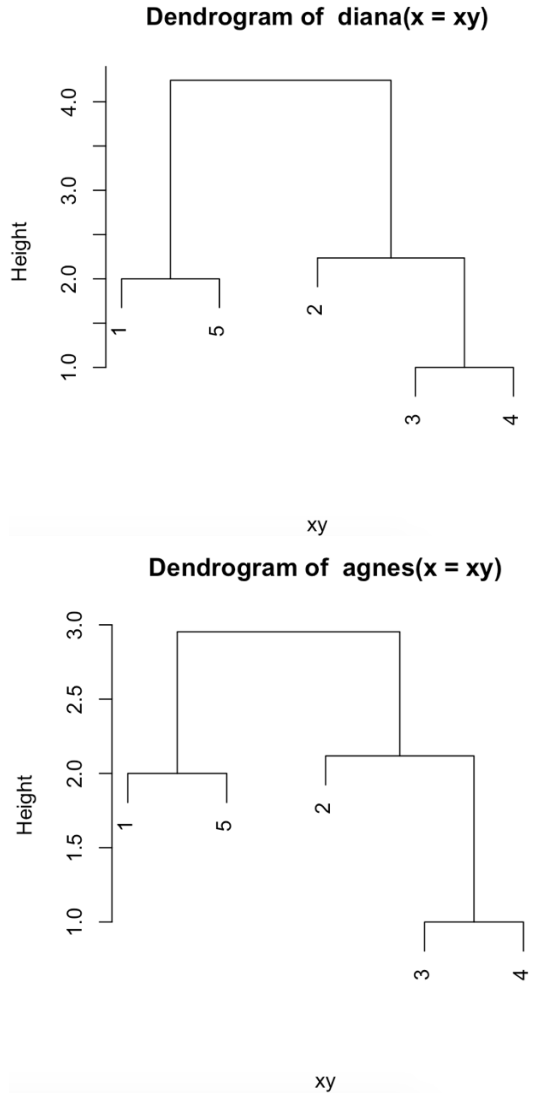


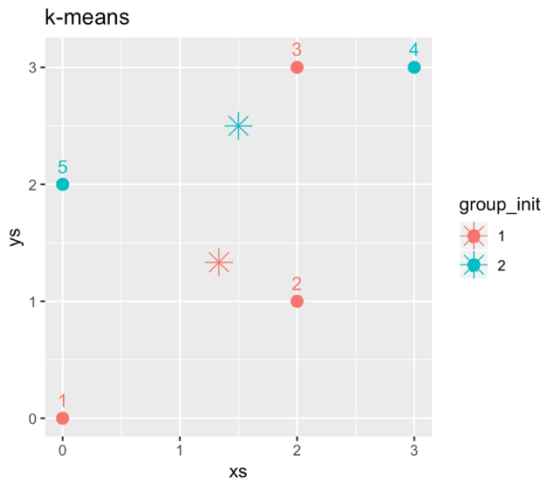
Figure 2.3 – Example dendrograms using the previously used dummy data (generated using the cluster package in R).

2.2.5.2.2 Partitioning methods

Another type of clustering method, which produce discrete cluster sets, as opposed to nested sets such as described in the previous paragraph, is referred to as partitioning clustering. Two common partitioning algorithms, K-means and partitioning around medoids (PAM), are presented in the following subsections.

2.2.5.2.2.1 K-means

K-means[69] is one of the most commonly used partitioning clustering algorithm. One parameter, k, is required to run the algorithm and must be chosen by the user based on previous knowledge, graphical interpretation or comparison with results produced using other values of this parameter. K represents the number of clusters that will be identified by the algorithm. This is an iterative process composed of two steps, a centroid calculation step and an assignment step. First, each sample is assigned to a random cluster (the number of clusters being defined by k), the algorithm will then compute the centroids of each one of the clusters (Figure 2.4, first panel). Once the centroids are computed, the samples are assigned to the group with the closest centroid (as defined by the Euclidean distance, Figure 2.4, second panel). This is repeated until convergence (Figure 2.4, third panel). The algorithm aims at minimising a function defined by the sum of within-cluster variation and will identify a local minimum for this optimisation function. As it will identify a local and not a global minimum, it can be useful to run the algorithm with different initial conditions, as defined by the random assignments used at the first step of the process. K-means clustering is frequently used in health-related studies. For example, one study clustered fat mass changes in obese subjects[70] using this algorithm.



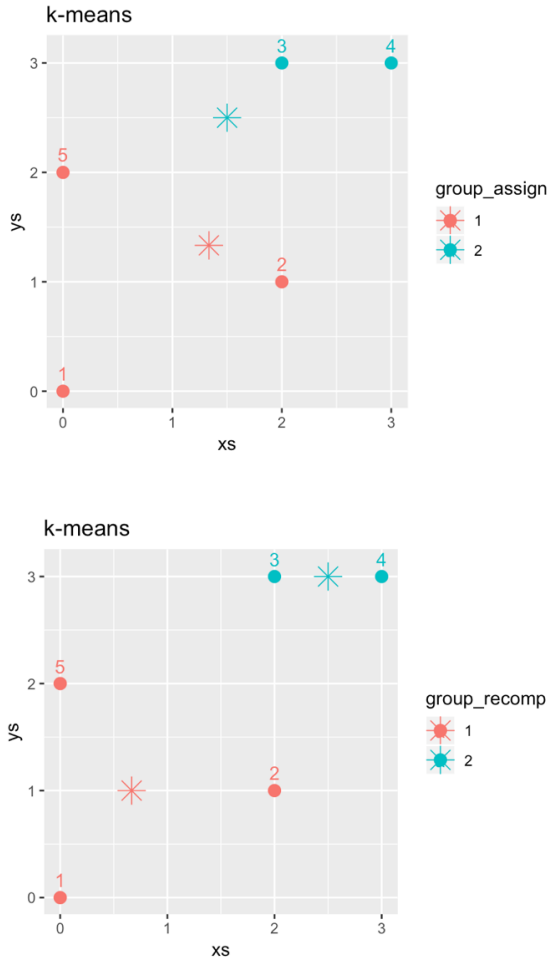


Figure 2.4 – K-means using dummy data and k=2,the three steps are represented. The initial random assignment and computation of centroids, represented as star markers, is illustrated in the first panel. The assignment to the closest centroid is illustrated in the second panel. Finally, the re-computation of centroids is done. Here, in this simplistic example, convergence is reached. (plots generated using ggplot2 package in R)

2.2.5.2.2.2 Partitioning Around Medoids (PAM)

Partitioning Around Medoids[63], usually referred to as PAM, is the most commonly used k-medoids algorithm and has been used to study disease clusters[71]. A medoid is a representative object of dataset or cluster.

The principle behind PAM is similar to that of K-means clustering, indeed, it aims at highlighting clusters that minimise the average dissimilarities between the samples and their associated cluster representation. The difference with K-means is that PAM uses medoids, which are actual samples, as opposed to

K-means which computes centroids, corresponding to the mean vector for a dataset or cluster.

After an initialisation step, for which medoids are randomly selected among the set of samples, given a value of k selected by the user, the main algorithm is organised around two steps. The first one, called the build step, during which the average dissimilarity between each sample and their corresponding medoid, as defined by the closest centre, is computed (Figure 2.5).

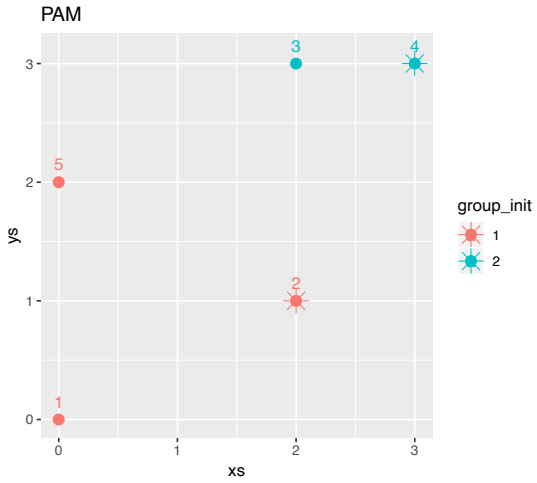


Figure 2.5 - PAM using dummy data and k=2, initiation and build steps, medoids are represented as star markers. (plot generated using ggplot2 package in R)

The second step consists of the swap phase, during which one of the selected centroids is randomly swapped for another object of the same cluster (Figure 2.6).

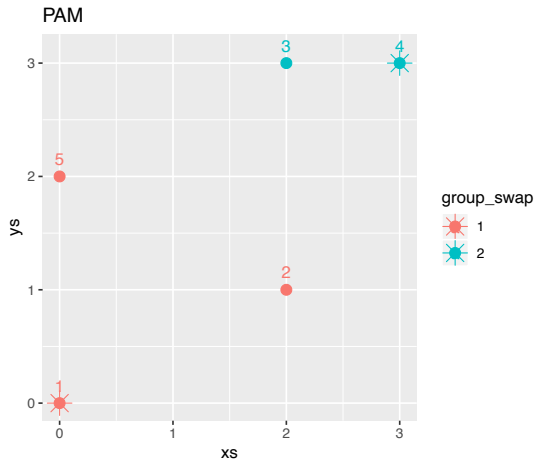


Figure 2.6 - PAM using dummy data and k=2, swap phase, group 1 medoid is swapped from point 2 to 1, medoids are represented as star markers (plot generated using ggplot2 package in R), here, in this simplistic example, convergence is reached.

Average dissimilarity is computed as previously and compared to the last obtained value, if this value is lower the swap is maintained, if the dissimilarity is greater the swap is dropped. This is repeated until no improvement can be obtained.

2.2.5.2.3 Density-based methods

Density-based methods encompasses strategies identifying clusters of points based on their density. This family of methods is able to identify clusters regardless of their shapes and deals well with outliers.

2.2.5.2.3.1 DBSCAN

Density-Based Spatial Clustering of Applications with Noise[72] (DBSCAN) is a density-based clustering method. The number of clusters will be inferred from the data. Density is estimated by looking at neighbouring points and given a radius defined by the user. A second parameter, the minimum number of points, also defined by the user, is used to defined clusters.

Three categories of points are defined in DBSCAN:

- **Core points**: any point with a number of neighbouring points (as defined using the radius value) containing at least the minimum number of points
- **Border points**: a point being reachable from a core point but with less than the minimum number of points in its neighbourhood.
- **Outliers**: a point not falling in any of the two previous categories

The algorithm then goes through three steps. First, a point not assigned to a cluster or defined as an outlier is chosen at random. If this point is not identified as core point, then it is defined as an outlier. If the point is a core point, then it will serve as a basis for a cluster. The cluster is defined by adding points that are in the neighbourhood. This is repeated for all added points. Finally, these two steps are repeated until all points are assigned to a cluster or defined as outliers. An example output is represented in Figure 2.7.

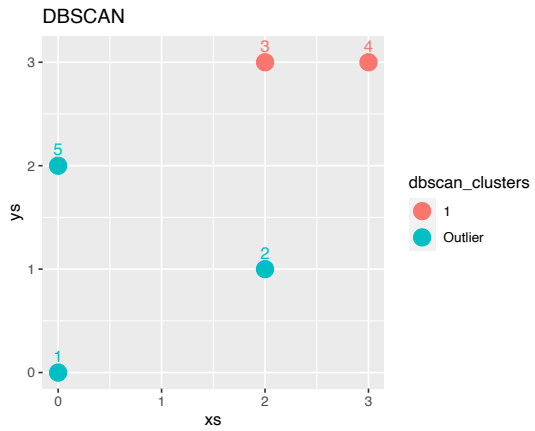


Figure 2.7 - DBSCAN using dummy data, radius=1 and minimum number of points=2. One cluster comprising two points is identified (represented in red) and three outliers (represented in blue) are highlighted. (plot generated using ggplot2 package in R)

2.2.5.2.3.2 OPTICS

Ordering Points To Identify the Clustering Structure[73] (OPTICS) is a clustering method similar to DBSCAN but palliating one of its weaknesses. Indeed, DBSCAN may have trouble identifying clusters of different densities. To do so core and reachability distances are defined.

The core distance corresponds to the minimum radius value which would result in a point being classified as a core point. The reachability distance is defined between a point and a core point and is either the core distance (if the point is within the area defined by the core distance) or the distance between the two points (if the point is further away from the core point). Once a point is processed, the next closest point will be processed. Reachability distances are used along points ordering to produce the reachability plot.

As opposed to DBSCAN, no clusters are produced are must be extracted by the user using the reachability plot produced. An example reachability plot is presented in Figure 2.8.

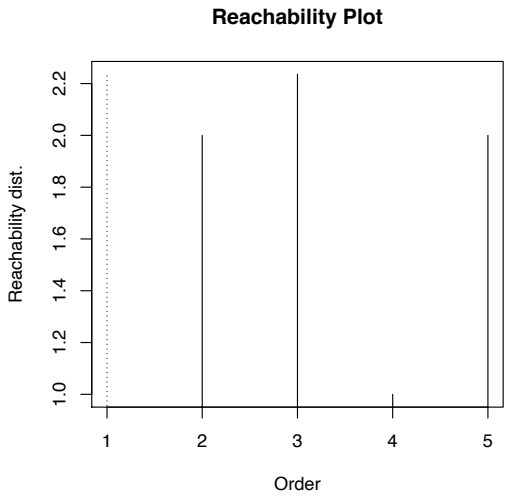


Figure 2.8 – Reachability plot produced using dummy data and a minimum number of points=2. Reachability distance is represented on the y axis and ordered points are represented on the y axis.

Using a reachability plot, clusters can be identified by looking at “hills” and “valleys”. For example, here, one cluster might be identified and would be composed of the third and fourth points (respectively identified as 3 and 4 in

Figure 2.9) as there is a steep decrease in reachability distance between them. Others would be identified as outliers. This process can be automated by applying a threshold on steepness values.

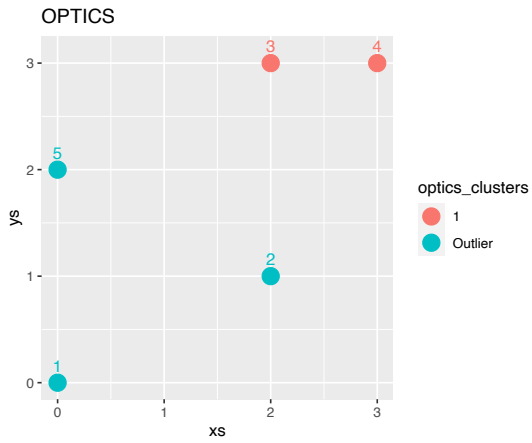


Figure 2.9 – Dummy data used to produce Figure 2.8. Coordinates of the 5 points are represented on axes x and y and identified clusters are reported as well. (plot generated using ggplot2 package in R).

2.2.5.2.4 Model-based methods

Model-based clustering methods are a family of methods based on finding a mathematical definition, or equation, to represent the data. Examples of model-based methods include Gaussian Mixture Models (GMM)[74] and Self-Organising Maps (SOM)[75].

2.2.5.2.4.1 GMM

GMM will try and model the data as several Gaussian components, one for each cluster. It is hypothesised that each sample has been generated by one of these Gaussian components (Figure 2.10).

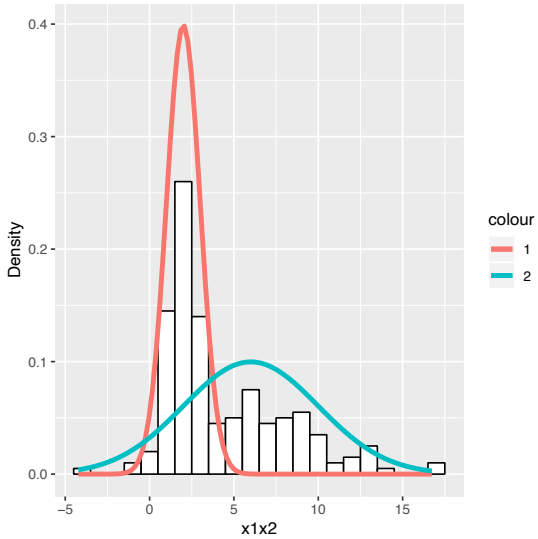


Figure 2.10 -Density histogram for a dummy variable (generated using two Gaussians, x1 and x2, with respective means 2 and 6 and standard deviations 1 and 4 with rnorm function in R) with Gaussian distributions overlaid (figure generated using ggplot2 library).

Parameters will be obtained by maximising the likelihood of the observed data. In other words, the likelihood of observing such a data distribution, given it has been generated by the current model, should be maximised. Omics data can be clustered using GMM, for example using gene expression data[76].

One intuitive way to define GMMs is to compare them to the k-means algorithm as they share common concepts. The former is a generalisation of the latter, indeed, GMM initial conditions will be randomly chosen. They consist of cluster location and cluster shape in the feature space. This will then be followed by an expectation-maximisation (EM) procedure, the expectation step (E) will define probabilities of membership for each sample and the maximisation step (M) will update the location and shape of all clusters given these probabilities in order to have the highest possible likelihood. This step will be repeated until convergence. The clusters shape can be constrained to a lower set of dimensions, or to the full set of dimensions, a lower set resulting in lower processing requirements.

This provides two main advantages over k-means, the cluster shape is not constrained to a sphere and the assignments are probabilistic (each sample

will have a confidence value for its cluster assignment). Not only it allows clustering but provides a mathematical way of describing the clusters.

Similar to the k-means algorithm, using GMMs with several different initial conditions can help find the most likely solution in a given dataset.

2.2.5.2.4.2 SOM

A SOM is a neural network-based model that can be used as part of a clustering task. SOM tries to capture as much as possible of the high-dimensional structure of a dataset, in a low-dimensional surface (also called the map). The main advantage over traditional dimensionality reduction techniques, such as principal component analysis (PCA)[77], is its ability to capture non-linear relationships. Applications are numerous and can help for example in deciphering gene expression patterns[78].

The starting point consists of a grid composed of neurons. The algorithm works by iteratively updating the neuron centres positions given samples positions. Initial conditions must be chosen, one option is to use PCA coordinates. For each data point, the closest centre is identified (it is referred to as Best Matching Unit, or BMU) using smallest Euclidean distance. SOM then looks for all centres that are within a given distance (that will decrease as the algorithm progresses) of the centre identified in the previous step using the distances in terms of the SOM surface. Finally, the positions of these centres are updated by ‘pulling’ the centres towards the current data point. One way to perform this consists of adding the weighted vector difference between the data point and cluster centre. The chosen weight is referred to as the learning rate and decreases for each iteration. This is then repeated until convergence or until the maximum number of defined iterations is reached. A schematic mapping between input data and a 2-dimensional grid is presented in Figure 2.11.

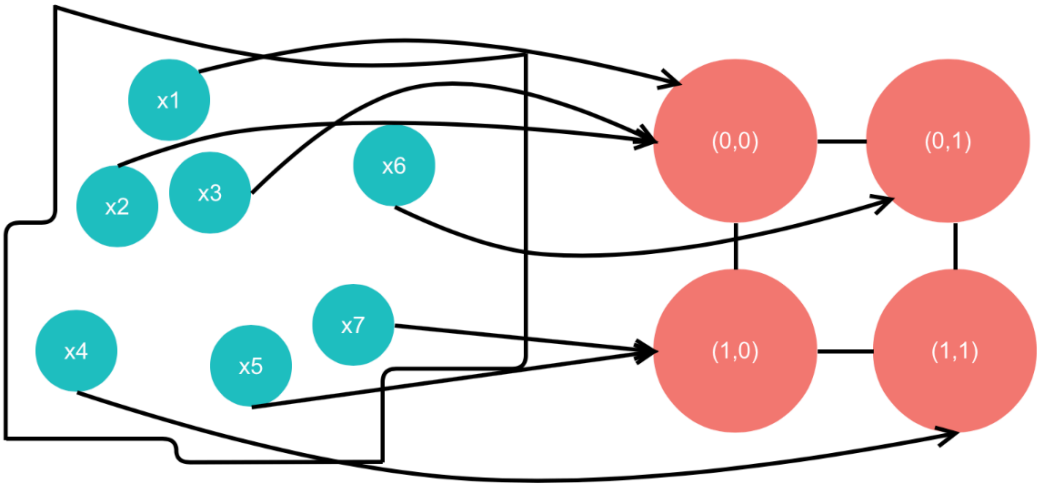


Figure 2.11 - Projection example between some input data and a 2-dimensional SOM (the input space is represented on the left-hand part of the figure and the SOM space on the right hand part of the figure).

One can identify clusters by looking at the final map density and the similarities between the different nodes.

In this section, different clustering methods were defined. All methods will provide many different results which must be assessed to define which should be chosen.

2.2.6 Assessment

2.2.6.1 Introduction

Using different methods and parameters, many different clustering solutions can be obtained. To select and/or compare results from clustering procedures one must prove the relevance and value of a solution.

A solution can be evaluated in terms of various statistical properties but also, depending on the context, in terms of meaningfulness for the field[79]. For example, in biology, one might want to test for biological processes highlighted by a clustering solution.

Finally, it is of utmost importance for a solution to be replicable. More specifically, a solution should not be specific to the dataset at hand but should be valid for other datasets as well.

2.2.6.2 Statistical properties

To validate a clustering solution, statistical properties of the partition can be looked at, using either pre-defined measures or by assessing the stability of a solution under induced change in the input dataset. Such measurements can be used to compare solutions and/or to find the optimal one but also to give a numeric value to the quality of a partition.

2.2.6.2.1 Internal indexes

As presented in section 2.2.4.2, silhouette scores can be used to assess the homogeneity of obtained clusters. However, silhouette scores alone are not sufficient to prove the validity of a partition (a set of clusters) and one must look at the characteristics of the obtained clusters to determine their value.

2.2.6.2.2 Stability

When looking at the results of a clustering algorithm, it is essential to make sure that the solution does not rely too much on one or a few samples. In other words, when clustering a set of samples, one would expect the global structure to remain roughly the same when a small proportion of samples is excluded as compared to when including the whole set.

Bootstrapping can be used to perform such task. It consists of excluding a proportion (user-defined) of the input samples, replacing them by copies of included samples and re-performing the clustering task. The resulting clusters can then be compared to the initial solution.

The Jaccard index can be used to compare two bootstrapped versions of a clustering solution[80]. This metric computes the overlap of two sets as defined by the following formula.

$$JI(C_i, C_j) = \frac{|C_i \cap C_j|}{|C_i \cup C_j|}$$

Jaccard index values range between 0 (no overlap) and 1 (perfect overlap) and can be used as a measure of similarity between clusters. When dealing with several clusters Jaccard index values can be averaged to provide a single value per partition.

2.2.6.3 Biological properties

When studying biological datasets, it is important to assess whether the results obtained are in line with known biology. The hypothesis arises from the fact that random clusters will not highlight any known biological processes as defined by pathway maps for example.

Depending on the data, several techniques can be employed to assess the biological relevance of the solution. Depending on the data type, pathway analysis or gene set enrichment analysis can be used to highlight pathways or biological terms of interest such as KEGG pathways[81] or gene ontology (GO) terms[82]. There is no gold standard to determine the relevance of a clustering solution and a careful examination of the results will be required.

Results from this kind of analyses can also be used for clusters interpretation as they provide insight into functional properties of the clusters.

2.2.6.4 Replication

For a solution to be valid, one would aim to reach the same conclusions when repeating an experiment.

For example, a different input dataset, collected at a different hospital and processed differently could be used to test the validity of a clustering solution. If valid, a similar structure should be observed.

The replication step generally aims at showing that obtained results are of sufficient quality and that our understanding of the studied system is correct[83].

2.2.7 Interpretation

2.2.7.1 Aims

As mentioned in the previous section, insights about the clustering results can be obtained using pathway analysis or GO terms enrichment analysis for example. However, there are many more ways to extract information from clusters. Some of them are presented in the following sections.

2.2.7.2 Proportions comparison

When looking at clustering results for patient data, if observational categorical variables, for example age or gender are available (they can be confounding variables as well), a simple comparison between proportions can be performed. It can be formalised by using a Chi-Square test to compare frequencies and then compute a p-value. An example would be to compare the proportion of males in a cluster and compare it to the proportion of males in another cluster, this would help determine if the proportions of males in the two groups are significantly different or, in other words, determine if the group labels are independent from the gender variable. This can also be done when there are more than two categories for both variables.

2.2.7.3 Analysis of variance

Average values per variable per cluster can be computed and compared using ANOVA (ANalysis Of Variance). One-way ANOVA consists of a statistical test

of whether averages are equal between independent groups. Post-hoc tests can be used to determine which groups were different for a given variable. When only two groups are present, a t-test can be a suitable alternative.

2.2.7.4 Prediction models

Another strategy consists of creating models that predict sample allocations given all or part of the input data. By doing so, one can extract variables contributions and thus highlight discriminant variables.

When more than two groups were highlighted, models can be generated using several designs. Some algorithms support multi-class classification problems, but it can sometimes be beneficial, in terms of accuracy, to divide the problem into smaller ones. For example, binarization has been successfully applied in many settings and can be subdivided into one-vs-one and one-vs-all approaches[84–86], the former referring to pairwise comparisons between groups and the latter to one group being compared to all others. Moreover, this can help highlight specificities inherent to each group rather than global differences.

Some examples are provided in the sections below.

2.2.7.4.1 Partial Least Squares-Discriminant Analysis (PLS-DA)

2.2.7.4.1.1 The algorithm

Partial Least Squares-Discriminant Analysis (PLS-DA) is a PLS problem, an approach maximising the covariance between the variance of predictors and a categorical outcome. PLS-DA is useful for dimensionality reduction, feature selection and classification[87], thus helping to decipher the properties of a cluster. It is especially well adapted to deal with high-dimensional, noisy and collinear (when variables can be predicted from linear combinations of others) data. A parallel can be made with principal components analysis (PCA), the main difference being that PLS-DA aims at maximising the covariance

between the data and group labels whereas PCA aims at maximising the represented variance, with no knowledge of group labels. PLS-DA can be seen as a supervised version of PCA.

The algorithm will project the data into a low dimensional space (the number of spaces being defined by the user) whilst representing as much covariance between the data and the group labels as possible.

The first step consists of extracting a weight vector, W given X, the input data and y, the outcome vector.

$$W = X'y$$

Secondly, a score vector is computed using the input data, X and the weight vector, W, computed during the previous step.

$$t = \frac{XW}{\sqrt{\sum W^2}}$$

The next step consists of computing the X and Y-loading vectors. The X-loading vector, consisting of a vector with a number of elements equal to the number of features, will describe how the different variables relate to each other. Correlated variables will have similar weights on the X-loading vector. Similarly, the y-loading vector will describe the relationships between the groups and will have a length equal to the number of samples. For example, variables with similar loadings will have a similar contribution towards the separation of different groups as part of the PLS-DA model.

X-loadings are computed as described in the equation below.

$$p = \frac{t'X}{\sqrt{\sum t^2}}$$

Similarly, Y-loadings are computed as follows.

$$q = \frac{y't}{\sqrt{\sum t^2}}$$

The equations previously outlined will define the first component of the PLS-DA model. If the user has chosen more than one component, the residuals (the variation not accounted for at this stage) will be used to compute other components using the equations defined but replacing X and y by the residuals, as defined in the two equations below.

$$res_X = X - tp$$

$$res_Y = y - tq$$

Once all defined components are computed, the prediction model will be created. Regression coefficients, one for each variable, are calculated as shown in the following equation, one for each component.

$$b = W(pW)^{-1}q$$

To obtain the predicted value $\hat{y}$ for an input sample X_{test}, the input data is multiplied by the matrix of regression coefficients B as shown below.

$$\hat{y} = X_{test}B$$

2.2.7.4.1.2 Variable importance for PLS-DA models

From PLS-DA models, variable importance can be computed. The importance of a variable is referred to as VIP (Variance Importance in Projection) scores and is the contribution of a variable to the PLS-DA model. The scores are computed from the correlations between PLS-DA components and each one of the variables. They can thus be used to rank variables and/or perform feature selection by dropping the variable with the smallest associated VIP score or a given proportion of variables with the smallest VIP scores. This can help find a small subset of variable explaining the variation between groups and give a functional description of each group.

2.2.7.4.2 Random forests

2.2.7.4.2.1 The algorithm

Random forests[88] is a classification algorithm that can work with both continuous and categorical variables, such as a group label, as outcomes. Random forests consists of constructing a set of decision trees, each of them consisting of a series of logic rules determining the label of a sample. A voting strategy is used to aggregate the results from all the trees and give a final label to a sample.

The first step of the algorithm consists of creating randomised samples of the data with replacement, some samples will then be left out (referred to as out-of-bag samples) and used to compute accuracy measures.

For each one of these randomised samples, a defined number of trees will be built. The number of trees will be defined by the user. Each tree will start from the root and the samples will be split using a subset of randomly selected variables (usually of size equal to the square root of the initial variables number), finally, only the best one will be used to perform the split. Then, another layer of nodes is added and this process is repeated until the defined number of layers is reached.

When performing each split, considering the chosen variable is continuous, several thresholds will be considered and the best one will be selected according to Gini impurity, a method for quantifying homogeneity for a node. Briefly, it computes the misclassification frequency of a random sample if its class was chosen solely given the distribution of sample labels after performing the split.

Each tree will produce a group allocation that may be different according to other trees of the forest, a voting strategy can be applied to allocate a sample to a group, for example using the majority vote.

Robustness is ensured by bootstrapping applied to the samples and the random selection of variables.

Errors are estimated using out-of-bag samples that are used as a testing set and averaged for all trees.

2.2.7.4.2.2 Variable importance for Random Forest

As with PLS-DA, variable contributions can be extracted, making random forests a useful tool for group characterisation.

There are two main ways of computing variable importance in random forests. Out-of-bag samples can be used, the importance will then be computed using the accuracy decrease when the variable of interest is shuffled as opposed to when using the original order of values.

Decrease of Gini impurity can also be used to estimate variables importance, if the Gini impurity at a split decreases consequently then it means that the used variable contributed greatly in helping classifying the samples. This can be averaged over all the nodes including this variable to generate a global value over all trees.

2.3 Clustering time-series multi-omics datasets

In practice, most of the presented methods in their original or adapted form can be used on any type of data. However, any analysis performed on multi-omics datasets present a unique set of challenges that must be addressed to ensure the validity of the analysis and get a full picture of involved processes. Some issues are described in the following sections as well as potential directions to address them.

2.3.1 Biological and technical variability

2.3.1.1 Challenges and specificities

Because of the nature of the different experiments within an omics dataset, there will be technical variation between the different omics but also within each omics type[89]. When combining different types of data, this must be taken into account so that highlighted variation is not irrelevant to the analysis carried out or biased by a confounding variable (leading to a false association between the output and one or more input variables). A commonly seen example of a confounding variable is batch effect, which can result from the samples being analysed in different locations and/or processed by different technicians.

Not all biological variation will be relevant to an experiment. For example, if the gene expression profile of two conditions are compared but the samples were collected on different tissues, not all observed differences will be associated with the disease, as gene expression will vary across tissues. Variation in cell type will also impact such experiments. More specifically, when collecting peripheral blood (a medium of choice because of its accessibility, other samples such as organs often requiring more complex and/or invasive procedures to collect), the sample will be composed of different cell types and the measured effect will be the average effect measured over all the subpopulations of different cells. Cell composition can vary, and this will impact the observed results. These are important factors to account for.

2.3.1.2 Dealing with biological and technical variability

From the first step of any project, such variations should be minimised and will be crucial to the results obtained.

Confounding variables, batch effects or differences in cell composition, must be corrected for, depending on the data type there are specific tools designed to do so.

For example, for unknown variation in high-throughput experiments, such as RNA-Seq, the R package sva[90] uses surrogate variable estimation to correct for unknown and unwanted variation. Another popular too, also available as an R package, ComBat[91], uses empirical Bayes framework to correct for known technical batch effect.

Other techniques, that can be used regardless of the data type, can be used to correct for unwanted variations such as linear models to only extract the variation which is not explained by the defined factors/confounders.

2.3.2 Data types heterogeneity and relationships

2.3.2.1 Challenges and specificities

2.3.2.1.1 Data types heterogeneity

Because of the difference in technologies used to produce different omics data, the data type associated to each set might differ. For example, RNA-Seq data, after pre-processing, might be available as counts (being relative to the expression of each gene) whilst metabolomics data might be available as area-under-the-curve values from MS spectra. Such differences in data types will prevent direct comparisons from being made as data distributions of each data type will have different properties.

2.3.2.1.2 Relationships between different omics

As briefly mentioned in the introductory paragraph of section 2.1, the different omics data are different pieces of the same puzzle, we need all of them to paint a complete picture of a system and they are related to each other. Indeed, they all have complementary roles but cannot be deduced (at least completely) from one another, in part because the cascade of reactions might not be happening at the same rate at every level of the system and because markers are not solely influenced by genetic factors but also by environmental factors[89].

2.3.2.2 Dealing with data heterogeneity and relationships

Data types heterogeneity can be accounted for using different strategies. One of the simplest consists of scaling the variables (data type permitting) so that they can contribute equally and fairly to the analysis by applying for example a standard scaling to the variables. However, this is not always possible, and a more complex approach might be required. Data set can be modelled using different distributions suited to their type. For example, read counts might be modelled using a negative binomial distribution[92,93]. Negative binomial distribution is especially suitable because read counts are positive integers, describing rare events (as the number of genes is very high, the probability of having a read mapping to a given gene is low) and present a variance which can be much higher than the mean.

Correlations might identify related biomarkers from different omics layers but only a moderate proportion of variance has been shown to be shared across omics layers[94].

Several tools have been specially designed to account for the heterogeneity in data types but also to exploit the relationships between the different omics sets.

One of them, Multi-Omics Factor Analysis[95] (MOFA) models each data type using a distribution suited to its type. It then returns features explaining the variance present in the data and highlights variance unique to each data type and variance shared between different types of omics data.

Another tool, SNF[96] (Similarity Network Fusion) will generate a summary view of a dataset as a network of patient nodes. It will first generate a network for each data type separately and finally produce a consensus network taking into account all data types, their unique and shared information.

Relationships between genetic and epigenomic, transcriptomic, proteomic and metabolomic markers can be studied by QTL[97] (Quantitative Trait Loci) studies. Such studies can help discovering which regions of the genome are

important for a continuous trait (or an omics marker). Such approaches require a rigorously controlled setting that cannot be reproduced in human studies and thus, have shown their limits when applied to human subjects. As a result of the breadth of available markers and their genotyping in human subjects, alternative approaches can be taken to identify QTLs[98].

2.3.3 High-dimensionality

2.3.3.1 Challenges and specificities

It is often required or advised to have a number of variables smaller than the number of analysed samples. However, multi-omics data inherently tends to have many more variables than samples. This is an asset, as much information will be present in the dataset but can also be a problem as the complexity will increase and it will be more difficult to separate the signal from the noise present in the dataset. This is referred to as the curse of dimensionality.

Similarly, when building a model describing a dataset with many features, one would want a model with a small number of parameters that would help in interpretability and also would prevent overfitting (when a complex model becomes too dependent on the values of the training dataset and performs poorly on testing/new data).

2.3.3.2 Dealing with high dimensionality

Dimensionality reduction is a crucial step in any multi-omics data analysis and provides many advantages. It can remove irrelevant and/or redundant variation thus leading to improved models with higher accuracies, interpretable models and the required processing time/storage will be lesser.

In most cases the structure of a high-dimensional dataset can be represented in a space with a small number of dimensions while conserving most of the variation between samples.

Dimensionality reduction can be very useful way to represent a dataset with many dimensions, PCA[77] is a widely used technique for dimensionality reduction. PCA produces principal components constructed from linear combinations of the input variables and aiming at representing as much variance as possible. PCA can be heavily influenced by outliers and thus one must be cautious when applying PCA to a dataset.

Another class of dimensionality reduction is based on feature selection (briefly described in paragraph 2.2.2) where features will be either discarded from the analysis or kept for further analysis. However, this can only be applied when the outcome variable is known. It can be used after clustering has been performed, to select features that will be used to generate the classification/description models.

There are four major steps to a feature selection process[99,100]. The first one, the generation step, consists of generating a model using all features (referred to as recursive feature selection) or one feature (referred to as forward feature selection). Once a model is generated it is evaluated according to a chosen criterium, such as accuracy. The model features can be ranked given their respective importance in the model (see 2.2.7.4.1.2 and 2.2.7.4.2.2) and the corresponding features can be discarded/kept according to the results. This is repeated iteratively by removing/adding variables and generating the model again. Once an optimum has been reached, the procedure is stopped, and the final model kept. A validation step can be performed by comparing the reduced model to the full model for example.

2.3.4 Time-series data

2.3.4.1 Challenges and specificities

One single time point collected for a set of samples might not always be enough to finely study a biological process. Many omics studies involve the collection of multiple time points to gain insight into the dynamic processes

involved in the studied system. Both linear and cyclic processes can be studied as well using time-series data.

However, there are challenges related time-series data. Indeed, as different points will not all be independent from each other, this will violate assumptions for many tests and/or models. In many study settings, especially when looking at humans, time-series might be shifted as not all patients would be recruited at the same stage of their disease and this must be accounted for in the model.

2.3.4.2 Dealing with time-series data

There is a wide range of methods that can be used to analyse time-series omics data[101–103].

Different time points can be analysed separately and then summarised or compared between one another or to a reference time point.

A time-series might be summarised using a linear model or by computing the area under the curve. Both of these strategies will result in a reduced number of values to integrate.

Another option would consist of using classical statistical methods, for example by modelling the data using functions referred to as splines.

To compare patterns followed by the markers over time, machine learning approaches might be used.

Network-based approaches can be employed too and exploit correlations between the different markers/samples.

2.4 Conclusions

In this chapter I have laid the foundations for the work presented in this thesis in terms of data and analysis. This literature review was crafted as part of the

first section of this project and was updated throughout to take into account new developments and aspects which were deemed of interest to the study. Sections of this chapter aimed at introducing the subject and also provide a view of the work achieved in the area. It does not constitute an exhaustive view of the omics or clustering fields. Instead, it presents a broad overview of the current landscape and lays a foundation upon which my work is built.

The first section introduced major types of omics data, namely, genomics, transcriptomics, proteomics and metabolomics. Common collection methods as well as preparation, and bioinformatics processing were described to give an overview of what is commonly used in multi-omics studies and to give a basis when describing datasets analysed as part of this thesis project.

It then presented main aspects of a clustering analysis, initial considerations and parameters choice. Popular algorithms and/or of potential interest to the subject were described. Finally, validation of a clustering and specificities inherent to the area of biology were introduced. General directions for interpretation were given as well as some specific to omics data.

The final section concluded by presenting specificities and potential issues of clustering analysis when applied to multi-omics datasets and gave directions that could be exploited to palliate them.

In the next chapter, the AP datasets used for the main project will be described.

3. Chapter 3 – Acute Pancreatitis (AP), datasets and results

This chapter introduces the main project. It is divided in four sections. The first section presents background and context information related to acute pancreatitis (AP) and the associated disease model. This provides essential context for the description of executed analyses and results presented in sections two and three. More specifically, section two presents the two cohorts used to perform the analyses, data acquisition, processing, chosen methods and adopted strategies into more details. Results obtained along with interpretation are illustrated as part of section three. The fourth section is focused on conclusions arising from this project and how they could apply to diseases other than acute pancreatitis. This work has been submitted as a preprint on bioRxiv[104] (doi: 10.1101/539569) in which all contributing co-authors are listed.

3.1 Introduction

3.1.1 Acute pancreatitis

Acute pancreatitis (AP) is an inflammatory condition affecting the pancreas[105], an essential organ which is of major importance as it plays a crucial role in both the digestion process and the control of sugar level in blood. The worldwide incidence of AP is of 34 per 100 000 person-years[36] and it is the most common gastrointestinal cause for emergency hospital admission in the United States[106]. Etiologies are diverse and include choledocholithiasis (gallstones in the bile duct), excess ingestion of alcohol, trauma, pancreatic manipulation at endoscopy, viral infections, some venoms, and specific drugs. Inflammation of the pancreas can cause extrapancreatic damage and propagate to other organs such as lungs or kidneys resulting in multiple organ dysfunction syndrome (MODS)[107,108] in 1 in 4 AP-patients, requiring the admission to a specialised unit (intensive care unit or high dependency unit).

MODS results from systemic organ dysregulation[37] and one fifth of MODS cases will be fatal[38] (Figure 3.1).

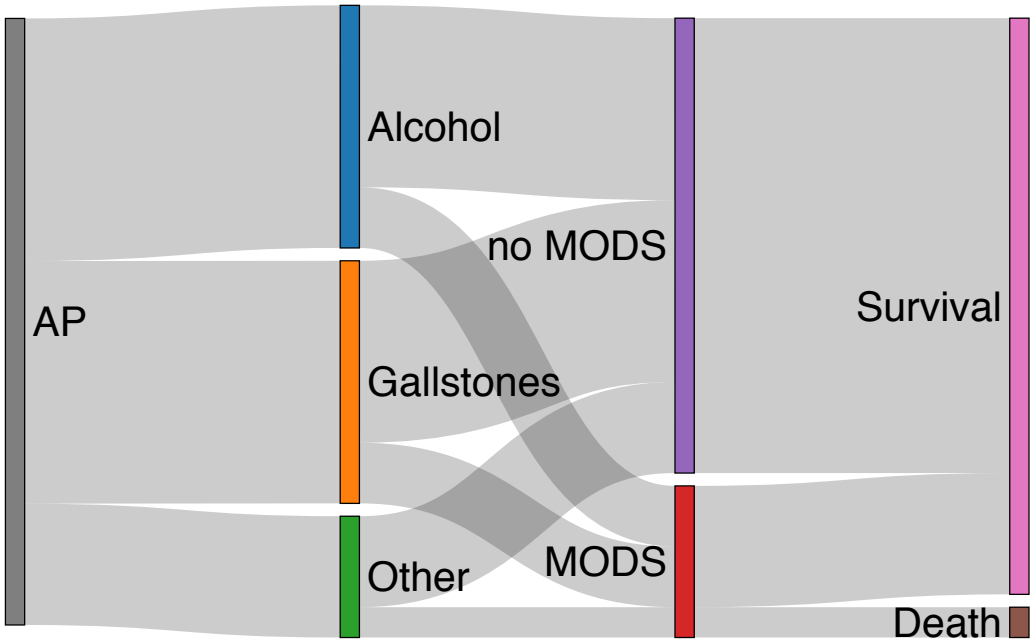


Figure 3.1 – AP, etiologies and outcomes (the widths of items are representative of reported percentages).

The molecular mechanisms underlying AP-MODS are not yet clearly understood and there is no treatment directly targeting AP. The current standard of care is supportive treatment only and can include pain control, ventilation and fluid resuscitation[109].

Recently, work was undertaken to develop a drug against AP[108]. Researchers identified an enzyme (kynurenine-3-monooxygenase, KMO) from the kynurenine pathway, involved in tryptophan catabolism and resulting in the production of NAD+ (nicotinamide adenine dinucleotide). This enzyme was shown to be a crucial element in the pathogenesis of AP-MODS in mice. Indeed, when mice lacking KMO activity were induced with AP, crucial organs (kidneys, liver and lungs) were protected against dysfunction. Following this a series of KMO specific inhibitors was developed with the goal of obtaining the same protection for patients with AP.

3.1.2 Hypothesis

Currently, the AP disease model is convergent, where the different etiologies lead to acinar cell damage (the acinar cells are the exocrine cells of the pancreas, producing digestive enzymes) and the resulting inflammatory responses is classified into one of three different levels, namely mild, moderate and severe, according to the extent of each individual patient's local complications and organ failure.

However, the amount of pancreatic damage is not directly linked to the occurrence of organ failure, nor its severity[11,12].

Developing the current disease paradigm further, it is highly likely that there is much more heterogeneity than the current model would suggest. Indeed, severity in AP cannot be predicted by simply considering the amount of pancreatic damage. Moreover, no clear pattern can be highlighted from routinely collected clinical data, and together, the lack of robust predictors makes individualised risk assessment difficult. Current prognosis scores include[110]:

- Ranson score[111], including 11 parameters in total, some of them requiring a measurement at 48 hours post admission.
- Glasgow-Imrie[112] score, including 7 lab measurements and age. It is similar to the Ranson score and requires some measurements to be taken post-admission.
- CRP[113] level can be used for severity stratification.
- APACHE II[111] score which is non-specific to AP and integrating 11 lab measurements, age and medical history.

In order to reconcile the shortcomings of current state-of-the-art, we hypothesized the existence of AP endotypes or molecular subtypes. We predicted that describing AP as a collection of endotypes would be relevant for the understanding of AP as well as for potential therapy strategies.

3.2 Materials and methods

3.2.1 The cohorts

3.2.1.1 IMOFAP

The IMOFAP cohort (Inflammation, Metabolism and Organ Failure in AP)[40] was a prospectively collected time-series cohort of samples and clinical data collected in a previous project by members of my group. As part of the IMOFAP cohort, 79 patients with suspected AP were recruited. Emergency attendees were recruited at the Royal Infirmary of Edinburgh between September and December 2013 using an alert system triggered when the following criteria were met, and a clinical verification was performed:

- Sudden abdominal pain with nausea and/or vomiting
- Serum amylase measurement value above the threshold of 100 IU/L (in order to capture those who were on the upslope of their serum amylase rise to meet the standard threshold of 300 IU/L)

Later on, the diagnosis was confirmed using the revised Atlanta criteria[114] (amylase > 300 IU/L with a clinical presentation consistent with AP) for 57 of the 79 initial patients. This allowed the recruitment of individuals as early as possible on their disease trajectory and sample consecutive time points shortly afterwards.

For these patients clinical and routine cytokine measurements were collected as well as peripheral blood samples at different time points between admission to hospital and up to 7 days after recruitment into the study (0, 3, 6, 12, 24, 48, 72 hours and 7 days). Peripheral blood was used (see 2.3.1.1) as a basis to perform omics measurements, namely metabolomics, proteomics and transcriptomics. Details are provided in Figure 3.2.

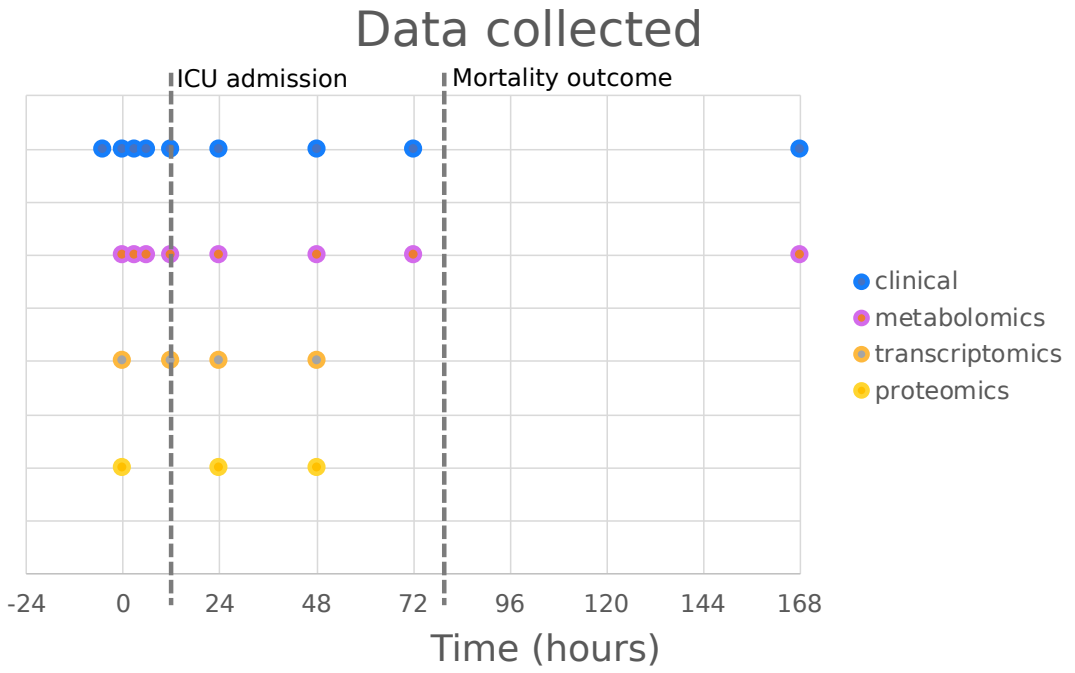


Figure 3.2 - Collected data details. Dashed lines indicate median time from admission to intensive care transfer when required (12 hours) and median time from admission to death for fatalities (82 hours)[38].

Every effort was made for all recruited individuals to have samples taken for all described time points, it was not always possible (due to patient refusal) and thus some samples could not be collected (Figure 3.3).

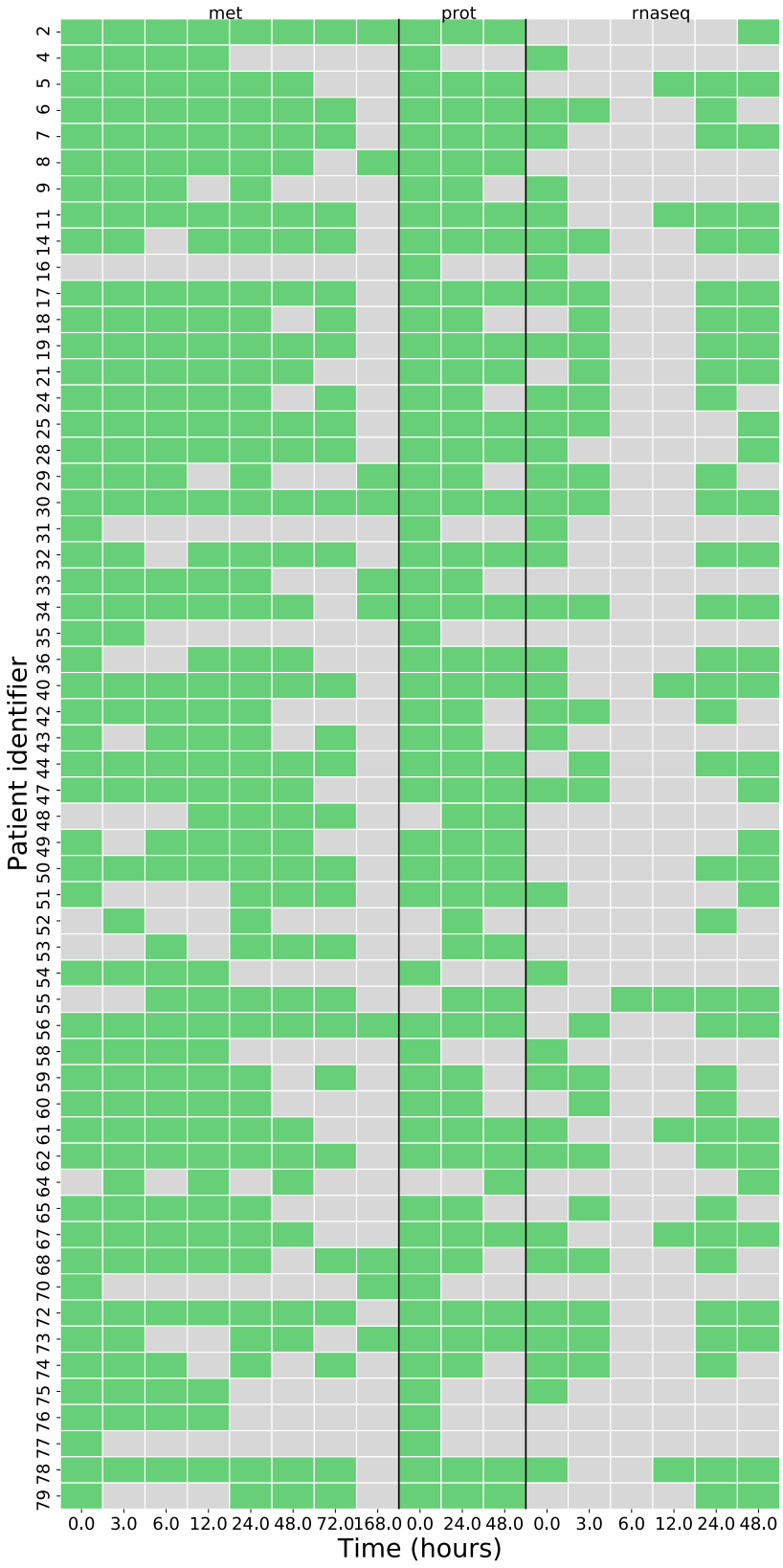


Figure 3.3 – Generated data. For each time point and data type, a green cell indicates that data was generated, a grey cell shows when data was not generated. 'met' refers to metabolomics, 'prot' to proteomics and 'rnaseq' to transcriptomics.

Out of the 57 patients, 54 had at least two metabolomics measurements (one measurement only was available for two individuals), 45 had at least two proteomics measurements (one measurement was available for 12 individuals) and 37 had at least two transcriptomics measurements (for 12 individuals only one measurement was performed).

3.2.1.2 KAPVAL

Data from a second AP cohort, KAPVAL (Kynurenine pathway in AP, VALidation), consisting of 312 AP-confirmed[114] individuals from the Royal Infirmary of Edinburgh (not overlapping with IMOFAP) recruited between February 2016 and January 2017 was available to me for the analyses. A serum amylase level above 300 IU/L was used to identify potential candidates which were then confirmed by specialist review of electronic health records. For those, metabolomics data was generated from serum and samples annotated with clinical and physiological data for a single time point corresponding to hospital admission.

3.2.2 The data

For both IMOFAP and KAPVAL cohorts, several data types were collected. Clinical and physiological measurements, and omics data were generated. To generate omics data peripheral blood samples were used.

Because of its accessibility, peripheral blood is a sample of choice in many analyses and can be used to identify and quantify biomarkers of interest. Plasma, serum and leukocytes can be extracted from such samples and used to sequence the genome and measure the transcriptome, proteome and metabolome, as described in section 2.1 of chapter 2.

3.2.2.1 Transcriptomics data

Using collected peripheral blood samples for four time points (0, 12, 24 and 48 hours after recruitment), the transcriptome of 49 individuals was measured using a rRNA depletion strategy (to measure total RNA) for a subset of samples (n=41) and a polyA selection strategy (to obtain mRNA) for the other subset (n=8). Samples were then sequenced for corresponding RNA using Illumina HiSeq 4000 system and generating 75 paired-end reads libraries. Reads were stored as FASTQ files. Quality control was performed using FASTQC (v0.11.2) and reads were filtered and trimmed accordingly using cutadapt (v1.4, 3' end trimming with a cutoff of 20 and reads shorter than 25 were discarded). Reads were aligned against the hg38[115] version of the human genome assembly using STAR[116] (v2.5.0a). Gene expression was estimated using read counts as a proxy and previously aligned reads using featureCounts (v1.5.2)[117].

To account for the difference in library preparation between the two subsets of samples (total RNA vs mRNA), a two-step strategy was applied. The strategy consisted of filtering the counts matrix to only keep protein-coding elements and applying a batch removal algorithm (using NOISeq[118] R library). PCA plots of counts before and after batch effect are shown in Figure 3.4 and Figure 3.5. A normalisation procedure, converting the filtered corrected counts to FPKM (Fragments Per Kilobase of transcript per Million reads mapped), was applied and consisted of a normalisation by sequencing depth followed by a normalisation by gene length. A Z-score scaling was finally applied to permit comparisons between samples.

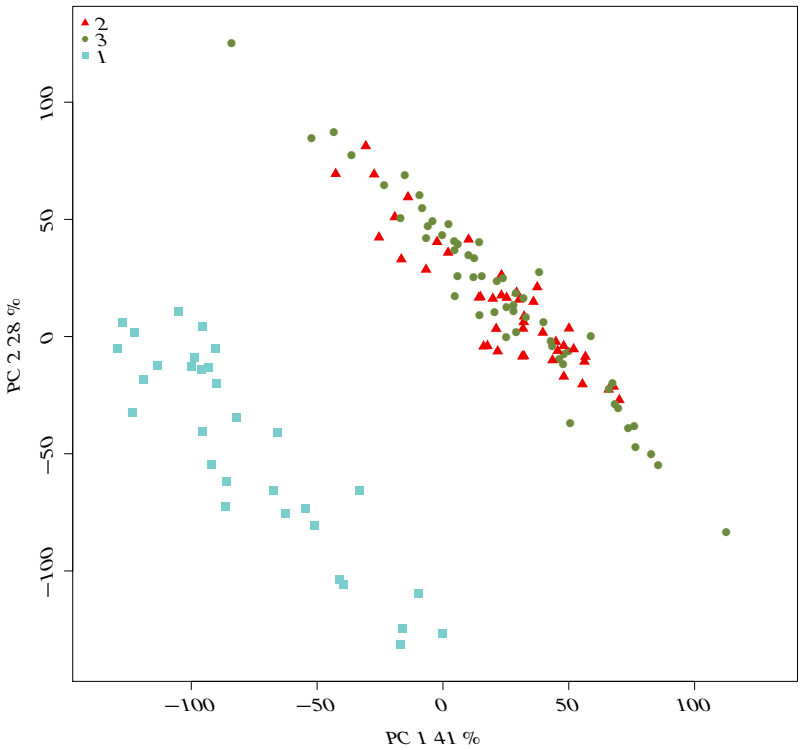


Figure 3.4 – RNA-Seq counts values from featureCounts output, representing only protein-coding genes. The different shapes/colours represent the batches (1 correspond to mRNA samples and 2 and 3 to total RNA samples).

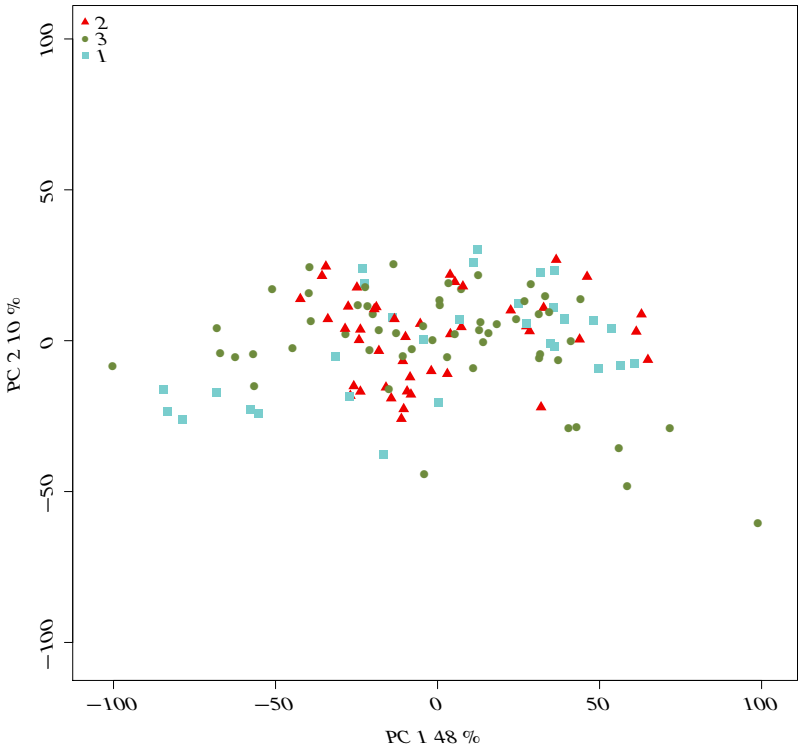


Figure 3.5 – RNA-Seq counts values from featureCounts output, representing only protein-coding genes, batch-corrected and FPKM-normalised. The different

shapes/colours represent the batches (1 correspond to mRNA samples and 2 and 3 to total RNA samples).

The final gene set, after pre-processing, consisted of 19 766 genes.

3.2.2.2 Proteomics data

Serum extracts were obtained from blood samples at time points 0, 24 and 48 hours after recruitment into the study in order to measure the proteome for 57 IMOFAP samples. Abundant proteins in samples were depleted (to reduce the complexity of serum samples which will be dominated by a small number of proteins not of interest here) then samples were denatured, alkylated and digested. Several samples were analysed simultaneously using tandem mass tags labels and RPLC-MS/MS/MS and spectra were produced. Proteins were identified with MaxQuant, performing a search against a UniProt-based human proteome, abundances were estimated from generated spectra. Protein species with 90% or more missing values for the cohort were discarded. Remaining missing values were imputed using the minimum value for each compound as missing values would imply values below the detection limit. During data visualisation, it was noted that samples clustered according to run groups, to prevent run bias, values were corrected using ComBat to remove non-relevant variation between samples. Measurements were then transformed into Z-scores, as with RNA-Seq data. The final set consisted of 371 protein variables.

3.2.2.3 Metabolomics data

Serum extracts were obtained from collected peripheral blood samples in the IMOFAP cohort for all time points between recruitment and up to 7 days after and used for metabolomics measurements for 56 IMOFAP individuals. Protein depletion was performed. They then underwent UPLC-MS/MS analysis to identify and quantify metabolites present. Metabolon’s proprietary software was used to identify compounds and abundance was estimated using area-

under-the-curve. As previously, metabolites with 90% or more missing values across the sample set were not retained. Remaining missing values were imputed using a minimum-value strategy and thus replacing missing values by the minimum values (detection limit) for that metabolite. A Z-score scaling was finally applied to all metabolites. In total, 651 metabolites were retained for further analysis for the IMOFAP cohort. For the KAPVAL cohort, data was processed as with the IMOFAP cohort, however, the aim was to use KAPVAL as a validation set and thus, we retained only metabolites that were in common between both datasets, resulting in 426 metabolites for this cohort (n=312 individuals).

3.2.2.4 Clinical measurements

Collected blood samples were annotated for clinical and measurements for all time points, when possible. Variables collected include:

- Age
- APACHE II score
- BMI
- Cause of pancreatitis
- Charlson index
- Critical care admission status
- eGFR
- Ethanol consumption status
- Ethanol excess
- Gender
- Inhospital mortality
- Length of stay
- Mean arterial pressure
- Modified MODS score
- Mortality at 30 days
- Organ dysfunction occurrence

- PaO2/FiO2 ratio
- Previous AP episode occurrence
- Recruitment source
- Smoking status
- Systolic blood pressure

3.2.2.5 Blood measurements

Collected samples were used to quantify blood markers. Variables collected (some of which were used to compute scores listed above) include:

- 3-hydroxyanthranilic acid (nanograms per millilitre)
- 3-hydroxykynurenine (nanograms per millilitre)
- Alanine aminotransferase (units per litre)
- Albumin (grams per litre)
- Alkaline phosphatase (units per litre)
- Amylase (units per litre)
- Asparate aminotransferase (units per litre)
- B7H1 (picograms per millilitre)
- Base excess (mEq per litre)
- Basophils ($x10^9$ per litre)
- Bicarbonate (millimoles per litre)
- Bilirubin (micromoles per litre)
- Calcium (millimoles per litre)
- Cancer antigen 15-3 (picograms per millilitre)
- Cardiac Troponin (picograms per millilitre)
- CD 163 (nanograms per millilitre)
- CD40 ligand (nanograms per millilitre)
- Chemerin (nanograms per millilitre)
- Creatinine (micromoles per litre)
- CRP (milligrams per litre)
- D dimers (micrograms per litre)
- Eosinophils ($x10^9$ per litre)

- Free T4 (picomoles per litre)
- Gamma glutamyl transferase (millimoles per litre)
- Glucose (millimoles per litre)
- Haematocrit (percent)
- Haemoglobin (grams per litre)
- High density lipoprotein (millimoles per litre)
- IL-10 (picograms per millilitre)
- IL-17a (picograms per millilitre)
- IL-22 (picograms per millilitre)
- IL-6 (picograms per millilitre)
- IL-8 (picograms per millilitre)
- IL1 beta (picograms per millilitre)
- Insulin (picograms per millilitre)
- Insulin C-peptide (picograms per millilitre)
- Interferon gamma (picograms per millilitre)
- Kynurenic acid (nanograms per millilitre)
- Kynurenine (nanograms per millilitre)
- Lactate (millimoles per litre)
- Lactate dehydrogenase (units per litre)
- Low density lipoprotein (millimoles per litre)
- Lymphocytes ($x10^9$ per litre)
- Magnesium (millimoles per litre)
- Mean corpuscular haemoglobin (picograms per cell)
- Mean corpuscular volume (femtolitres)
- Monocytes ($x10^9$ per litre)
- Neutrophils ($x10^9$ per litre)
- Partial pressure of carbon dioxide (kPa)
- Partial pressure of oxygen (kPa)
- pH
- Phosphate (millimoles per litre)
- Platelet count ($x10^9$ per litre)
- Potassium (millimoles per litre)

- RAGE (nanograms per millilitre)
- Red cell count (x10^9 per litre)
- SDF 1 alpha (picograms per millilitre)
- Sodium (millimoles per litre)
- TFF3 (nanograms per millilitre)
- Total plasma cholesterol (millimoles per litre)
- TRAIL (picograms per millilitre)
- Triglycerides (millimoles per litre)
- Tryptophan (nanograms per millilitre)
- TSH (mU per litre)
- Tumour necrosis factor alpha (picograms per millilitre)
- Urea (millimoles per litre)
- White cell count (x10^9 per litre)

3.2.3 Methods

We chose to apply unsupervised techniques to highlight subgroups in our datasets. This allowed to analyse data with no hypothesis as to which mechanisms might be involved or which data types might drive the variation in the dataset and to prevent this from biasing the results.

3.2.3.1 Tools and data

To carry out analyses, Python (version 3.5) and R (version 3.3.2) were used. Libraries used included dtwclust in R, numpy, pandas, rpy2, scipy, sklearn and statsmodels in Python.

The script used to compute distances between patients is available as appendix A.1.

As mentioned in section 3.2.2, Z-scores were used as input before running chosen analysis methods.

3.2.3.2 Clustering

3.2.3.2.1 Methods used to generate distance matrices

The following methods were used to generate distances between all samples and allowed to generate pairwise distance matrices.

3.2.3.2.1.1 Single time point Euclidean distances

First, we selected a single time point, here time point 0, corresponding to the recruitment time point and time points 24 and 48 to compute distances between included samples. Using pre-processed data, Euclidean distance was used to obtain a measure of dissimilarity between samples.

3.2.3.2.1.2 Area Under the Curve and PCA (AUC-PCA)

To obtain a single value per variable for a time series in a selected sample, we computed area-under-the-curve (AUC, Figure 3.6) values using the trapezoidal rule.

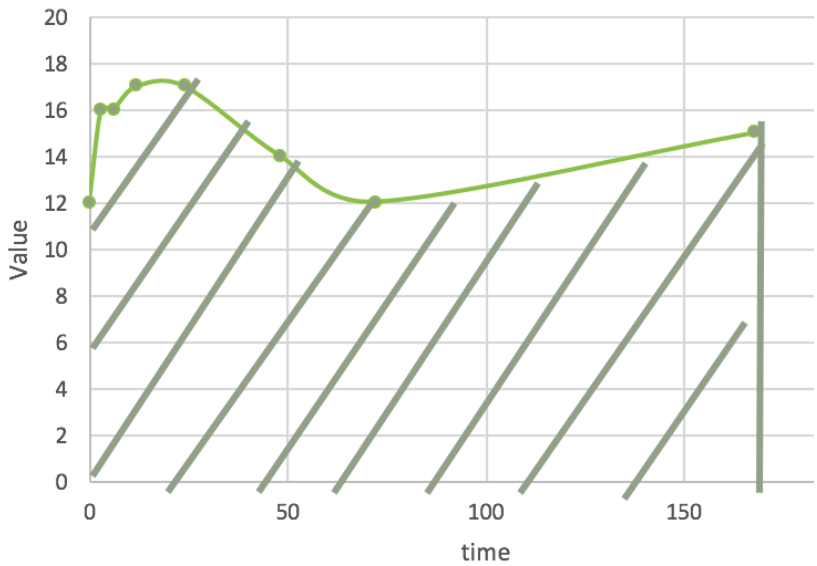


Figure 3.6 – AUC for a given time-series. Data values are represented by grey points and the hashed area represents the AUC.

This was repeated for all variables and samples. Computed AUC values allowed to summarise each time series as a single value expressing the cumulative magnitude over time. This allowed us to process values as being independent, broadening the analysis strategies that could be used, independence being a common assumption in many statistical analyses. Values were normalised to take into account the differences in length between some of the time series. Obtained values were represented using principal components as shown in Figure 3.7.

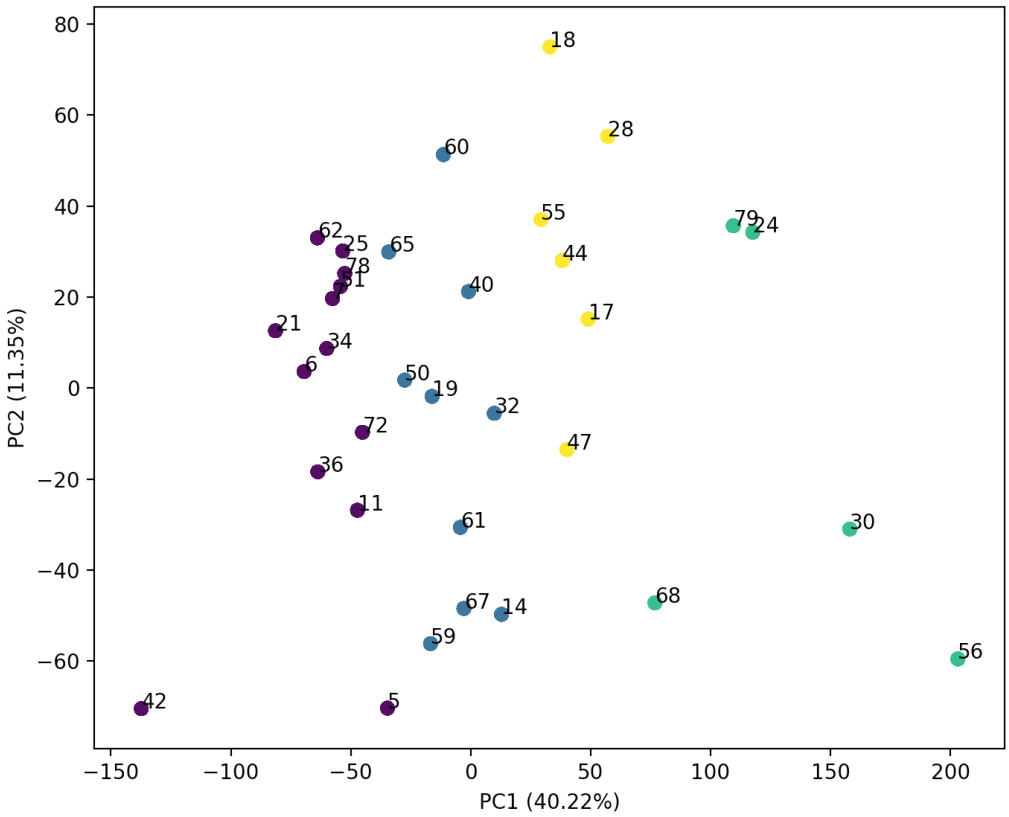


Figure 3.7 – Example PCA plot with potential clusters identified using different colours. Represented variance is reported for each axis.

More specifically, the first two components, representing the greatest part of the variance, were selected. Using coordinates in this 2-dimensional space, Euclidean distances were computed between all pairs of samples and were weighted according to the represented variance for each one of the principal components. This permitted to give more weight to a distance on principal component 1 as compared to principal component 2. Principal component

analysis (PCA) was chosen because of its ability to represent data in a lower dimensional space than the initial one and because it is hypothesis free[77]. Indeed, PCA does not make assumptions related to the stratification of the data. PCA is solely sensitive to the correlation structure present in the data.

3.2.3.2.1.3 Trajectory through PCA space

To integrate in more details patients trajectories over time, I used an approach described in a published study[119] which demonstrated that trajectories of samples through selected components could be helpful in clustering individuals. Using all selected time points from all selected individuals, a projection into a 2-dimensional PC space was done. If data was missing between two time points for an individual, I performed linear interpolation. The aim was to characterise their trajectory through this space and use this to compute distances between individuals. To define the trajectory of a patient through the PC space I defined the direction taken between each pair of consecutive time points for this specific patient and coded the direction using integer values of 1 to 4 corresponding to a space division into four quadrants. This is illustrated in Figure 3.8.

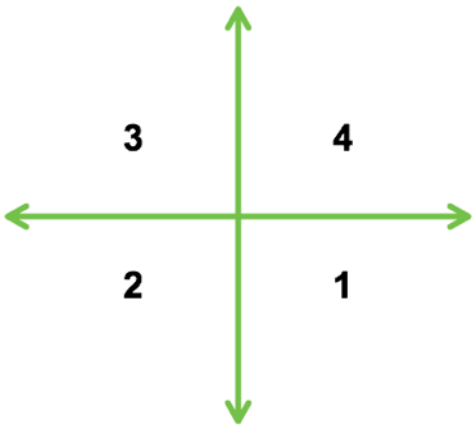


Figure 3.8 – Possible directions and associated values.

This procedure was repeated for every patient resulting in one direction vector per individual. Hamming distances between all pairwise combinations of

direction vectors were computed by counting the number of different values, element-wise. These distances were used as a proxy for dissimilarity in trajectories. We thus combined the advantages of PCA and trajectory analysis.

3.2.3.2.1.4 Dynamic time warping

Our last tested strategy consisted of using dynamic time warping distances[120] as distance measure between individuals. Such distances were computed using dtwclust in R. The algorithm started by considering each pair of samples, for each variable, a matrix was generated and reported the difference in magnitude, without considering the time axis, between all possible pairs of time points. For each pair of individuals, a single matrix was obtained by summing all variable matrices, element-wise. This matrix was then used to perform the warping procedure, during which the time axis would be warped in order to minimise the distances between the two series. It consisted of finding a path in each matrix so that the summed number was minimal. The path started necessarily from the matrix element corresponding to the first points of each series and ended when the matrix element corresponding to the last points of each series was reached. An example of warping is represented in Figure 3.9.

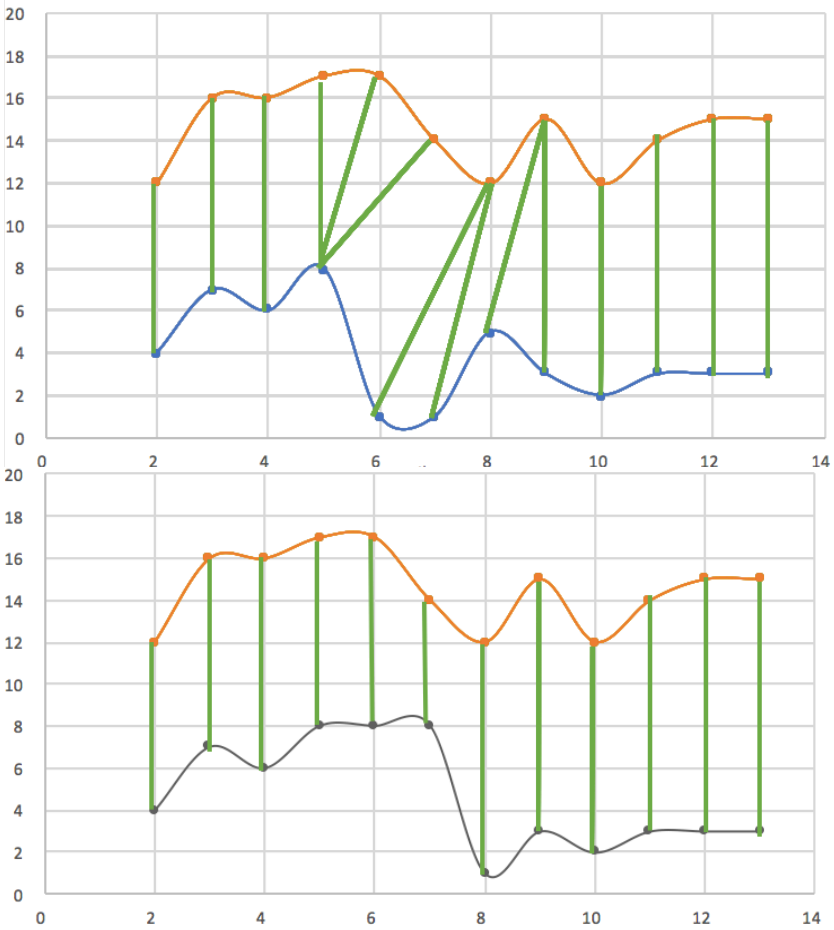


Figure 3.9 – Example of alignment produced using the dynamic time warping algorithm. The orange and grey/blue curves represent two patients for which the values of a variable were measured and are represented on the y axis. The top figure represents the original data and the bottom figure the alignment produced (using the orange curve as the reference).

This was done on a summarised matrix, in other words, the optimum matrix, as defined by the warping procedure, represented a consensus alignment minimising the summed differences in magnitude between the two compared individuals, for all variables, rather than one alignment per variable. The summed number can then be used as a distance measure between two patients. To allow a fairer comparison we chose to compare equally long time series with equally spaced time points and thus linear imputation was performed when required.

3.2.3.2.1.5 Advantages and disadvantages of presented methods

A summary of the main advantages and disadvantages are presented for the four different strategies in Table 3.1.

Table 3.1 – Methods advantages and disadvantages.

	Single time point Euclidean distances	**AUC-PCA**	**PCA + Trajectory analysis**	**Dynamic time warping**
+	Results are easier to interpret, and it can be done for any time point	Results are easier to interpret	Dynamic profiles can be compared	Dynamic profiles can be aligned and compared
-	Only one time point is used	Dynamic dimension reduced to one value, and similar values can be obtained from very different curves	Shift in time-series will cause bias	Chosen time-shift is the same for all variables

All presented methods were selected to generate dissimilarity values between pairs of samples and we aimed at comparing the results obtained in order to select the most relevant.

3.2.3.2.2 Clustering strategy

Once the different dissimilarity matrices were obtained, I performed clustering to highlight potential subgroups of interest. Hierarchical clustering and Ward's method were used. Ward's method forms groups by minimising the sum-of-

squares within each group and is commonly used for Euclidean-based distances. Not all presented methods are Euclidean-based, however, it has been used successfully for other types of distances[121–123]. For consistency, it was thus used for all distance matrices obtained.

Any number of clusters between 1 and the number of elements clustered can be extracted from the type of clustering analysis described here. For this reason, an optimum number was chosen according to the stability of each solution. As one cluster would not have been informative and too many would have resulted in singletons, for which only little information could have been extracted, we restricted the number of clusters between two and twenty.

Stability was assessed using bootstrapping combined with a Jaccard index (as defined in section 2.2.6.2.2) to estimate the results quality[80,124]. For each number of clusters between two and twenty, we sampled the original samples with replacement to create one hundred new input datasets. Each one of the new datasets was processed as the original dataset to obtain a partition with the corresponding number of clusters. I then compared the newly generated solution with the original solution, assessing the similarity by computing the overlap between pairs of most similar clusters and using the Jaccard index. An average value was then computed to assess the stability of the results. The optimum number of clusters was chosen by maximising the average Jaccard index. To maximise interpretability, solutions with one or more groups presenting less than three individuals was not retained for further analysis.

3.2.3.3 Evaluation

To evaluate the different solutions we defined a priori criteria which are described in this section.

3.2.3.3.1 Assessment strategy

3.2.3.3.1.1 Statistical robustness

Stability results obtained from the bootstrapping strategy presented in the previous section were used to select solutions for further analysis. If a solution had an associated Jaccard index higher than others, its structure was deemed to be more robust to change and of greater interest.

3.2.3.3.1.2 Biological plausibility

To quantify biological plausibility, we ran compound set enrichment analyses on partitions having passed the stability testing. We hypothesised that the highlighted partitions could be, to some extent, detected using solely time point 0 and thus was used to perform the enrichment analyses. Moreover, for the partition to be of maximum utility, groups would have to be detected as early as possible.

Compounds identifiers were converted, when required, using the biomaRt package in R.

We obtained compound sets from two R packages, GAGE[125] for gene and protein data (KEGG-based data[81]) and MetaboAnalystR[126] for metabolite data. FANTOM5 co-expression gene sets, describing genes expressed in different cell types, were also downloaded from the project's data[127,128].

Using the optimum partition, the aim was to determine whether a subset of compounds (with a similar function or part of a same biological process) presented an association with the group labels obtained from the clustering. To test this, we used generalised linear models and a p-value fusion strategy.

For each element of a given compound set, we fitted a model using the group label as a fixed effect and the values of the corresponding compound as the response variable. We then compared a given model to the null model (using only the intercept) and performing a likelihood ratio test (using the *anova*

function with test='LRT' in R) to assess the effect of the group label on the variable values. This returned a single value that allowed us to classify the variable as statistically associated with the group labels and thus the partition of interest.

For each compound set, composed of several variables, we combined the p-values obtained using the likelihood ratio tests to obtain a single p-value. We gave weights to every compound according to the number of other sets it was present in. The weight was computed using the inverse of the number of sets a compound was part of. This allowed to minimise common sets from biasing the results.

We used this summarised p-value as a way to quantify enrichment of different compound sets. For every type of compound set, both gene/protein and metabolite sets, we counted the number of significant elements (using a threshold FDR-corrected p-value of 0.05) and used this number to quantify the biological plausibility of a solution.

Moreover, FANTOM5 data was used to identify involved cell types as it reported cell types gene signatures. The same enrichment strategy and cut-off value were used to identify and count significant elements.

This provided us an overview of the biological processes that may be involved and select the most meaningful clustering partition.

3.2.3.3.2 Enrichment analysis

3.2.3.3.2.1 Variable selection

To describe biological processes specifically associated to each group, rather than looking at the problem globally, as described in the previous section, Partial Least Square Discriminant Analysis (PLS-DA) was used. PLS-DA is a dimensionality reduction algorithm which can also be used to perform feature selection and classification.

Here, we used PLS-DA to highlight candidate biomarkers and biological processes uniquely associated to each one of the identified clusters. The strategy consisted of creating *k* models, one for each cluster, to highlight differences between each one of them and all others, regardless of their label. As described in section 2.2.7.4.1, data was projected onto a new space, with a number of components chosen by the user. Components were generated by maximising covariance with group labels. This allowed to generate discriminant components. Using calculated components, Variable Importance in Projection (VIP) scores can be computed using weights and will be related to how much each variable is involved in group discrimination. Variables can then be ranked and selected using a VIP threshold[129]. Selected elements were finally used to perform compound set enrichment analysis.

3.2.3.3.2.2 Enrichment procedure

To analyse VIP-filtered lists, a Reactome gene set was downloaded from Reactome's website (https://reactome.org/download-data, lowest level pathway files). We generated compound sets containing all integrated data types, namely, transcriptomics, proteomics and metabolomics. These sets were then filtered to only include those containing at least 10 elements and no more than 500 elements, others were deemed uninformative and would have produced less robust results.

To determine if a compound set was significantly represented in the VIP-filtered lists, Fisher's exact test was run using the number of matches in the list and the total number of compounds from the original set (which was used as a background set). P-values were generated and corrected for multiple comparisons using an FDR-based correction and applying a threshold of 0.001 to limit the number of elements for visual representation and inspection.

This was reproduced for time points corresponding to 24 and 48 hours sampling times.

3.2.3.3.3 Data visualisation

An interactive webpage (available at http://baillielab.net/pancreatitis/, username: pancreas and password: review) was built to allow data visualisation. The webpage was created using D3[130]. It permitted to visualise AUC values per group per variable and average Z-score values between time points 0 and 48 per group. To limit the number of displayed variables, different datatypes were displayed separately, using a dropdown selection box and only a subset of variables was displayed. Selected variables were chosen according to their associated maximum VIP value across all four groups using a threshold of 2. As clinical variables and cytokines measurements had no VIP values, as they were not integrated in the PLS-DA models nor in the clustering procedure, we filtered them according to ANOVA results using a threshold of 0.05.

3.2.3.4 Reproducibility

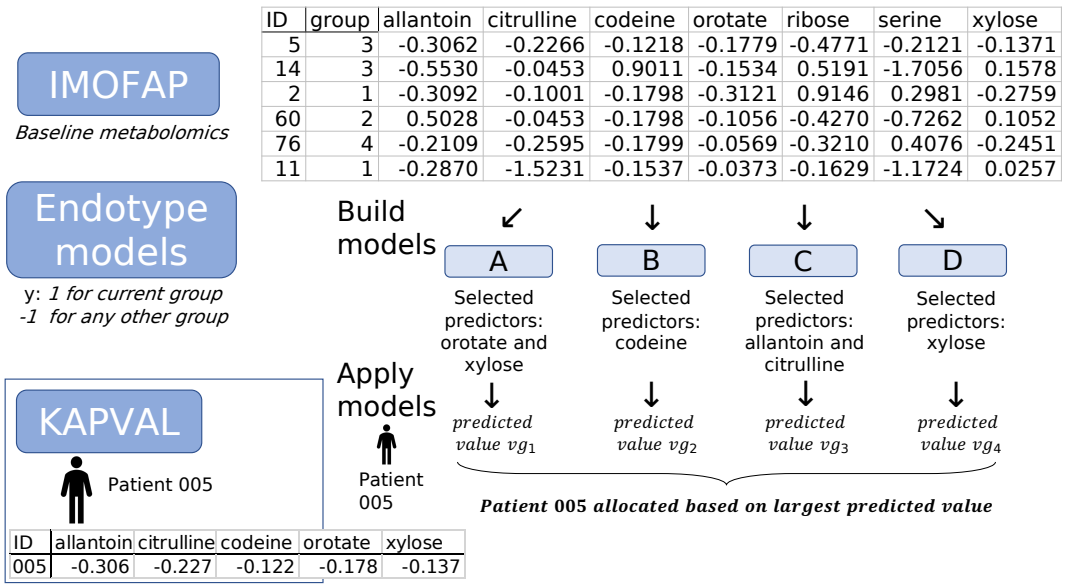
3.2.3.4.1 Reproducibility in an independent dataset (KAPVAL)

To test whether groups highlighted were relevant we aimed at demonstrating that they could be identified from an independent AP dataset. To do so we used the KAPVAL (Kynurenine pathway in AP, VALidation) cohort comprising 312 AP-confirmed individuals and as described in section 3.2.1.2.

After discarding drug metabolites and using the 413 metabolites in common with the IMOFAP cohort, we generated four PLS-DA models, similarly to the procedure described in section 3.2.3.3.2.1. Maximising accuracy when training the models with IMOFAP data, we chose to integrate 3 components. To maximise interpretability and to prevent over-fitting, we limited the number of predictor variables to 25. The optimum number, between 3 and 25, was chosen according to accuracy values.

To classify KAPVAL individuals we applied each one of our four models to each individual and allocated them to the closest matching group, given the highest predicted value.

An overview of the process is described in Figure 3.10.



ID	group	allantoin	citrulline	codeine	orotate	ribose	serine	xylose
5	3	-0.3062	-0.2266	-0.1218	-0.1779	-0.4771	-0.2121	-0.1371
14	3	-0.5530	-0.0453	0.9011	-0.1534	0.5191	-1.7056	0.1578
2	1	-0.3092	-0.1001	-0.1798	-0.3121	0.9146	0.2981	-0.2759
60	2	0.5028	-0.0453	-0.1798	-0.1056	-0.4270	-0.7262	0.1052
76	4	-0.2109	-0.2595	-0.1799	-0.0569	-0.3210	0.4076	-0.2451
11	1	-0.2870	-1.5231	-0.1537	-0.0373	-0.1629	-1.1724	0.0257

ID	allantoin	citrulline	codeine	orotate	xylose
005	-0.306	-0.227	-0.122	-0.178	-0.137

Figure 3.10 - Schematics representing the assignment process for KAPVAL samples to one of the four endotypes identified in IMOFAP cohort using PLS-DA models.

As any sample would have been allocated to a group, it was necessary to check that the signal between the groups identified in IMOFAP was similar to the corresponding groups formed of allocated KAPVAL samples. I calculated, for each group, the average value of every variable not included in the models (369 metabolites). I compared, using a strategy inspired by Sweeney *et al*[131], Spearman's correlation coefficients (computed by comparing ranks) between corresponding groups and associated p-values (based on a t-distribution).

Additionally, in-group proportion (IGP) values and associated FDR-corrected p-values for the 413 included metabolites were calculated. The IGP of a subgroup is the proportion of samples having their nearest neighbour allocated to that same subgroup. The nearest neighbour is determined using Pearson's correlation coefficient.

Both strategies aimed at determining if the same signal was present in both cohorts.

3.2.3.4.2 Comparison with an external dataset

To compare obtained groups to data from another condition, related to AP but distinct (severe cases of AP can result in ARDS, other causes included sepsis, trauma and major surgery[132]), we chose Acute Respiratory Distress Syndrome (ARDS) endotypes described in another study[39]. Out of our 57 AP-confirmed patients, 6 (no measure was available for one individual) met the Berlin definition of ARDS[133] for the recruitment time point. Two ARDS endotypes were described as part of this study, for each one of them a ranking of variables were available and consisted of routinely measured variables. These rankings were used and compared to rankings for identified AP endotypes using Spearman's correlation. 19 variables out of 31 (8 physiological, 9 clinical biochemical, and 2 cytokine variables that were not used to produce the clusterings) from the study could be matched to variables available as part of the IMOFAP cohort. For each ARDS cohort, Spearman's correlation coefficients, along with FDR-corrected p-values were generated using scipy and statsmodels in Python.

3.2.3.4.3 Comparison with results from an independent tool (MOFAtools)

To assess the validity of our strategy, we compared the optimum stratification obtained to the solution obtained using a multi-omics data analysis tool, MOFA[95]. Briefly, MOFA highlights, in multi-omics datasets, variables explaining variation and variation patterns using factor analysis.

Using the generated factors, clustering can be performed using R package MOFAtools with variable values computed (single time point or AUC data). We used default parameters along with an additional filter of 1% applied to factor explained variance for all omics, below which, factors were not kept for further

analysis. To perform a fair comparison, the number of clusters was chosen equal to that of the optimum solution. Overlap between partitions was computed using the Jaccard index.

KAPVAL labels, as obtained using PLS-DA models (described in section 3.2.3.4.1), were compared to the ones obtained using MOFA. The overlap was computed using Jaccard index as well. The aim was to determine if a similar structure could be highlighted using MOFA.

3.3 Results

3.3.1 Clinical cohorts and measurements

3.3.1.1 IMOFAP

For the 57 Atlanta-confirmed AP individuals, before running the analysis, a further exclusion criterium was applied and consisted of excluding individuals with an interval between symptom onset to recruitment greater than 200 hours. This was done to prevent the introduction of bias as individuals in this situation would have been more likely to be on the late phase of their disease trajectory. The applied filtering of samples, describing exclusion criteria, is presented in Figure 3.1.

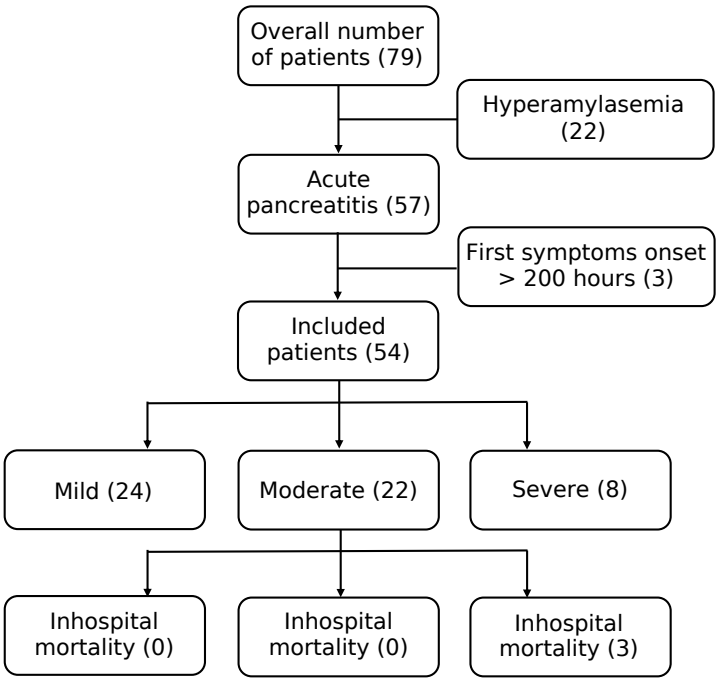


Figure 3.11 - Study flow chart for included patients from the IMOFAP study showing filtering process, reasons for exclusion and some demographics.

For the 54 pre-selected patients from the IMOFAP cohort, 24 had mild AP, 22 had moderately severe AP and 8 had severe AP requiring critical care. The number of deaths was equal to 3.

The median time interval between symptom onset and recruitment, for the 54 pre-selected individuals, was 21.3 hours (IQR 40.8 hours, Q1-Q3 13.5-54.4). More detailed demographics are presented in Table 3.2.

Table 3.2 - IMOFAP demographics. Summary clinical data for included participants (n=54).

Number of patients		54
Gender	Male	55.60% (n=30)
Age (years)	Median IQR (Q1-Q3)	56.95 28.98 (47.40-76.38)
BMI	Median IQR (Q1-Q3)	27 7.75 (23-30.75)
Source of recruitment	A&E Other	88.89% (n=48) 11.11% (n=6)
Length of hospital stay (days)	Median IQR (Q1-Q3)	5 4.75 (3-7.75)
Aetiology	Gallstones Alcohol Other	44.44% (n=24) 33.33% (n=18) 22.23% (n=12)
Charlson index (time point 0)	Median IQR (Q1-Q3)	2 3 (1-4)
Inhospital mortality (binary)	1	5.56% (n=3)
Time onset recruitment (hours)	Median IQR (Q1-Q3)	21.29 40.84 (13.54-54.38)
Alcohol use	Current Previous None	57.41% (n=31) 5.56% (n=3) 37.04% (n=20)
Smoking	Current	48.15% (n=26)

	Previous None	18.52% (n=10) 33.33% (n=18)
Critical care admission (binary)	1	7.41% (n=4)
APACHE II day 1	Median IQR (Q1-Q3)	10 5 (8-13)
Previous AP	0 1 2 3 or more	68.52% (n=37) 22.22% (n=12) 5.56% (n=3) 3.70% (n=2)
CRP (mg/L) (time point 0)	Mean SD	77.39 93.54

In terms of time points, as the median time interval from admission into hospital to intensive care transfer for those who needed it was 12 hours and the median time interval from admission to death for fatalities was 82 hours[38], we chose to focus on the time points between recruitment into the study and up to 48 hours after that.

As we were especially interested in the dynamic dimension of the IMOFAP cohort, when analysing data for more than one time point, we selected samples given the completeness of the multiomic set for an individual, across different time points. Logically, we also discarded samples with less than two time points for a data type. This resulted in a cohort with 34 patients (consisting of 16 mild, 13 moderate and 5 severe AP cases).

For single time point data analysis, we selected individuals with a complete set for the selected time point. For time point 0, this resulted in 40 patients being selected for analysis (consisting of 22 mild, 14 moderate and 4 severe AP cases).

3.3.1.2 KAPVAL

All samples from the KAPVAL cohort consisted of AP-confirmed individuals and thus there was no need for filtering. Symptom onset data was not available for this cohort.

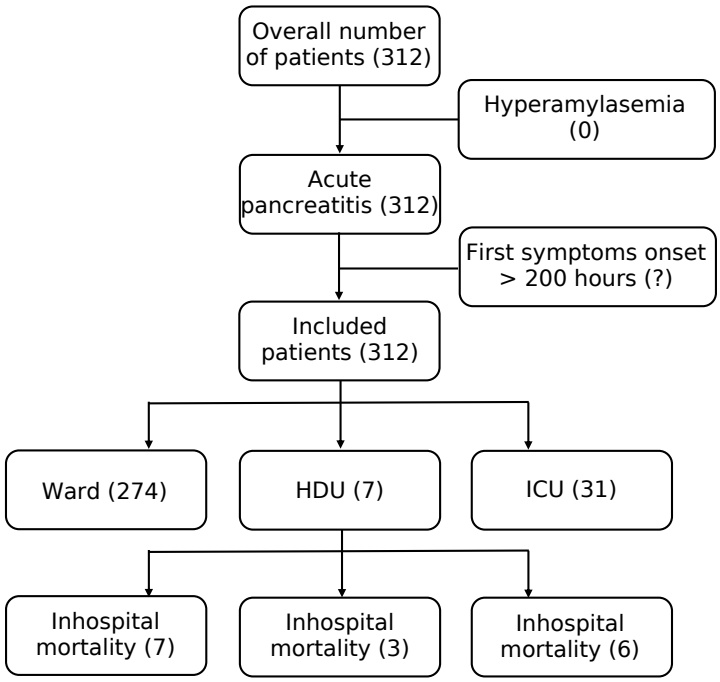


Figure 3.12 - Study flow chart for included patients from the KAPVAL study showing filtering process, reasons for exclusion and some demographics.

For included participants and at admission time point, 274 were in wards, 7 in high-dependency units and 31 in intensive care units. The number of deaths was equal to 16, of which 7 corresponded to patients initially in wards, 3 in high-dependency units and 6 in intensive care units.

Some demographics for the KAPVAL cohort are presented in Table 3.3.

Table 3.3 – KAPVAL demographics. Summary clinical data for included participants.

Number of patients		312
Gender	Male	46.79% (n=146)
Age (years)	Median IQR (Q1-Q3)	56.00 30.25 (40.75-71.00)
Inhospital mortality (binary)	1	5.13% (n=16)
Critical care admission (binary)	1	12.18% (n=38)
CRP (mg/L)	Mean SD	47.62 79.85

3.3.2 Evaluation of results

AUC-PCA produced the optimum result, the total percentage of variance explained by the two selected components was 51.5% (40.2% for principal component 1 and 11.3% for principal component 2). Main results for all three dynamic-based methods are presented in Figure 3.13. Results based on a single time point (using Euclidean distances) were not presented in this figure as the clusters obtained using these presented a poor stability (Jaccard indexes from bootstrapping, as explained in section 3.3.2.1, never exceeded 0.75) and were not carried out for analysis. This confirms that time series data was here necessary to highlight clusters. The dendrogram obtained for the analysis of time point 0 is presented in Figure 3.14.

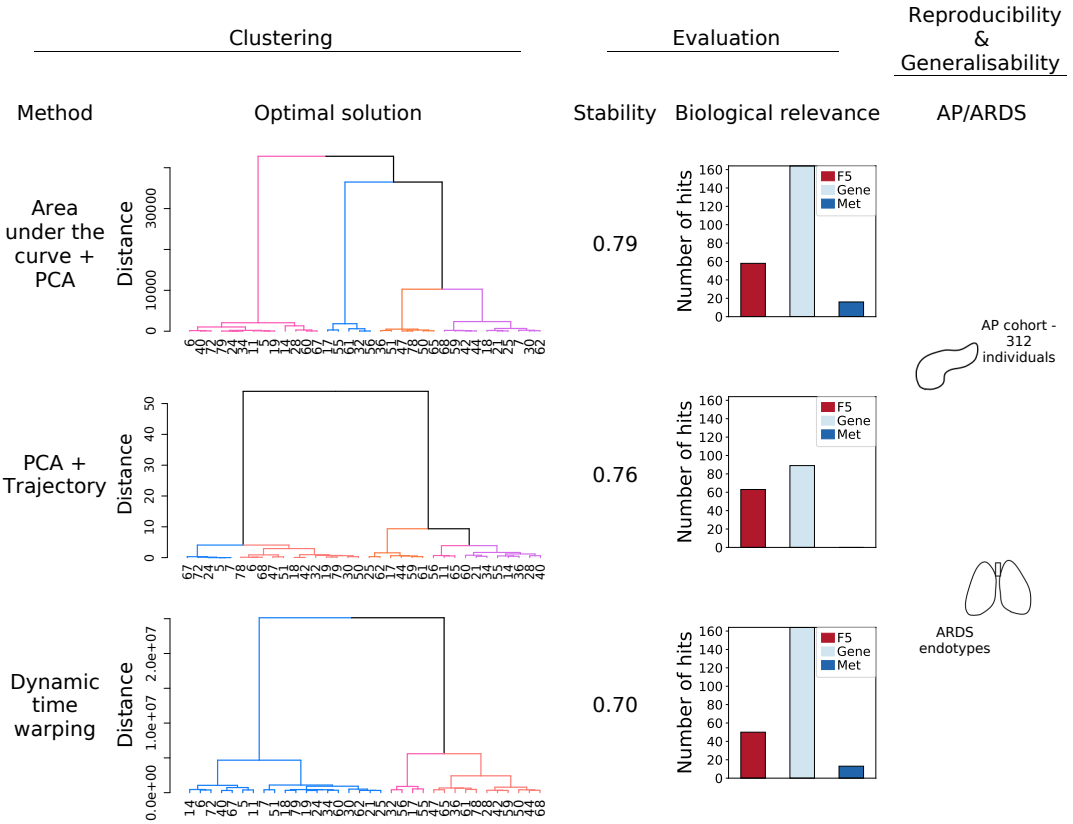


Figure 3.13 - Pipeline overview using the 34 pre-selected IMOFAP individuals (individuals with less than 2 time points were not included in the analysis, n=20). Hierarchical trees for each time series-based clustering method are presented along with the optimal solution. Each of the clustering stability measures is reported (average Jaccard index) and a summary of the number of compound sets significantly enriched is shown for each category (respectively "F5" for FANTOM5 results, "Gene" for gene-based results and "Met" for metabolic compound results). For each one of the three methods based on time series, the best solution, equivalent to the optimal number of clusters (choice based on highest Jaccard index and represented using different colours in the dendrograms), is presented along with stability, as defined by the Jaccard index, and compound set analysis results summary. Reproducibility and Generalisability corresponds to two cohorts which were external from the main cohort.

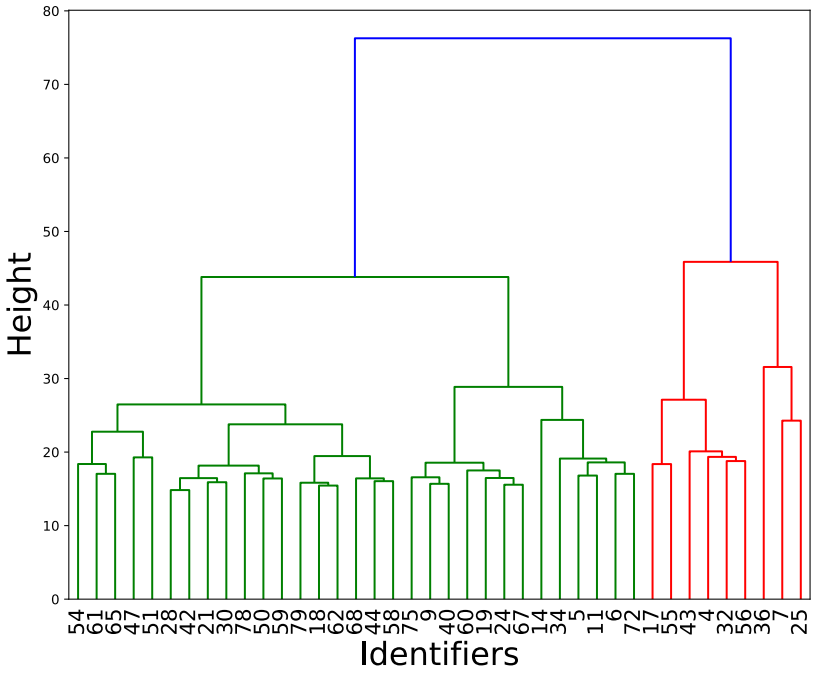


Figure 3.14 – Hierarchical clustering results for time point 0 using Euclidean distances and Ward's algorithm. Number of clusters chosen arbitrarily.

We also looked at silhouette scores[62] for the identified solution, with limited success, to assess the groupings. The current chosen solution resulted in an average silhouette score of 0.39 (individual clusters ranging from 0.23 to 0.57).

3.3.2.1 Internal validity

After performing 100 iterations of our bootstrapping strategy and computing the average Jaccard indexes, the AUC-PCA approach produced the best stability result, with a 4-cluster partition. The corresponding average computed Jaccard index was 0.79, showing that the obtained groups were stable. For the PCA-based trajectory approach and dynamic time warping approach, the average Jaccard index values were respectively 0.76 and 0.70.

We compared the overlap for results obtained using the three methods, for a same number of clusters of four, results are reported in Figure 3.15.

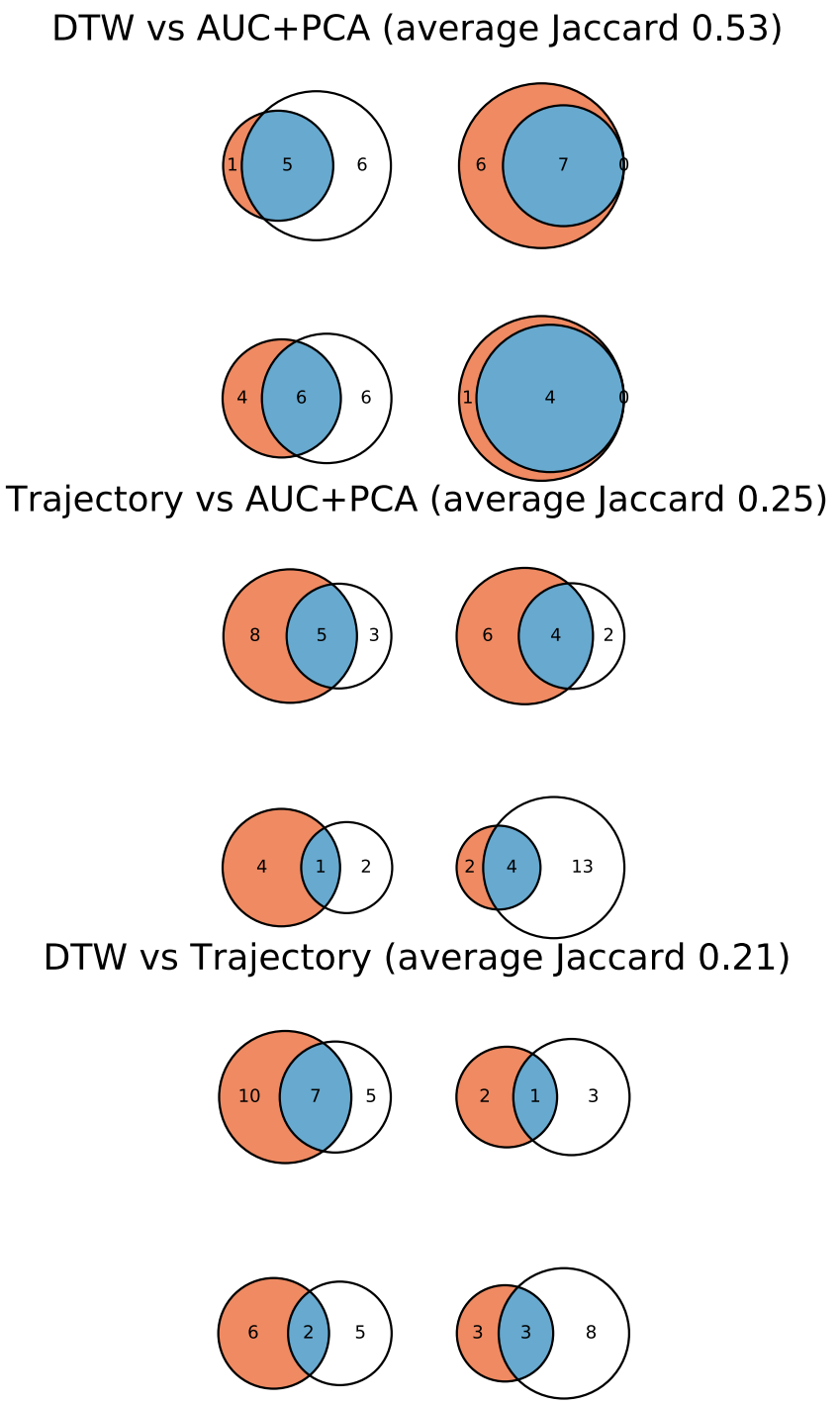


Figure 3.15 – Overlap between clustering solutions for 4 clusters. Numbers in blue areas represent individuals in common between the two groups being compared. Average Jaccard index values are reported for each pairwise comparison. DTW refers to dynamic time warping, AUC+PCA to area-under-the-curve combined to principal component analysis and trajectory to trajectory in principal component analysis space.

The chosen clustering showed moderate overlap (JI = 0.53) when compared with dynamic time warping, and lower similarity (JI = 0.25) when compared with the trajectories in PCA space strategy.

A similar comparison was performed but using 3 and 5 clusters (respectively the optimum number for dynamic time warping and PCA-based trajectory analyses), results are reported in Table 3.4 and Table 3.5.

Table 3.4 - Overlap between clustering solution for 3 clusters. Average Jaccard index values are reported for each pairwise comparison.

Average Jaccard index	**AUC+PCA**	**PCA+Trajectory**	**Dynamic time warping**
AUC+PCA	/	/	/
PCA+Trajectory	0.31	/	/
Dynamic time warping	0.63	0.27	/

Table 3.5 - Overlap between clustering solution for 5 clusters. Average Jaccard index values are reported for each pairwise comparison.

Average Jaccard index	**AUC+PCA**	**PCA+Trajectory**	**Dynamic time warping**
AUC+PCA	/	/	/
PCA+Trajectory	0.24	/	/
Dynamic time warping	0.46	0.21	/

3.3.2.2 Biological validity

For all three analysis strategies, compound set analysis produced significant results, showing the biological validity of highlighted partitions. Compound set analysis resulted in AUC-PCA being identified as the best clustering method. Histograms presenting the results for each one of the three methods are

represented in Figure 3.13. For the selected AUC-PCA partition, the top 20 compound sets are reported in Table 3.6.

Table 3.6 - Using likelihood ratio test, top 20 pathways (using KEGG data for gene, protein and metabolite data and FANTOM5 data for gene and protein data) for the AUC combined with PCA method. FDR-corrected p-values obtained are reported (as computed in R, any value smaller than 2.225074e-308 displayed as 0) along with pathway names/identifiers. Time point 0 used as input.

Pathway	FDR-corrected p-value
hsa00190 Oxidative phosphorylation	<.001
hsa00230 Purine metabolism	<.001
hsa00240 Pyrimidine metabolism	<.001
hsa00510 N-Glycan biosynthesis	<.001
hsa00970 Aminoacyl-tRNA biosynthesis	<.001
hsa03008 Ribosome biogenesis in eukaryotes	<.001
hsa03010 Ribosome	<.001
hsa03013 RNA transport	<.001
hsa03015 mRNA surveillance pathway	<.001
hsa03018 RNA degradation	<.001
hsa03040 Spliceosome	<.001
hsa04010 MAPK signaling pathway	<.001
hsa04110 Cell cycle	<.001
hsa04120 Ubiquitin mediated proteolysis	<.001
hsa04141 Protein processing in endoplasmic reticulum	<.001
hsa04142 Lysosome	<.001
hsa04144 Endocytosis	<.001

hsa04146 Peroxisome	<.001
hsa04660 T cell receptor signaling pathway	<.001
hsa00280 Valine, leucine and isoleucine degradation	<.001

3.3.3 Endotypes description

As per the Stratified Medicines Framework[3], which provides guidelines for attempting to stratify patient groups, we aimed to highlight distinct functional and/or pathophysiological mechanisms to confirm that these groups represented disease endotypes.

3.3.3.1 Endotypes characterisation

Ultimately, groups were highlighted, using the AUC combined with PCA method, given the criteria defined in section 3.3.2. More specifically, a four-group partition produced the best results in terms of stability and biological relevance, compared to partitions of different sizes obtained with the same method.

To select variables strongly associated to one of the four groups, we built four PLS-DA models, using time point 0, and ranked the variables using VIP scores. Indeed, endotypes, to be of maximum clinical utility, would need to be identifiable as soon as possible, ideally when an individual is admitted to the hospital.

Top 10 variables for each group, along with associated VIP values, are represented in Figure 3.19 and consisted of gene or metabolite compounds.

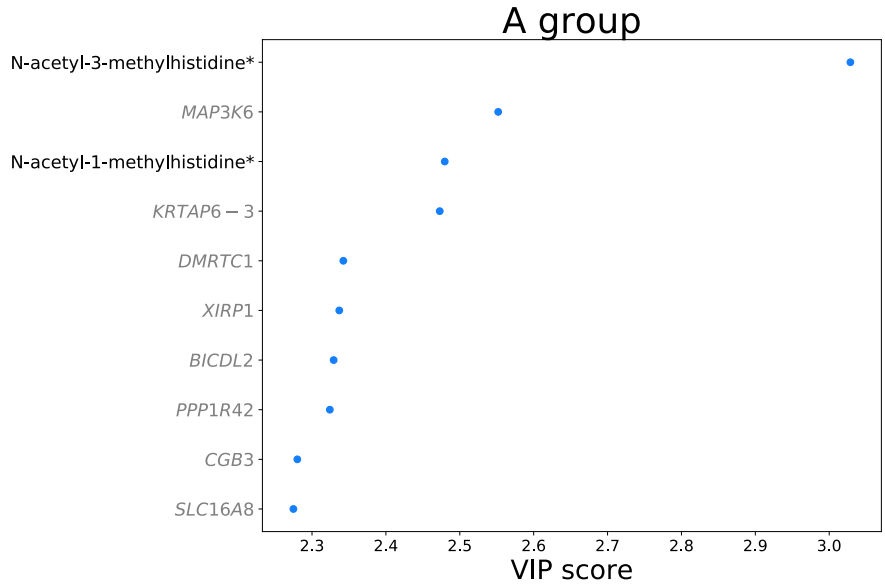


Figure 3.16 -Top 10 variables from the endotype A PLS-DA model using VIP values. Names on the y axis refer to gene (in grey italic) or metabolite compounds.

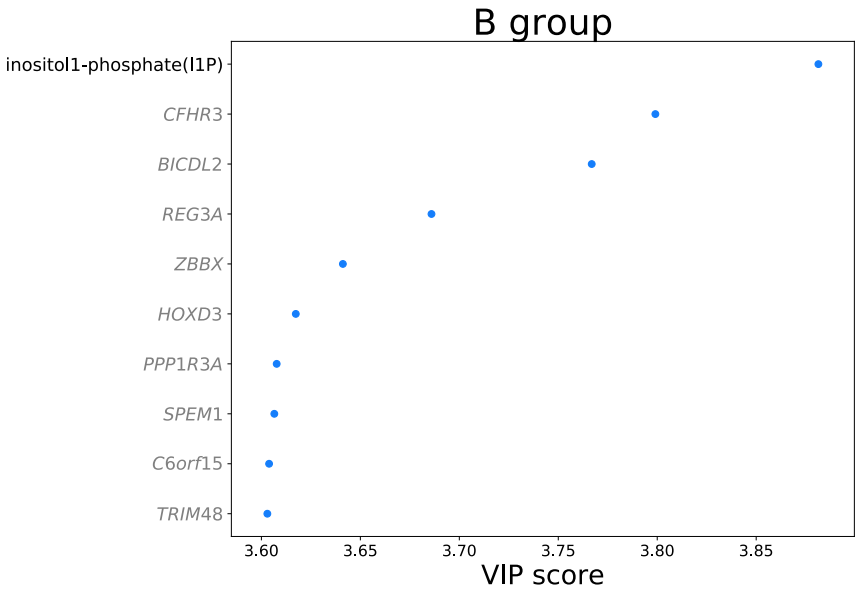


Figure 3.17 - Top 10 variables from the endotype B PLS-DA model using VIP values. Names on the y axis refer to gene (in grey italic) or metabolite compounds.

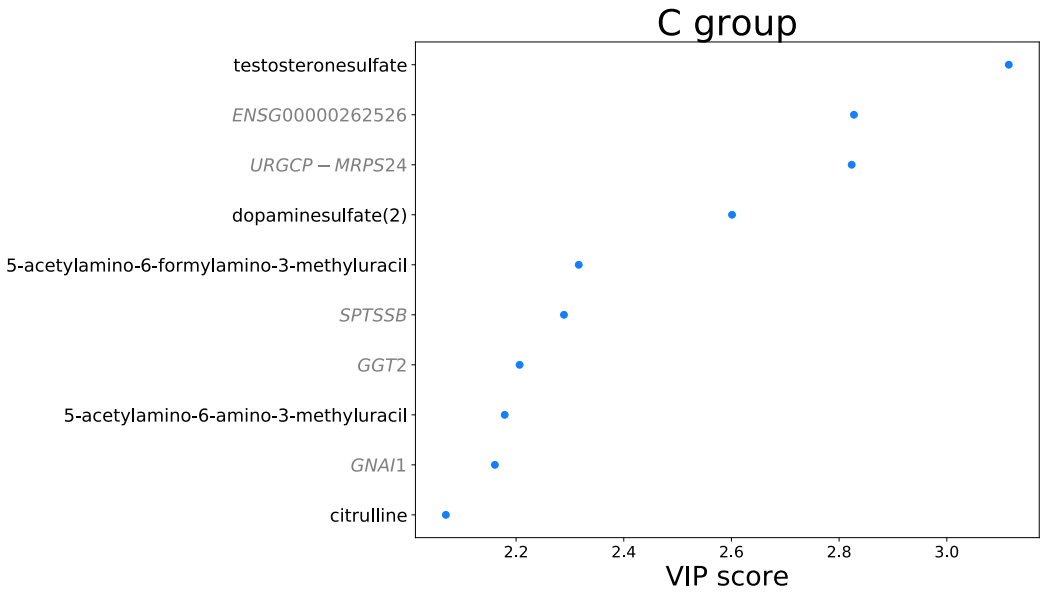


Figure 3.18 - Top 10 variables from the endotype C PLS-DA model using VIP values. Names on the y axis refer to gene (in grey italic) or metabolite compounds.

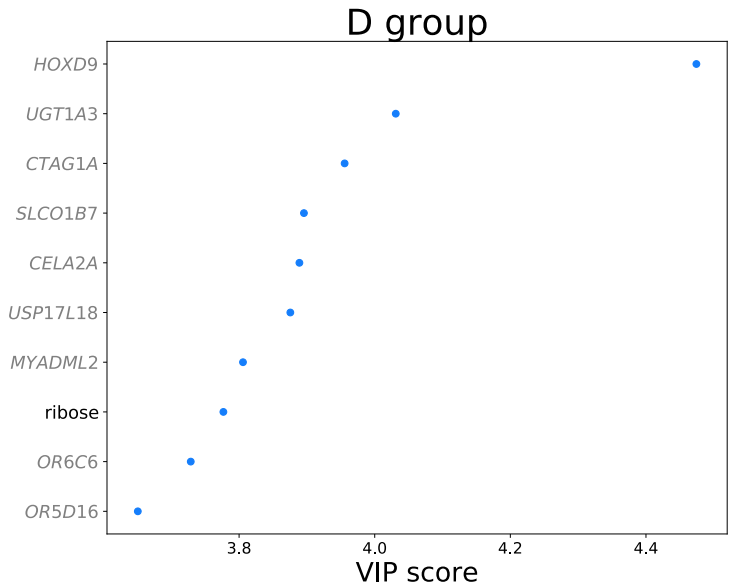


Figure 3.19 - Top 10 variables from the endotype D PLS-DA model using VIP values. Names on the y axis refer to gene (in grey italic) or metabolite compounds.

The group characterisation, using top VIP variables, was done by cross-referencing with publicly reference online resources, namely, GeneCards (Weizmann Institute of Science), HUGO Gene Nomenclature Committee, NCBI EntrezGene, UniProtKB, Ensembl, NCBI PubChem, NCBI PubMed and Google Search. Prominent features, their complete names and associated processes are described in Table 3.7.

Table 3.7 - Compounds detailed table for the top 10 elements for each identified endotype. Complete gene names were fetched using the GeneCards resource and additional information using online resources as described in the previous paragraph.

Heatmap compound	Endotype	Complete name	Additional information
DMRTC1	A	DMRT Like Family C1	
CGB3	A	Chorionic Gonadotropin Subunit Beta 3	
N-acetyl-1-methylhistidine*	A	/	Amino acid metabolism; Rhabdomyolysis; Renal failure
N-acetyl-3-methylhistidine*	A	/	
PPP1R42	A	Protein Phosphatase 1 Regulatory Subunit 42	
SLC16A8	A	Solute Carrier Family 16 Member 8	Lactate transporter; Ketone body transporter
KRTAP6-3	A	Keratin Associated Protein 6-3	Muscle-specific actin binding protein upregulated during muscle injury
XIRP1	A	Xin Actin Binding Repeat Containing 1	
MAP3K6	A	Mitogen-Activated Protein Kinase Kinase Kinase 6	Apoptosis signaling
BICDL2	A/B	BICD Family Like Cargo Adaptor 2	
ZBBX	B	Zinc Finger B-Box Domain Containing	

CFHR3	B	Complement Factor H Related 3	Heparin-binding; Complement regulation
Inositol 1-phosphate (I1P)	B	/	Inositol biosynthesis
HOXD3	B	Homeobox D3	Increases immune cell adherence; Overexpression upregulates glycoprotein IIb/IIIa
SPEM1	B	Spermatid Maturation 1	
C6orf15	B	Chromosome 6 Open Reading Frame 15	Putative heparin/fibronectin binding
TRIM48	B	Tripartite Motif Containing 48	Interferon-γ signalling (oxidative stress/apoptosis signal-reducting kinase 1)
REG3A	B	Regenerating Family Member 3 Alpha	Bactericidal C-type lectin; Known as pancreatitis-associated protein
PPP1R3A	B	Protein Phosphatase 1 Regulatory Subunit 3A	Genetic association with type 2 DM and familial partial lipodystrophy 3
GNAI1	C	G Protein Subunit Alpha I1	N-acetyl transferase activity

SPTSSB	C	Serine Palmitoyltransferase Small Subunit B	Tricarboxylic acid cycle
Citrulline	C	/	Sphingolipid biosynthesis
Dopamine sulfate (2)	C	/	Gastrointestinal dopamine metabolism
Testosterone sulfate	C	/	
5-acetylamino-6-amino-3-methyluracil	C	/	Caffeine metabolism
5-acetylamino-6-formylamino-3-methyluracil	C	/	
GGT2	C	Gamma-Glutamyltransferase 2	γ-glutamyl transferase; Glutathione homeostasis
URGCP-MRPS24	C	URGCP-MRPS24 Readthrough	
ENSG00000262526	C	/	Protein coding
OR5D16	D	Olfactory Receptor Family 5 Subfamily D Member 16	
CTAG1A	D	Cancer/Testis Antigen 1A	
MYADML2	D	Myeloid Associated Differentiation Marker Like 2	
Ribose	D	/	

CELA2A	D	Chymotrypsin Like Elastase Family Member 2A	Pancreatic elastase-2
HOXD9	D	Homeobox D9	
OR6C6	D	Olfactory Receptor Family 6 Subfamily C Member 6	
UGT1A3	D	UDP Glucuronosyltransferase Family 1 Member A3	Associated with Gilbert-type hyperbilirubinemia
SLCO1B7	D	Solute Carrier Organic Anion Transporter Family Member 1B7 (Putative)	Cysteine-type endopeptidase
USP17L18	D	Ubiquitin Specific Peptidase 17-Like Family Member 18	Liver-specific organic anion transporter; Bile secretion

To describe our endotypes in a systematic way, we also performed a compound set enrichment analysis.

We filtered variables deemed significant for the classification task using a VIP threshold value[129] of 1 for each one of the models (Table 3.8).

Table 3.8 – VIP-selected variables summary.

	Number of variables with VIP>1
A	10,216
B	6,584
C	9,112
D	7,037

Enrichment results using these filtered lists are presented in Figure 3.20 and Table 3.9.

Figure 3.20 - Significant pathway terms (adjusted p-value threshold of 0.01, red indicates a significant item) from enrichment results for each identified group based on variables lists selected using VIP scores. Pathway data extracted from Reactome database. Results for time points 0, 24 and 48 are reported for all four endotypes.

Table 3.9 – Full names of significant pathway terms.

HSA identifier	Full pathway name
HSA-6807505	RNA polymerase II transcribes snRNA genes
HAS-4570464	SUMOylation of RNA binding proteins
HSA-191859	snRNP Assembly
HSA-5419276	Mitochondrial translation termination
HSA-6809371	Formation of the cornified envelope
HSA-72163	mRNA Splicing - Major Pathway
HSA-2408557	Selenocysteine synthesis
HSA-72187	mRNA 3'-end processing
HSA-1799339	SRP-dependent cotranslational protein targeting to membrane
HSA-975956	Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC)
HSA-159230	Transport of the SLBP Dependant Mature mRNA
HSA-5368286	Mitochondrial translation initiation
HSA-3108214	SUMOylation of DNA damage response and repair proteins
HSA-4551638	SUMOylation of chromatin organization proteins
HSA-1268020	Mitochondrial protein import
HSA-156827	L13a-mediated translational silencing of Ceruloplasmin expression
HSA-975957	Nonsense Mediated Decay (NMD) enhanced by the Exon Junction Complex (EJC)
HSA-72689	Formation of a pool of free 40S subunits
HSA-6791226	Major pathway of rRNA processing in the nucleolus and cytosol
HSA-72165	mRNA Splicing - Minor Pathway

HSA-9010553	Regulation of expression of SLITs and ROBOs	HSA-72702	Ribosomal scanning and start codon recognition
HSA-5389840	Mitochondrial translation elongation	HSA-159231	Transport of Mature mRNA Derived from an Intronless Transcript
HSA-159236	Transport of Mature mRNA derived from an Intron-Containing Transcript	HSA-6790901	rRNA modification in the nucleus and cytosol
HSA-8951664	Neddylation	HSA-6805567	Keratinization
HSA-159227	Transport of the SLBP independent Mature mRNA	HSA-72764	Eukaryotic Translation Termination
HSA-72649	Translation initiation complex formation	HSA-109688	Cleavage of Growing Transcript in the Termination Region
HSA-72695	Formation of the ternary complex, and subsequently, the 43S complex	HSA-72706	GTP hydrolysis and joining of the 60S ribosomal subunit

3.3.3.2 Data visualisation

To visualise some of the discriminant variables, heatmaps were generated using top-10 variables for each category, as defined using VIP values from previously generated PLS-DA models and are represented in Figure 3.21 and Figure 3.22, respectively for the group average and individual variable values.

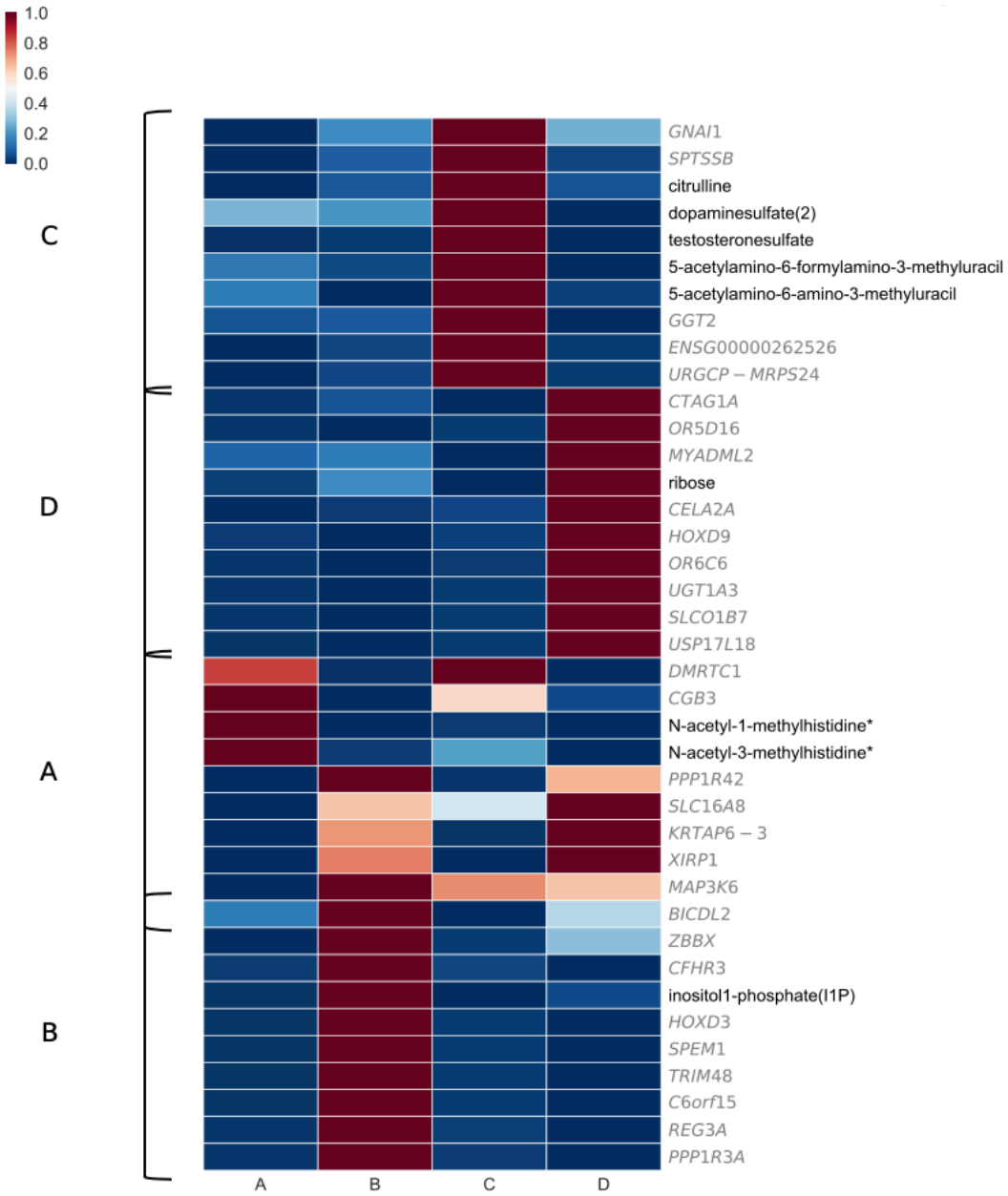


Figure 3.21 - AP endotypes. The top 10 VIP-selected variables, average values (normalised and scaled) for each identified group are displayed. For visualisation purposes row values were scaled between 0 and 1. Colours are representative of the range of observed values. Values were clustered based on expression patterns considering average values per variable per group.

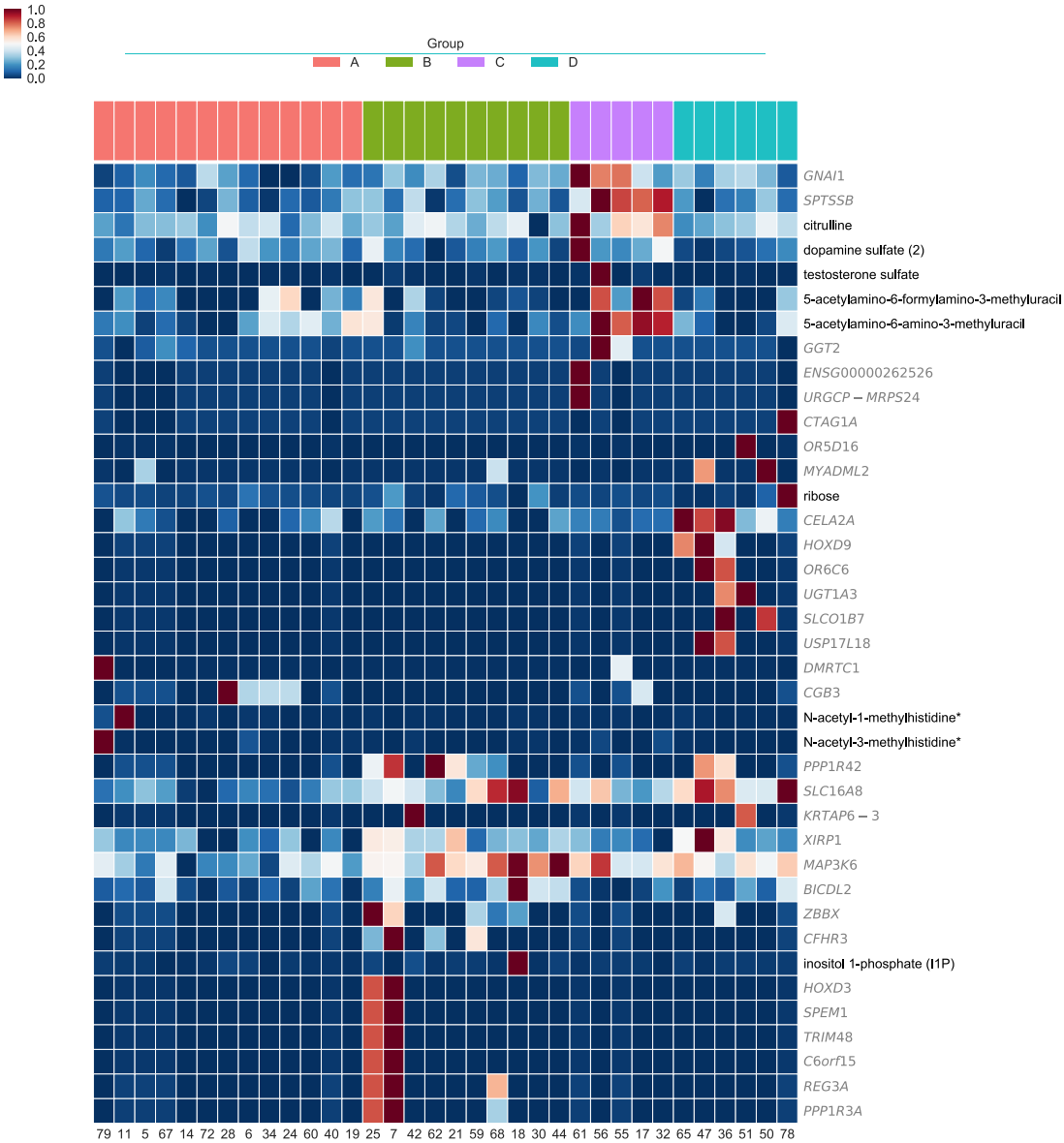
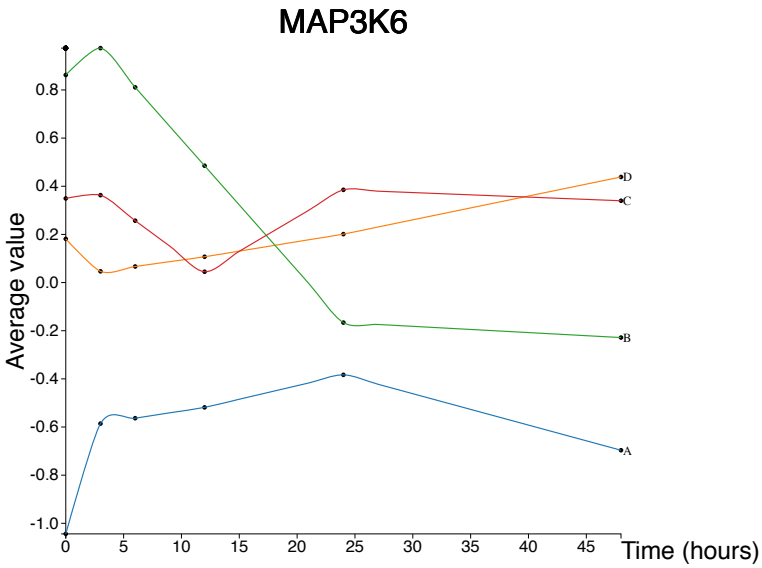
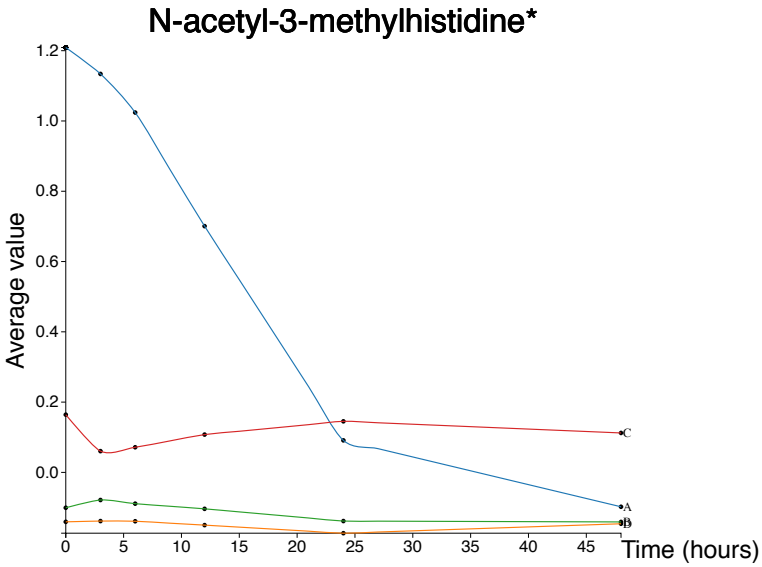
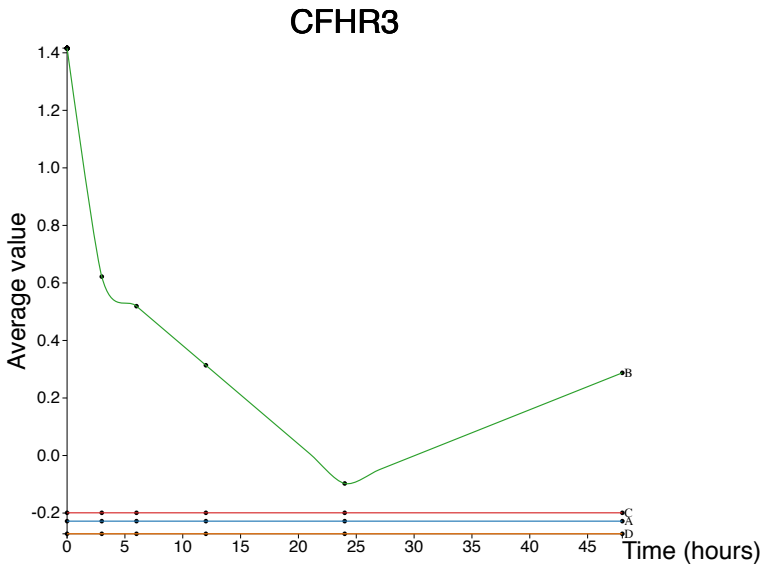


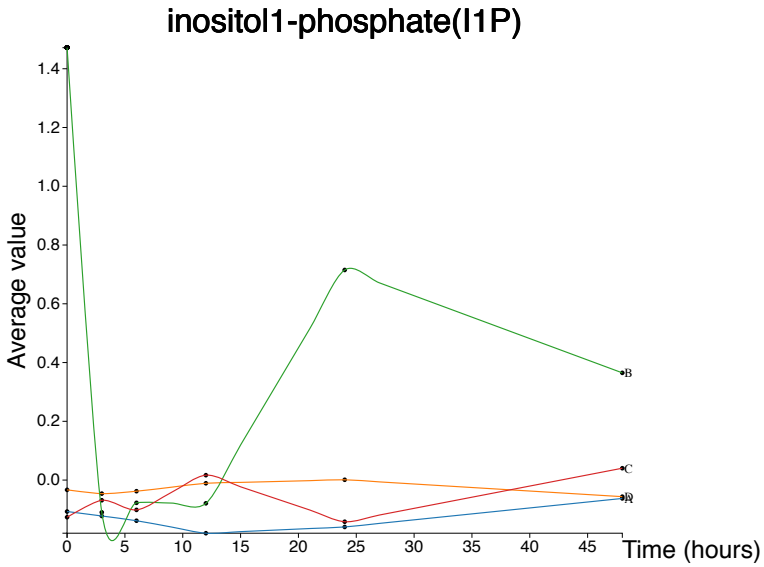
Figure 3.22 - For the top 10 variables, normalised and scaled average values for each identified patient are displayed. For visualisation purposes row values were scaled between 0 and 1. Colours are representative of the range of values observed. Values were clustered based on expression patterns considering average values per variable per group. Patients 5 and 11 did not survive.

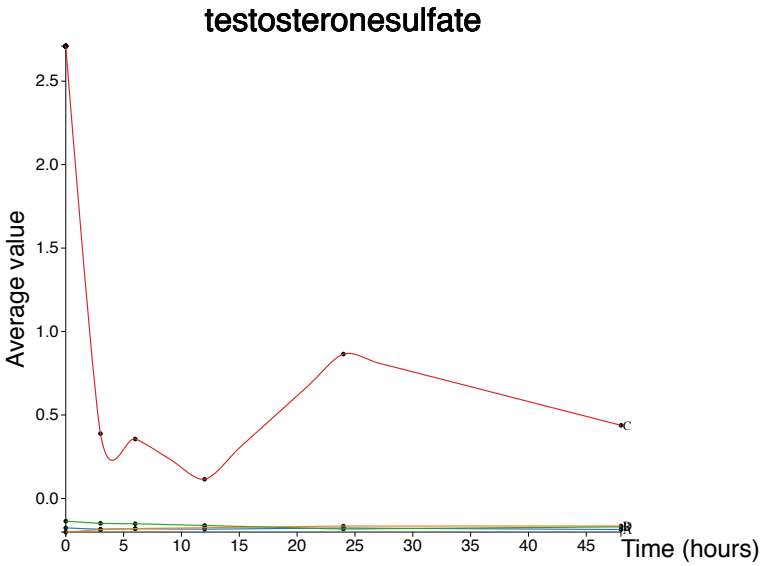
Time profiles for the top-2 VIP-selected variables are represented in Figure 3.23 and allowed to compare their evolution over time for the four identified groups.

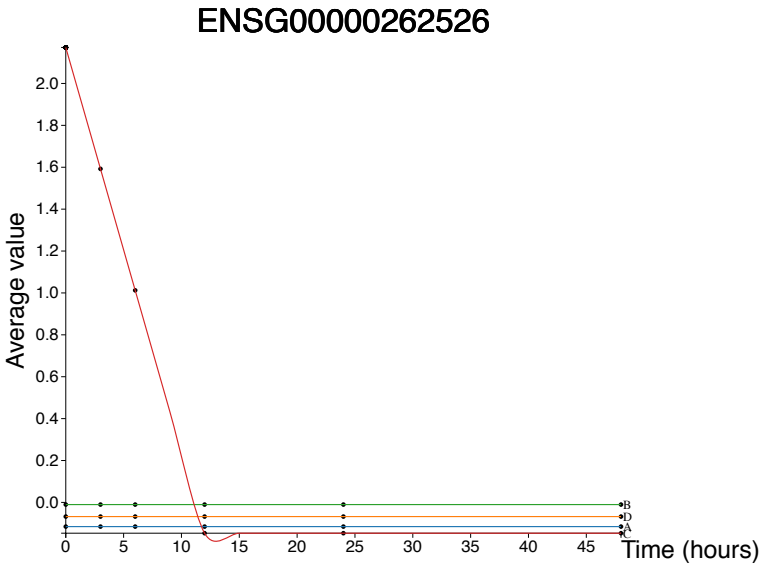

MAP3K6
Average value
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
0
5
10
15
20
25
30
35
40
45
Time (hours)
D
C
B
A


N-acetyl-3-methylhistidine*
Average value
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
30
35
40
45
Time (hours)
C
A


CFHR3
Average value
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
0
5
10
15
20
25
30
35
40
45
Time (hours)
B
C
A
D


inositol1-phosphate(I1P)
Average value
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
30
35
40
45
Time (hours)
B
C
D
A


testosteronesulfate
Average value
2.5
2.0
1.5
1.0
0.5
0.0
0
5
10
15
20
25
30
35
40
45
Time (hours)
C
B
D
A


ENSG00000262526
Average value
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
30
35
40
45
Time (hours)
B
D
A
C

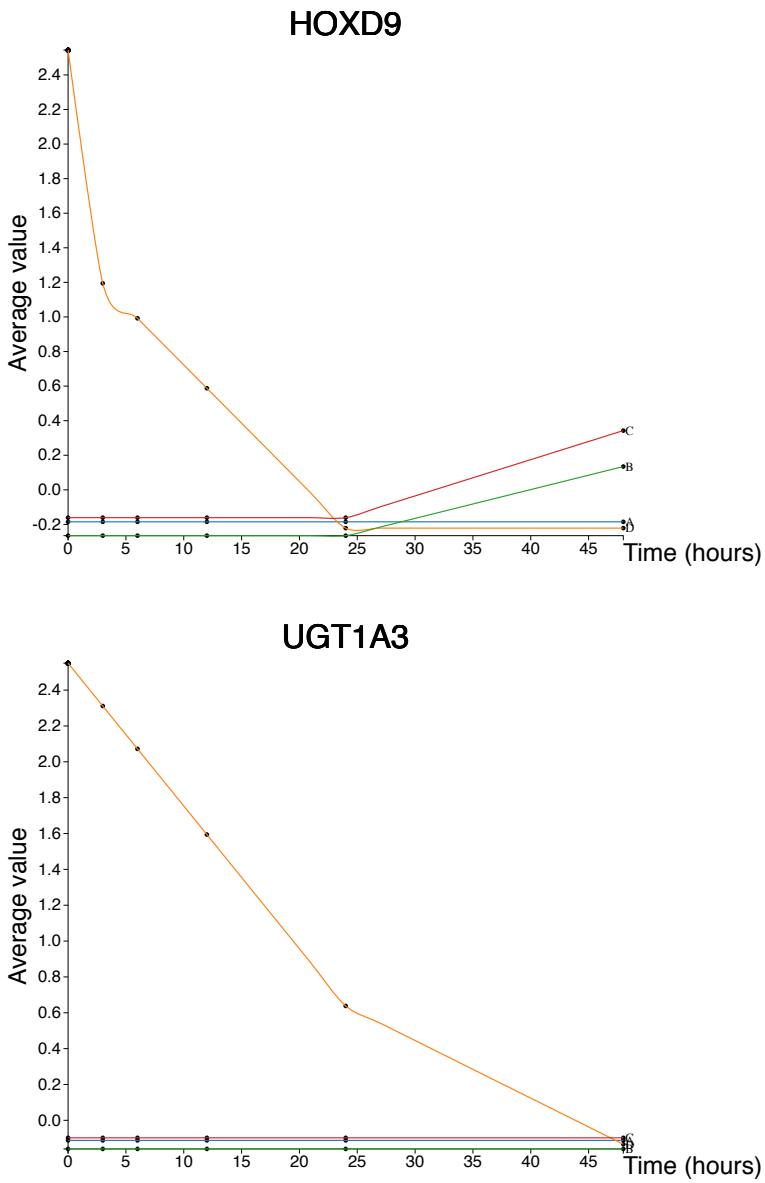


Figure 3.23 – For best two VIP-selected variables across endotypes, time profiles are represented. Values were generated as average z-score value per time point per group identified. Graphs generated using [http://baillielab.net/pancreatitis/.](http://baillielab.net/pancreatitis/)

Clinical data, for selected variables of interest, were then inspected. The distribution of severity levels (mild, moderate and severe) across our four endotypes is represented in Figure 3.24.

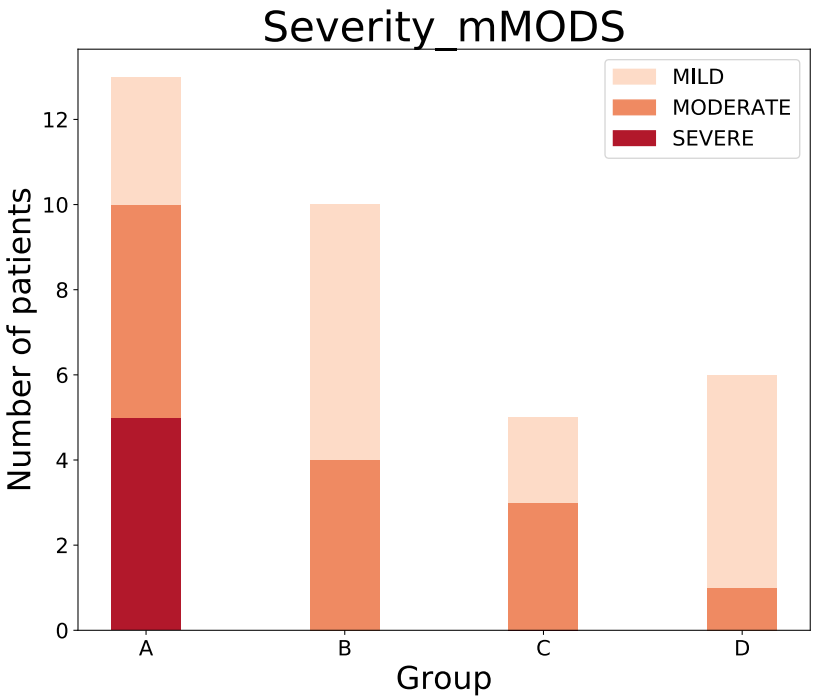


Figure 3.24 - For comparison purposes, distribution of clinical severity categorised by mMODS score in each identified endotype.

When the proportion of severe versus non-severe cases was compared between the groups, independence between the group labels and severity was rejected (Fisher's exact test, p=0.038). All individuals with a severe form of AP clustered in the A group, showing that our clustering was relevant in terms of disease severity.

The distribution of etiologies was also represented in Figure 3.25 and was independent of groups labels (Fisher's exact test, p=0.97).

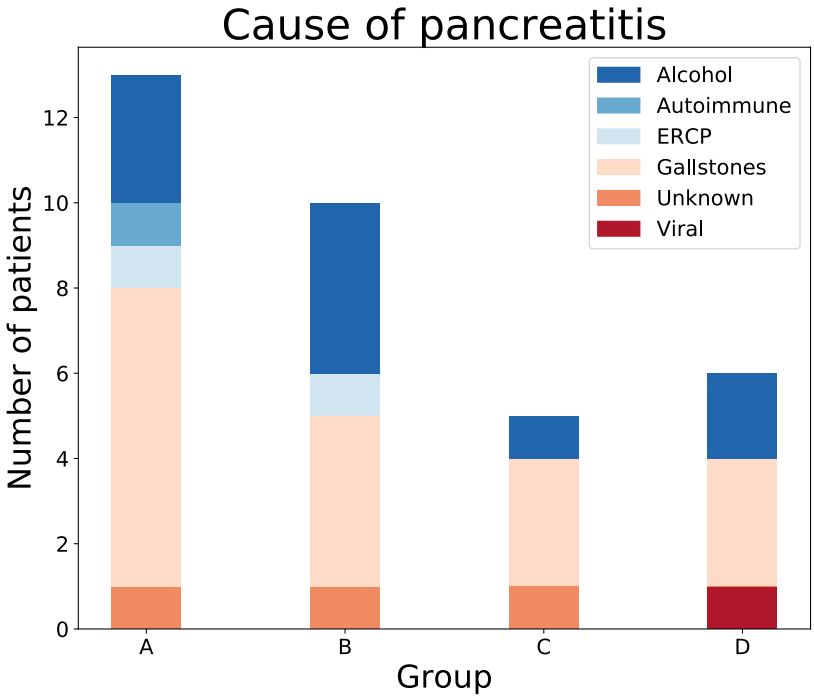


Figure 3.25 - Distribution of etiology in each endotype, for comparison purposes. For each identified endotype, the number of patients is shown.

We inspected gender distribution as well (Fisher's exact test, p = 0.67) and time of onset of symptoms, using one-way ANOVA (p = 0.97) but could not reject independence with group labels. This confirmed that our groups did not reflect differences due to gender or symptom onset time.

Systemic inflammatory response syndrome (SIRS) was not significantly associated with group labels when comparing SIRS versus no-SIRS (Fisher's exact test, p=0.097). SIRS distribution across endotypes was represented in Figure 3.26.

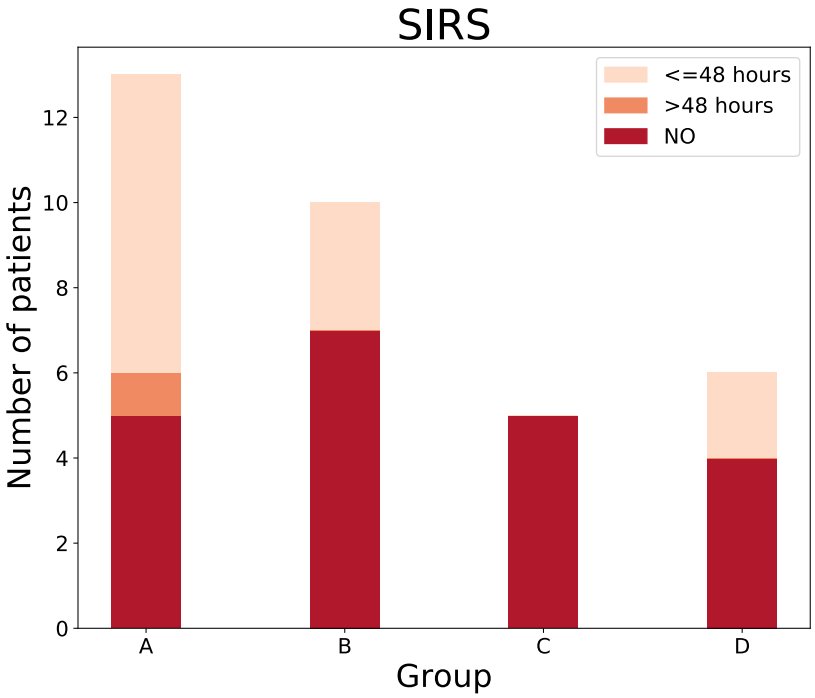


Figure 3.26 - SIRS distribution per endotype. For each identified endotype, the number of patients is shown. 'NO' corresponds to no SIRS.

To confirm that clusters structure was not solely determined by gender, age or time of onset of symptoms, we re-performed the clustering strategy, with the AUC-PCA method, using residuals from a linear model including gender, age and time onset as predictors. In other words, we only looked at variation that could not be linearly explained by any of these variables and performed the clustering using residuals only. We compared the 4-cluster partition obtained to our chosen clustering and observed a high level of similarity (Jaccard index of 0.82), moreover, distance matrices obtained showed a high correlation (0.91, Mantel test p-value = 0.01).

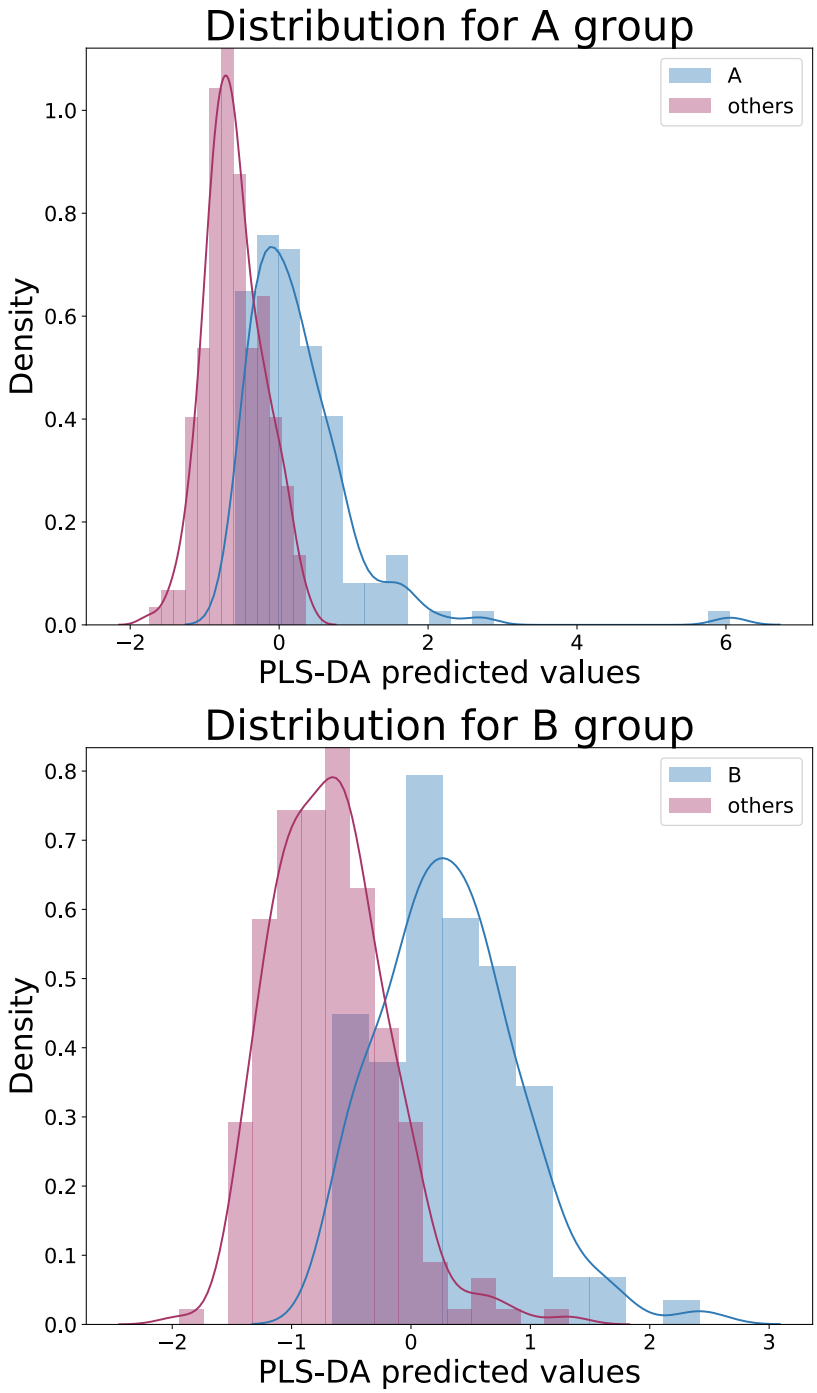
3.3.4 Validation results

3.3.4.1 External validity

3.3.4.1.1 Allocation results

After applying our four PLS-DA models (one for each of our endotypes) we inspected the distributions of PLS-DA predicted values. They are represented in Figure 3.27. We observed that the predicted values associated with

allocated samples were higher than, and not overlapping with, values associated with unallocated samples.



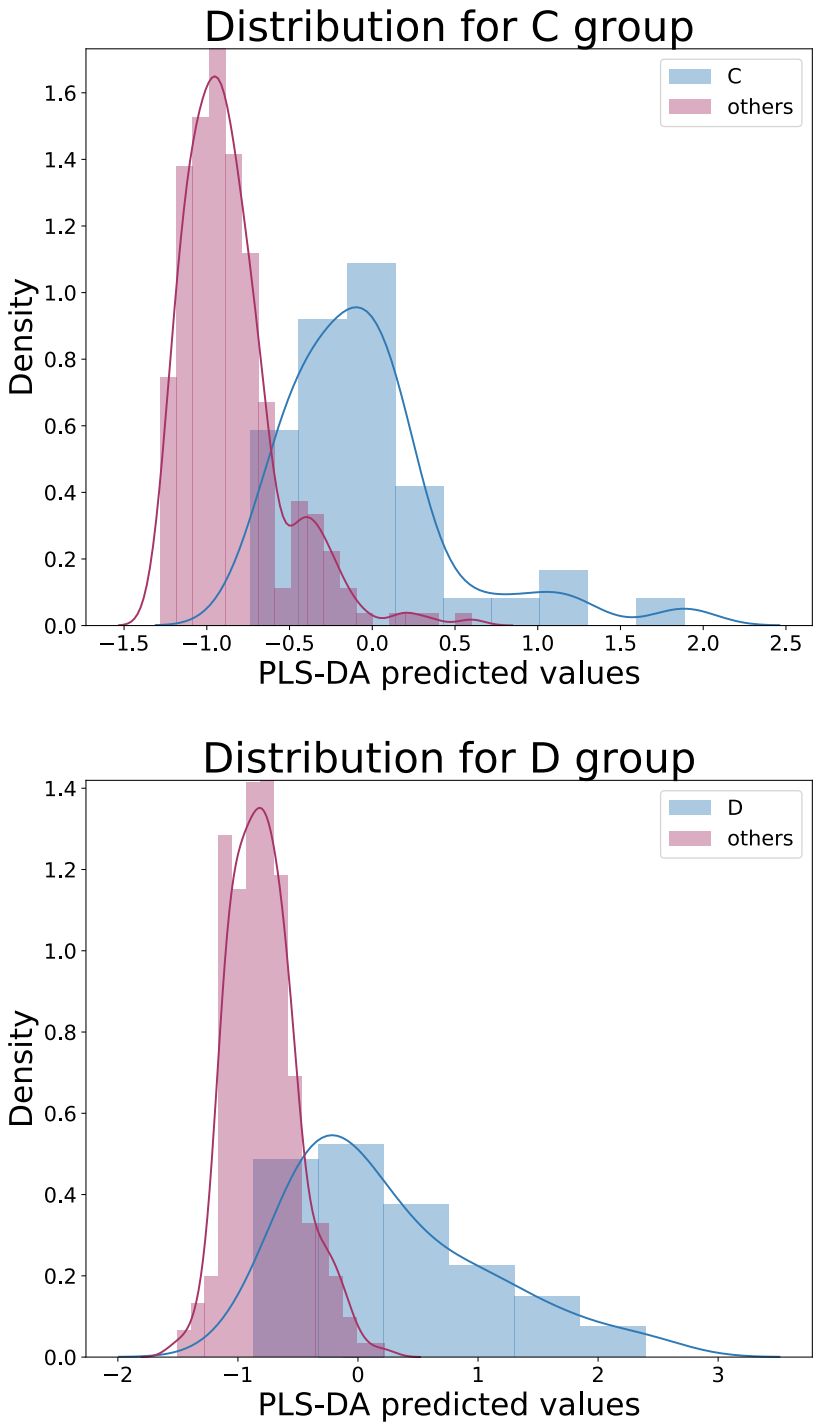


Figure 3.27 - For each endotype, distribution of PLS-DA predicted values for assigned (if the current endotype was the 'best fit') and unassigned KAPVAL individuals are represented. -1 is the target value for samples not from the current endotypes and 1 is the target value for samples from the current endotype.

3.3.4.1.2 Inspection of allocated samples

When comparing KAPVAL allocated samples to IMOFAP samples for the corresponding groups, using only variables not included in the PLS-DA models, we obtained significant Spearman's correlation coefficients. For groups 1 to 4, correlation coefficients ranged from 0.38 to 0.65 with FDR-corrected p-values <0.001. Results are presented in Figure 3.28.

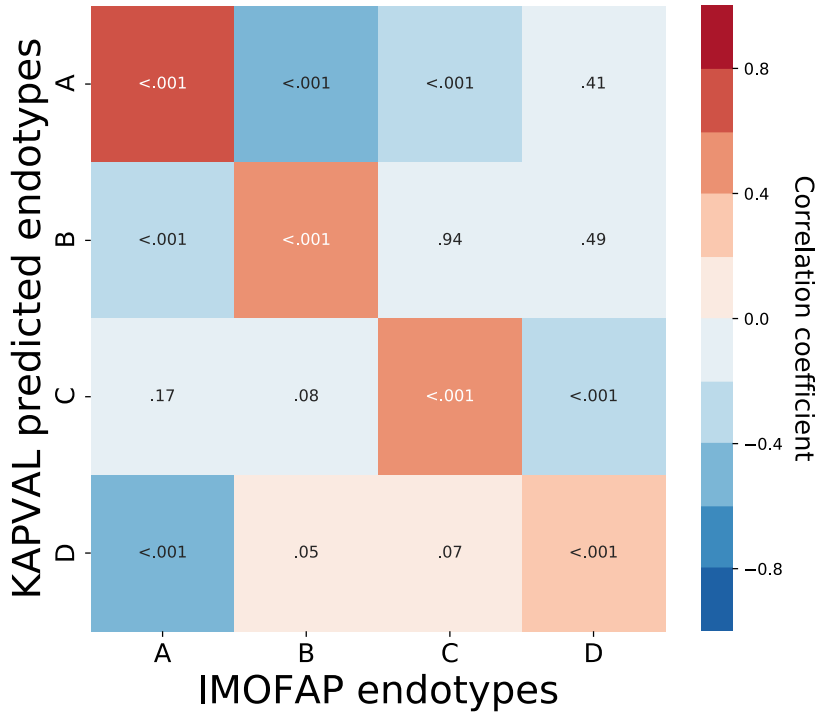


Figure 3.28 - Spearman's correlation results, reported using colours, for pairwise comparisons between variable average values from training set (IMOFAP) and testing set (KAPVAL). FDR-corrected p-values associated to each correlation coefficient, reported within each cell of the heatmap, were calculated as well.

This confirmed that the endotype separation signal identified in the IMOFAP dataset could also be observed in KAPVAL, and was unlikely to be observed by chance.

Distributions of in-hospital mortality and care level for allocated KAPVAL samples are represented in Figure 3.29 and Figure 3.30.

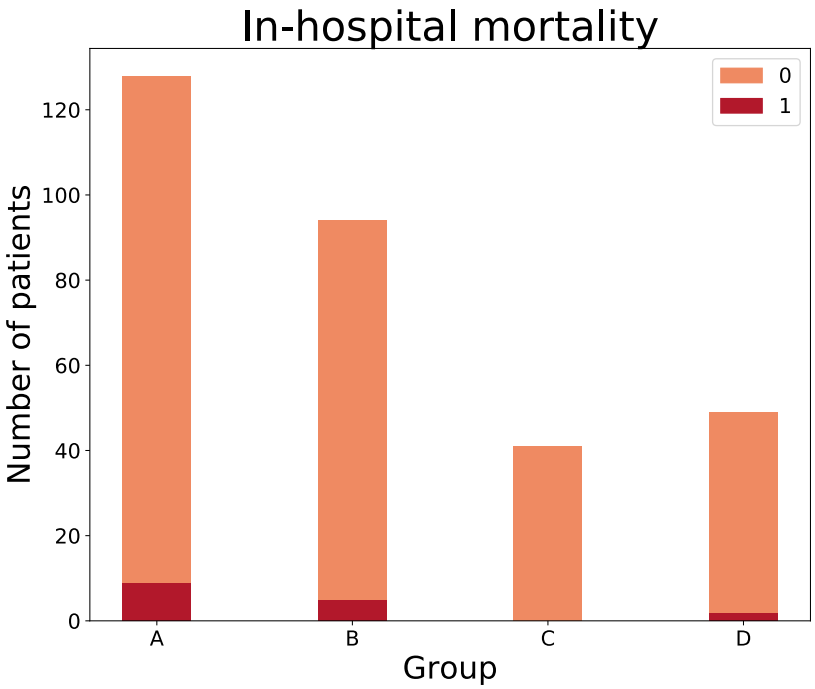


Figure 3.29 - Distribution of in-hospital mortality for KAPVAL-allocated individuals. 1 represents a death event and 0 corresponds to no in-hospital death reported.

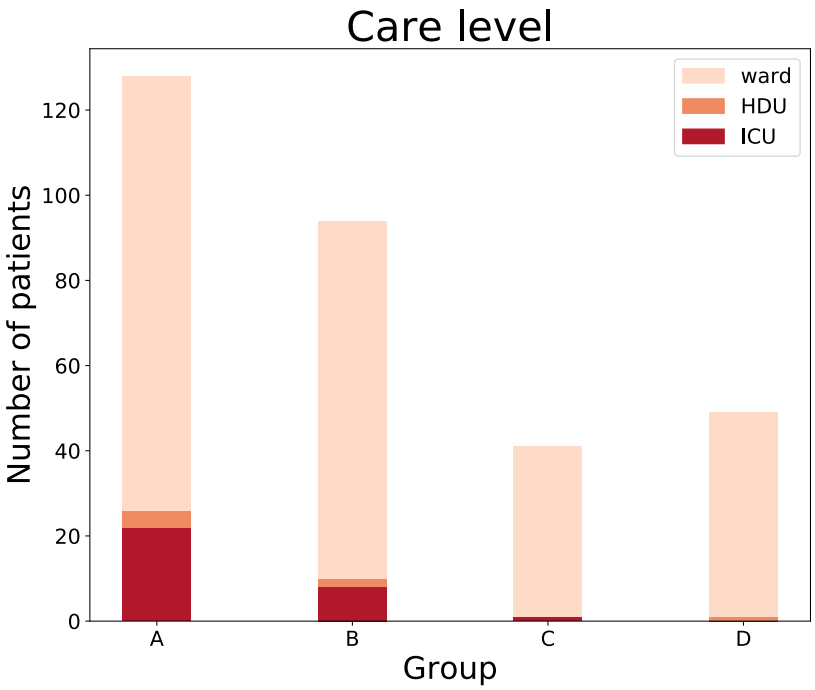


Figure 3.30 - Distribution of care level (war, HDU or ICU) for KAPVAL-allocated individuals.

Independence between in-hospital mortality and group labels was not rejected (Fisher's exact test, p=0.39), but admission to critical care was (Fisher's exact test, $p < 0.001$, comparing ward stay versus HDU or ICU)

Length of stay per group of KAPVAL-allocated samples were illustrated in Figure 3.31.

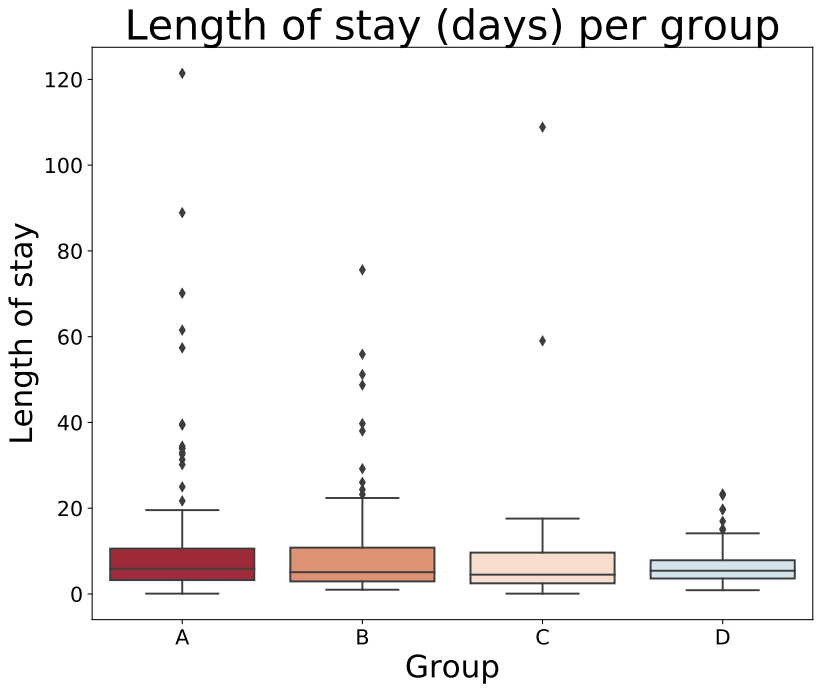


Figure 3.31 – Per endotype, boxplots representing length of hospital stay, in days, for KAPVAL-allocated samples. Bars represent 95% confidence intervals.

Median length of stay values in days were respectively 5.9, 5.1, 4.5 and 5.3 and corresponding interquartile ranges 7.3, 10.9, 5.5 and 4.7 days (with Q1-Q3 3.2-12.8, 2.8-10.8, 2.5-9.7 and 3.3-7.8 days).

IGP values were computed for all endotypes at recruitment. We obtained values of 0.73, 0.51, 0.64 and 0.63 respectively for endotypes A, B, C and D. Associated p-values were smaller than 0.001 for endotypes A and D and equal to 0.01 for endotypes B and C. In practice this meant that the cluster quality, as measured with the IGP, was higher than one would expect by chance, for all four endotypes.

3.3.4.2 Generalisability in ARDS

We sought to compare our AP endotypes to two ARDS endotypes, which were reported in the Calfee *et al* paper, because we observed similarity in the orders of importance for both studies.

Using the nineteen matched variables (as listed in Figure 3.32 and Figure 3.33) between our dataset and the ones from the ARDS endotypes study (and their corresponding ranks, as reported in the paper[39]), we compared ranks using Spearman's correlation coefficients and obtained significant results when comparing our A endotype with both ARDS cohorts (FDR-corrected p = 0.046 for both the ALVEOLI and ARMA cohorts). We obtained similar findings when comparing our endotype C with the ALVEOLI and ARMA cohorts (p = 0.046 and p = 0.046 respectively). Results are illustrated in Figure 3.32 and Figure 3.33.

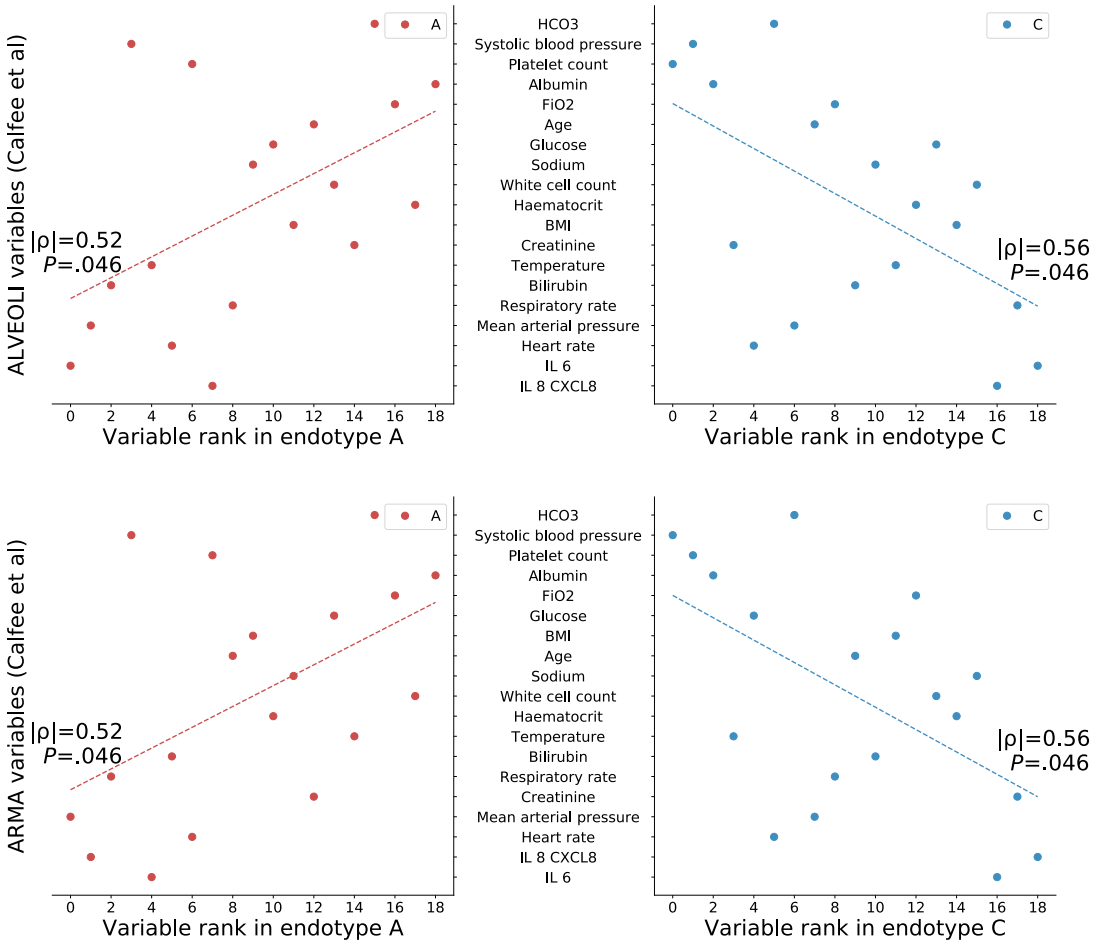


Figure 3.32 – Ranks of ordered average normalised values represented for A and C endotypes on the x axis. Variables that occur in common with those reported in the ARDS study of Calfee et al are presented on the y axis. Linear trends were computed and represented using the ALVEOLI and ARMA cohort results. FDR-corrected p-values are reported for each.

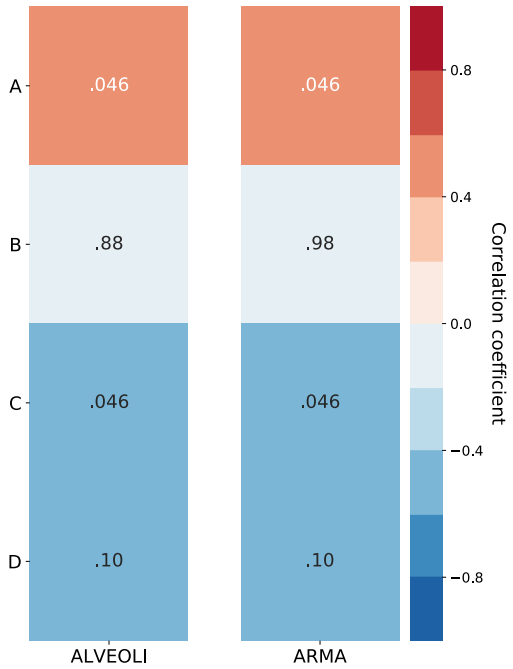


Figure 3.33 - Spearman correlation coefficients between the four identified groups (on the y axis) and the two ARDS cohorts (on the x axis). FDR-corrected p-values are reported for all pairwise comparisons.

3.3.4.3 MOFA results comparison

To compare our results with those obtained using a validated tool, we chose to run MOFA on our dataset using area-under-the-curve values. Using the first two latent features, we extracted four clusters and compared them to ours using Jaccard index to estimate the overlap and the information in common between our selected method and MOFA. We obtained a Jaccard index of 0.88, confirming the validity of our solution.

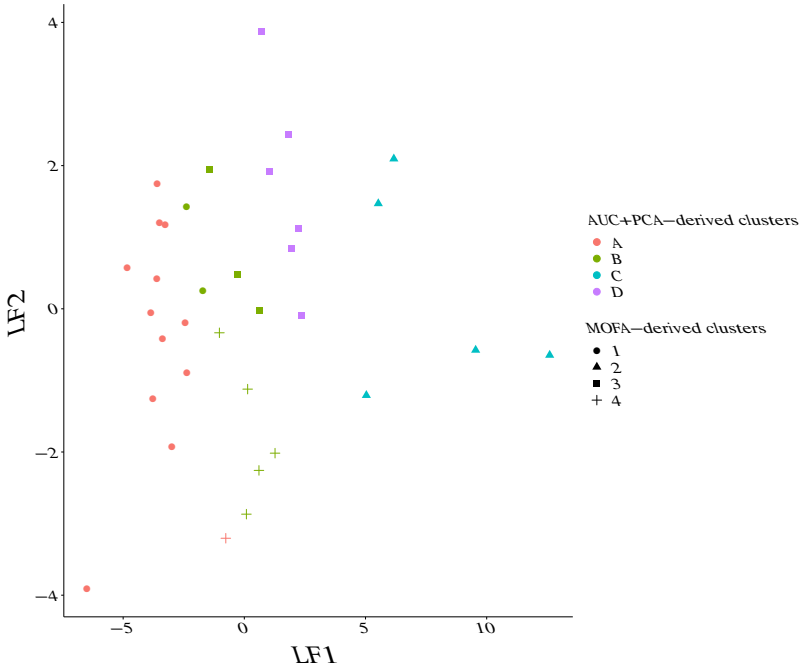


Figure 3.34 - Comparison of clusters obtained using MOFAtools with identified clusters. AUC values were used as input and a 4-cluster solution was extracted from MOFA results using the first two latent features based on explained variance. Colours are representative of clusters described as part of this project and shapes of MOFAtools predicted allocations.

We then decided to compare the allocations of KAPVAL samples obtained previously using PLS-DA models with groups obtained from MOFA. The results are illustrated in Figure 3.35.

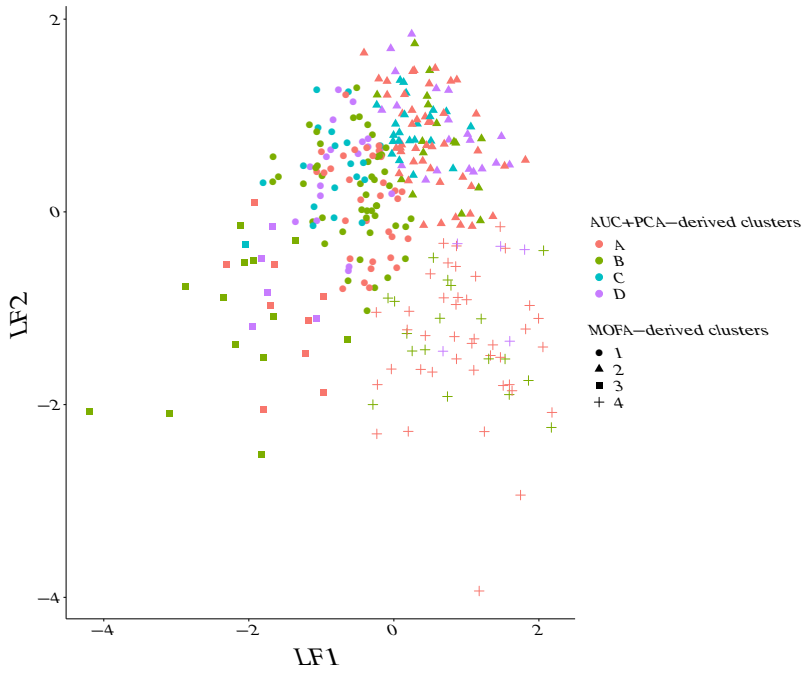


Figure 3.35 - Using KAPVAL metabolomics data and selecting a 4-cluster solution, comparison of results obtained with MOFAtools and results arising from PLS-DA models. Colours indicate results obtained using PLS-DA models and shapes show MOFAtools results.

The two results presented structures which were quite dissimilar, the Jaccard index was equal to 0.22. We hypothesised that, as only one time point was available for the KAPVAL cohort, a single time point was not sufficient to highlight groups using MOFA and that we needed some knowledge of the dynamics (expressed through the PLS-DA models which were trained using IMOFAP sample allocations arising from area-under-the-curve values).

3.4 Conclusions

This analysis confirms the existence of molecular subtypes in AP. In itself, this is an important and novel observation. Additionally, the discovered endotypes go partway to explaining some of the heterogeneity of AP, and its consequences. The AP endotypes could be identified using multiomics data. More specifically, transcriptomics, proteomics and metabolomics, measured for different time points across a time course, defined endotypes that could not otherwise be identified using standard clinical and laboratory measurements.

The four identified groups were proved to be statistically stable and are likely to be biologically relevant, although the precise direct clinical relevance of the groups will need to be uncovered in future work. More specifically, group A was identified as a potentially higher risk group but no other correlations with clinical variables were identified at this point.

Importantly, I could also find similar molecular endotype signatures in an independent validation dataset of AP patients, and also in a distinct but probably pathologically overlapping syndrome, ARDS. Statistically significant similarities were highlighted between our IMOFAP AP dataset and ARDS endotypes, described using clinical and cytokine measurements, for a single time point, in two cohorts of 549 and 473 ARDS-affected individuals, even though used the statistical approaches used were fundamentally different. This is illustrated in Figure 3.36.

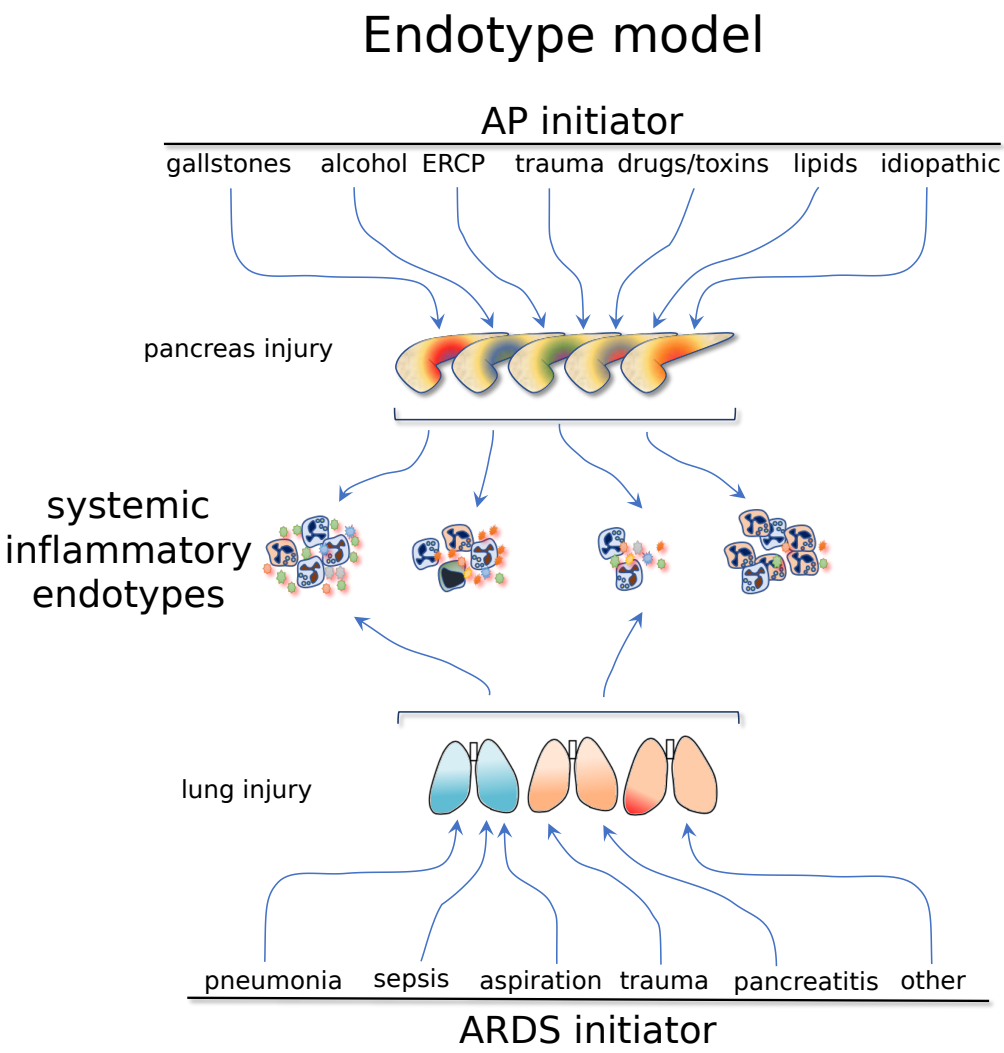


***Figure 3.36 – Endotype model summary figure. Our final model consists of a systemic inflammatory endotypes model. Endotypes highlighted here are represented alongside ARDS endotypes identified in the Calfee et al* paper.**

Using an unsupervised approach, we could re-discover these two ARDS endotypes in a distinct clinical syndrome, with almost no overlap.

This similarity between our AP endotypes and previously reported ARDS endotypes was not expected and show that the different signals defined by endotypes A, B, C and D were not specific to AP. Such similarities could be used to look at critical illnesses from a new perspective which could be beneficial for the understanding and treatment of other diseases.

We conclude that these patterns reflect generalisable features of the host response to critical injury and that they may be observed in other illnesses. To demonstrate this, attempts will be made to highlight identified signals in other critical illnesses such as sepsis or trauma.

4. Chapter 4 – Generalisability of critical illness endotypes

This chapter builds upon findings presented in chapter 3, in which four Acute Pancreatitis (AP) endotypes were identified, two of which were also detected in Acute Respiratory Distress Syndrome (ARDS). Here, the context and hypothesis are presented in the first section to lay the basis for the second part of this chapter, presenting methods that were considered and then employed to test our hypothesis. Results are presented in the following section. Finally, the main findings are discussed and summarised.

4.1 Context

4.1.1 Starting point

When responding to critical injury, the host response can vary greatly between individuals. Many recent studies have focused on explaining this heterogeneity by describing endotypes, also referred to as molecular subtypes, within critical illness syndromes such as sepsis[131,134–137], trauma[138], inluenza[139] and acute respiratory distress syndrome[39]. Usually, variables, such as gene expression values, are used to describe affected patients who are then divided into groups using a clustering algorithm. Potentially, patients from different subgroups may respond differently to the same therapeutic strategy and this could help to identify personalised treatments.

Endotype description always occurs in cohorts with the same illness, however, endotypes may share similarities across different syndromes, meaning these endotypes may represent parallel processes underlying different diseases.

This is what was described previously in chapter 3 which looked at acute pancreatitis data. While studying the results obtained in the main project (chapter 3. Results) similarities in inflammatory signatures between AP and

ARDS were highlighted (section 3.3.4.2). In this chapter, the endotypes will be referred to as 1, 2, 3 and 4, as to not induce confusion between the different analyses carried out.

4.1.2 Hypothesis and aims

Starting from our findings in AP, we hypothesised that similar signal might be observed in other critical illness syndromes. We looked at a combination of datasets from varied sources such as public repositories and collaborating labs in order to test our hypothesis (Figure 4.1). As data could not always be shared directly with us, we liaised with the different research groups to determine the data format and created scripts allowing to compute metrics for those datasets without having the actual data onsite. Specifically, I am very grateful to Dr. Tracy Chew and Dr. Benjamin Tang from the University of Sydney, Dr. Brendon Scicluna from the Academic Medical Centre in Amsterdam, Dr. Tim Sweeney from Inflammatix, and Dr. Justin Whalley and Dr. Julian Knight at the University of Oxford for their contribution.

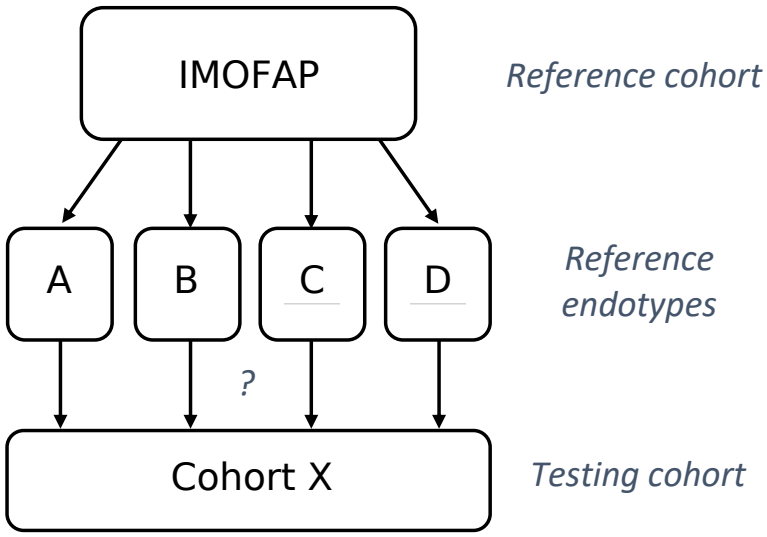


Figure 4.1 – Hypothesis overview. We aimed at testing if the endotypes highlighted in the IMOFAP cohort (identified as A, B, C and D in chapter 3 and referred to here as 1, 2, 3 and 4 respectively) could be detected in other cohorts from individuals with various illnesses.

4.2 Materials and methods

4.2.1 Datasets

4.2.1.1 Summary of used data

*Table 4.1 –Data overview. (*metabolomics data were available for 33 individuals and transcriptomics data for 30 individuals.)*

Cohort name/label	Condition	Number of samples	Controls	Data type	Variables in common
IMOFAP	AP	33/30*	/	Metabolomics/Transcriptomics (RNA-Seq)	/
KAPVAL	AP	312	/	Metabolomics	432
AP 2	AP	87	/	Transcriptomics (RNA-Seq)	19,734
MARS	Sepsis	522	42	Transcriptomics (microarray)	14,104
Sepsis 2	Sepsis	700	/	Transcriptomics (microarray)	8,355
Sepsis 3 and Sepsis 4	Sepsis	403+130	/	Transcriptomics (microarray)	19,766
MOSAIC	Flu	109	130	Transcriptomics (microarray)	10,481
GSE33828	/	/	881	Transcriptomics (microarray)	17,220

4.2.1.2 IMOFAP

The IMOFAP cohort and associated data were described in detail in sections 3.2.1.1 and 3.2.2. We used identified reference clusters as described in chapter 3. In total, 34 individuals were clustered into four subgroups, highlighted using metabolomics, proteomics and transcriptomics data and which were then tested for in distinct datasets.

To perform the comparisons between different datasets, we used transcriptomics or metabolomics data, using single time points (all collected close to recruitment) as no dynamic data was available for the testing cohorts. Moreover, for all testing cohorts, samples collected early in the disease trajectory were available and thus, for consistency and a fairer comparison, we chose early time points from the IMOFAP cohort.

Out of the 34 IMOFAP samples clustered (as described in chapter 3), 33 had metabolomics data available at time point 0. However, adding time point 3 did not permit to include the 34^{th} individual as no data was available.

For the transcriptomics data, using time points 0 (corresponding to recruitment) or 3, when no data was available for the former, we could include 30 samples out of the initial 34.

For the metabolomics data we used the pre-processing applied to the IMOFAP metabolomics data, as described in section 3.2.2 but performed the quantile normalisation and standard scaling on selected time points only, rather than on all time points simultaneously. This was done to have datasets as comparable as possible. Similarly, for the transcriptomics data, we performed standard scaling after time points selection.

This pre-processed data consisted in our reference dataset and each one of the following described datasets was used to test for the four AP endotypes.

4.2.1.3 KAPVAL

The KAPVAL cohort was described in detail in section 3.2.1.2 and consisted of 312 acute pancreatitis-affected individuals with single time point metabolomics data available.

The KAPVAL data was pre-processed as the IMOFAP metabolomics data and as described in section 3.2.2.2.

To be able to compare the two datasets, we filtered them both to only keep variables in common between the two sets. 432 metabolites were retained following this filtering.

4.2.1.4 Pancreatitis data from Benjamin Tang's lab (AP 2 cohort)

Pancreatitis data from another cohort was also available and consisted of gene expression data measured for 87 individuals within 24 hours of admission to hospital. Expression values used to run the analyses consisted of single-end RNA-Seq data sequenced using a HiSeq2500 and normalised using RPKM. The data was annotated using Ensembl identifiers (version 94).

Variables in common were selected before computing the in-group proportion and consisted of 19,734 genes.

4.2.1.5 MARS sepsis data

Sepsis data, provided by Brendon Scicluna, was also available to test our hypothesis and consisted of expression data measured using microarrays[136] (Affymetrix U219 or Illumina HTV3/HTV4). For this dataset, 522 sepsis and 42 control samples were available. Available gene expression values consisted of log2-transformed Robust Multi-array Average-normalised values.

Both our IMOFAP reference cohort and this sepsis datasets were generated using different genome annotation versions (Ensembl version 86 and 87,

respectively), as Ensembl gene identifiers refer to the same entities between different releases, we converted gene symbols from the sepsis dataset to Ensembl gene identifiers using the corresponding biomaRt Ensembl version (87). We then selected genes in common between the two datasets and variables with non-null variance in our reference dataset, this resulted in 14,104 genes. The overlap between the two gene sets is represented in Figure 4.2.

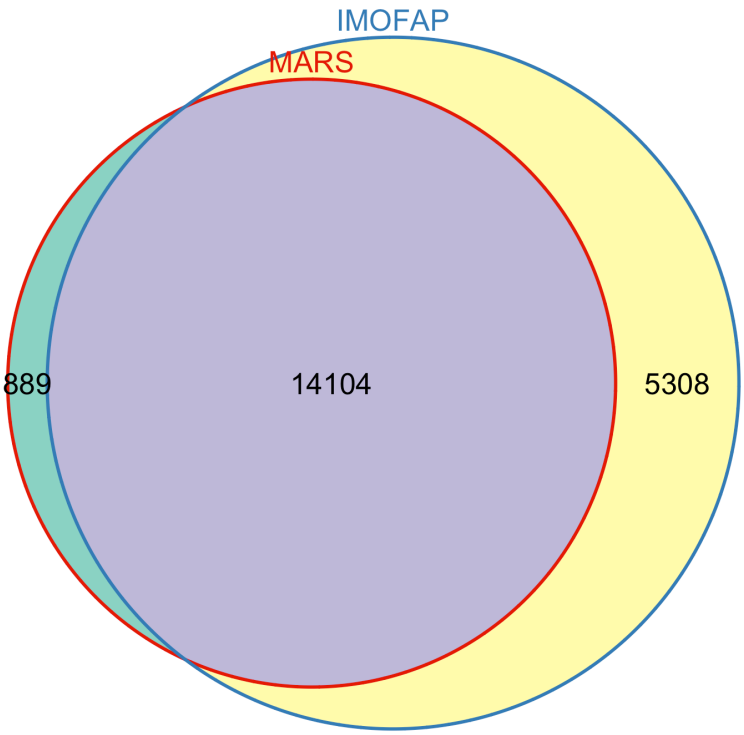


Figure 4.2 – Venn diagram of genes in common between the MARS and IMOFAP gene datasets.

The dataset was filtered to only keep sepsis samples (and thus we dropped the 42 control samples) and allow a fairer comparison with the IMOFAP dataset. As the original data consisted of probe measurements from microarrays, we combined probe sets targeting transcripts of a same gene using the probes with the highest variance values. Indeed, we hypothesised that probes with higher variance were more likely to help in identifying the cohort heterogeneity. The sepsis data was then scaled using Z-scores to maximise the comparability between the datasets.

4.2.1.6 Sepsis data from a pooled dataset from Tim Sweeney's lab (Sepsis 2 cohort)

To test for the four IMOFAP-based AP endotypes, bacterial sepsis data[131] from 14 datasets, consisting of 700 individuals, was used. The different datasets (all from ArrayExpress or Gene Expression Omnibus) were co-normalised as described in the corresponding publication[131]. Transcriptomics data was available for these individuals which was annotated with HGNC names. To compare the IMOFAP dataset to this dataset, we converted our Ensembl identifiers (version 86) to gene symbols using biomaRt and Ensembl's latest version (97).

8,355 genes were retained after filtering variables which were not in common between our reference (IMOFAP) dataset and this dataset.

4.2.1.7 Sepsis data from J. Knight's lab (Sepsis 3 and Sepsis 4 cohorts)

Two cohorts of individuals with sepsis were used to perform the analysis. The first cohort of 403 individuals consisted of community acquired pneumonia sepsis patients. The second cohort was composed of 130 faecal peritonitis sepsis patients.

Gene expression was measured for these patients using microarrays (HumanHT-12 v4, from Illumina).

For these two cohorts, we selected variables in common with the IMOFAP dataset. This resulted in 19,766 genes.

4.2.1.8 Flu data from the MOSAIC cohort

Transcriptional profiles were measured for individuals admitted to hospital with influenza[139] and were available from GEO (identifier GSM3029333). Log2-normalised expression data was available for 130 healthy controls and 229

samples with influenza (199 H1N1, 24 B, 2 A and 4 H3N2) and was measured using microarrays (HumanHT-12 v4, from Illumina). As repeated measurements were available for some individuals, we chose to use the first time point (denoted as T1 and corresponding to enrolment) only, resulting in 109 individuals with flu being selected.

BiomaRt was used to annotate the probes and to extract Ensembl gene identifiers using Ensembl's last version (97). We matched variables in common and 10,481 genes were selected for analysis. An overview of the overlap between the gene sets of the IMOFAP and MOSAIC cohorts is represented in Figure 4.3.

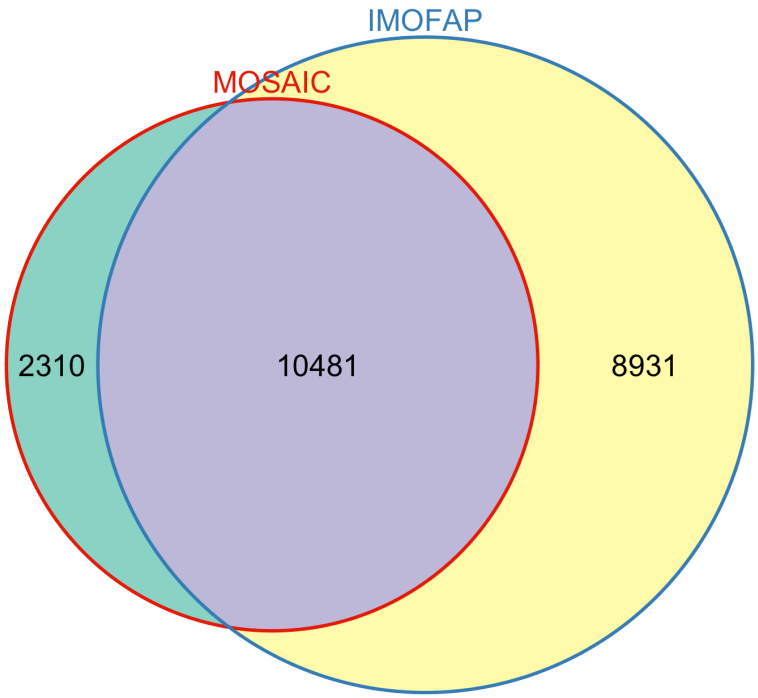


Figure 4.3 - Venn diagram of available genes in common between the MOSAIC and IMOFAP gene datasets.

As described for the MARS cohort, we selected a single probe per Ensembl gene identifier by selecting the one with the highest variance across all our samples (both cases and controls). Finally, a standard scaling applied to each gene was performed.

4.2.1.9 Control data from GEO (GSE33828)

Transcriptional profiles were measured for individuals of different ages to study the variations related to aging. To do so whole blood samples were collected for 881 individuals aged 45 and over as part of the Rotterdam study[140]. Gene expression values were measured using a microarray platform (HumanHT-12 v4, from Illumina).

Probes from the microarray chip were annotated using BiomaRt and Ensembl gene identifiers were extracted using Ensembl's last version (97). We matched variables in common and 17,220 genes were kept for further analyses. An overview of the overlap between the gene sets of the IMOFAP and GSE33828 cohorts is represented in Figure 4.4.

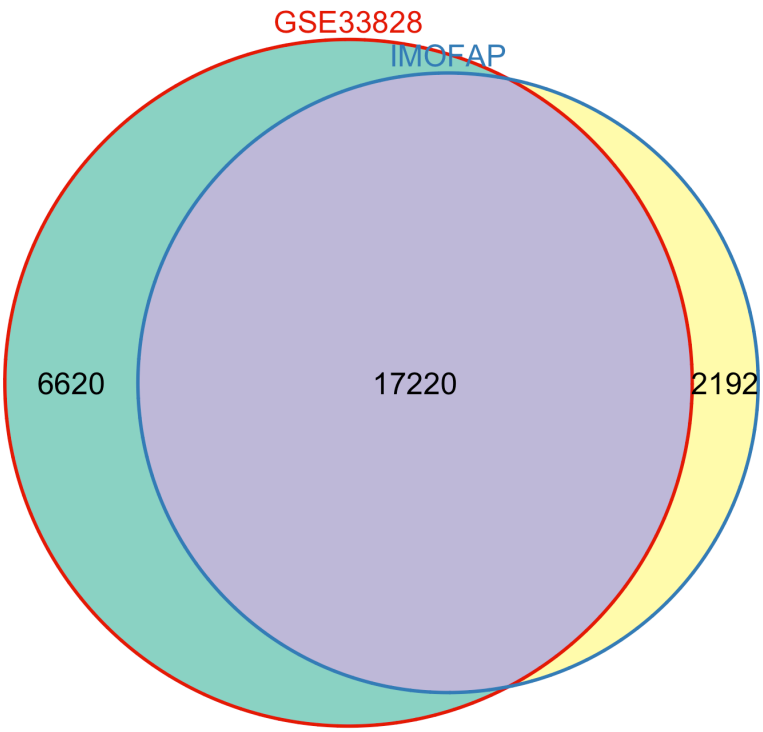


Figure 4.4 - Venn diagram of genes in common between the GSE33828 and IMOFAP gene datasets.

4.2.2 Methods

4.2.2.1 Considered strategies

4.2.2.1.1 PLS-DA-based strategies

To reproduce our findings in a different cohort, we built Partial Least Squares Discriminant Analysis (PLS-DA) models defining each one of our reference endotypes which could then be used to predict allocation values for new samples.

4.2.2.1.1.1 PLS-DA models

To define our endotypes we chose to generate one PLS-DA model for each one of our endotypes. Details and advantages of using PLS-DA models with high-dimensionality data are reported in section 2.2.7.4.1.

For each dataset, we identified variables in common with our acute pancreatitis dataset, as described previously. Once this subset of variables was identified we generated one-vs-all PLS-DA models. For each PLS-DA model, we oversampled/undersampled (explained below) according to the number of samples available for the current group. This was done to prevent one group from dominating the dataset and driving all the differences.

For each model (corresponding to a cluster, one-vs-all design), regardless of the number of cluster elements, we kept all elements which were part of the current (elements labelled as ‘one’) cluster.

If the current number of samples labelled as ‘one’ was greater than the average number of elements per cluster, we performed random undersampling for elements which were not part of the current cluster (not labelled as ‘one’), keeping a number of elements per cluster equal to the size of the smallest cluster. For example, 11 samples were available for group label 1 and 20 for other labels (respectively 10, 5 and 5 for group 2, 3 and 4), as 11 was greater than the average number of samples per cluster (7.75), we kept the 11 samples from group 1 and undersampled the rest (for the 20 samples from the

three other clusters) by keeping 5 elements of each group (randomly selected if more than 5 elements were available for a group).

If the number of samples labelled as 'one' was smaller than the average number of elements per group, we performed undersampling for other classes (not labelled as 'one'), as described in the previous paragraph, and oversampling for the current class (labelled as 'one'). To oversample, we used SMOTE[141] with k=4, being the number of neighbours to use for each point of the current group when generating new data points. One data point was synthesised per sample originally in the group, resulting in a doubled number of samples.

To generate the models and have a robust estimate of both parameters and performance, we used repeated cross-validation with the R package caret[142].

Once the dataset was selected, we reduced the number of variables used in each model to maximise interpretability and prevent overfitting. First, a model with all pre-selected variables was generated and we obtained an estimate of its performance. Importance scores for all variables were generated to rank the variables. The ranking was finally used to drop 20% of the variables with the lowest scores and a new model was generated. This was repeated until one variable remained. For each model, the accuracy value was computed. We then selected the number of variables with the highest accuracy value, if the optimum value was obtained for several variable sets, we selected the one with the lowest number of variables.

4.2.2.1.1.2 Predicted values distributions

Once the four PLS-DA models were generated, we applied them to samples from our new dataset, resulting in four predicted probabilities of belonging to each group for each individual using Bayes' method.

The strategy was to inspect the distributions of these predicted probabilities. We hypothesised that if the signal associated to one of our endotypes was also present in another dataset then we would obtain a bimodal distribution: one peak around 0 for samples not likely to belong to the tested group and another close to 1 for samples likely to be part of the tested group, identifying samples corresponding to that signal.

4.2.2.1.1.3 Spearman's correlations using allocated samples

Our second strategy consisted of using the probabilities generated from the PLS-DA models, as described in the previous two paragraphs. The aim was to allocate new samples to the best matching endotype (corresponding to the highest probability) and compare the average expression profile of allocated samples to the expression profile of each endotype in our AP dataset. To do so we used Spearman's correlation coefficients and computed p-values using a t-distribution.

4.2.2.1.2 In-group-proportion strategy

4.2.2.1.2.1 In-group-proportion

In-group proportion[143] (IGP) was considered as a method to determine if clusters identified in chapter 3, using acute pancreatitis data, were present in independent datasets. The relationship between reproducibility and prediction accuracy is exploited in the IGP strategy. Indeed, a cluster defined in a dataset can be validated in another if predictions are accurate. IGP corresponds to the proportion of elements of a cluster for which their nearest neighbour (using Pearson correlation coefficient) is also classified in the same group. An example is shown in Figure 4.5 in which 5 samples are represented given the values of 3 variables (x, y and z). 3 of these samples (identifiers 2, 3 and 4) are allocated to cluster 2 (represented in blue) and 2 (identifiers 1 and 5) are allocated to cluster 1 (represented in red). Pairwise correlation coefficients are presented in Figure 4.6. To compute the IGP for the cluster 2, we look at

samples 2, 3 and 4 and their nearest neighbours, samples 1, 4 and 3 respectively. 2 of the samples from cluster 2 have their nearest neighbour in the same cluster (samples 3 and 4) but sample 2 does not (sample 1 is in a different cluster), the IGP value is then 0.67 (2/3).

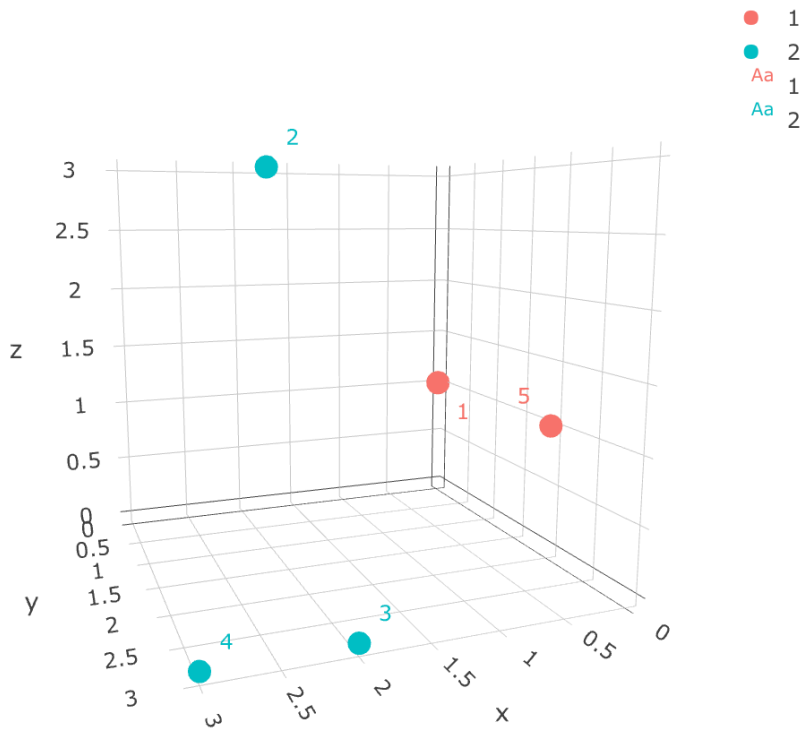


Figure 4.5 – Cluster examples. Colours denote of cluster assignment. Three variables are represented and are referred to as x, y and z.

```
> cor(rbind(xs,ys,zs),method='pearson')
          [,1]       [,2]       [,3]       [,4]       [,5]
[1,]  1.0000000  0.8660254 -0.9449112 -1.0000000  0.0000000
[2,]  0.8660254  1.0000000 -0.9819805 -0.8660254 -0.5000000
[3,] -0.9449112 -0.9819805  1.0000000  0.9449112  0.3273268
[4,] -1.0000000 -0.8660254  0.9449112  1.0000000  0.0000000
[5,]  0.0000000 -0.5000000  0.3273268  0.0000000  1.0000000
```

Figure 4.6 – Pearson correlation coefficients between all pairs of samples.

To compute IGP values for a test dataset, centroids for the reference clusters must be computed. Once computed, samples from the test dataset are allocated to one of the reference centroids. This is done by looking at Pearson correlation coefficients and allocating a sample to its closest centroid. Once all samples are allocated to a centroid, IGP values for the test cohort are calculated.

To determine if IGP values are higher than one would expect by chance only, a reference IGP distribution can be generated by randomly generating centroids and computing IGP values, using for example 1,000 iterations. As some genes will not be independent, centroids are generated using permutations within the principal components-defined space so that they would be plausible data points without being too similar to the original centroids. P-values will then be the proportion of IGP values from our IGP reference distribution that are higher than the obtained IGP values using the real data.

4.2.2.1.2.2 Binomial confidence intervals

P-values generated from permutation distributions might vary depending on the number of permutations used to compute them. As the process of extracting p-values from permutation distributions can be assimilated to counting the number of successes (corresponding to the number of times the permuted samples were equal to or greater than the obtained value) out of a certain number of draws (the number of permutations in total, used for the computation of the p-value), binomial confidence intervals can be used[144]. Especially, for large n (>=40), the Agresti-Coull[145] interval has been recommended[144] and is described below. $qnorm$ describes the corresponding quantile (or the boundary value) of a standard normal distribution and $conf_{level}$ is the confidence level required for the confidence interval (95% for example).

$$z = qnorm(1 - 0.5 * (1 - conf_{level}))$$

$$x' = x + 0.5 * z^2$$

$$n' = n + z^2$$

$$p' = \frac{x'}{n'}$$

$$CI = p' \pm z\sqrt{\frac{p'(1-p')}{n'}}$$

4.2.2.1.3 Network density analysis strategy

Network density analysis[146] (NDA) is an algorithm quantifying density within a subset of nodes part of a wider network. For example, as illustrated in Figure 4.7, if we consider a subset of nodes (represented in red) and wish to quantify the density within this subset, as opposed to the rest of the network, NDA can be used.

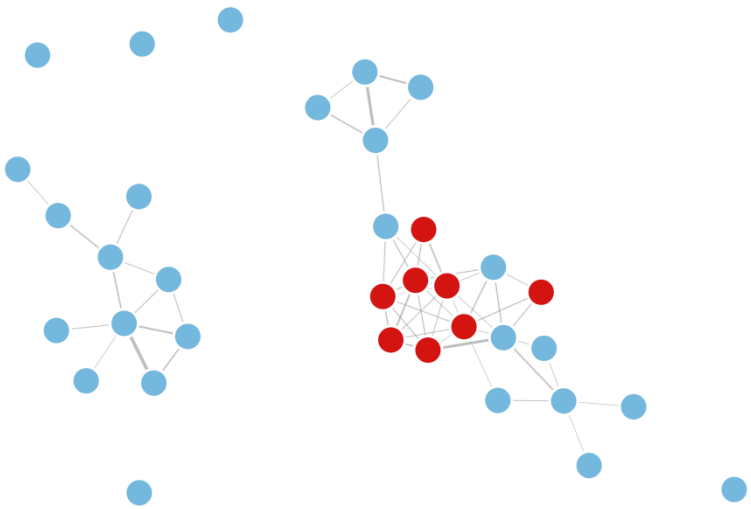


Figure 4.7 – Example network. Red nodes correspond to the subset of nodes we are interested in and edges to the correlations between the different nodes.

Pairwise correlation values between samples are used as input for the NDA algorithm which then, given a subset of the samples, quantifies the subnetwork density. A final value is generated and consists of the sum of $-log_{10}$ probabilities of a relationship at least that strong occurring between two nodes of the current network. To determine if the value is higher than one would expect by chance, this is repeated using random sets of nodes from the same network. Generated values are used to estimate the distribution and a final p-value is computer by comparing the value obtained for the original network with values obtained from randomly generated sets.

For example, we will consider a network composed of 10 nodes (labelled as GSM) and its matrix of pairwise correlations Figure 4.8.

	GSM151369	GSM151370	GSM151371	GSM151372	GSM151376	GSM151373
GSM151369	1.000000	0.894356	0.867166	0.832580	0.000000	0.000000
GSM151370	0.894356	1.000000	0.832979	0.000000	0.788083	0.000000
GSM151371	0.867166	0.832979	1.000000	0.799193	0.893383	0.831689
GSM151372	0.832580	0.000000	0.799193	1.000000	0.000000	0.000000
GSM151376	0.000000	0.788083	0.893383	0.000000	1.000000	0.000000
GSM151373	0.000000	0.000000	0.831689	0.000000	0.000000	1.000000
GSM151374	0.771235	0.832823	0.895209	0.000000	0.000000	0.000000
GSM151375	0.000000	0.808934	0.892203	0.000000	0.000000	0.000000
GSM151377	0.770592	0.931300	0.945329	0.865538	0.783899	0.000000
GSM151378	0.816109	0.000000	0.910047	0.000000	0.910047	0.000000

	GSM151374	GSM151375	GSM151377	GSM151378
GSM151369	0.771235	0.000000	0.770592	0.816109
GSM151370	0.832823	0.808934	0.931300	0.000000
GSM151371	0.895209	0.892203	0.945329	0.910047
GSM151372	0.000000	0.000000	0.865538	0.000000
GSM151376	0.000000	0.000000	0.783899	0.910047
GSM151373	0.000000	0.000000	0.000000	0.000000
GSM151374	1.000000	0.000000	0.000000	0.000000
GSM151375	0.000000	1.000000	0.000000	0.000000
GSM151377	0.000000	0.000000	1.000000	0.000000
GSM151378	0.000000	0.000000	0.000000	1.000000

Figure 4.8 – Correlation coefficients between all pairs of samples.

Now we consider the subnetwork consisting of nodes *GSM151369, GSM151370* and *GSM151371.* For the sake of brevity, 5 permutations will be run. We thus generate 5 permuted sets of nodes: [GSM151369, GSM151376, GSM151370], [GSM151374, GSM151376, GSM151370], [GSM151369, GSM151374, GSM151373], [GSM151373, GSM151369, GSM151371] and [GSM151371, GSM151373, GSM151377]. For each one of the sets of permuted nodes, we compute NDA values. For our original set, we consider correlation values between a pair of nodes and all other nodes of the network. First, we compare node *GSM151369* to node *GSM151370* which have a correlation value of 0.894356. We then extract all correlation values between node *GSM151369* and all other nodes and count the number of values at least this high (2 values meet the criterium and we obtain a value of 2/9). This is repeated for all pairs of nodes (including the reverse comparison between the same two nodes). Once this is done for all nodes of the subnetwork, all probabilities are transformed with $-log_{10}$ and summed to obtain a summary value. This process will be repeated for all 5 permuted sets and obtained values will be used to determine if density of the subnetwork is higher than one would expect by chance.

To test for subnetworks here, we allocated samples from a dataset to IMOFAP endotypes using the closest centroid as described in 4.2.2.1.2.1.

As p-values are obtained using permuted sets, binomial intervals, as described in section 4.2.2.1.2.2 are relevant as well and can be used here.

4.2.2.2 Chosen strategy

We did not choose to use any of the PLS-DA-based strategies. Indeed, even though the idea of generating models was attractive and promising, many tuning steps were involved and had the potential to influence results obtained. The tests we performed were not entirely satisfactory, bimodal probability distributions could be obtained with new data but also with randomly generated data, indicative of an inadequate model. Regarding the allocation strategy, significant correlation coefficients were obtained for case data. However, even though random data did not produce significant coefficients, healthy data did, showing this strategy could identify biologically meaningful samples regardless of their disease status, which was not the topic of interest here.

The NDA-based strategy was not chosen either because of its sensitivity to detect subnetworks. When running tests using the NDA method, p-values obtained were significant in most cases when the input network consisted of a real dataset (as opposed to a randomly generated dataset), whether the input consisted of individuals presenting critical illnesses or not. This could be explained by the allocation strategy, which may not have been accurate enough. It may also be explained by the noise generated by the high dimensionality of the datasets used, which would result in the algorithm detecting commonalties corresponding to widespread biological processes for example.

The IGP-based strategy was selected to perform the analyses. It has been designed for the type of problems we are trying to solve[143], has been tested and is much less likely to be biased as there are no other parameters or filters

involved as part of required pre-processing or the calculation of the metric. Moreover, when generating permuted datasets, the algorithm uses axes of variation from the dataset, thus resulting in plausible samples, which would help in identifying disease samples versus controls.

4.3 Results

IGP results are presented for the different cohorts (two AP cohorts, four sepsis cohorts, one flu cohort and one control cohort) in the following sections. For each cohort, we used 10,000 permutations, unless stated otherwise, to generate the IGP reference distributions. Confidence intervals reported correspond to the Agresti-Coull confidence intervals[144,145] (95%). Reported p-values correspond to the probability of obtaining at least this IGP value given that the group tested is not present in the other dataset.

4.3.1 Case results

4.3.1.1 KAPVAL results

To compare the signals present in metabolomics data between IMOFAP and KAPVAL, we previously used PLS-DA models and Spearman's correlation coefficients (section 3.2.3.4.1). Here, the in-group proportion was used with 10,000 permutations, results are reported in Table 4.2.

Table 4.2 – IGP results for the KAPVAL cohort. Significant p-values (0.05 threshold) are represented in bold characters.

Endotype number	IGP value	Number of allocated samples	p-value	Confidence interval
1	0.73	98	**<0.001**	[0,0.019]
2	0.48	62	**0.025**	[0.014,0.045]
3	0.60	86	**0.026**	[0.018,0.038]

4	0.68	66	**<0.001**	[0,0.006]

4.3.1.2 AP 2 data results

In-group proportions results for the AP 2 cohort are reported in the following table.

Table 4.3 –AP 2 lab data IGP results. Significant p-values (0.05 threshold) are represented in bold characters.

Endotype number	IGP value	Number of allocated samples	p-value	Confidence interval
1	0.65	17	**0.029**	[0.023,0.037]
2	0.62	21	0.122	[0.11,0.134]
3	0.94	31	**0.003**	[0,0.011]
4	0.72	18	**0.017**	[0.013,0.023]

4.3.1.3 MARS results

In-group proportion was used with the MARS cohort as well. IGP values and associated p-values are reported in Table 4.4.

Table 4.4 –IGP results for the MARS cohort. Significant p-values (0.05 threshold) are represented in bold characters.

Endotype number	IGP value	Number of allocated samples	p-value	Confidence interval
1	0.61	131	**0.0498**	[0.035,0.071]
2	0.80	118	**<0.001**	[0,0.006]
3	0.88	168	**<0.001**	[0,0.017]
4	0.61	105	**<0.001**	[0,0.009]

4.3.1.4 Sepsis 2 pooled cohort results

Results (in-group proportion values and associated p-values) for the pooled sepsis data from a previous publication[131] are reported in the following table.

Table 4.5 –Sepsis 2 cohort IGP results. Significant p-values (0.05 threshold) are represented in bold characters.

Endotype number	IGP value	Number of allocated samples	p-value	Confidence interval
1	0.73	188	**<0.001**	[0,0.014]
2	0.67	124	**<0.001**	[0,0.036]
3	0.86	236	**<0.001**	[0,0.041]
4	0.65	152	**<0.001**	[0,0.009]

4.3.1.5 Sepsis 3 and Sepsis 4 data results

CAP cohort:

Table 4.6 – Sepsis 3 cohort IGP results for CAP cases. Significant p-values (0.05 threshold) are represented in bold characters.

Endotype number	IGP value	Number of allocated samples	p-value	Confidence interval
1	0.83	104	**<0.001**	[0,0.007]
2	0.73	79	**<0.001**	[0,0.008]
3	0.87	130	**<0.001**	[0,0.016]
4	0.64	90	**0.04**	[0.029,0.055]

FP cohort:

Table 4.7 - Sepsis 4 cohort IGP results for FP cases. Significant p-values (0.05 threshold) are represented in bold characters.

Endotype number	IGP value	Number of allocated samples	p-value	Confidence interval
1	0.80	35	**0.038**	[0.03,0.047]
2	0.68	25	0.088	[0.075,0.103]
3	0.85	40	**0.014**	[0.008,0.022]
4	0.67	30	0.18	[0.177,0.161]

4.3.1.6 MOSAIC results

We computed in-group proportion values for the MOSAIC data. IGP values and associated p-values are reported in Table 4.8.

Table 4.8 – IGP results for the MOSAIC data. Significant p-values (0.05 threshold) are represented in bold characters.

Endotype number	IGP value	Number of allocated samples	p-value	Confidence interval
1	0.57	30	0.26	[0.24,0.277]
2	0.50	10	**<0.001**	[0,0.251]
3	0.98	41	**<0.001**	[0,0.022]
4	0.61	28	0.13	[0.118,0.144]

4.3.2 Summary of case results

P-values for tested endotypes in the different cohorts are summarised in the table below.

Table 4.9 – Summary of IGP result for tested case datasets. Significant p-values (threshold 0.05) highlighted in bold.

Cohort name/lab	Endotype 1	Endotype 2	Endotype 3	Endotype 4

KAPVAL	**<0.001**	**0.025**	**0.026**	**<0.001**
AP 2	**0.029**	0.122	**0.003**	**0.017**
MARS	**0.0498**	**<0.001**	**<0.001**	**<0.001**
Sepsis 2	**<0.001**	**<0.001**	**<0.001**	**<0.001**
Sepsis 3 and Sepsis 4: CAP cohort FP cohort	**<0.001** **0.038**	**<0.001** 0.088	**<0.001** **0.014**	**0.04** 0.18
MOSAIC	0.26	**<0.001**	**<0.001**	0.13

An overview of significant matches between IMOFAP endotypes and the presented cohorts can be seen in Figure 4.9.

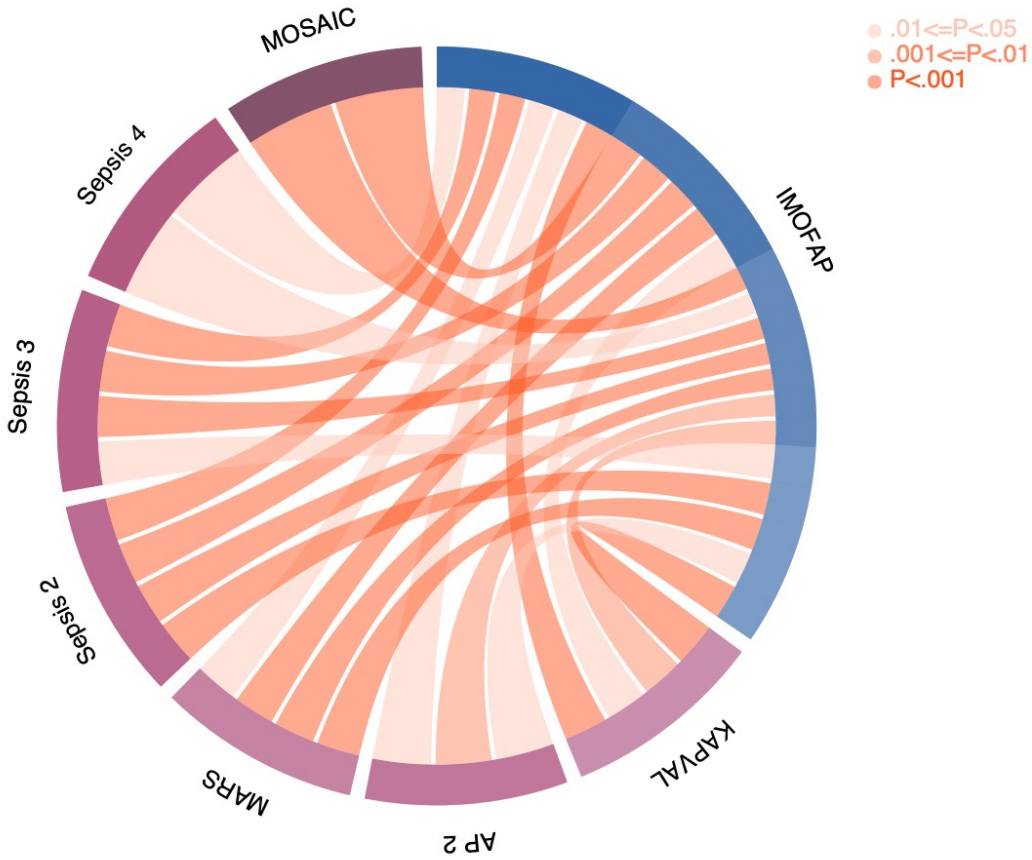


Figure 4.9- Circos view of significant comparisons between the four IMOFAP endotypes and the seven tested case cohorts. The name of each cohort is represented on the outer circle of the figure. The different shades of blue represent the four IMOFAP endotypes. Chords represent significant comparisons. The darker the chord, the more significant the comparison.

4.3.3 Controls results

IGP values were computed for cohorts of control samples. Results are reported in the following tables.

Controls (MARS):

Table 4.10 - IGP results for the MARS control data. Significant p-values (0.05 threshold) are represented in bold characters.

Endotype number	IGP value	Number of allocated samples	p-value	Confidence interval
1	0.50	10	0.233	[0.222,0.244]
2	0.58	12	0.21	[0.199,0.222]
3	0.94	16	**0.007**	[0.003,0.016]
4	0	4	1	[0.978,1]

Controls (MOSAIC):

Table 4.11 - IGP results for the MOSAIC control data. Significant p-values (0.05 threshold) are represented in bold characters.

Endotype number	IGP value	Number of allocated samples	p-value	Confidence interval
1	0.64	33	**0.023**	[0.018,0.03]
2	0.42	19	0.071	[0.03,0.15]
3	0.92	48	**<0.001**	[0,0.039]
4	0.60	30	**0.019**	[0.016,0.026]

Controls (GSE33828):

Table 4.12 - IGP results for the MOSAIC control data. Significant p-values (0.05 threshold) are represented in bold characters.

Endotype number	IGP value	Number of allocated samples	p-value	Confidence interval
1	0.71	255	**<0.001**	[0,0.02]
2	0.72	173	**<0.001**	[0,0.03]
3	0.86	326	NA	/
4	0.50	127	NA	/

4.3.4 Summary of control results

P-values for tested endotypes in the different control cohorts are summarised in the table below.

Table 4.13 - Summary of IGP result for tested control datasets. Significant p-values (threshold 0.05) highlighted in bold. Significant p-values (0.05 threshold) are represented in bold characters.

Cohort name/lab	Endotype 1	Endotype 2	Endotype 3	Endotype 4
MARS	0.233	0.21	**0.007**	**1**
MOSAIC	**0.023**	0.071	**<0.001**	**0.019**
GSE33828	**<0.001**	**<0.001**	NA	NA

4.4 Conclusions, discussion and future direction

4.4.1 Conclusions

We hypothesised that identified subgroups in the IMOFAP cohort were not specific to AP and could be detected in other critical illness syndromes. We

confirmed this using different datasets from multiple sources and in-group proportion measurements. Significant results were highlighted for other AP cohorts. More specifically, for the KAPVAL cohort, all four subgroups were significantly identified, and three subgroups were found significantly matching data from the AP 2 cohort. Similar results were obtained for the sepsis cohorts. For the MARS cohort, Sepsis 2 cohort data and the CAP samples from the Sepsis 3 cohort all four comparisons were found to be significant. For the FP samples of cohort Sepsis 4, two groups were detected in this cohort with significant values. Some overlap was also identified between our four groups and a cohort of individuals with flu with two out of four comparisons being significant.

When performing a validation to make sure that the same results were not obtained when looking at control datasets, we looked at control samples from the MARS and MOSAIC cohorts, and the GSE33828 dataset. For the MARS cohorts, we obtained one significant comparison with group 3. For the MOSAIC cohort, three groups were significant detected (groups 1, 3 and 4). Moreover, for the GSE33828 dataset, we obtained two significant matches (plus two which might have been significant if more permutated sets had been generated). Thus, some common signal was highlighted.

4.4.2 Discussion

When testing for IMOFAP data-based subgroups in other cohorts, some comparisons did not produce significant matches. A possible reason for this result could be that, some samples might be mildly affected by a disease and are actually closer to healthy samples than case samples, thus resulting in smaller IGP values. It could also be explained by small cohort sizes, thus preventing the computation of a reliable IGP value. Another explanation would simply consist of some subgroups not existing for some of the studied diseases or individuals of a subtype not sampled for a given cohort.

The results obtained suggest that critical illnesses share common molecular signatures, but also highlight that there are also likely to be mechanistic and clinically-relevant differences between critical illness responses. This is rational, given that for example AP is a paradigm of sterile systemic inflammation, and faecal peritonitis is clearly due to microbial contamination. We also must be cautious as some of the signal was also detected in healthy samples. This could have several causes. For example, it could be that the ratio of signal/noise is too low or that some shared signal is detected which is in fact due to biological processes common to most samples, whether healthy or affected by a disease. To address this potential issue, variables could be filtered so that only relevant variables are selected, and the true signal can be tested for. Another issue which could arise and cause disease subgroup signals to be detected in healthy cohorts is the mislabelling of some samples. Indeed, some control samples might not be considered healthy because of the study setting or because of mislabelling. Lastly, as illustrated in Figure 4.2, Figure 4.3 and Figure 4.4, the overlap between our study set (IMOFAP) and the MARS/MOSAIC/GSE33828 datasets is quite different. This could lead in variables driving the differences between the different endotypes being lost. Furthermore, although samples from the GSE33828 data are healthy, a good proportion (43.4%, 382 out of 880 samples, one value being unknown) was above 60 years old and might have driven the cohort towards a less healthy gene expression signature.

Using matched cases and controls, or having cases and controls for all studies, would have potentially been an asset as healthy could have been used to perform a normalisation on all case datasets. This could have been done for example using COCONUT[147] (COmbat CO-Normalisation Using conTrols).

Ensemble methods, consisting of combining the results of several algorithms applied to a same problem, might help in palliating the drawbacks of different methods and in reaching an optimal result.

4.4.3 Future directions

Ultimately, this approach could be used to test for our four identified endotypes (IMOFAP cohort) in other available critical illness datasets. This could also be used to test for endotypes of other diseases and not only the subtypes identified in this study.

Our findings show that the study of omics patterns, and particularly transcriptomics, is a potentially promising and novel approach to study severe systemic injury.

5. Chapter 5 – Endotypes in inflammatory bowel disease

This chapter presents a new analysis of existing data focusing on genomic data in inflammatory bowel disease (IBD) affected individuals. The dataset was previously published (doi: 10.1371/journal.pcbi.1005934)[146]. The chapter is divided in five sections. The first section will present background and context information related to the starting hypothesis and data. In section two, the data and methods used to answer our hypothesis will be laid out. Preliminary results will be presented in section three. Discussion and conclusion will consist of sections four and five, respectively. An idea of future work that could be carried out given the results obtained here will also be presented throughout the different sections of this chapter. The input and strategy which could be adopted will be detailed as well as the expected output of the analysis and the impact it may have on the understanding of IBD.

5.1 Introduction

5.1.1 Background

GWAS (genome-wide association studies) combined with transcription data has allowed the identification of loci of interest linked to given phenotypes[146]. FANTOM5[127] cap analysis of gene expression (CAGE) data has allowed us to describe, with high accuracy, promoter specific activity and thus, to quantify precisely shared transcriptional regulation (coexpression) related to specific diseases by looking at associated loci lying within regulatory regions[146]. In short, the study of coexpression patterns has allowed us to identify loci of interest for the study of diseases.

Here, we take two diseases: Crohn's disease (CD) and ulcerative colitis (UC). CD and UC are the two main forms of inflammatory bowel disease (IBD). Spearman correlation-derived p-values from GWAS data and corresponding

coexpression values from the dataset described above were analysed and allowed the separation of loci into distinct clusters. For both CD and UC, identified loci were organised around two components with distinct expression profiles. Selection based on coexpression values permitted us to create two lists of loci for each disease, based on the identified components, that could be further studied. The hypothesis is that each one of these lists is in fact related to distinct forms of the diseases (endotypes), with specific underlying mechanisms, the stratification of which would greatly impact the care provided to affected patients.

5.1.2 Aims and objectives

In order to determine the mechanisms behind these disease components (two for CD and two for UC), corresponding to the four different lists of loci, we analysed genomic data and corresponding clinical data for CD and UC individuals.

Two main strategies were considered. First, to understand why we observed distinct patterns of coexpression, we chose to use BUHMBOX[148]. BUHMBOX allows the user to discern pleiotropy from heterogeneity, that is, to distinguish a same locus affecting different diseases from the presence of subtypes, presenting similar phenotypes but being caused by distinct mechanisms. Here, BUHMBOX could also allow to distinguish two-hit mechanisms from heterogeneity. For example, if we want to distinguish if two groups of independent mutations are required to cause a disease from each one of these mutation groups being linked to a different disease subtype.

Genetic burden can be defined as the relative risk an individual has of developing a disease given his or her genotype. The second considered strategy consisted of quantifying the genetic burden of individuals given their single nucleotide polymorphisms (SNPs) at the loci of interest for each one of the lists and correlate this burden with measured clinical features such as response to treatment.

5.2 Materials and Methods

All analyses were run within a UNIX environment. We also used R version 3.3.2 and Python 3.4.3 for some of the operations.

5.2.1 SNP lists origin

The lists of SNPs were extracted from our previous work[146] using GWAS results obtained as part of distinct studies[149–151]. However, the coexpression analysis was performed again using newer GWAS results[152,153], as the number of significant loci reported was much higher for the latter. I will present only the newest results for both BUHMBOX and the genetic burden analysis.

Network figures corresponding to SNP subgroups for both CD and UC, as described in the paper[146], are presented in Figure 5.1 and Figure 5.2. Corresponding lists of SNPs are detailed in Table 5.1 and Table 5.2.

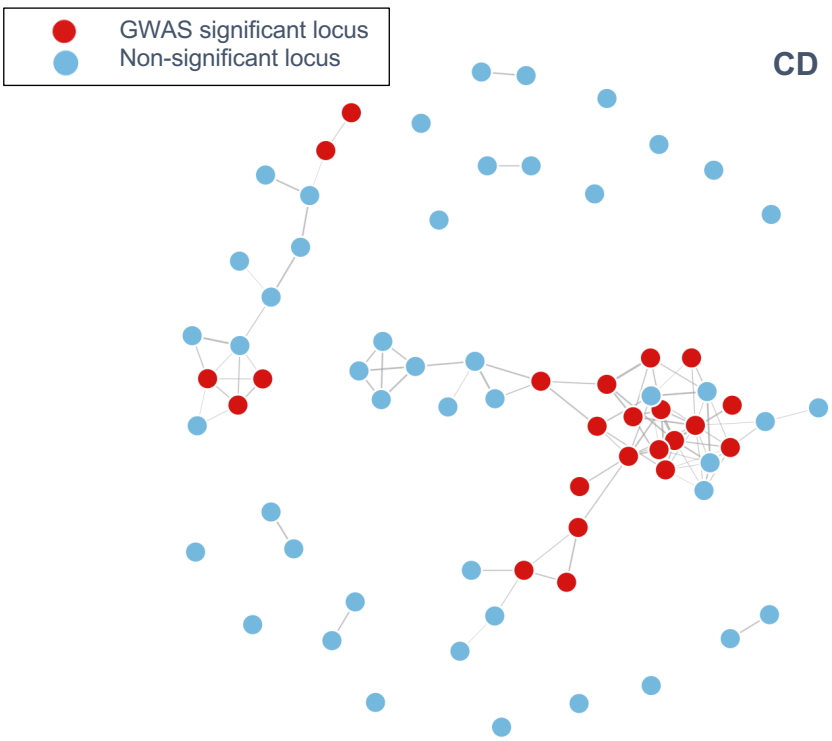


Figure 5.1 – CD coexpression network (available at https://baillielab.net/coexpression/view_results.php?id=cd-meta-remapped_first_db138thresh5e-06_complete_BACKCIRC_pj0.1_f5ep&specialdir=publish4). Each node represents a

SNP (red if significant at a 5e-6 threshold in the GWAS study) and each edge the coexpression values. The edge threshold (-log10 p-value=1.5) was chosen visually so that compact groups could be observed.

Table 5.1 - SNP identifiers for the two main clusters. Only significant SNPs are reported. SNPs located in the same region (given distance and correlation p-value) were merged and are represented between square brackets.

Cluster 1 SNPs	Cluster 2 SNPs
[rs3024505]	[rs9368699]
[rs7900536]	[rs1058207]
[rs8005161]	[rs10065570]
[rs17294280]	
[rs6545835]	
[rs2838522]	
[rs713875]	
[rs7720838]	
[rs3762313|rs3762314]	
[rs1057108]	
[rs2236262]	
[rs9909593]	
[rs1322]	
[rs2070727]	
[rs7759127]	
[rs3135395]	
[rs241448|rs241447|rs241452|rs17034|rs241451|rs241449]	
[rs2351010]	

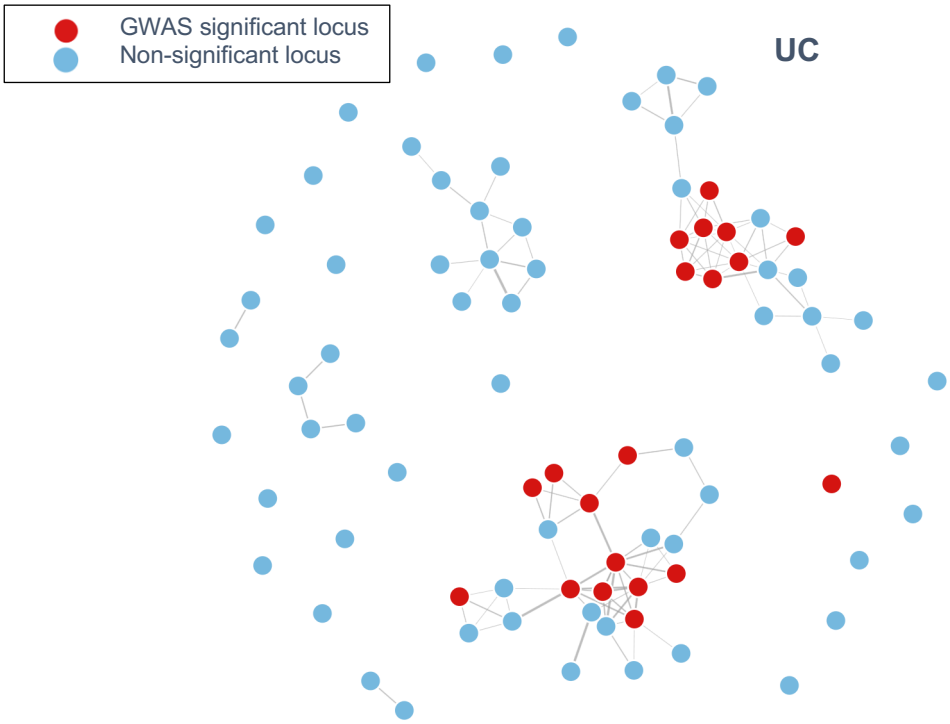


Figure 5.2 - UC coexpression network (available at https://baillielab.net/coexpression/view_results.php?id=uc_db138thresh5e-06_complete_BACKCIRC_pj0.1_f5ep&specialdir=publish4). Each node represents a SNP (red if significant at a 5e-6 threshold in the GWAS study) and each edge the coexpression values. The edge threshold (-log10 p-value=1.56) was chosen visually so that compact groups could be observed.

Table 5.2 - SNP identifiers for the two main clusters. Only significant SNPs are reported. SNPs located in the same region (given distance and correlation p-value) were merged and are represented between square brackets.

Cluster 1 SNPs	Cluster 2 SNPs
[rs949969]	[rs3024493]
[rs10839564]	[rs1886730]
[rs12936231]	[rs2382817]
[rs2427533]	[rs9272426]
[rs9261467]	[rs907611]
[rs7554511]	[rs3135391]
[rs661946]	[rs1058026]
[rs10883371|rs10883373]	[rs4934730]

[rs1048709] [rs3812584] [rs12064796]	

In both cases we can distinctly see two distinct subgroups of SNPs.

Networks obtained using the latest GWAS summary statistics[152,153] are presented in Figure 5.3 and Figure 5.4. Significantly coexpressed SNPs which are part of node subgroups are represented in Table 5.3 and Table 5.4 respectively. In this situation, however, we can see two subgroups for the CD coexpression network but only one for the UC coexpression network.

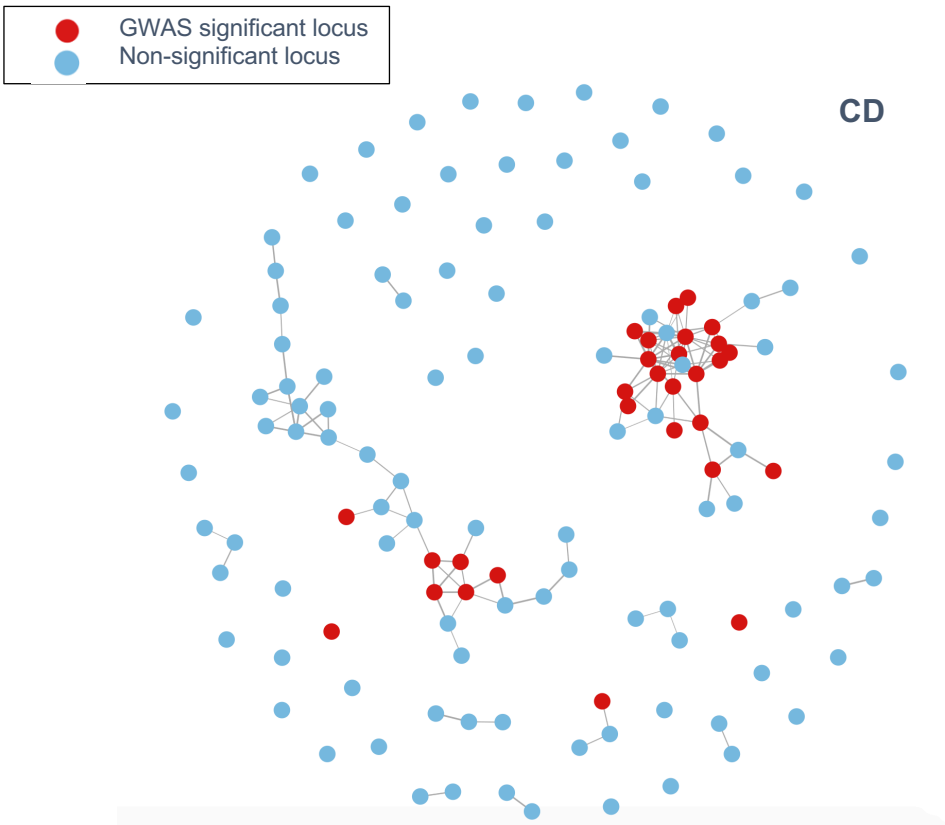


Figure 5.3 – CD coexpression network using updated summary statistics. Each node represents a SNP (red if significant at a 5e-6 threshold in the GWAS study) and each edge the coexpression values. The edge threshold (-log10 p-value=2.08) was chosen visually so that compact groups could be observed.

Table 5.3 – SNP identifiers for the two main clusters. Only significant SNPs are reported. SNPs located in the same region (given distance and correlation p-value) were merged and are represented between square brackets.

Cluster 1 SNPs	Cluster 2 SNPs
[rs3762314|rs3762313]	[rs6887599]
[chr11_61602453_D]	[rs9909593]
[chr19_10491352_I]	[rs9656588]
[chr16_28834254_D]	[rs6545835]
[rs61813286]	[rs2351010]
[rs41293856|rs185786231]	[rs1561925]
	[rs8005161]
	[rs11738827]
	[rs1250566|rs1250567]
	[rs2149090|rs2149091|rs2149092|rs2149093]
	[rs2066848]
	[rs1057108]
	[rs4713555]
	[chr13_99934479_I]
	[rs11679791]
	[rs7734434]
	[rs4256159|rs4257569]
	[rs28510097]
	[rs9370774]
	[rs1422877|rs12651787]

Figure 5.4 - UC coexpression network using updated summary statistics. Each node represents a SNP (red if significant at a 5e-6 threshold in the GWAS study) and each edge the coexpression values. The edge threshold was chosen at random as only two significant loci were present.

Table 5.4 – SNP identifiers for the two main clusters. Only significant SNPs are reported. SNPs located in the same region (given distance and correlation p-value) were merged and are represented between square brackets.

Cluster 1 SNPs
[chr6_31240916_I] [rs9271171|rs9271150|rs9271151|rs536810|rs56245106|rs660895|rs117043483|rs111463829|rs9271148|rs9271149|rs661330|rs9271152|rs9271153|rs9271155|rs9271156|rs535852|rs36233208|rs13205658|chr6_32577644_D|rs17840121|rs35406945|rs9271161|rs9271162|rs9271164|rs9271165|rs13206219|chr6_32577873_D|rs9271170|rs9271147|rs9271163]

As we aimed to use the latest summary statistics files, and only CD subgroups could be highlighted using the coexpression analysis, we chose to focus on CD for all analyses presented below.

5.2.2 Summary statistics

5.2.2.1 Data description

GWAS summary statistics, containing details for each tested SNP and association results, were retrieved from the International IBD genetics consortium' website (https://www.ibdgenetics.org) for both CD and UC. As in the previous section, two sets of summary statistics files were available, corresponding to different GWAS results. We used the most recently generated dataset for our analysis. A sample of the CD summary statistics file is presented in Table 5.5.

Table 5.5 – CD summary statistics sample from the ibdgenetics' website. Values correspond to the latest version GWAS results[152,153]. Chr is the chromosome number where the SNP is located. SNP is the accession number and base pair position is the position in the b37 version of the human genome. A1 is the minor/risk allele and A2 the reference allele. Odds ratio correspond to the association between the phenotype (here, CD or control) and the tested alleles. P is the corresponding p-value for the odds ratio.

Chr	SNP	Base pair position	A1	A2	A1 frequency in cases	A1 frequency in controls	Odds ratio	P
10	rs185339560	2392426	T	C	0.00954	0.00942	0.91	0.55
9	rs11536848	141037798	T	C	0.42	0.419	1.08	0.03
10	rs7894567	1153222	A	G	0.746	0.739	1.04	0.12

These summary statistics were used for the extraction of significant SNPs during the coexpression analysis (using a p-value threshold), to run BUHMBOX and as a reference to compute genetic burden values (using allele codes, odds ratios and p-values).

A potential alternative would consist of re-running the coexpression analysis using association files generated from the available genotyping data (see 5.2.3.1), rather than the ones described here as they corresponded to different datasets and did not fully overlap with the available genotyping data. Moreover, BUHMBOX has been tested using genotype data and association files which were generated using a same dataset and doing otherwise might result in a decreased power.

5.2.2.1 Data specifications for the different analyses

5.2.2.1.1 BUHMBOX required input

To run BUHMBOX, summary statistics for the SNPs of interest are needed. Namely, the SNP identifier, the risk allele, its corresponding frequency in the control cohort, and the odds ratio are required. As BUHMBOX aims to detect heterogeneity and is based on correlations between loci, it is crucial to filter out correlated SNPs beforehand. Such SNPs could distort the actual structure that BUHMBOX is trying to detect and bias the entire analysis, thus predicting an incorrect result. These calculations can be performed on genotyping result files, as described below in 5.2.3, or using a reference panel. Both association files (from different GWAS results versions) can be used for the analysis, in order to compare the results. It would provide insights into the choice of suitable input when running this kind of analyses. However, we did not do it as part of this project.

5.2.2.1.2 Genetic burden analysis input

To compute genetic burden scores, a target association file is required as well. Usually, the association file used to compute the genetic burden scores does not originate from the data on which we aim to quantify genetic burden. Indeed, if there are samples in common between the source (corresponding to the summary statistics) and the target (corresponding to the genotyping data on which we wish to compute the genetic burden scores) it could cause an inflation of the association between the disease and the scores. From the chosen association file, the same fields as the ones required for the BUHMBOX analysis will be needed. Some filtering, which will be described in section 5.2.3, will be performed to choose a SNP subset. Only the newest association file will be used here.

5.2.3 Genotype data

5.2.3.1 Data description

The available genotype data consisted of PLINK-formatted files which we processed using the version 1.90p of the software. PLINK-formatted files come in a variety of different format but one of the most commonly used consists of a set of three files:

- A bim file corresponding to a list of the different markers tested and containing the chromosome identifier, positional information, the SNP identifier, and the corresponding minor and major alleles.
- A bed file which is a binary ped file. A ped file contains 6 fields plus 2 fields per genotyped SNP. The first 6 fields correspond to family identifiers, sex and phenotype information. Each pair of following fields constitute allele calls.
- A fam file containing the same 6 first fields as the bed file.

The PLINK data we had available consisted of IBD cases (both CD and UC) and controls from a previous study[152] which aggregated data from different cohorts of individuals of European descent. We chose not to add other ethnic

groups as they might cluster due to population structure and 'hide' our subgroups of interest. Genotyping was performed using the Immunochip, a custom Illumina Infinium microarray platform consisting of 196,524 SNPs and indels. The loci selection is based on GWAS results from diseases characterised by immune response dysregulation. 33,977 controls samples were available as well as 17,897 CD and 13,768 UC samples (some of which were already pre-filtered).

5.2.3.2 Data pre-processing

Along with the PLINK data, quality control results were available and consisted of a series of filters, some of which had been applied prior to data sharing. These filters are described in details elsewehere[154].

5.2.3.2.1 Filtering of individuals

Individuals not meeting the following criteria were not considered for the rest of the analysis:

- There should be 2% or less missing data (or genotype calls) per sample
- The heterozygosity rate of the sample should not be an outlier (based on F coefficient estimates with an FDR threshold of 0.01)
- There should not be duplicates or related samples (based on pi-hat values with a threshold of 0.4)
- The phenotype information should be available

The missing phenotype and duplicated/related samples filter was applied on the pre-filtered set.

5.2.3.2.2 Filtering of SNPs

SNPs were filtered according to the following criteria:

- SNPs should not be on allosomes
- There should be 2% or less missing data for a SNP across all batches
- There should be 10% or less missing data for a SNP in each batch
- The Hardy-Weinberg equilibrium, stating that the allele frequency should be equal to the genotype frequency, should be respected in all batches (for the batches containing more than 100 samples and with an FDR threshold of 1E-5).
- SNPs should also be present in the 1000 Genomes project panel (in order to fetch relevant statistics and perform imputation if need be)
- Allele frequencies between the batches should be homogeneous in all batches (for the batches containing more than 100 samples, using a chi-square test with an FDR threshold of 1E-5).
- SNPs should not be monomorphic
- The missingness between cases and controls should be of the same order (with a threshold of 1E-5)

5.2.3.3 Data formatting

There are many versions of the human reference genome, therefore it is important to check that the versions between the association and genotyping results file match. Here, our association file corresponded to the b37 version of the genome, but the genotyping results were mapped to a previous version, namely b36. To update the SNP identifiers and coordinates from b36 to b37, we used LiftOver[155].

Whether it is for the BUHMBOX analysis or for the genetic burden analysis, it is important to filter correlated SNPs, as mentioned in 5.2.2.1. To identify correlated SNPs, linkage disequilibrium (LD) can be used and illustrates non-random relationships between different loci. PLINK can be used to compute correlation values and perform pruning using the *indep-pairwise* option. More specifically, for the BUHMBOX analysis, we used a 50kb window, a variant count of 5 (used to shift the window) and a correlation threshold of 0.1, meaning that correlated SNPs within a same 50kb window will be pruned

successively, in a pairwise fashion, until no such correlations remain. The window will then be shifted by 5 variants and the above step will be reproduced. These parameters, which are relatively stringent, are suggested by BUHMBOX's developers[148].

For the genetic burden analysis, such correlations will have a lesser impact and thus we chose a window size of 200kb, with a variant count of 50 and a threshold of 0.25. This allowed a faster analysis and it permitted to retain more variants, which was especially important as the original SNP lists were quite small.

5.2.4 BUHMBOX

BUHMBOX v0.38 was used to perform the analyses presented in this section. For the power calculations, the version 0.1 of the corresponding script was used.

5.2.4.1 Power calculation

BUHMBOX's main script is provided with a power calculation script. This script takes into account the number of cases and controls, the number of loci used in the analysis, the risk allele frequencies of these alleles and their corresponding odds ratios, the estimated proportion of samples expected to be stratified given the current list of loci and the desired significance threshold. A number of simulations will then be run and will provide the user with a power value which can be expected from the analysis. In other words, this will constitute a measurement of our power to detect the heterogeneity, assuming it is present. Under low power, if we obtain a non-significant p-value we cannot say with confidence that the alternative hypothesis is true but we cannot reject the null hypothesis either. To increase the power of the analysis, one can increase the number of samples and/or loci included in the analysis, where possible.

Using proportions of heterogeneity equal to 0, we can compute the false positive rate (FPR). This will correspond to the probability of detecting heterogeneity within a cohort when there is actually none.

5.2.4.2 Analysis

We applied BUHMBOX to determine if a subgroup of patients with CD presented independent genetic characteristics based on each defined list of SNPs that could uncover heterogeneity.

BUHMBOX computes correlations between the different lists of independent loci and patients with CD to identify excessive positive correlations between input loci and a subgroup of patients. If the output p-value is significant then it can be inferred that a distinct subgroup exists within the studied cohort and that the list of imputed SNPs can be used for stratification. If this procedure is repeated using the second list of loci for the same phenotype and a significant p-value is produced, then it can be inferred that at least two subgroups are present in the cohort and that they can be characterised by the input lists of loci.

Input data consists of a list of SNP identifiers, risk allele, minor allele frequency and odds ratios on one side, with imputed GWAS data on the other side to analyse the structure of the selected population, given these loci. These values were extracted from the summary statistics files, as explained previously.

5.2.5 Genetic burden

5.2.5.1 Polygenic risk score

Polygenic risk scores (PRS) are especially relevant because most diseases are polygenic in nature which means that many loci will contribute differently to the trait studied and it can be hard to have an idea of the associated risk.

PRS are a way to summarise the effect of several loci at the same time. It is usually computed, using the information of a summary statistics file, with a weighted sum of the risk alleles. The simplest way of computing the PRS for an individual is shown in the following equation:

$$PRS_i = \sum_{k=1}^{n} OR_k * SNP_{ik}$$

OR_k corresponds to the odds-ratio for the kth SNP and SNP_{ik} is the number of risk alleles for that SNP (which can be 0, 1 or 2), as described in the summary statistics file in section 5.2.2. We consider here a list of n SNPs.

We chose to use Plink to compute the scores. In Plink the formula used to compute polygenic risk scores is slightly different and is as follows:

$$PRS_i = \frac{\sum_{k=1}^{n} OR_k * SNP_{ik}}{P * M_i}$$

Here P corresponds to the ploidy and will be equal to 2 in this case, as we are looking at human data. M_i is the number of non-missing SNPs for individual i. The advantage of adding this denominator is that scores will be scaled and thus it will be easier to compare scores computed using different SNP lists. It is important to remember that PRSs provide relative risks and thus cannot be directly compared rigorously if computed using different SNP lists.

Usually, PRS are used to compare the genetic burden between individuals and identify those which are more (or less) at risk compared to others.

These scores would ideally be used in a clinical setting to identify patients who would, for example, benefit from a closer monitoring. In a 2018 study[156], researchers computed polygenic risk scores for a range of different diseases and tested them using the UK biobank data. They identified patients who had increased risk for these diseases. Such approaches could help the early detection and prevention of diseases.

Ideally, PRS scores are computed on a set of patients which is not overlapping with the set used to compute the summary statistics file. Indeed, this could artificially inflate the association between the disease or trait studied and the SNPs. Here, the available summary statistics were computed on a set of individuals overlapping with the individuals in the genotyping file and thus running the analysis using the genotyping and summary statistics data available would have resulted in inflated associations. The best option would be to get data from another source, if possible.

For binary traits (disease vs healthy for example), the odds-ratios can be log-transformed. Log-transformed values can then be used to weight the different alleles. A negative value will then correspond to a protective allele and a positive value to a risk allele.

Plink will be used to compute the PRSs.

5.2.5.2 Analysis plan

Here, we used the weighted sums to quantify the genetic burden for individuals given the different SNP lists. We used two lists of SNPs significantly associated with CD (Figure 5.3 and Table 5.3). For all CD individuals, we aimed at computing the genetic burden associated to each one of the two lists.

One of the final goals was to compare clinical features for individuals with a high genetic burden in one of the two lists, to individuals having a high genetic burden in the other list. To select them we considered computing the difference between the two PRSs and select the ones with extreme values (in the distribution tails, for example 5% from each side of the distribution). These two groups of individuals could then be compared in terms of their clinical features such as outcome or response to treatment.

Another potential aim would consist of correlating PRS values for both groups with continuous measurements such as blood measurements and or

continuous clinical features, if available. However, clinical measurements corresponding to genotyped samples were extremely sparse and would not have permitted reliable comparisons here.

5.2.5.3 Expected output and potential impact

Significant association and/or correlation with clinical measurements would help to identify potential discriminating features between the two subgroups of individuals highlighted using PRSs as described in the above section. Moreover, it might provide insights into which individuals might respond or not to specific treatments and how they could be identified from a genotyping array for example. Individuals at higher or lower risk might also be identified early on in their disease trajectory.

5.3 Preliminary results

5.3.1 Input data

After applying some additional filters to the pre-processed genotyping data, we retained 27,458 controls, 17,897 CD and 13,768 UC cases for the analyses. However, as mentioned previously in section 5.2.1, using the latest GWAS summary statistics files, we could only obtain distinct groups of SNPs when looking at CD data. Thus, all presented results pertain to CD individuals.

After applying the SNP filtering, 144,245 SNPs remained for the analyses.

Applying liftOver to update the SNP coordinates given the b37 version of the genome resulted in the loss of 7 SNPs and a final set of 144,238 SNPs.

5.3.2 BUHMBOX analysis results

Considering that our lists of SNPs ranged from 3 to 31 unique elements (see 5.2.1), we simulated several scenarios and computed power values using the

tool provided along with the main BUHMBOX script. At best we would have 26 SNPs and 17,897 CD case samples and 31 SNPs and 13,768 UC case samples. For both, 27,448 controls samples would be available. To mimic a case in which we would have performed pruning beforehand, we chose to select only one SNP per region, as defined by an r2 filter of 0.1 and a window of 100 kb (illustrated between brackets in 5.2.1). More specifically, we chose the most significant (given the p-value) SNP per region. This was less stringent than the pruning suggested by BUHMBOX and thus would result in an inflated power value. The aim was to get an idea of what could be done before running the analyses. The longest list of SNPs which could be produced this way corresponded to CD-associated variants as illustrated in Figure 5.3 and Table 5.3. After this filtering, 26 SNPs remained.

Using a significance threshold of 0.05 and proportions of heterogeneity (or proportion of the cohort expect to be stratified given a list of SNPs) ranging from 0.1 to 0.6, we obtained power values between 0.07 and 0.15. These values were low and indicated that, if there was indeed heterogeneity, the chance of a significant result would be below 0.15.

When running BUHMBOX on pre-processed SNP lists (as described in 5.2.3.3), no significant results were obtained.

5.3.3 Polygenic risk score

To generate polygenic risk scores, scripts allowing to compute PRSs from pre-processed PLINK data were written and are available as part of appendices A.2 an A.3.

5.4 Conclusion

The findings from our previous analysis[146], highlighting groups of SNPs lying in regulatory regions presenting similar patterns of expression, led to the

hypothesis that there might be subgroups within CD and UC. When using BUHMBOX and the latest GWAS data we could not reproduce these results, nor we could use them to try and determine if they corresponded to subgroups and to which clinical features they were related. To increase BUHMBOX's power, the number of loci examined could be increased. This could be done in different ways. For example, the coexpression analysis could be performed using different parameter values. Or a different set of input data, for example whole-genotype data (as opposed to microarray data here), could have been used to increase the number of variants available for the analysis. The PRS analysis could not be completed at this time.

However, this does not mean that, under different conditions, subgroups related to these lists of SNPs might have been confirmed using the two strategies presented here.

In theory, the proposed approaches could be applied to any disease, assuming adequate data is available. Results from such analyses could help inform physicians, scientists and patients about disease pathogenesis, the relative risks and the likeliness of a specific patient to respond to a particular treatment, as well as the most likely outcome.

5.5 Discussion

The growing field of precision medicine together with the ever-increasing amount of molecular data collected requires more and more sophisticated analysis strategies to harness the information present and move towards the identification of relevant patient subgroups. The ultimate goal is to improve the quality and outcome of care provided and move further away from the 'one fits all' approach.

Usually, endotypes are highlighted from types of omics data that are not genomic data. Using coexpression analysis and combining GWAS results with gene expression patterns highlighted subgroupings of interest.

To our knowledge, there are no validated methods nor tools allowing the analysis of results from coexpression analyses. Here, we attempted first to validate whether the identified subgroups of SNPs were associated with disease subgroups, using BUHMBOX. However, BUHMBOX gives an answer relative to whether or not the list of given SNPs stratifies the studied cohort but does not return information related to which individuals belong to which subgroups, or whether they are linked to clinical characteristics relevant for the disease under study. For this reason, we put together an analysis strategy to allocate individuals to different subgroups (as described in 5.2.5) and characterise them using available clinical measurements.

For the PRS analysis it could be argued that the odds ratios used for the score computation are biased because they were computed on a full dataset, and thus pertain to CD rather than the potential subgroups of interest. There are two potential alternatives to that premise: first, allocate samples to subgroups given their proportion of risk alleles carried and then perform a new association analysis from which the results could be used as weights for the PRS calculation; second, to use equal weights for all loci. However, the latter could also bias the results. Furthermore, SNP subgroups were highlighted given their association with the studied trait, that is CD, as a whole, and therefore it would be sensible to keep using the original values.

Subgroups in CD have been recently identified elsewhere. A study[157] published in 2016 demonstrated that there are two distinct subtypes of CD based on gene expression and regulation in colon samples. Moreover, these endotypes exhibited differences in immune response and metabolism, and correlated with disease behaviour as well. Along with the manuscript, the authors provided a list of differentially expressed genes and regions showing variation in chromatin accessibility. We naively compared both these results to our two lists of CD-associated SNPs but found no overlap. However, as our lists did not contain many SNPs this could have been expected and it would be interesting to compare the results of both analyses as it could lead to further insight.

6. Chapter 6 – Stratification in a Parkinson's disease dataset

In year 3 of my PhD project, I undertook a 12-week placement in GlaxoSmithKline, Stevenage, UK, in the computational biology group. The project I carried out in GSK is presented in this chapter, which is composed of four main sections. In the first section, after introducing the project, the objectives are stated. Then, methods considered because of their suitability for the type of data analysed are presented in the second section. Results obtained are presented in detail in the third section. Finally, the fourth part consists of a discussion around the results obtained and how they relate to the initial hypothesis and stated aims.

6.1 Introduction and aims

6.1.1 Parkinson's disease

Parkinson's disease (PD) is a progressive neurological condition. It was characterised for the first time in 1817 by physician James Parkinson. This disease results from the loss of neurons in parts of the brain and more specifically in a region called substantia nigra, which is involved in the production of dopamine. This neurotransmitter is responsible for body movement regulation and a reduction of its concentration in the brain will be the main cause of motor symptoms observed in PD cases. A combination of genetic and environmental factors[158] is believed to be linked to the loss of neurons in the substantia nigra. To this day, research has highlighted several gene mutations that appear to be causal in PD[159] but, in most cases, it seems that a combination of factors is involved. Environmental factors, such as head injury or pesticide exposure, have been associated to PD[160].

Between patients, clinical features and progression of PD cases can vary greatly and it is thus difficult to make predictions or understand the underlying

biology involved in this observed heterogeneity. In the interest of improved treatment, it is crucial to identify biomarkers allowing to characterise the course an individual with PD will take and the implications in terms of therapeutic strategy.

6.1.2 Context

The LRRK2 (Leucine-Rich Repeat Kinase 2) protein-coding gene was identified in individuals with PD more than a decade ago[161,162] and has since been a gene of great interest in the study of PD along with other genes such as GBA or SNCA[163].

The G2019S mutation, occurring in the LRRK2 gene, was identified as being the most common PD-related mutation[164–166] (respectively by G2019S genotyping, LRRK2 exons sequencing and LRRK2 exon 41 sequencing). Carriers are more likely to develop PD (reported odds ratio of 9.62 in a GWAS report[167]) over the course of their lives compared to non-carriers. Study of heterogeneity, in terms of disease progression, response to treatment and molecular signatures, in a population of PD individuals (both G2019S carriers and non-carriers) would allow characterisation of the differences between these subpopulations and identify potential non-carriers that have a similar molecular signature to carrier individuals. Moreover, this could shed some light on processes involved in PD, as well as on the reasons behind heterogeneity in PD.

For example, if the difference in progression rate was found to account for some of the heterogeneity, an association with some biomarkers could be tested for. This would help in understanding the differences between slow- and fast-progressing individuals and uncover potential therapeutic avenues for PD-affected patients.

6.1.3 Objectives

The objectives of this study are to characterise PD heterogeneity using several cohorts of PD affected individuals by integrating multiple data types (clinical observations and omics measurements) using unsupervised clustering. More specifically, stratification linked to the G2019S mutation will be highlighted. Moreover, non-carrier individuals with similar data signatures to G2019S carriers will be identified in order to detect patients who might benefit from a similar treatment approach. Other covariates will be tested for correlations with clustering results such as progression rate for example (which can be determined using a PD rating scale, the MDS-UPDRS[168]).

As a second objective, heterogeneity will also be linked to clinical observations. More specifically, we will try to answer the following question: can clinical differences be correlated with distinct molecular signatures based on the available data?

6.1.4 Parkinson's Progression Markers Initiative

To answer the unmet needs in this research area, The Michael J. Fox Foundation for Parkinson's Research (MJFF) has invested into PD biomarker research resulting into a collaborative project, The Parkinson's Progression Markers Initiative (PPMI)[169]. The aim of this project is to collect many different types of data (clinical, imaging and biological) for large numbers of PD patients to identify biomarkers of PD progression.

6.2 Materials and methods

6.2.1 Data overview

6.2.1.1 Selected cohorts

As part of the PPMI, several patient cohorts were available. We selected cohorts containing PD affected individuals only (namely PD, GENPD and

REGPD). Numbers of individuals in selected cohorts and selection criteria applied are detailed in the following table.

Table 6.1 – PD cohorts overview.

Cohort (enrolled participants)	**Description**	**PD diagnosis date**	**PD medication status**	**Mutation status**	**Family mutation status**
PD (423)	De Novo PD subjects	Two years or less	Not taking any PD medication	/	/
GENPD (250)	Genetic Cohort Subjects	/	/	Genetic mutation in LRRK2, GBA or SNCA	/
REGPD (204)	Genetic Registry Subjects	/	/	Genetic mutation in LRRK2, GBA or SNCA or a first-degree relative with a mutation in one of these genes	

6.2.1.2 Available data

An overview of available data for Individuals from the described cohorts is presented in the next figure. Data was organised along three main items: "Study data" containing clinical data and biospecimen measurements, "Imaging data" and "Genetic data" for high-throughput experiments results.

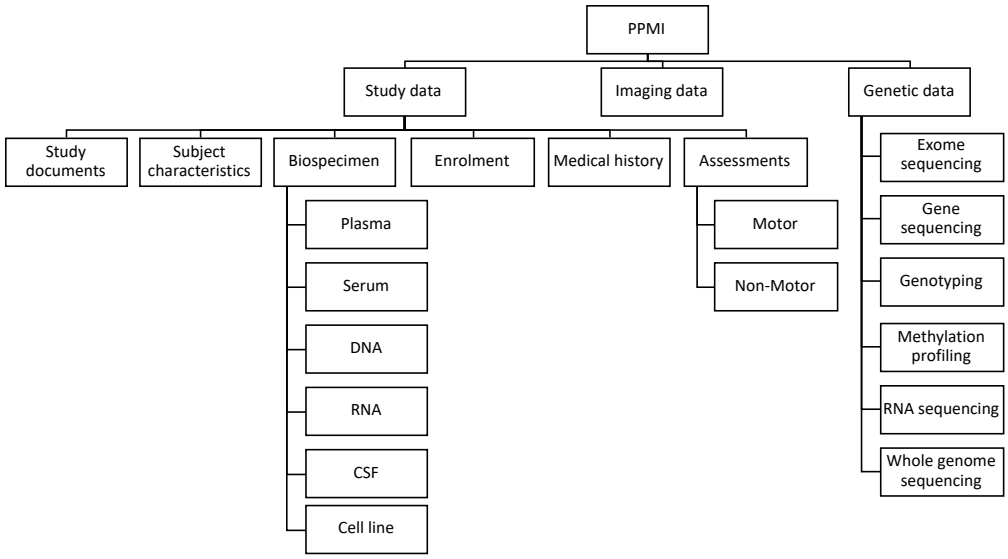


Figure 6.1 – PPMI's website available data. CSF: cerebrospinal fluid.

Long-term follow-up was carried out as part of the study. Presented data are available for different time points between screening visit and up until 5 years after baseline visit.

6.2.2 Data filtering and pre-processing

6.2.2.1 Multi-omics analysis

Individuals and measurements were selected based on several criteria as they could not all be integrated for practical and analysis-related considerations. Indeed, some data measurements had only been done for a limited number of individuals. To be able to look at different data modalities simultaneously, we chose to focus on baseline data as more data was available for this time point. Indeed, for some cohorts, recruitment was still ongoing and some data acquisitions remained to be done.

6.2.2.1.1 Samples filtering

Time between PD diagnosis and baseline visit could vary greatly across cohorts and individuals. This is illustrated in the following figure.

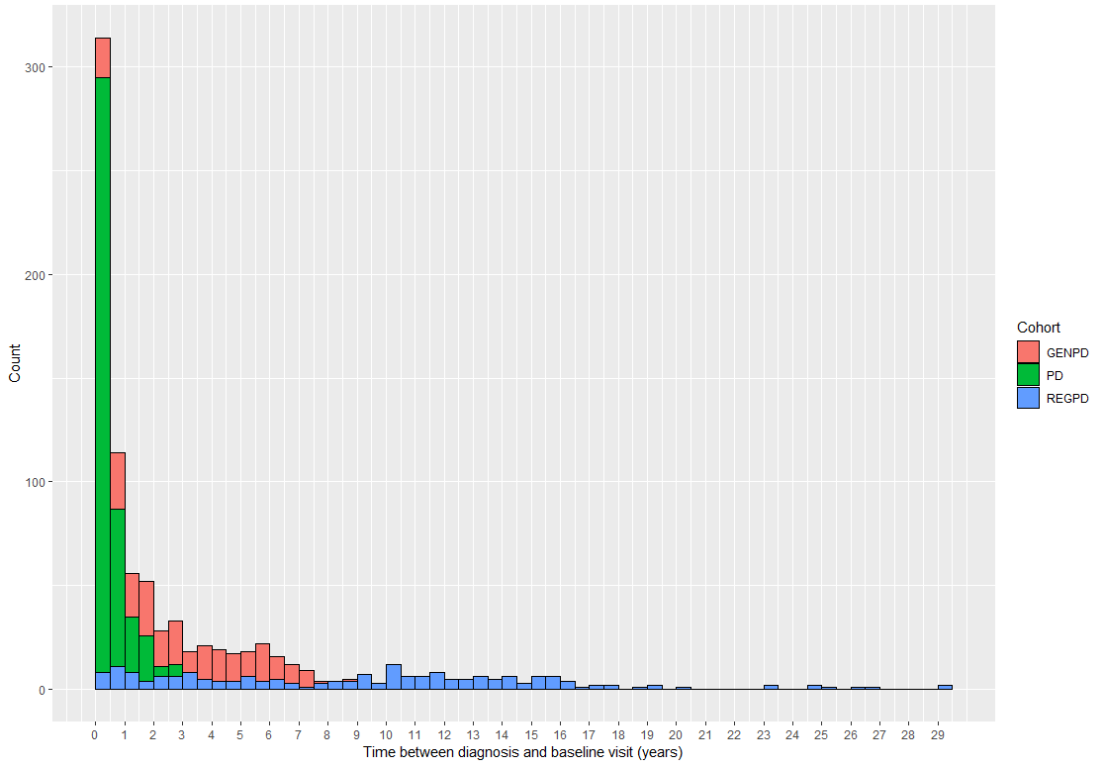


Figure 6.2 – Time between diagnosis and baseline visit in years (each bin represents 6 months and the colour shows the distribution per selected cohort).

To keep as many individuals as possible but also to eliminate a part of the bias that would arise from individuals having larger amount of time between diagnosis and baseline visit, we decided to retain patients who attended baseline visit within 7 years and a half of diagnosis.

We chose to integrate biospecimen measurements (performed on blood, cell lines and cerebrospinal fluid samples), RNA-Seq, DNA methylation and imaging data for which most individuals had measurements performed and available.

Patient filtering is summarised in Figure 6.3.

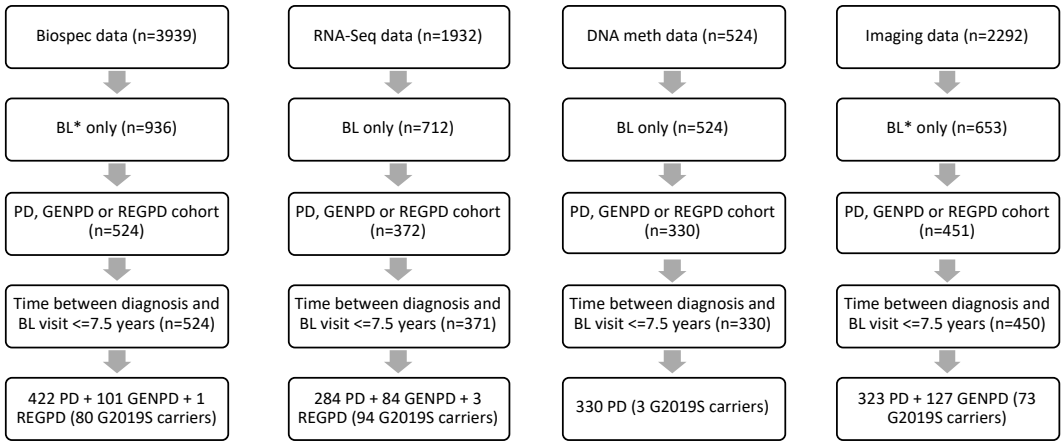


Figure 6.3 -Summary of filtering applied to individuals with resulting numbers for each data type (BL refers to baseline samples. Number of G2019S carriers retained after the filtering are reported as well).

In terms of overlap between the different data types, 129 subjects were reported as having data available for all four investigated data types. 98 individuals had a single data type recorded. In total 610 patients were pre-selected. Figure 6.4 gives an overview of the overlap between the different data types.

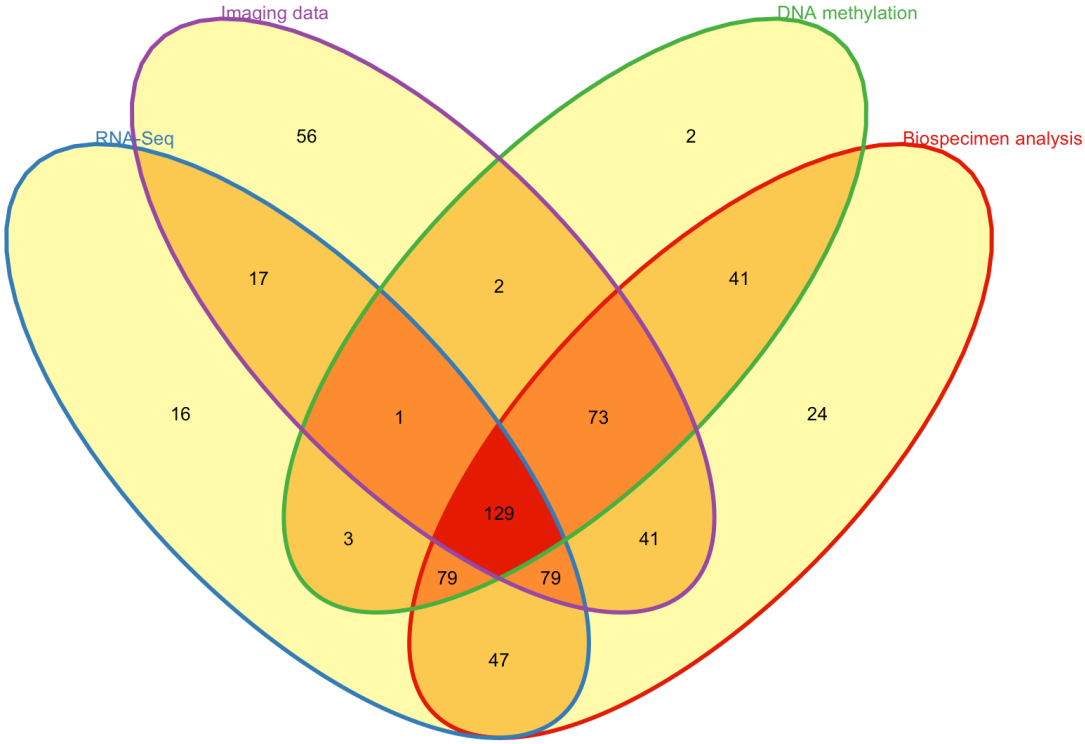


Figure 6.4 - Venn diagram of available data types for pre-selected individuals.

Data was then pre-processed before being analysed, as described in the following section.

6.2.2.1.2 Variables filtering

Regarding biospecimen analysis results and imaging data, some variables had many missing values, whenever this proportion was greater than 50%, the variable was not considered for further analysis and was subsequently dropped.

6.2.2.1.3 Data pre-processing

To account for the differences in time between diagnosis and baseline visit, corresponding to the time point analysed, we generated linear models for each one of the remaining variables. The time values were used as the sole

predictor and residuals were extracted to obtain data free from time-related variation. As some of the models used accepted missing data as input (such as MOFA) and some did not (such as SNF), we generated two sets, one kept unchanged and the other one imputed using k-nearest neighbours (obtained using the package VIM[170] in R). Finally, data was centred and scaled.

RNA-Seq count values were pre-processed using the DESeq2[92] package in R and consisted of count values. To start with, features with no counts across all samples were dropped. A variance stabilising transformation (vst in DESeq2), aiming at stabilising the variance along the range of mean values and involving library depth normalisation, was applied to the remaining data. Values were then log 2 transformed (using a prior count of 0.25) and adjusted for time between diagnosis and baseline visit as well as gender biases using the R package limma[171] (with the removeBatchEffect function). We selected the 5, 000 genes with the highest variance across the studied set.

DNA methylation data consisted of beta-values. After dropping probes located on chromosomes X and Y, values were transformed to M-values as they are more suited for statistical analyses[172]. Similarly to the processing applied to RNA-Seq data, we retained only the 1% probes with the highest variance (8,448 probes retained).

A summary of retained variables per data type is presented in the following table.

Table 6.2 - Multi-omics data overview.

Data type	Number of retained variables	Variables	Details
Biospecimen results	4	pTAu, tTau, ABeta 1-42 and Alpha-synuclein	Measured in cerebrospinal fluid
Imaging	4	Right and left putamen Right and left caudate	Dopamine transporter SPECT imaging
RNA-Seq	5,000	/	From whole-blood samples
DNA methylation	8,448	/	From whole-blood samples

6.2.2.2 Time-series analysis

In parallel to the integration of multi-omics data, to be able to study the dynamics of the disease, the focus was oriented towards time-series data. More specifically, on RNA-Seq data for which four time points (baseline, 12, 24 and 36 months visits) were available.

6.2.2.2.1 Samples filtering

For the analysis of time-series RNA-Seq data, individuals with at least two time points were retained, resulting in 329 distinct patients, all from the PD cohort. Filtering summary and overview of available individuals for each time point are summarised in Figure 6.5 and Figure 6.6.

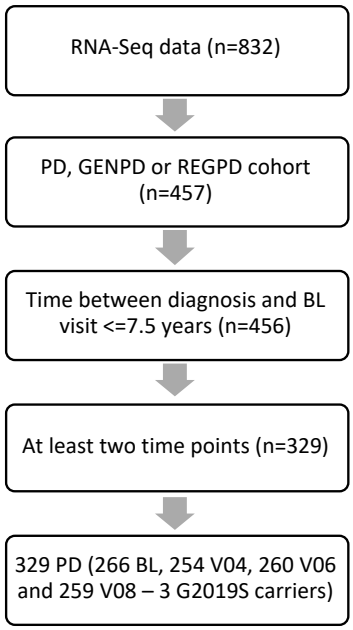


Figure 6.5 - Summary of filtering applied to individuals with resulting numbers for each data type (number of G2019S carriers retained after the filtering are reported as well).

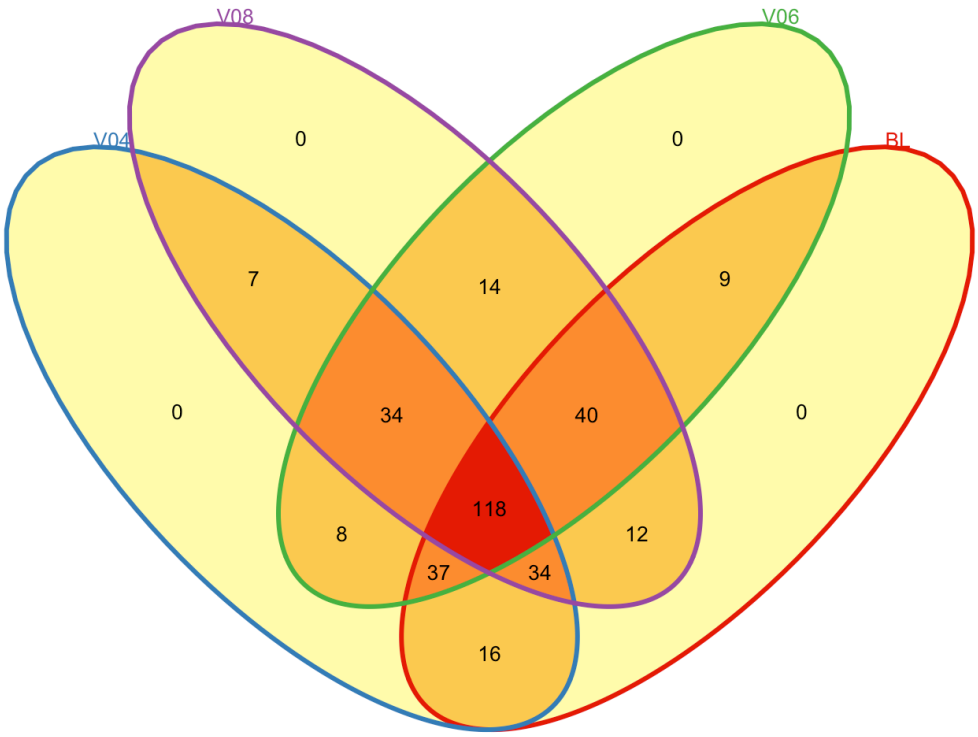


Figure 6.6 - Venn diagram of available time points for RNA-Seq data (BL, V04, V06 and V08 respectively for baseline, 12, 24 and 36 months).

6.2.2.2.2 Variables filtering and data pre-processing

The data was pre-processed using the same strategy that was applied to the baseline RNA-Seq data but for each time point separately. As we wished to study the evolution of genes across time points, we selected the same 5,000 genes for all four time points using the 5,000 genes with the highest variance for the baseline time point.

6.2.3 Methods

Two published methods were selected to analyse the data, namely MOFA[95] and SNF[96], both designed to integrate different data types measured in the same set of individuals.

6.2.3.1 MOFA

The first one, MOFA (Multi-Omics Factor Analysis), is based on multiple factor analysis and aims at identifying the main sources of heterogeneity from a dataset with multiple data modalities. Factors representing the variation, similar to principal components, will be inferred from the data. Each one of the obtained factors will represent an independent source of variation that can be unique to a data modality or shared between several/all. We chose to limit the algorithm to 10,000 iterations (to limit computational burden) or when no significant improvement was accomplished. The latter was defined as a gain equal to or smaller than 0.1 in the ELBO (Evidence Lower BOund) score, which is calculated relative to how well the model fits the data, a higher value was associated to a better model. The algorithm was deemed converging when one of the two conditions was met.

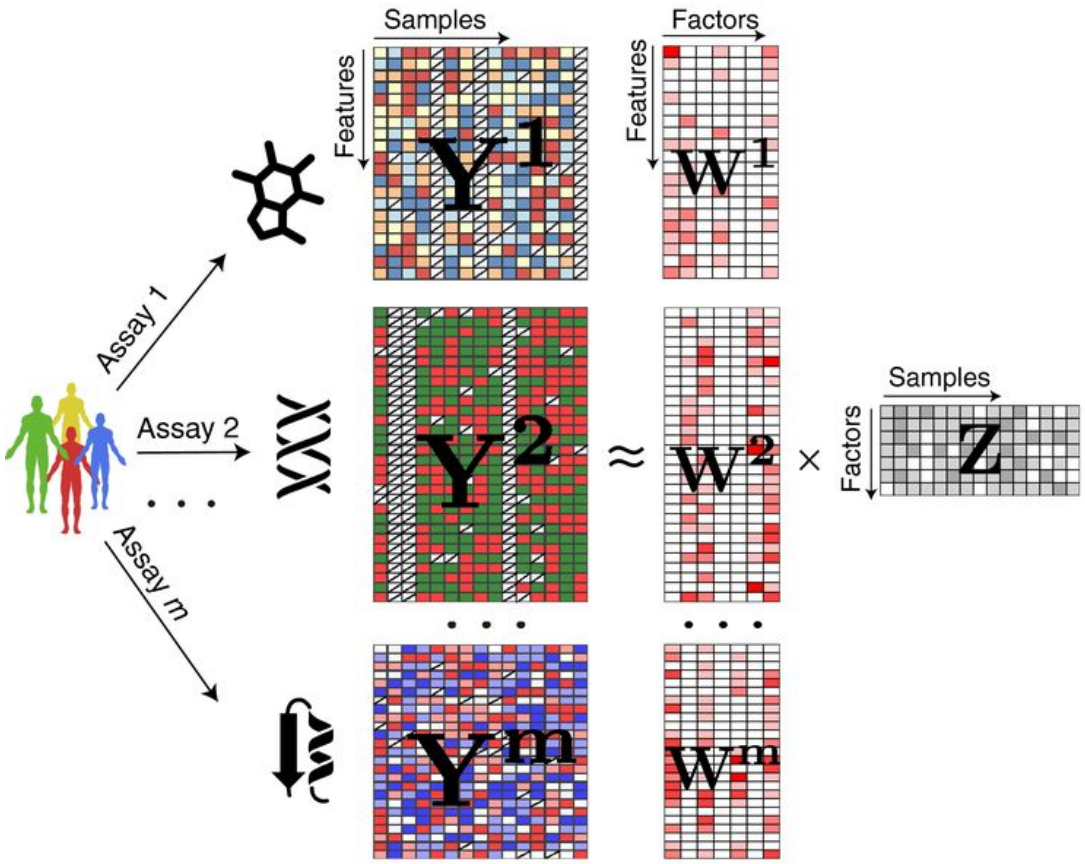


Figure 6.7 - MOFA overview from MOFA's manuscript[95]. Z, the factor matrix and the weights matrices (W) are obtained from the decomposition of the input matrices (Y for the different data modalities).

6.2.3.2 SNF

The second algorithm, SNF (Similarity Network Fusion), aims at producing one view of a dataset, given several data types. For each data modality/type, a pairwise distance matrix between all the individuals must be produced beforehand. This will be used as input for the SNF algorithm. Given each one of the distance matrices, patient networks will be produced. They will then be updated iteratively using data from all other networks to converge towards a final fused network giving an overview of the data. 20 iterations were used when running the fusion step, based on recommendations from the original manuscript[96] and trials.

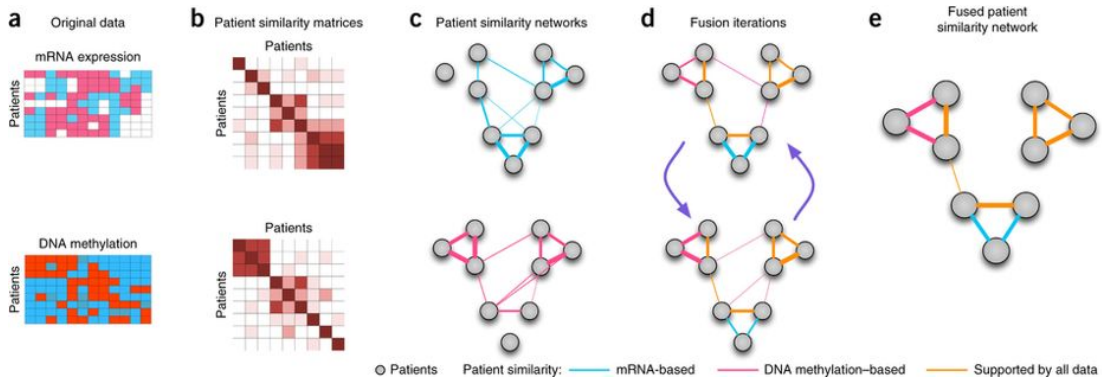


Figure 6.8 - SNF overview[96]. For each data type, patient similarity networks are used to generate the fused network.

6.2.3.3 Commonalities and differences between MOFA and SNF algorithms

MOFA and SNF, by design, take into account differences in input datasets (different number of variables and different distributions for example) and thus will prevent one data modality from dominating the models because of a greater number of variables or a different variance pattern.

The main difference between the two algorithms lies in missing data handling. Indeed, MOFA will deal with missing values within a variable but also missing modality within a sample, the caveat being a potential bias that could be introduced if the structure of missing data is not random. SNF will not accept missing values nor missing modalities and thus they have to be dealt with before using the model, resulting, in most cases, in heavily imputed or reduced datasets (patients with missing modalities will have to be dropped for example).

Batch effect can negatively affect the results obtained in both cases and should be corrected for beforehand or checked for.

Both these algorithms aim to, given different data modalities, produce representative overviews of the variation present in the data. Although similar in their aim, the methods use different strategies and will produce different outputs. We compared the results obtained using the two methods to

investigate their concordance and assess how relevant the information they generate was to our objectives.

6.2.4 Analysis

Two main strategies were adopted to explore the dataset and characterise its sources of variation. For each one of these strategies we applied MOFA and SNF, the two algorithms presented in the previous section. Both tools can be run using R packages respectively named MOFAtools (v0.99.0) and SNFtool (v2.3.0). Analyses were performed with R 3.5.1.

The first strategy consisted of looking at pre-selected biospecimen analysis results, imaging, RNA-Seq and DNA methylation data. At the time, DNA methylation had been performed solely on PD cohort individuals thus, the coverage for the available data was uneven across individuals and cohorts and two other options were considered. The first option was to exclude DNA methylation data and the second option focused solely on individuals from the PD cohort, as the available data was more homogenously collected for this cohort. We also hypothesised that this might reduce the bias linked to integrating data from different cohorts because of the pattern of missing values.

The second strategy focused on the time dimension. Data was collected across a time course, especially, RNA-Seq data was collected over four different time points. We chose to integrate each one of the time points as a different data modality to characterise the variation present in the data across the different measurements.

6.2.5 Clustering

MOFAtools and SNFtool packages allow generation of clusters based on the results of the analyses.

More specifically, using the factors (one or a combination of several) generated by the model implemented in MOFA, one can cluster samples using K-means (described in chapter 2, section 2.2.5.2.2.1 K-means). K-means is an unsupervised learning algorithm that will, using distances between individuals, generate K clusters (where K is a positive integer defined by the user). It first generates K random centroids that will be used as initial conditions. It will then work iteratively by assigning samples to their nearest centroid and by updating the centroids until convergence is reached.

The output of SNF being a similarity matrix, spectral clustering can be used to extract clusters. It is a graph-based clustering that will compute eigenvectors of the Laplacian matrix (matrix used to represent a graph) and use them to extract clusters, using for example, K-means[173]. This algorithm allows the user to capture the global structure present in the graph.

For both clustering strategies one must select a number of clusters that will result in the best partition of the data. As the ground truth, namely the cluster allocations, is unknown, validity and stability indices can be computed to select a 'best' solution among the set of partitions, as described in chapter 2, section 2.2.6.

A 'good' solution would consist of individuals within clusters to be more similar to each other (in terms of pairwise distances) than to individuals from different clusters.

Many validity indices exist and can be computed to assess a clustering solution, they all have strengths and weaknesses related to the way they are computed. To choose a solution, a majority vote can be taken using the suggested best number of clusters suggested by each one of these indices. NbClust (v3.0)[174], an R package, was designed to compute validity indices for different number of clusters and to return the best according to the values obtained. This allows the user to do a majority voting and choose a 'best' number of clusters.

As NbClust was not compatible with the spectral algorithm used with SNF results, a different strategy was adopted to choose the optimal solution for SNF-generated solutions. Within the SNFtool package, a function is available to estimate the optimal number of clusters using two validity indices, the eigengap index as well as one computed using the Laplacian matrix eigenvectors structure.

Stability is also an important feature to consider, indeed, a 'good' partition would be expected to change only slightly when under variation. To assess this, a nonparametric bootstrapping technique can be adopted, where a new set of individuals will be used to compare the new clustering solution to the partition obtained using the whole set of individuals. This new set will be the same size as the complete set and composed of a random draw with replacement from the same pool of subjects. This was be performed using the clusterboot function from the fpc (2.1-11.1)[80] package.

6.2.6 Downstream analyses

6.2.6.1 Extracting the results

For all generated models, data was clustered as described in the previous section. A visual inspection of the results was carried out to try to highlight a link between selected covariates not included in the models (such as G2019s carrier status or rate of progression, determined using parts I to III of the MDS-UPDRS, a PD rating scale, integrating non-motor and motor assessments) and identified clusters.

As the output of MOFA consisted of factors inferred from the data, positions of individuals along these factors could be extracted and visualised using, for example, violin plots (showing actual values as well as their distribution) or two or three-dimensional scatter plots.

The SNF algorithm produced a fused similarity matrix consisting of pairwise similarities between analysed individuals. The matrix could then be illustrated

as a network (the individuals being represented as nodes and the similarity between them as edges) and overlapped with covariates of interest.

Clusters obtained using the outputs of each one of the two previously described methods were compared as well to assess whether they conveyed similar information.

6.2.6.2 Enrichment analyses

To understand the biological processes driving the variation used to extract clusters, enrichment analyses were performed using the Reactome database[175].

As part of MOFAtools, the runEnrichmentAnalysis function was available to perform enrichment analysis using the results of MOFA and more specifically the factors, thus being independent of the chosen partition. This function is based on the idea of principal component gene set enrichment[176] and uses loadings (relative to the importance of each variable for a given component) from computed factors to quantify variables contributions.

From SNF output, for each gene, normalised mutual information (NMI) score was computed against SNF clustering results allowing to produce a ranked list of elements. NMI values are comprised between 0 and 1, the latter corresponding to a perfect correlation between the group labels and the variable of interest. ReactomePA[177] (1.24.0) package was then used to perform enrichment on selected genes, given an NMI value threshold or choosing the N genes with the highest NMI values for example.

6.2.6.3 Comparisons of results obtained with MOFA and SNF algorithms

To quantify the overlap between the results obtained with both algorithms, in order to assess if one of the algorithms was better at identifying a structure of interest, we needed to make sure they were indeed comparable. For the

comparison to be as fair as possible, MOFA was re-run using only the subset of individuals that was used to run SNF. We then extracted the best number of clusters, as previously described in section 6.2.5. Rand index was used to compute a value of the overlap between the two cluster solutions. It is calculated using the number of elements being in the same cluster in both solutions, as well as the number of elements being in different clusters. As it was possible to rank the variables in each one of the solutions, using factor loadings for MOFA results and NMI for SNF results, the ranks were compared by computing correlation values using Spearman's correlation coefficient. Input data, or a pre-selected subset, was also represented and groups obtained from MOFA and SNF results were added to the plot to visually identify potential trends and/or stratifying traits in the data.

6.3 Results

For each one of the presented strategies, only the best solution, as described in section 6.2.5, will be presented here.

6.3.1 Algorithm outputs

MOFA was applied to three different data sets, as described in the analysis section. Variances explained per data modality and per factor were extracted. The number of factors computed by MOFA was chosen as to drop any factor explaining less than 1% of variance in the dataset, across all data modalities.

To determine whether an approach seemed relevant in terms of G2019S mutation status or progression rate, patient values along MOFA identified factors were plotted and values for these variables of interest overlaid.

Similarly, SNF was applied to the same three datasets, filtered for some samples to avoid any missing values, as described previously. The distance metric used was the Euclidean distance. Whenever comparing results between MOFA and SNF, for consistency, MOFA was re-run on the same filtered set of individuals and variables. For each run, the final fused matrix will

be represented. On top of this same matrix, variables of interest will be overlaid to identify any possible stratifying trait. The contribution of each data type to the final matrix will also be reported.

6.3.2 Results using multi-omics data

6.3.2.1 All cohorts with RNA-Seq, imaging and biospecimen analysis data

6.3.2.1.1 MOFA results

Input data is presented in the following table. In total, this consisted of 466 individuals, 90 of them were G2019S carriers.

Table 6.3 - MOFA input data overview for PD cohort and multi-omics data.

MOFA input data	**Biospecimen data**	**RNA-Seq**	**DNA methylation**	**Imaging data**
Variables included	4	5 000	/	4
Individuals included	448	352	/	340

As represented in Figure 6.9, RNA-Seq seems to explain most of the variance (R2=74.70%) and imaging data only a small proportion (R2=6.31%). Biospecimen data contribution falls under 1%.

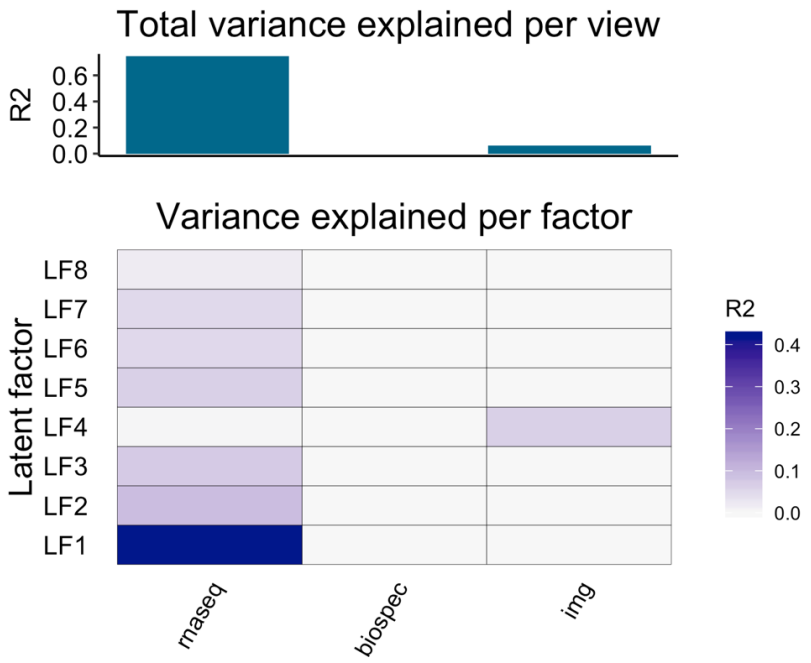


Figure 6.9 - Proportions of variance explained per factor and per data type.

Figure 6.10 represents coordinates of all individuals for each one of the first five factors (the 5 representing the most variance, LF1 to LF5). Visually, to check whether selected covariates were stratifying, values were represented as colours on the coordinates graphs. For example, progression rate was coded as 'fast' or 'slow' given the number of MDS-UPDRS points (parts I, II and III) gained on average per month. If 0.5 or more points were gained per month, on average, then the individual was characterised as fast progressor. Otherwise it was classified as slow progressor. To visualise the carrier status of samples, different shapes of points were used given the mutation status for G2019S.

No obvious grouping could be observed regarding progression rate or carrier status.

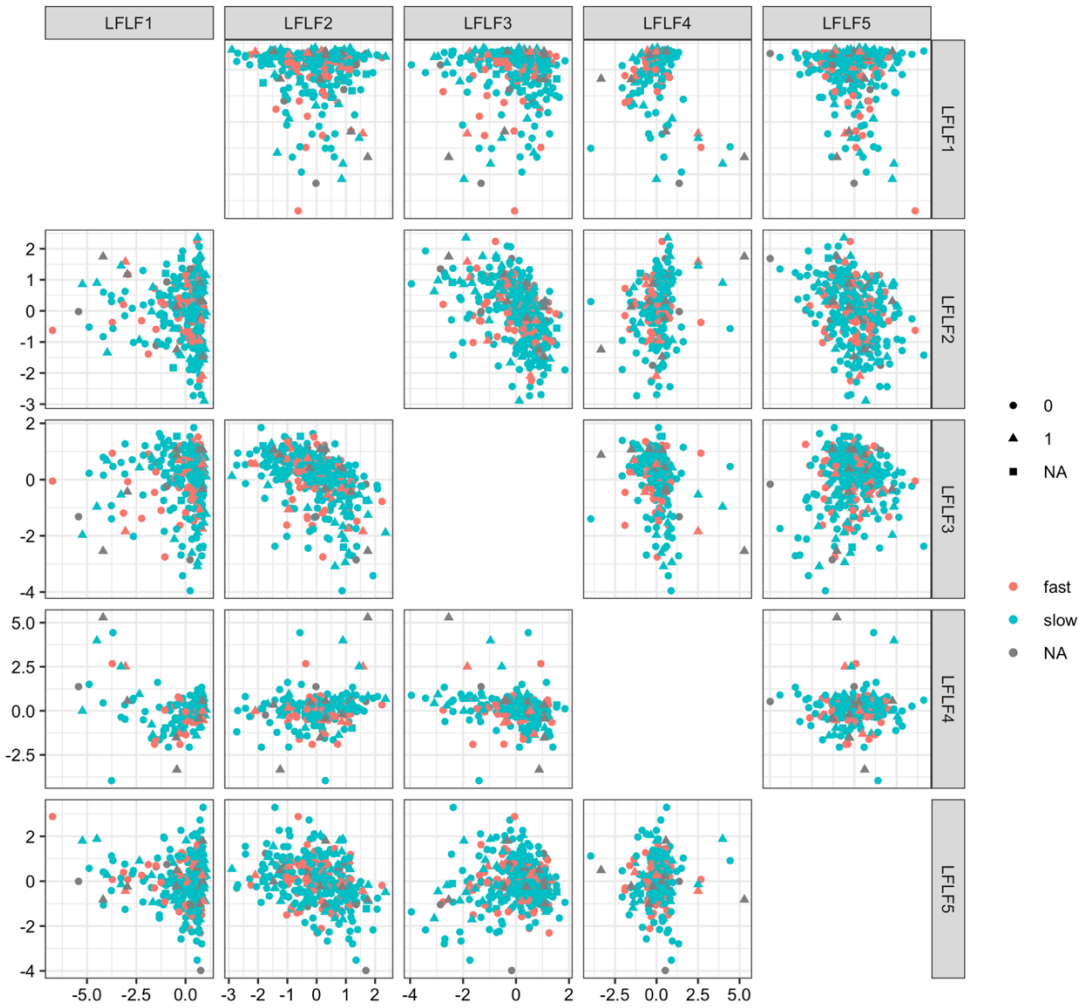


Figure 6.10 - Pairwise plots for top 5 MOFA factors (marker colour represents progression rate and shape represents G2019S carrier status). No obvious stratification can be observed.

Enrichment using Reactome pathway database was performed. Using an FDR threshold of 1%, differentially expressed pathways were identified for all eight factors. This is summarised in Figure 6.11.

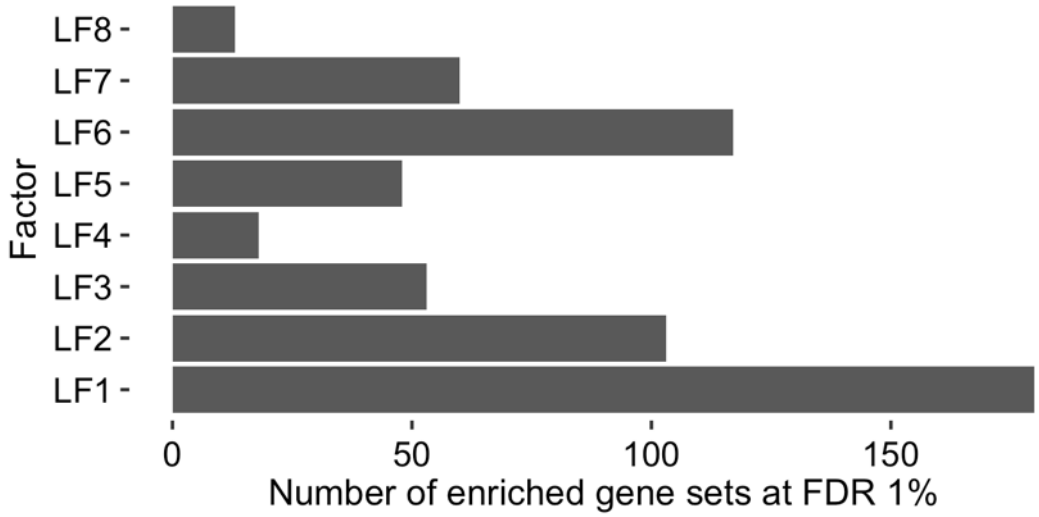


Figure 6.11 - Enrichment analysis results per factor.

6.3.2.1.2 SNF results

For this analysis, there were 208 individuals, 33 of them were G2019S carriers.

Table 6.4 - SNF input data overview for PD cohort and multi-omics data.

SNF input data	**Biospecimen data**	**RNA-Seq**	**DNA methylation**	**Imaging data**
Variables included	4	5 000	/	4
Individuals included	208	208	/	208

As part of the SNF algorithm, affinity matrices were computed and used to generate a fused matrix representing similarity values between pairs of individuals in the networks. Zero-values are attributed to non-neighbours. Each data type is represented in Figure 6.12 to Figure 6.14 and can be compared to the obtained fused matrix plotted in Figure 6.15. The order of samples represented in the matrices was defined by the labels from the 2-group spectral clustering results using the final fused matrix (2 was identified as the optimum number of clusters, as described in section 6.2.5) to allow visual comparison.

Progression rates (as described in 6.3.2.1) and carrier status were overlaid to the plot for visual inspection but no stratification was observed.

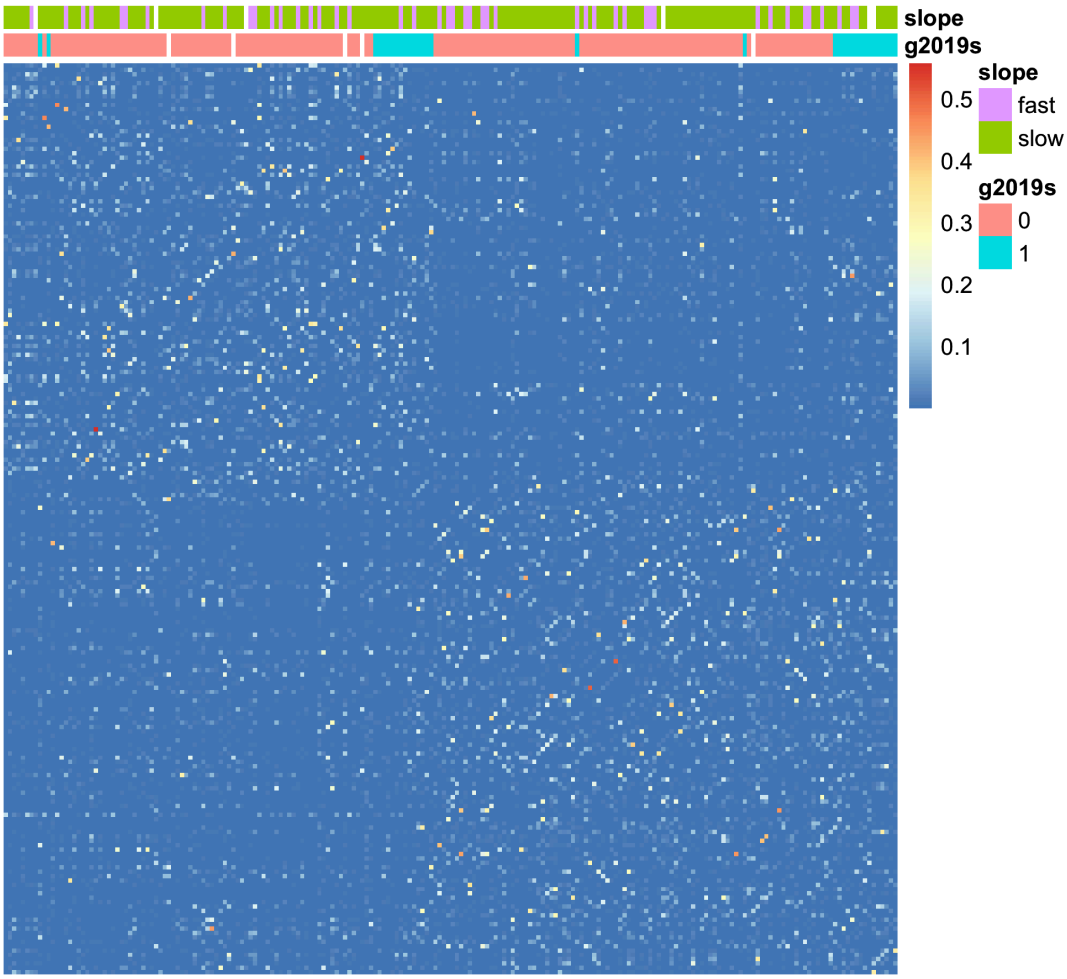


Figure 6.12 - Affinity matrix for biospecimen analysis results (ordered by SNF subgroups, progression rate and carrier status are represented). The two groups can be respectively seen in the top left corner and in the bottom right corner of the matrix.

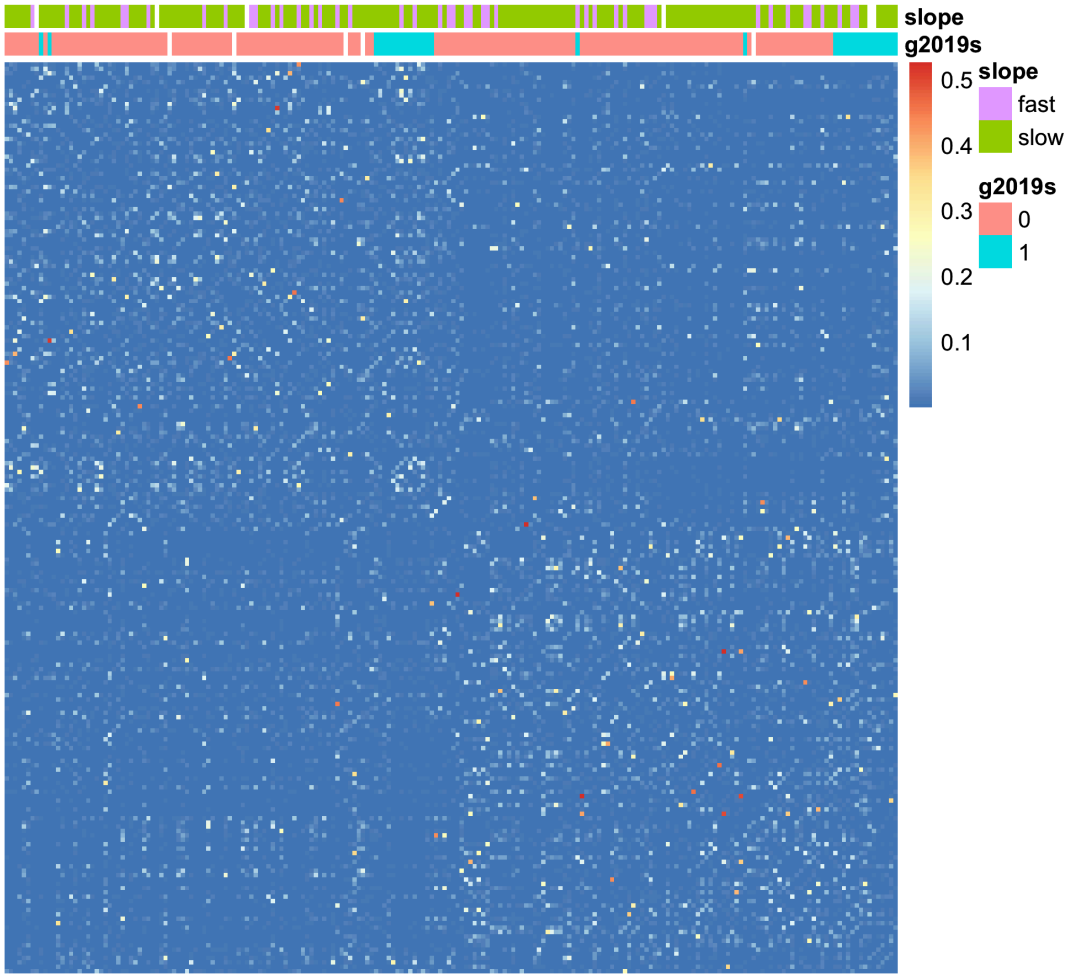


Figure 6.13 - Affinity matrix for imaging data (ordered by SNF subgroups, progression rate and carrier status are represented) The two groups can be respectively seen in the top left corner and in the bottom right corner of the matrix.

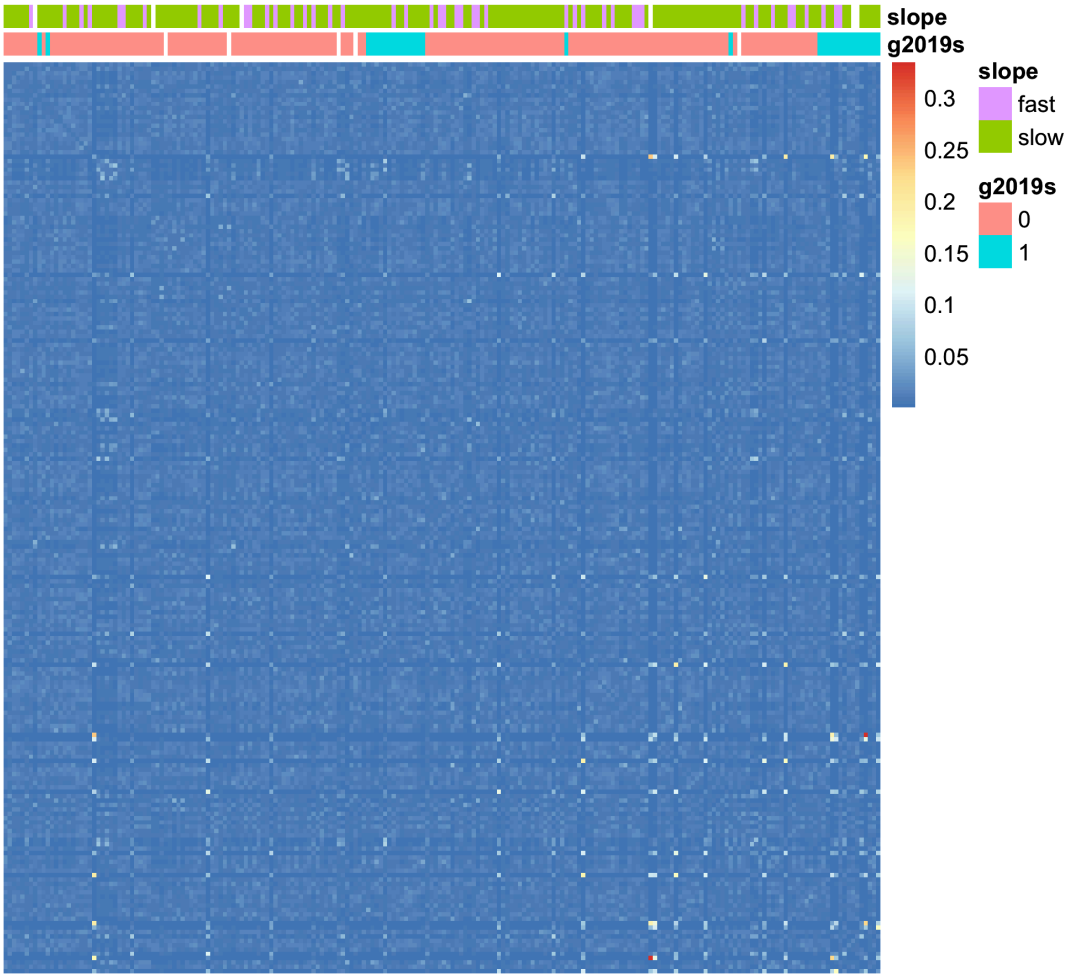


Figure 6.14 - Affinity matrix for RNA-Seq data (ordered by SNF subgroups, progression rate and carrier status are represented as well). The two groups cannot be identified for this graph.

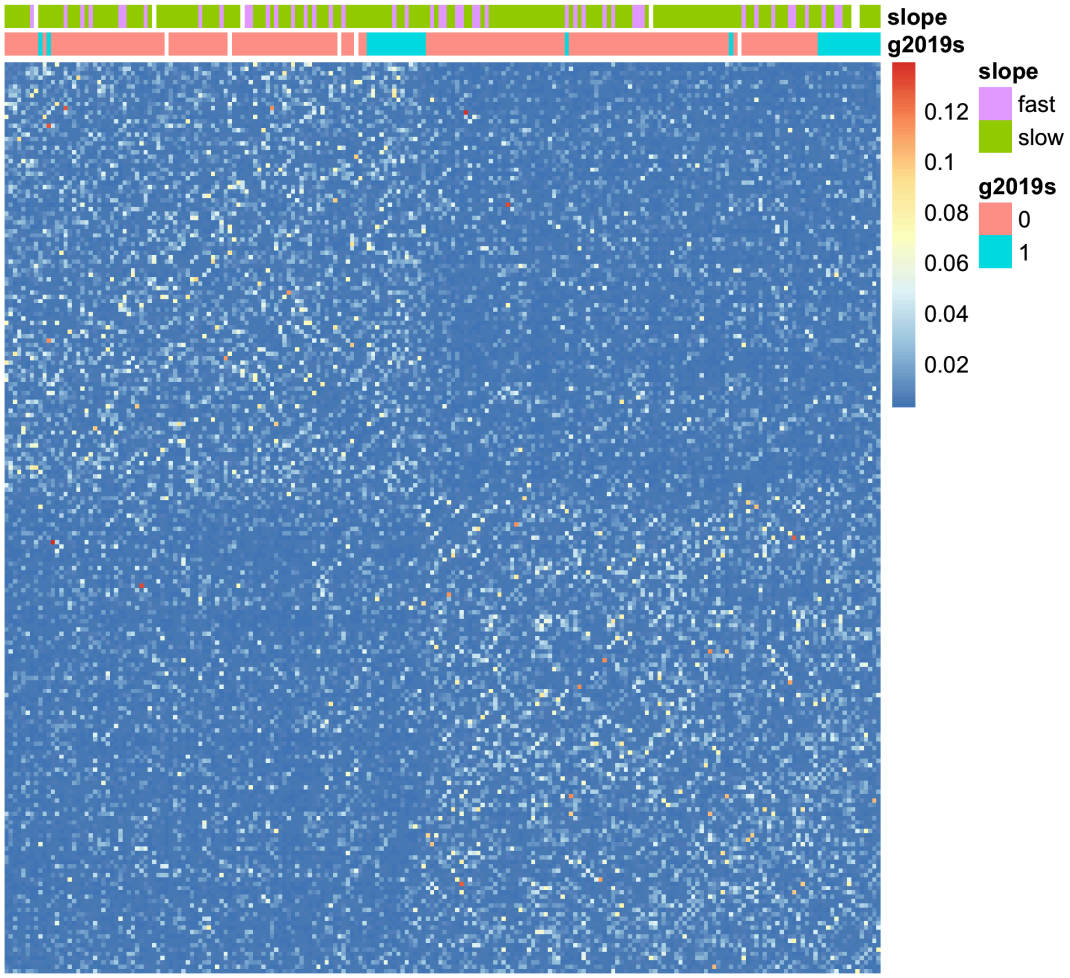


Figure 6.15 – Fused matrix ordered by SNF subgroups and representing progression rate and carrier status. Two distinct groups are observed in the top left corner and in the bottom right corner of the matrix.

NMI, measuring mutual dependence, was computed between all 2-cluster solutions extracted from affinity matrices (for individual data types and complete dataset) to assess the individual and shared contributions of each datatype to the clusters extracted from the fused matrix obtained using SNF. An NMI value of one would indicate a perfect cluster overlap. The NMI values are low (Table 6.5) and we can see that the fused matrix is quite different from the individual data types matrices. This indicates that, using individual datatypes, it would not be possible to highlight the structure extracted from the fused matrix.

Table 6.5 - NMI concordance matrix. (A value of one will be associated to two identical objects and a value of zero will indicate no mutual information)

	Fused results	**Biospecimen data**	**Imaging data**	**RNA-Seq**
Fused results	1			
Biospecimen data	0.02346255	1		
Imaging data	0.00991664	0.00020970	1	
RNA-Seq	0.02558910	0.01275553	0.01506157	1

6.3.2.2 PD cohort individuals with RNA-Seq, DNA methylation, imaging and biospecimen analysis data

6.3.2.2.1 MOFA results

For this analysis, in total, 412 individuals were available, 7 of them were G2019S carriers.

Table 6.6 -MOFA input data overview for four omics data types.

MOFA input data	**Biospecimen data**	**RNA-Seq**	**DNA methylation**	**Imaging data**
Variables included	4	5 000	8 448	4
Individuals included	405	283	328	269

As illustrated in Figure 6.16, RNA-Seq explains most of the variance (73.24%) and DNA methylation data a smaller proportion (3.29%). Biospecimen and imaging data contributions are both under 1%.

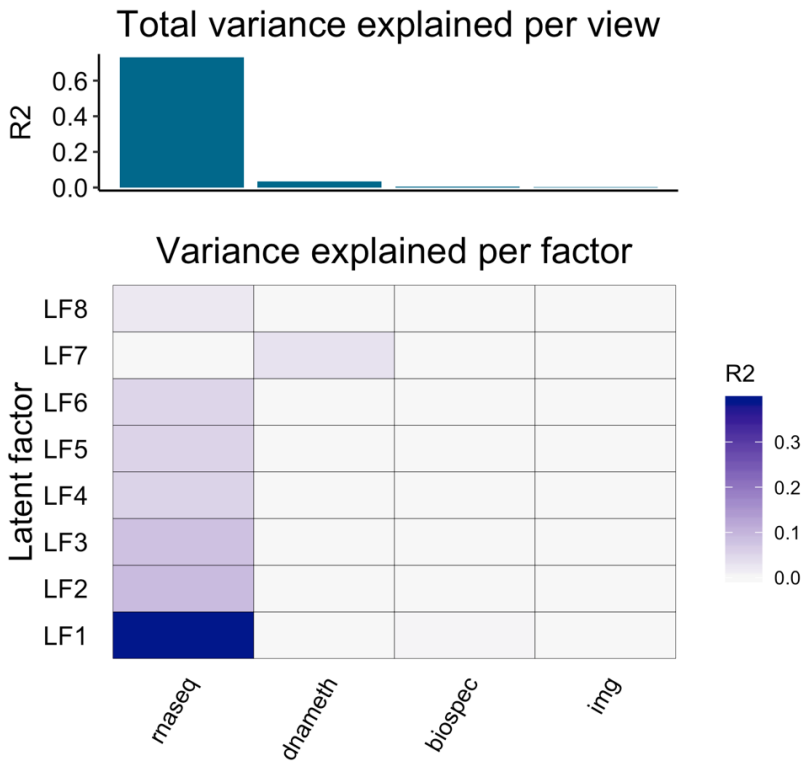


Figure 6.16 - Proportions of variance explained per factor and per data type.

Figure 6.17 illustrates the coordinates of each individual for each one of the first five factors. Similarly to previous analyses, no obvious grouping could be extracted in regard to progression rate or carrier status.

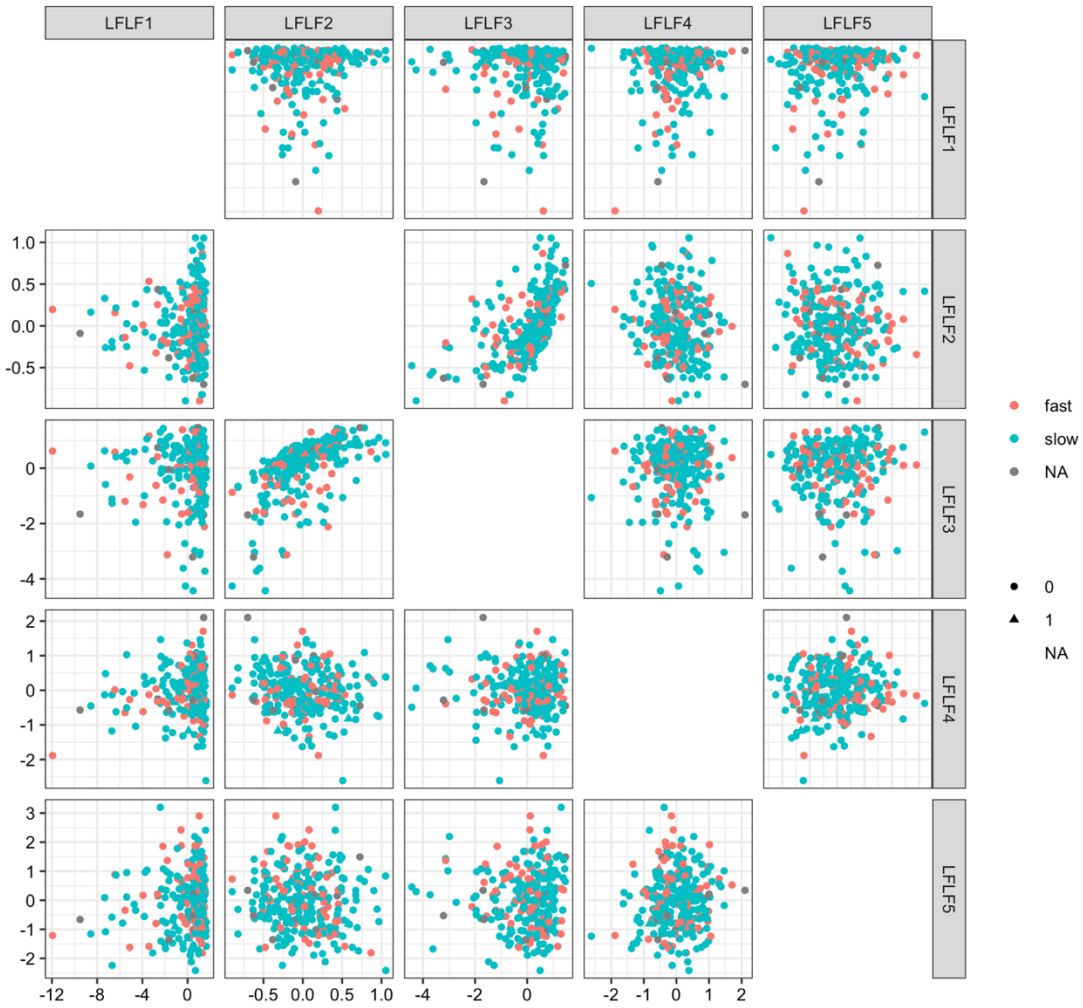


Figure 6.17 - Pairwise plots for top 5 MOFA factors (marker colour represents progression rate and shape represents carrier status). No obvious clustering could be identified.

Enrichment using Reactome gene sets was performed. With an FDR threshold of 1%, the following numbers of pathways were identified for each one of the eight factors identified using MOFA. As illustrated in Figure 6.18, the first factor was associated with the highest number of differentially expressed pathways.

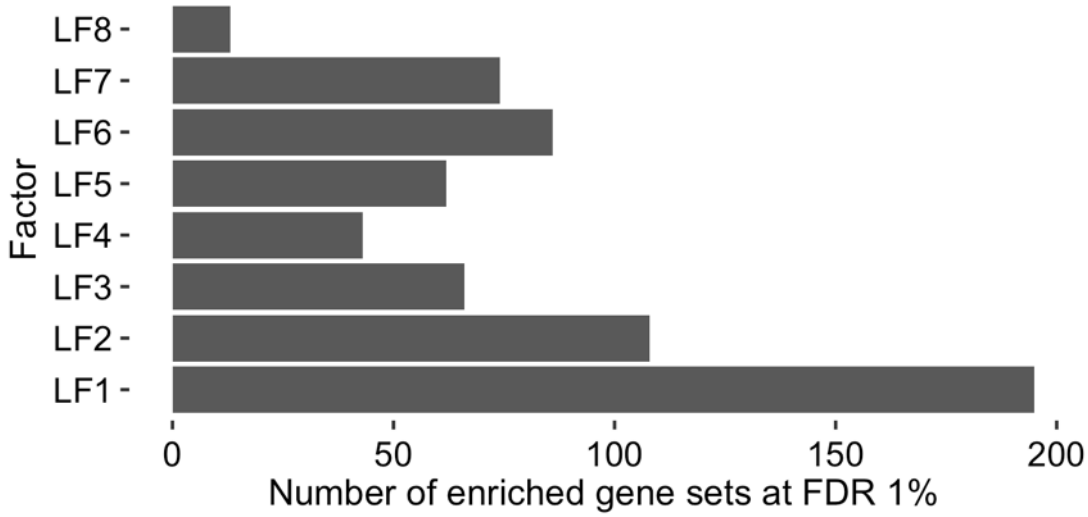


Figure 6.18 - Enrichment analysis results per factor.

6.3.2.2.2 SNF results

The data consisted of 129 distinct individuals. Among them, there were 2 G2019S carriers.

Table 6.7 - SNF input data overview for four omics data types.

SNF input data	**Biospecimen data**	**RNA-Seq**	**DNA methylation**	**Imaging data**
Variables included	4	5 000	8 448	4
Individuals included	129	129	129	129

Affinity matrices used as input for the SNF algorithm are represented in Figure 6.19 to Figure 6.22 for each data type integrated. They can then be compared to the obtained fused matrix plotted in Figure 6.23. Progression rate and carrier status are reported on each figure as well. As previously, 2 was identified as the optimal group number.

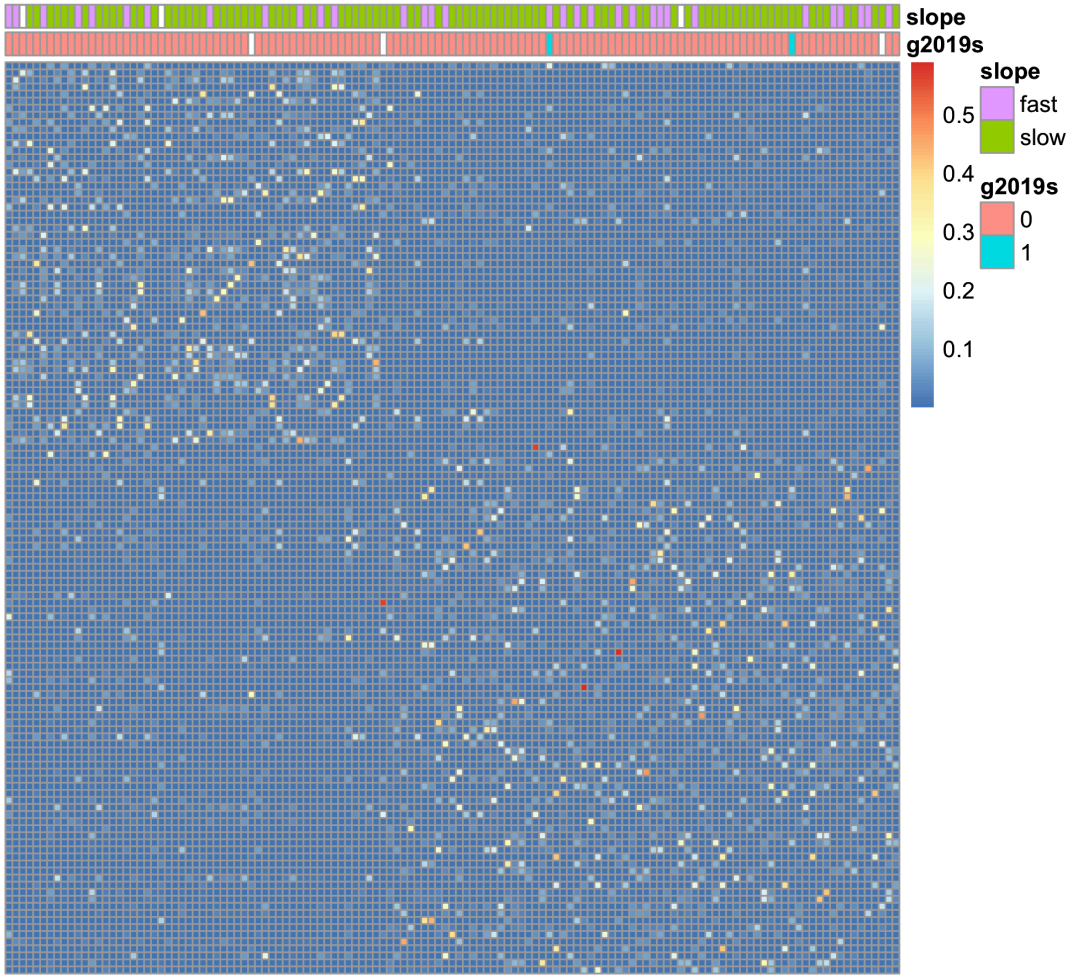


Figure 6.19 - Affinity matrix for biospecimen analysis results (ordered by SNF subgroups, progression rate and carrier status are represented). Highlighted subgroups can be respectively seen in the top-left and bottom-right corners.

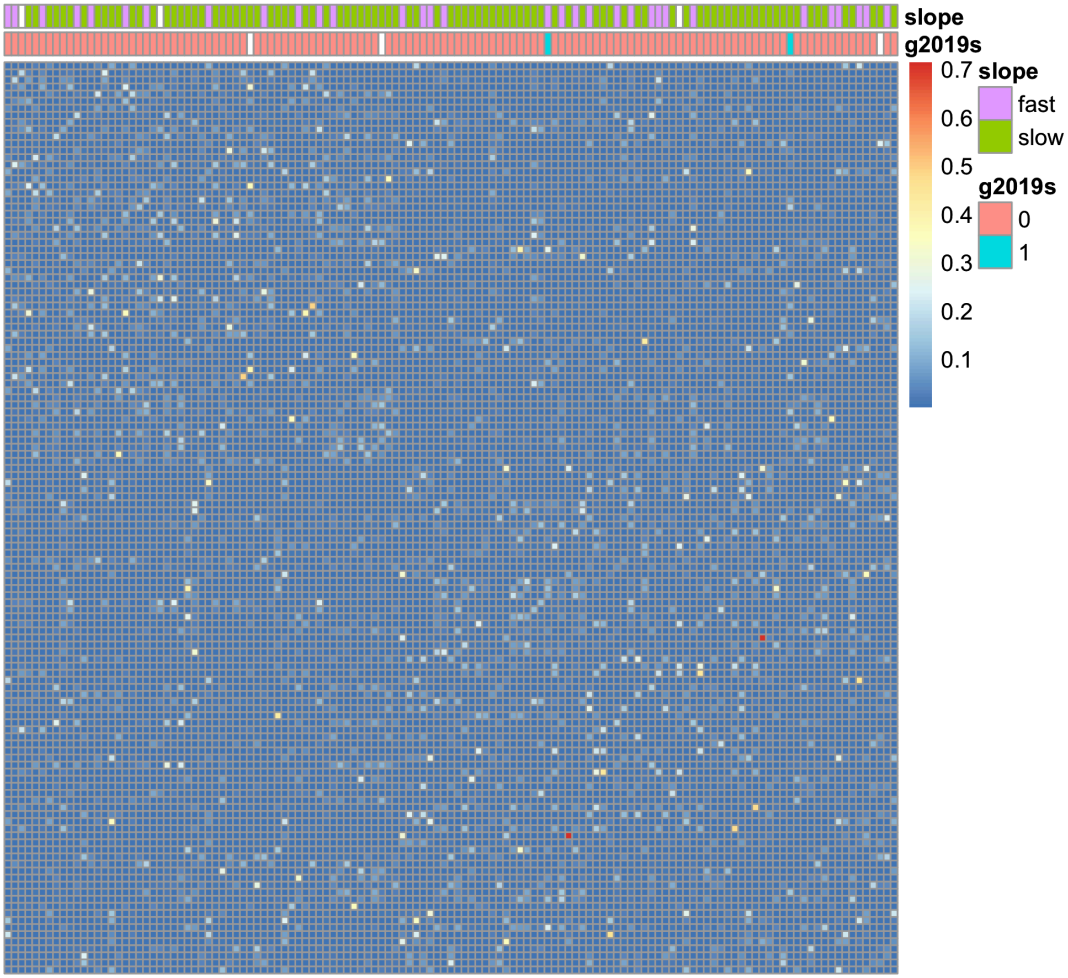


Figure 6.20 - Affinity matrix for imaging data (ordered by SNF subgroups, progression rate and carrier status are represented). Clusters identified from SNF cannot be obviously identified when looking at the imaging data alone.

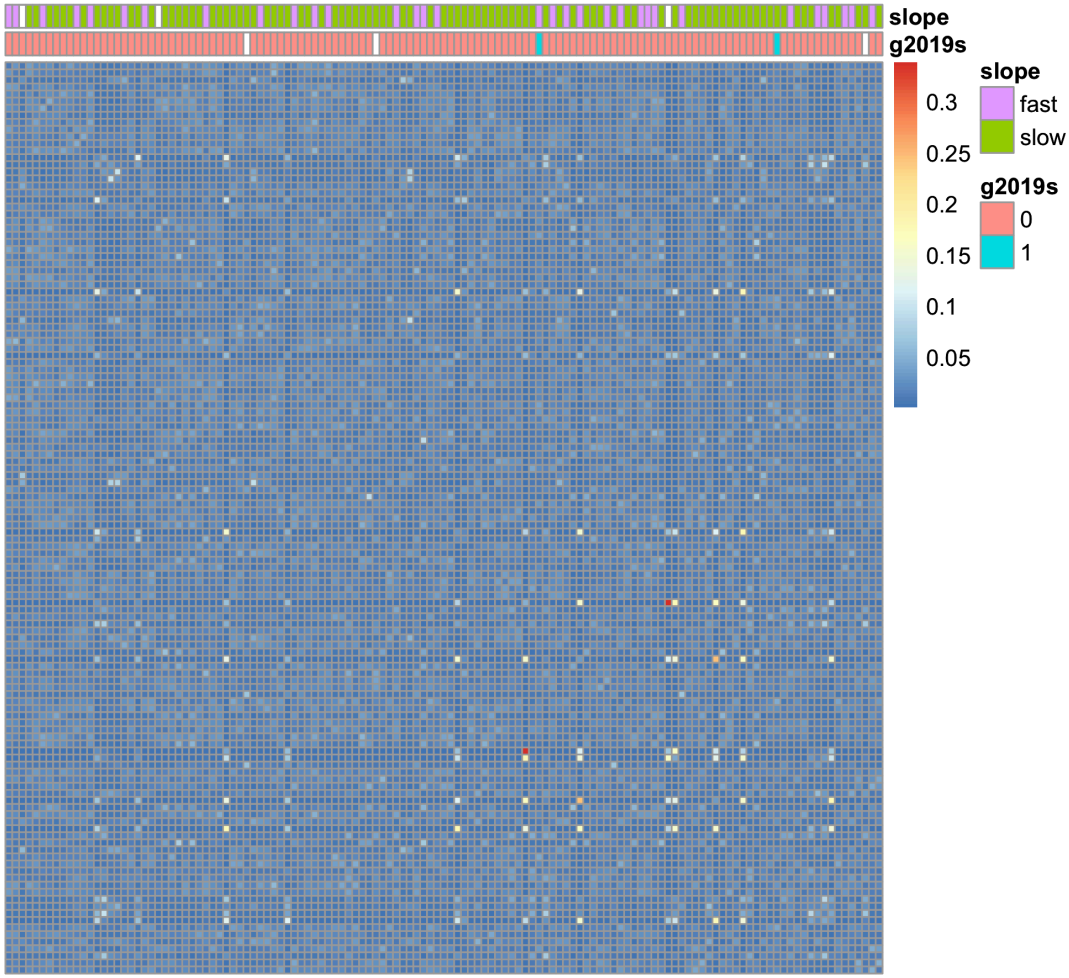


Figure 6.21 - Affinity matrix for RNA-Seq data (ordered by SNF subgroups, progression rate and carrier status are represented). Clusters identified from SNF cannot be seen when looking at the RNA-Seq data alone.

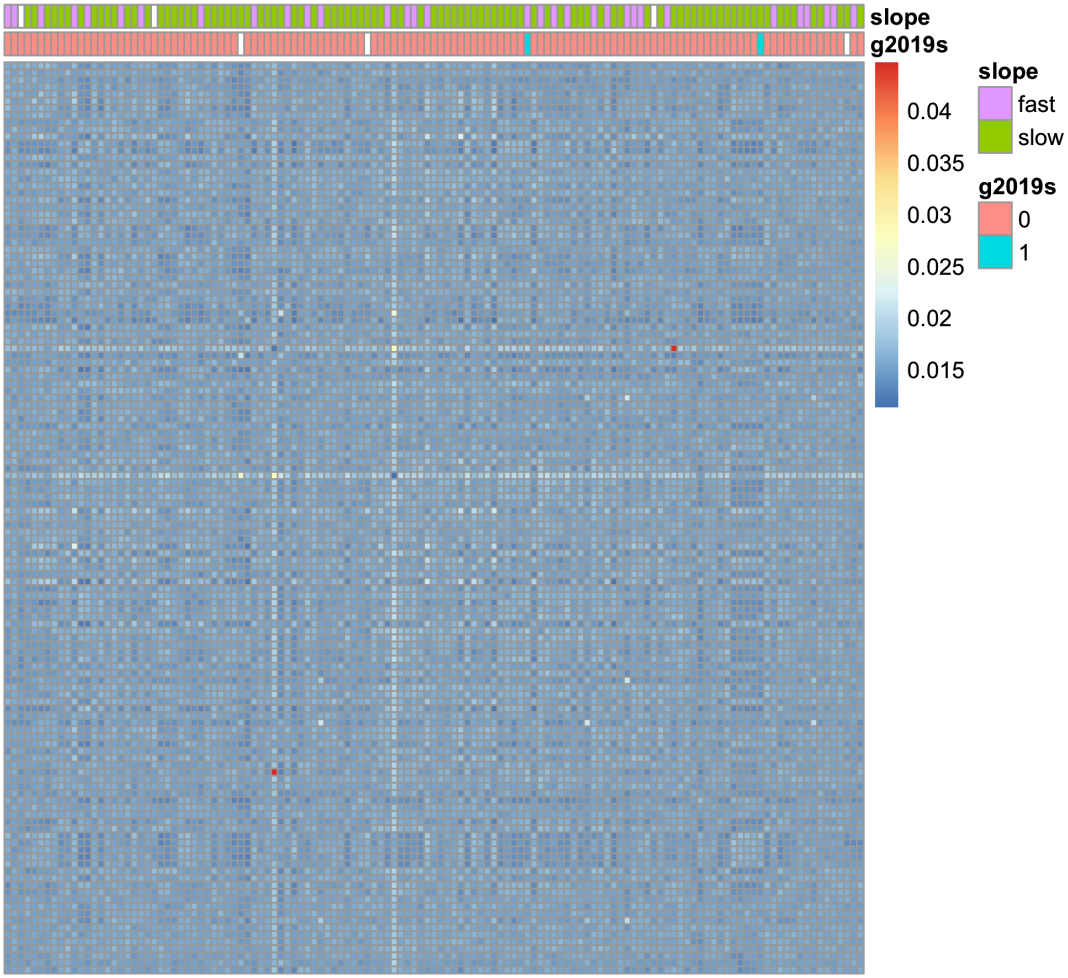


Figure 6.22 - Affinity matrix DNA methylation data (ordered by SNF subgroups, progression rate and carrier status are represented). Clusters identified from SNF cannot be seen when looking at this DNA methylation data alone.

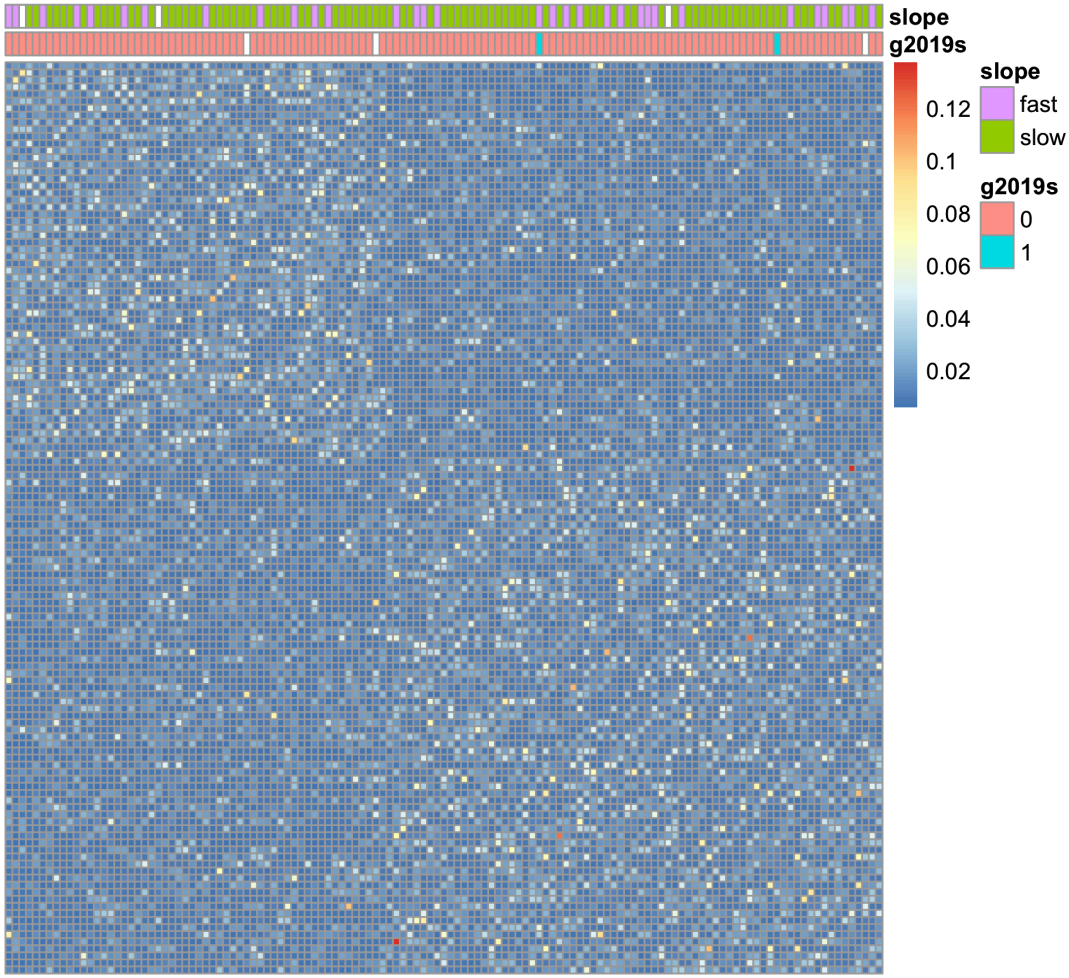


Figure 6.23 - Fused matrix ordered by SNF subgroups and representing progression rate and carrier status. Two distinct clusters are observed.

Normalised mutual information scores were computed between all affinity matrices to assess the proportion of mutual information between the different matrices and with the fused results matrix.

Table 6.8 - NMI concordance matrix (obtained from pairwise comparisons).

	Fused results	**Biospecimen data**	**Imaging data**	**RNA-Seq**	**DNA methylation**
Fused results	1				
Biospecimen data	0.13027931	1			
Imaging data	0.00238842	0.00022318	1		
RNA-Seq	0.01344134	0.00000875	0.00098431	1	
DNA methylation	0.02425110	0.00016880	0.00105544	0.01087827	1

6.3.3 Results using RNA-Seq time points data

6.3.3.1 MOFA results

For this MOFA run, 329 individuals had at least two data points for which RNA-Seq data was available. 6 of them were G2019S carriers.

Table 6.9 - MOFA input data overview for time-series RNA-Seq data.

MOFA input data	**BL**	**V04**	**V06**	**V08**
Variables included	5 000	5 000	5 000	5 000
Individuals included	266	254	260	259

As illustrated in Figure 6.24, all four RNA-Seq time points explained a similar proportion of the variance (70.93% for baseline data, 74.23% for V04 data, 79.28% for V06 and 72.41% for V08 data point).

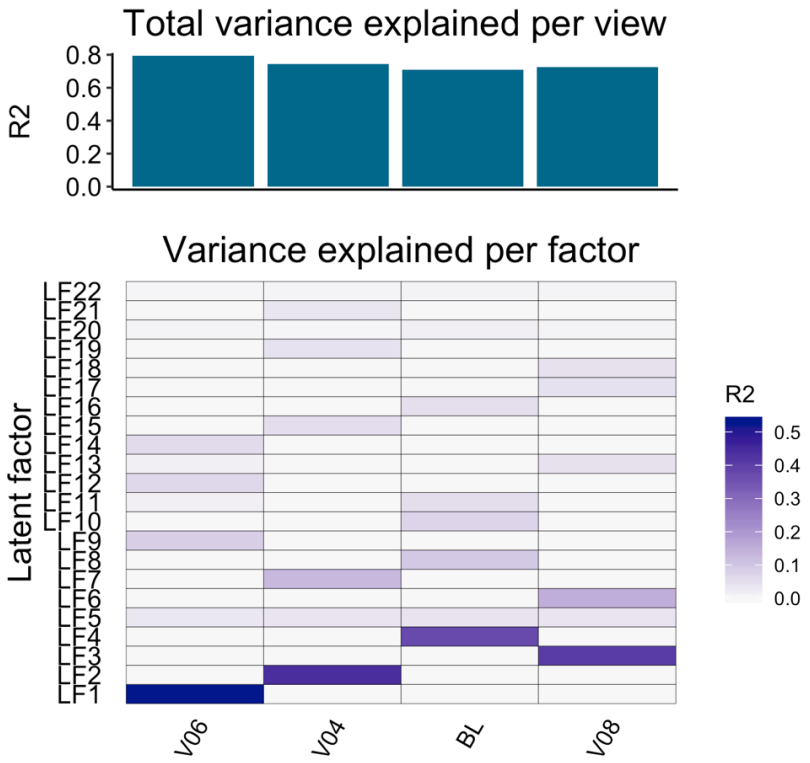


Figure 6.24 - Proportions of variance explained per factor and per data type.

The following figure shows coordinates of all included individuals for each one of the top five factors. Visually, there are no relevant groups that could be linked to progression rate or carrier status.

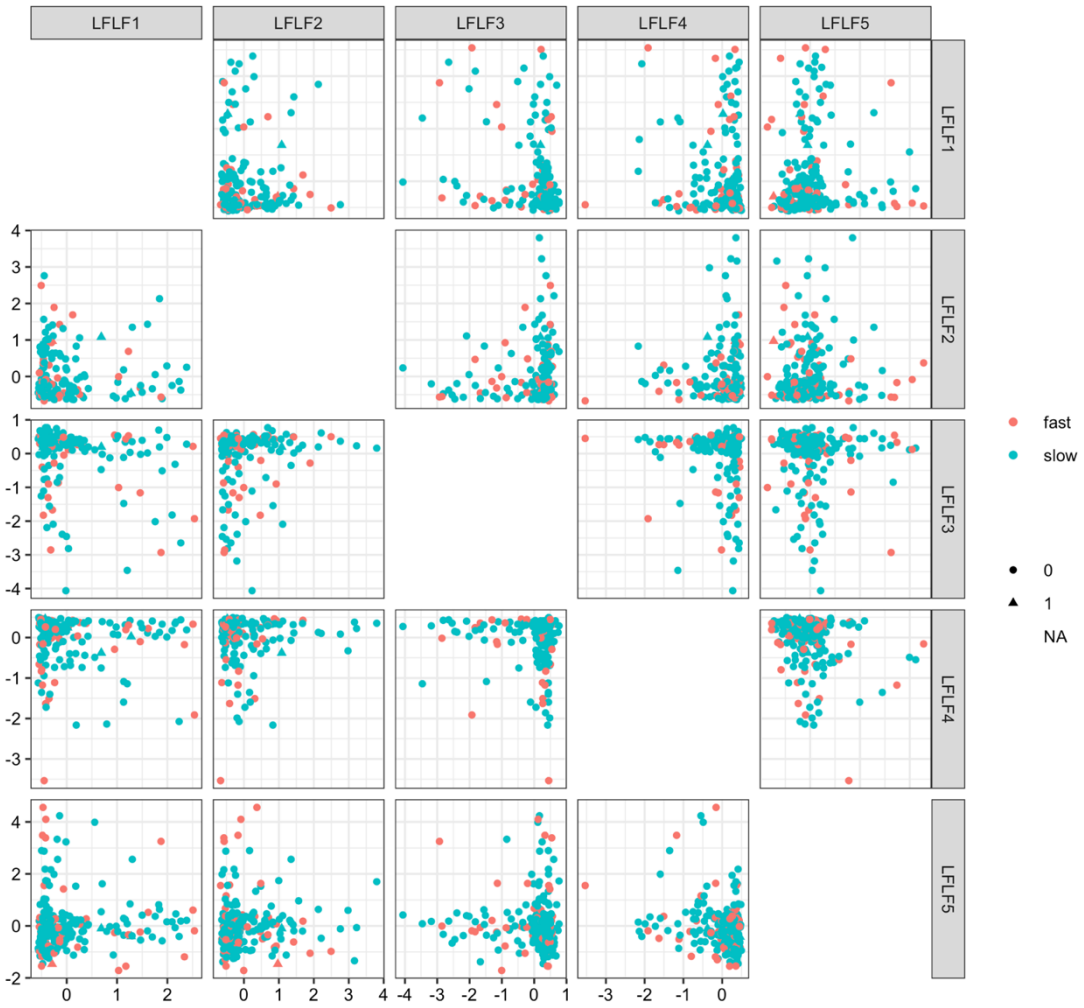


Figure 6.25 - Pairwise plots for top 5 MOFA factors (marker colour represents progression rate and shape represents carrier status). No clusters related to plotted covariates can be observed.

Gene set enrichment, using factor loadings, was performed as described in the methods section. With an FDR threshold of 1%, the following numbers of pathways were identified for each one of the factors extracted with MOFA (22 in total). Factors 4, 19 and 8 were associated with a higher number of differentially expressed pathways.

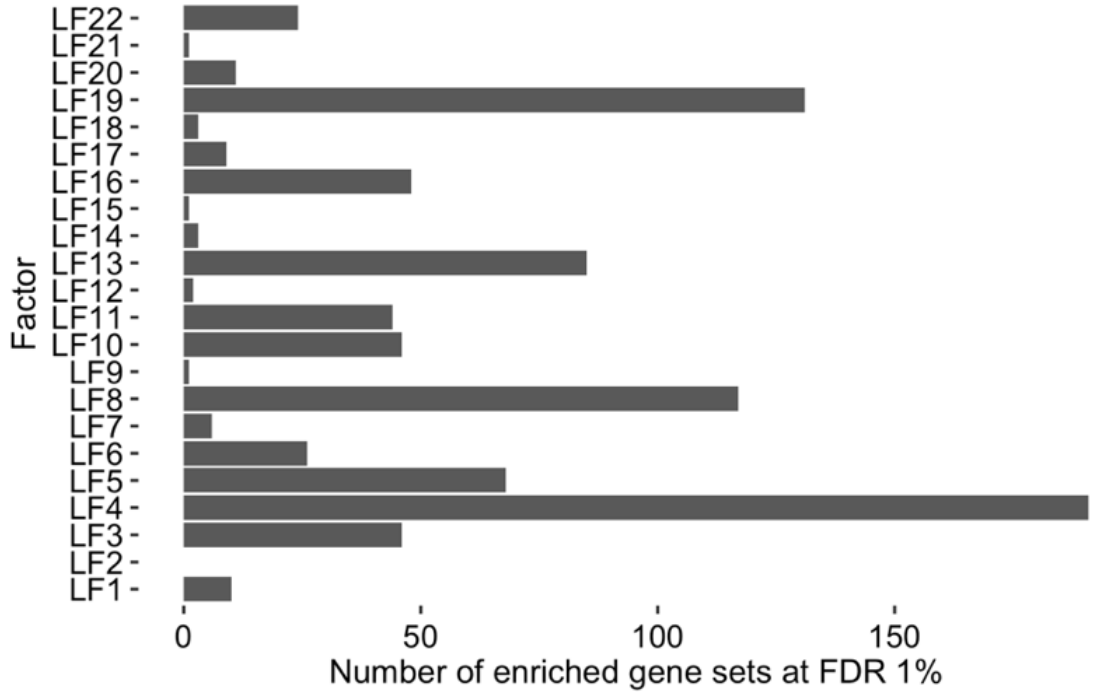


Figure 6.26 - Enrichment analysis results per factor.

6.3.3.2 SNF results

As part of this analysis, 118 individuals were used and 2 of them were G2019S carriers.

Table 6.10 - SNF input data overview for time-series RNA-Seq data.

MOFA input data	**BL**	**V04**	**V06**	**V08**
Variables included	5 000	5 000	5 000	5 000
Individuals included	118	118	118	118

Figure 6.27 to Figure 6.30 represent the affinity matrices used as input for the SNF algorithm for each RNA-Seq time point data. The output of the algorithm is represented in Figure 6.31 and is the final fused matrix. As for previous

analyses, samples are ordered given SNF-identified clusters and 4 was identified as the optimal cluster value (as explained in section 6.2.5).

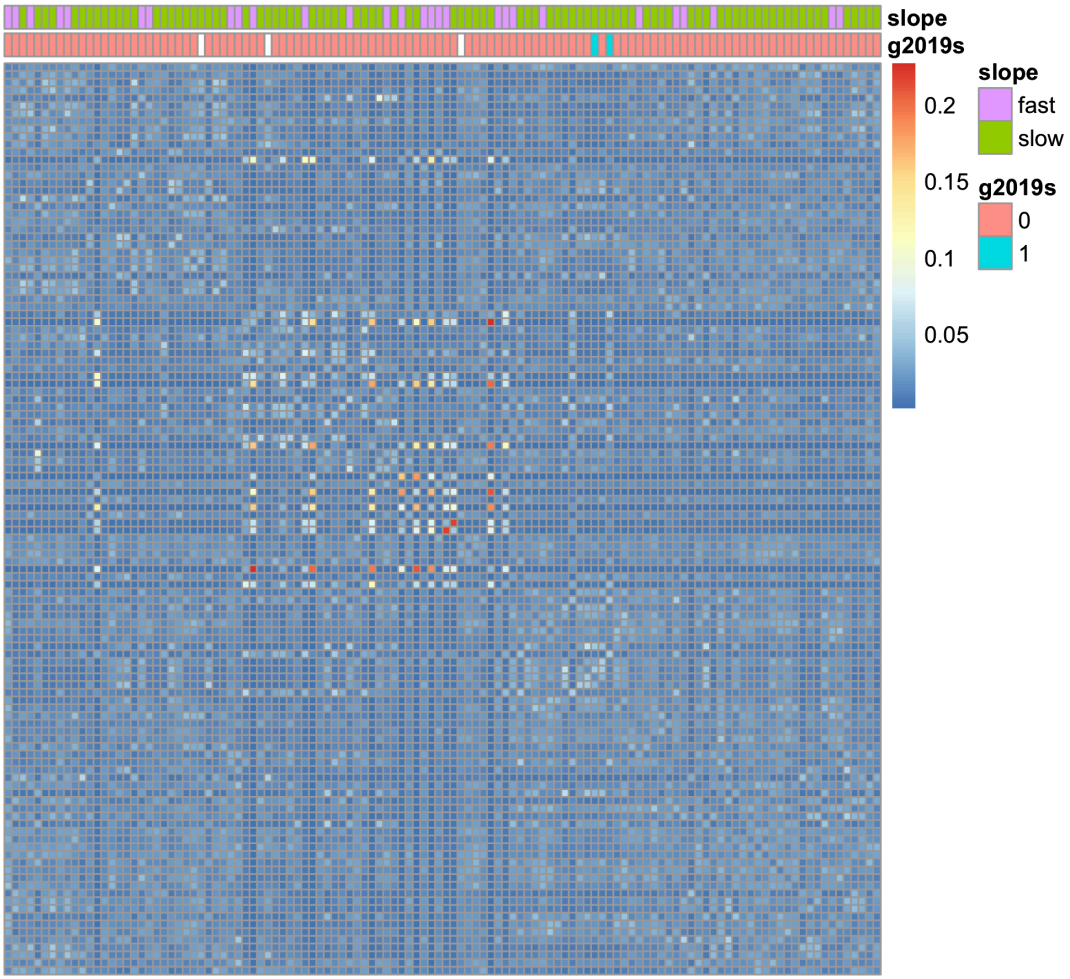


Figure 6.27 - Affinity matrix for RNA-Seq baseline data (ordered by SNF subgroups, progression rate and carrier status are represented). Some of the four clusters can be seen here, especially the third one from the top.

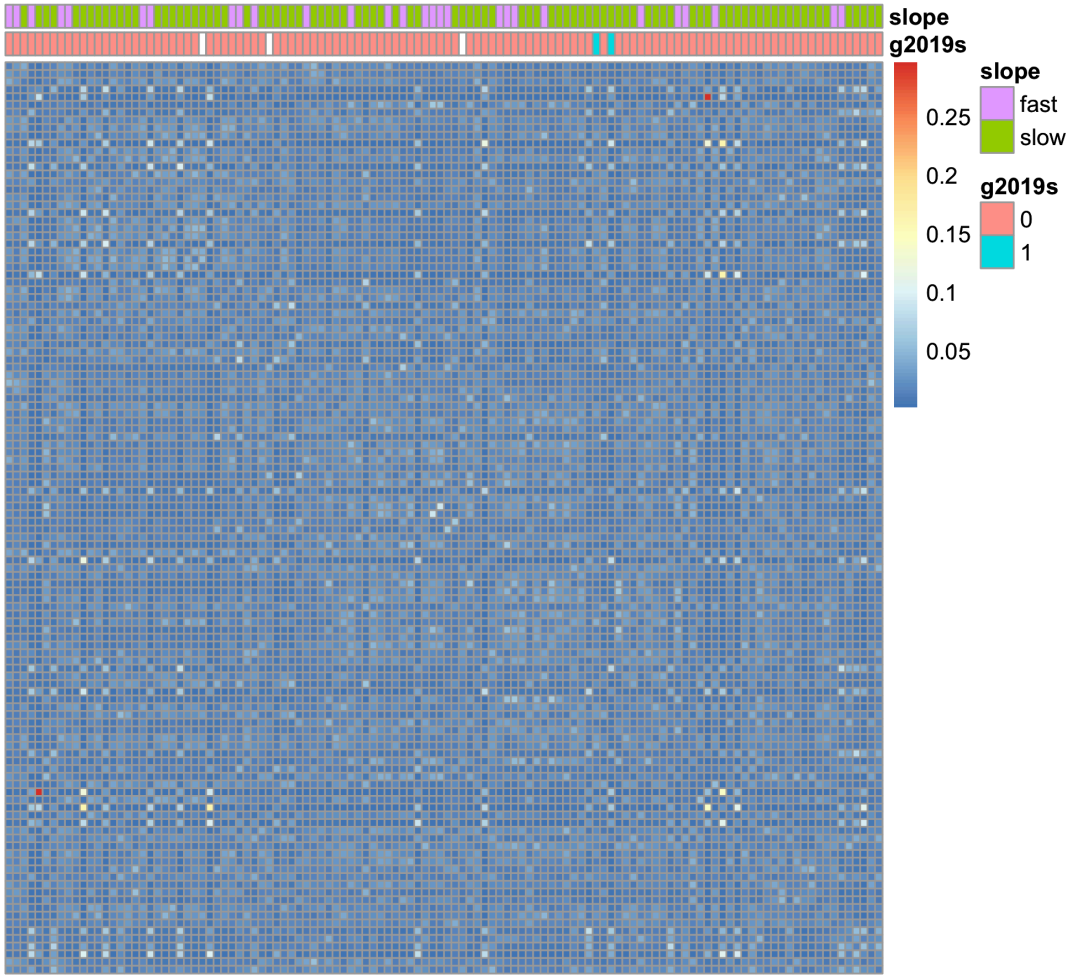


Figure 6.28 - Affinity matrix for RNA-Seq V04 data (ordered by SNF subgroups, progression rate and carrier status are represented). The four highlighted clusters cannot be obviously identified from this figure.

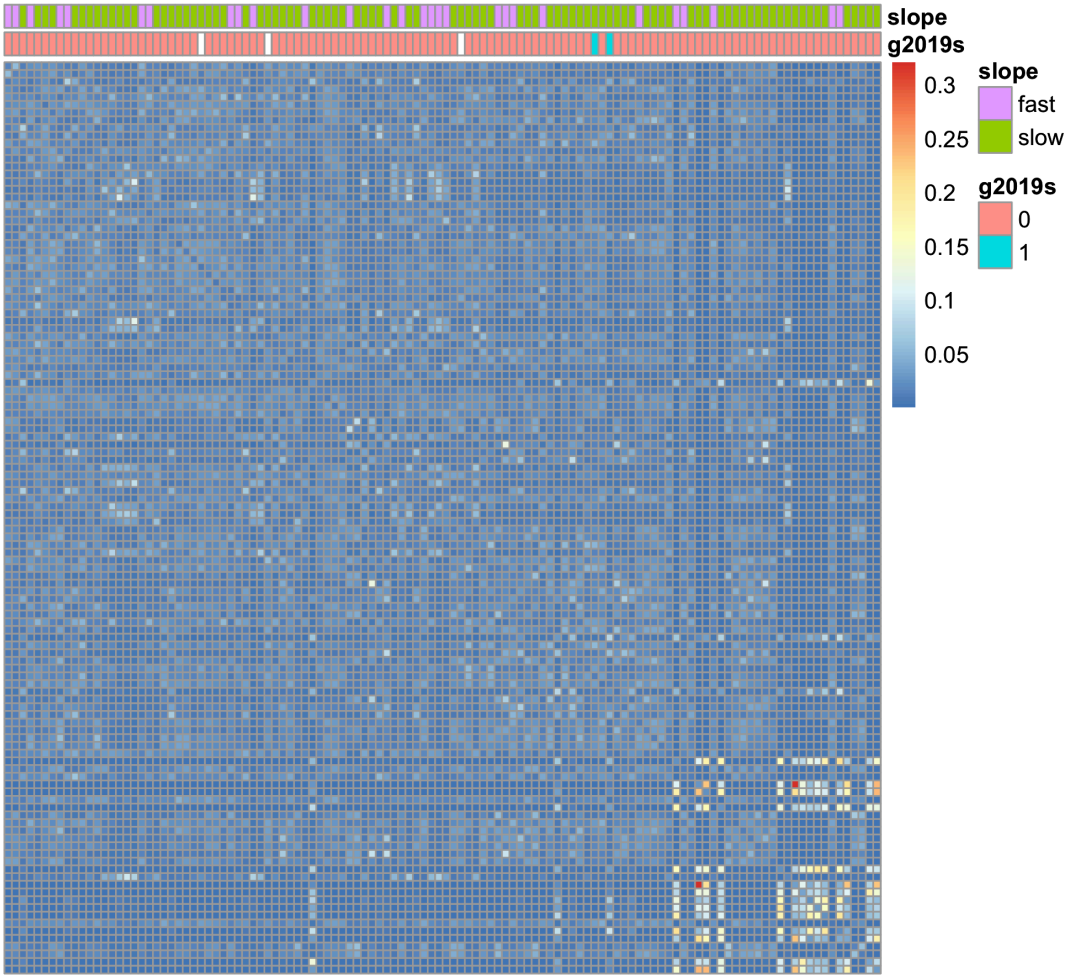


Figure 6.29 - Affinity matrix for RNA-Seq V06 data (ordered by SNF subgroups, progression rate and carrier status are represented). Cluster 4 can here be seen on the bottom-right part of the graph.

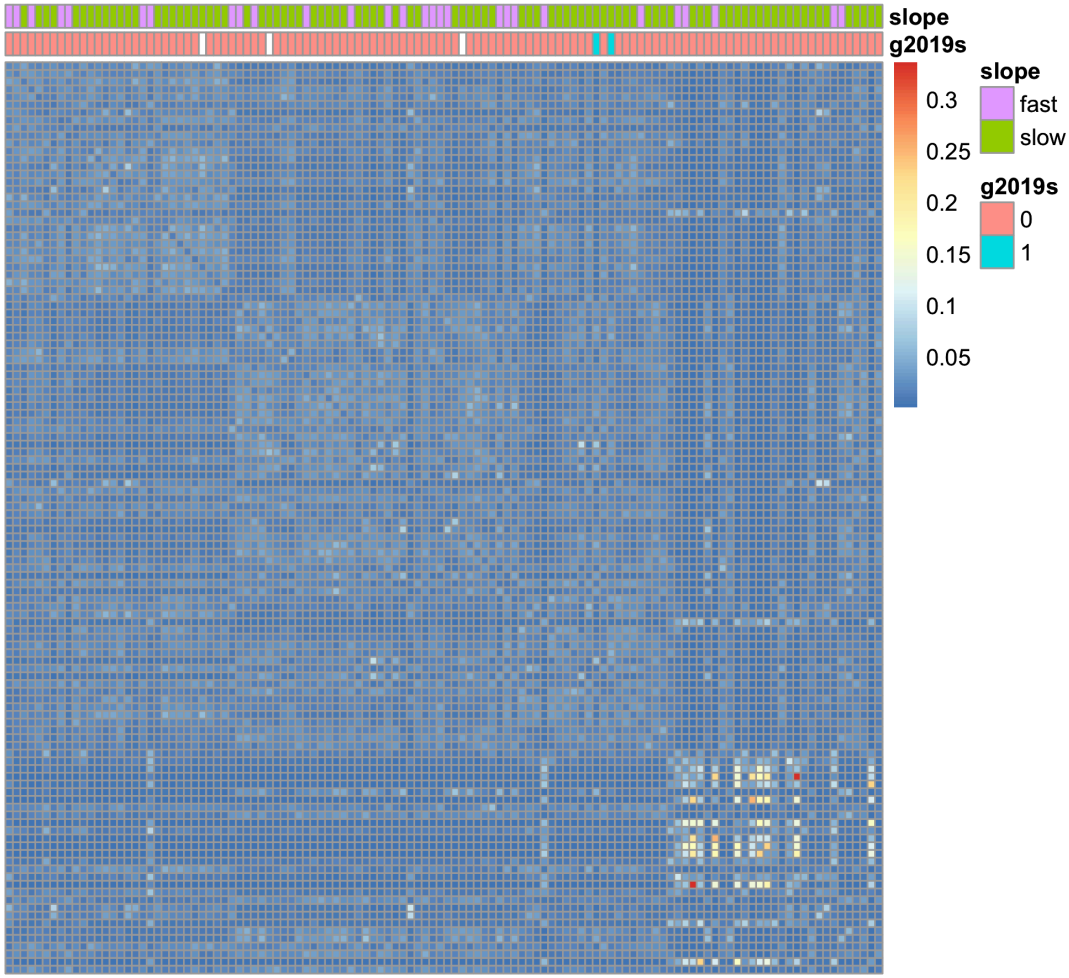


Figure 6.30 - Affinity matrix for RNA-Seq V08 data (ordered by SNF subgroups, progression rate and carrier status are represented). Cluster 4 can here be seen on the bottom-right part of the graph.

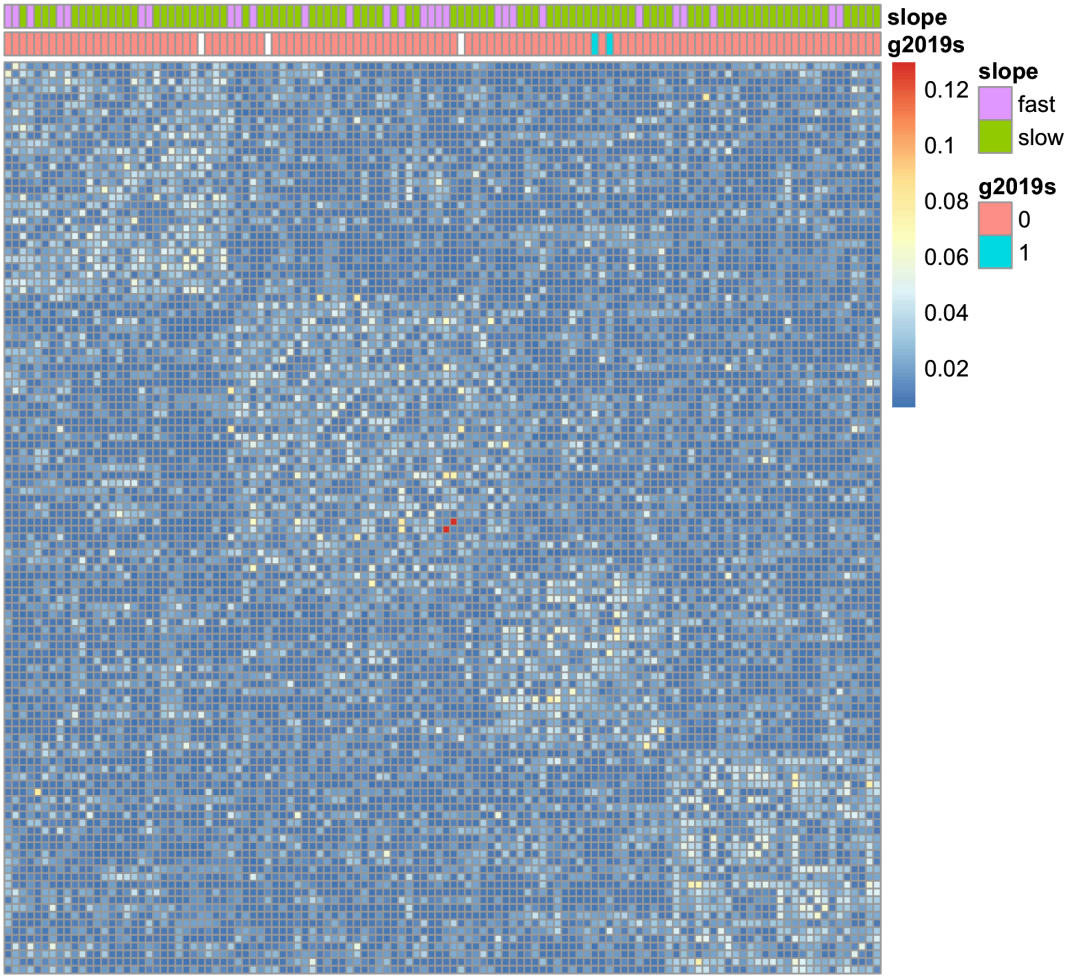


Figure 6.31 - Fused matrix ordered by SNF subgroups and representing progression rate and carrier status. Four distinct clusters are observed.

To compare the contribution of each data type to the final result, NMI scores were computed.

Table 6.11 - NMI concordance matrix (obtained from pairwise comparisons).

	Fused results	BL	V04	V06	V08
Fused results	1				
BL	0.13119198	1			
V04	0.02239540	0.00293057	1		
V06	0.00188194	0.00868078	0.00186331	1	
V08	0.05045202	0.04864873	0.00006263	0.01544828	1

Visual inspection of MOFA and SNF outputs did not highlight groupings correlated to G2019S carrier status or progression rate.

6.3.4 Example detailed results and comparisons between MOFA and SNF

6.3.4.1 MOFA model details

Although highlighted variations did not correlate with the two variables of interest, namely progression rate and G2019-carrier status, it was deemed of interest to further investigate the observed variation. To do so, the model generated using the MOFA method which used biospecimen analysis results, imaging, RNA-Seq and DNA methylation data, for PD cohort individuals, was analysed in more detail.

Specifically, we looked at Reactome enriched terms (as represented in Figure 6.19) more closely for all 8 factors. We also used all 8 MOFA-generated factors to cluster the selected population and visualise those using pairwise graphs. We aimed at investigating the cluster structures across selected factors.

For each factor, the top 25 terms with FDR values under a threshold of 1% were highlighted in Figure 6.32 to Figure 6.39.

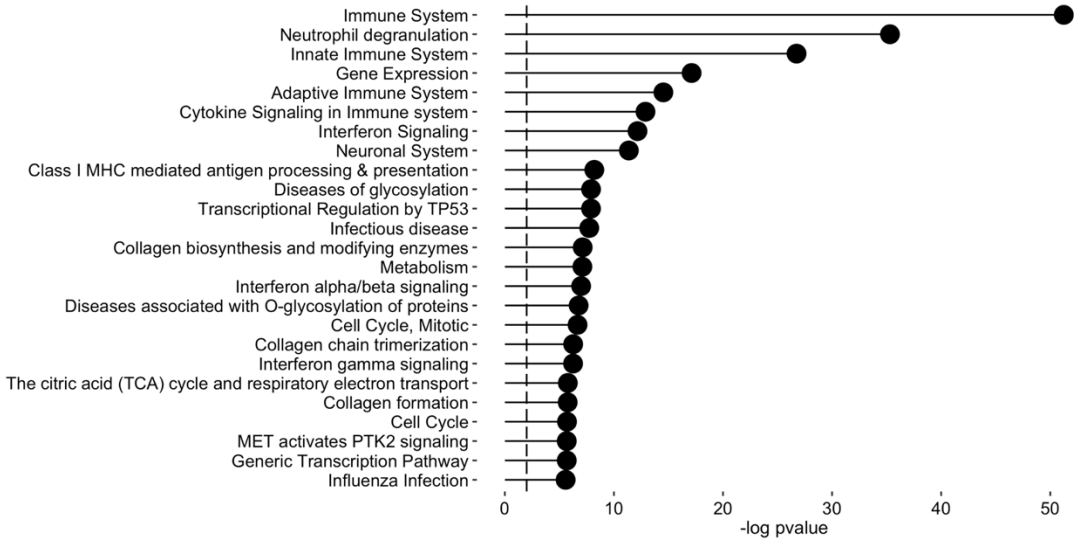


Figure 6.32 - Enrichment results for factor 1. Absolute values of log-transformed p-values are represented for top-25 pathway hits.

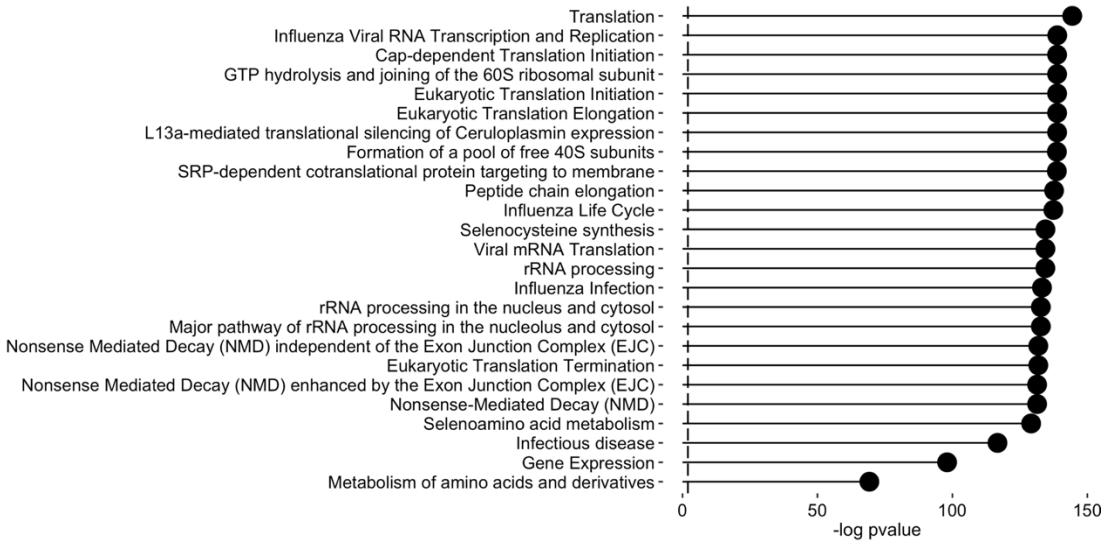


Figure 6.33 - Enrichment results for factor 2. Absolute values of log-transformed p-values are represented for top-25 pathway hits.

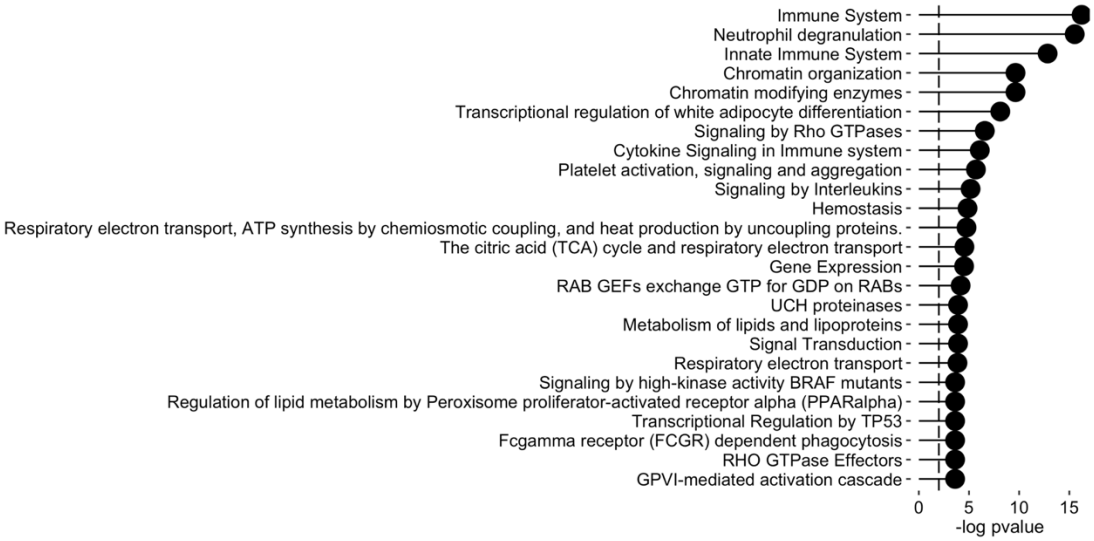


Figure 6.34 - Enrichment results for factor 3. Absolute values of log-transformed p-values are represented for top-25 pathway hits.

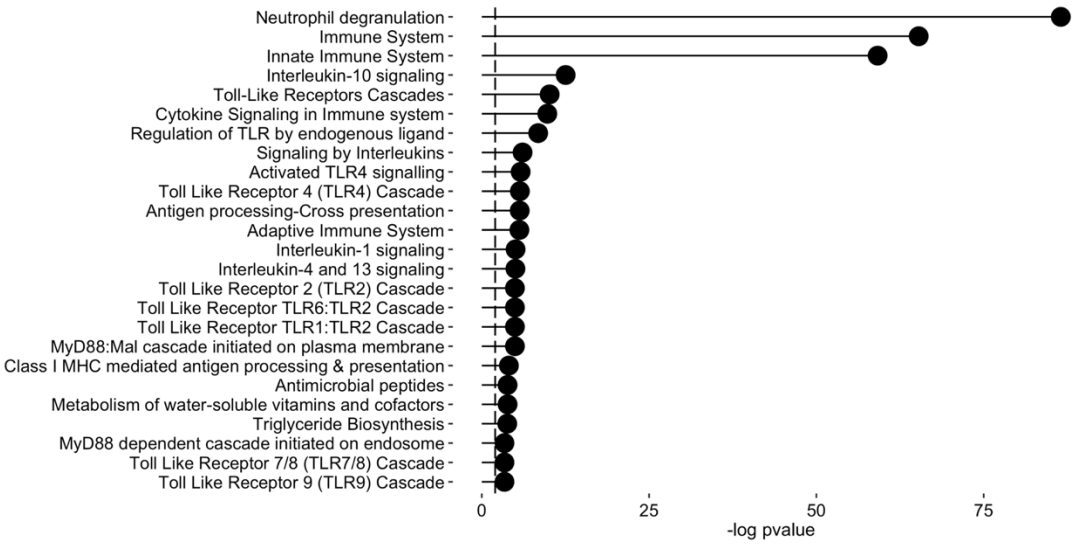


Figure 6.35 - Enrichment results for factor 4. Absolute values of log-transformed p-values are represented for top-25 pathway hits.

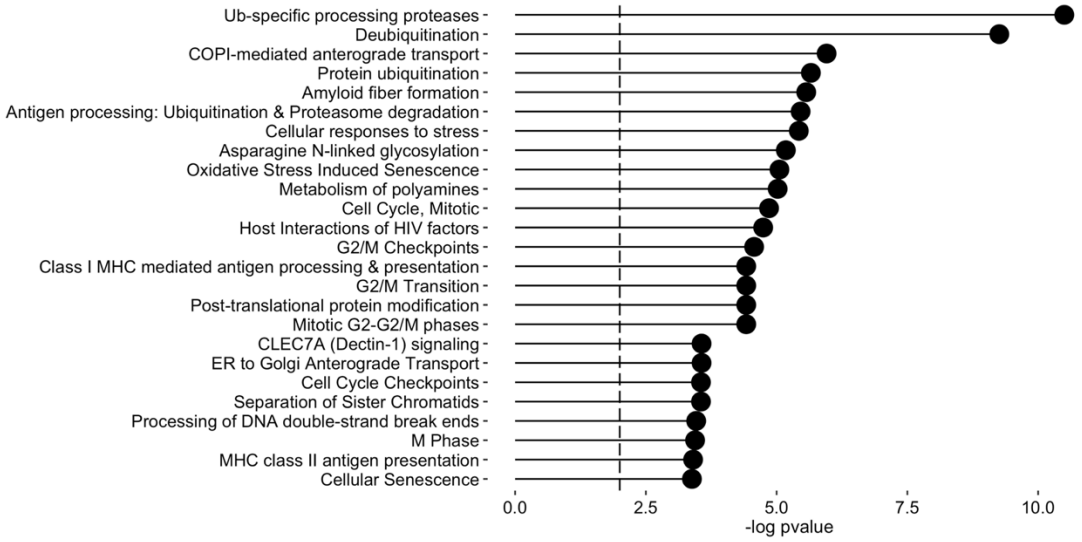


Figure 6.36 - Enrichment results for factor 5. Absolute values of log-transformed p-values are represented for top-25 pathway hits.

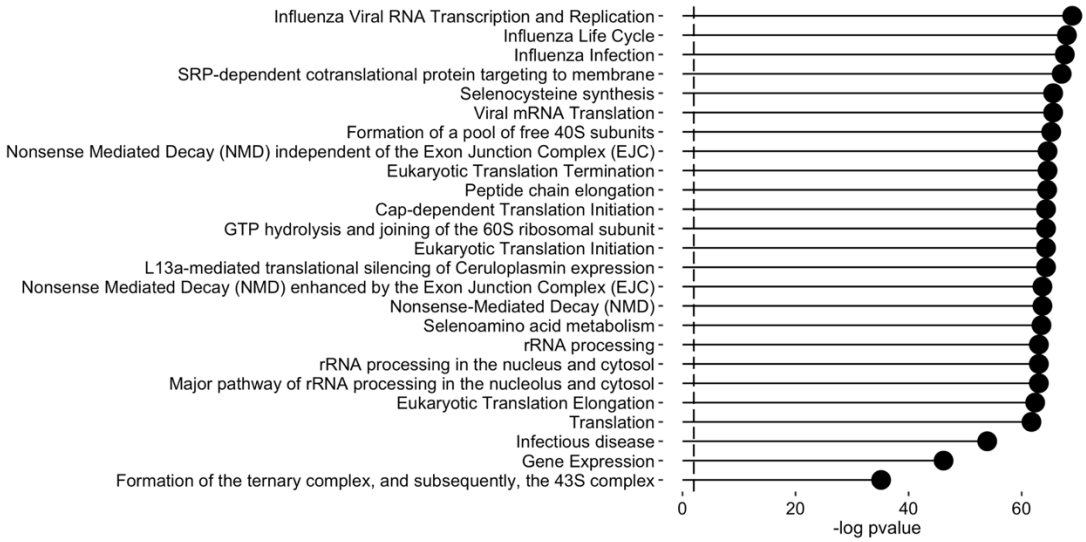


Figure 6.37 - Enrichment results for factor 6. Absolute values of log-transformed p-values are represented for top-25 pathway hits.

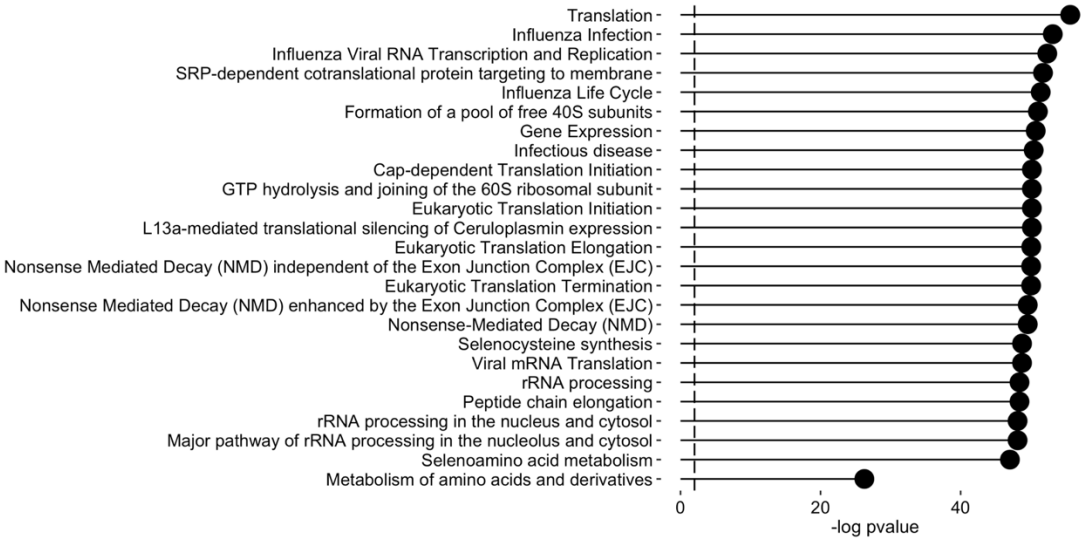


Figure 6.38 - Enrichment results for factor 7. Absolute values of log-transformed p-values are represented for top-25 pathway hits.

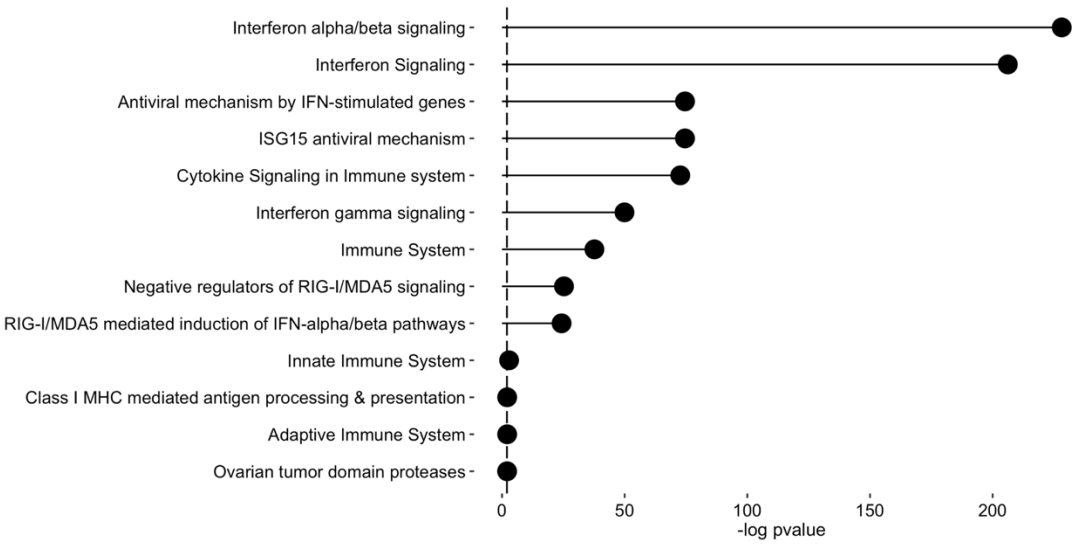


Figure 6.39 - Enrichment results for factor 8. Absolute values of log-transformed p-values are represented for top-25 pathway hits.

The optimal number of clusters for this solution was 2. This was determined based on internal validity indexes (most were optimal for 2 clusters, for example, the silhouette score was 0.45) and bootstrapping (for which stability was maximised when considering a 2-cluster solution, the average Jaccard score being 0.95).

As there were missing values, generated clusters did not involve all the samples, the two obtained clusters were respectively composed of 178 and 34 patients. Using all pairwise combinations of the 8 MOFA factors, clusters were represented in the following figure.

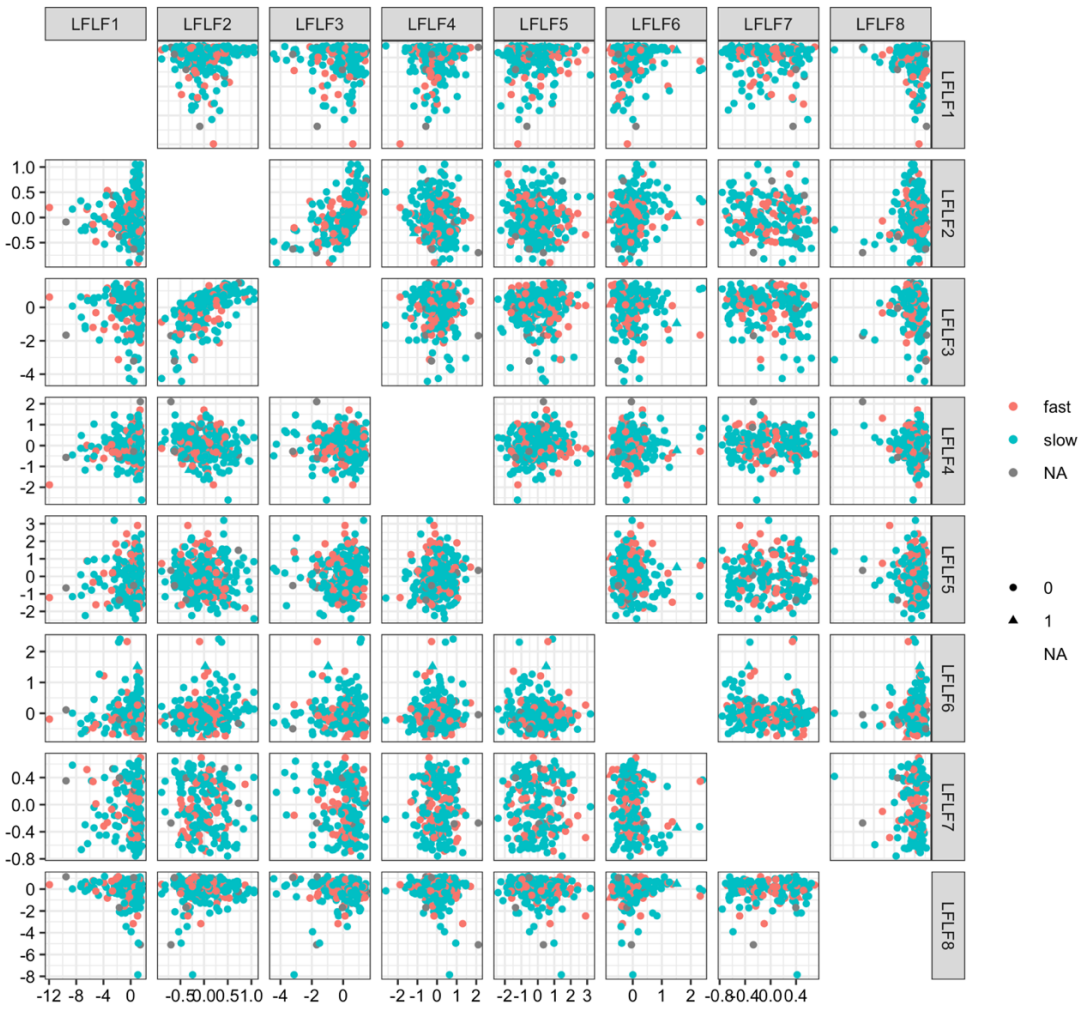


Figure 6.40 - Pairwise plots for top 8 MOFA factors (marker colour represents cluster allocations). No obvious clustering was extracted from this figure.

Clusters generated were highly correlated with factor 1. This is expected as, even though all 8 factors were used, factor 1 represented more variance when compared to other factors. Recurring enrichment terms linked to factor 1 were immune system, interferon signalling and glycosylation.

As RNA-Seq is the data type contributing the most to this factor, top absolute loadings for all 8 factors among RNA-Seq variables are represented in Figure 6.41.

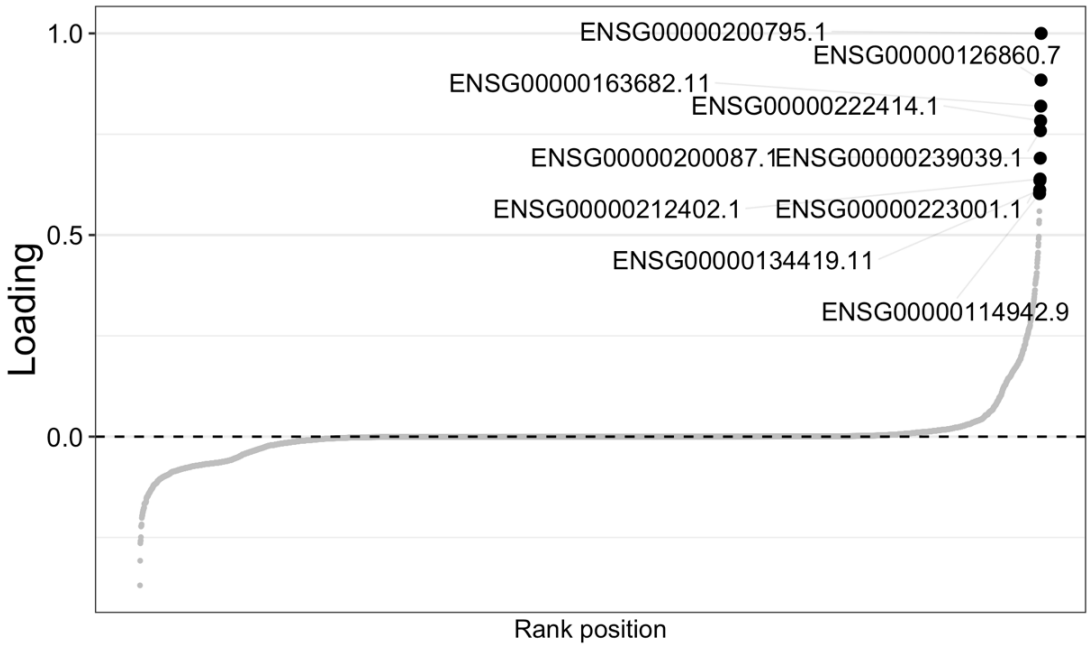


Figure 6.41 - Variables with highest absolute weights on factor 1.

Details for these genes are available in the following table.

Table 6.12 - Top 10 genes details.

ENSG gene identifier	Gene name	Gene type
ENSG00000200795	RNU4-1	snRNA
ENSG00000126860	EVI2A	Protein coding
ENSG00000163682	RPL9	Protein coding
ENSG00000222414	RNU2-59P	snRNA
ENSG00000239039	SNORD13	snoRNA
ENSG00000200087	SNORA73B	snoRNA
ENSG00000212402	SNORA74B	snoRNA
ENSG00000223001	RNU2-61P	snRNA
ENSG00000134419	RPS15A	Protein coding
ENSG00000114942	EEF1B2	Protein coding

6.3.4.2 SNF network details

SNF clusters were extracted (as described in section 6.2.5) and investigated, for the same set of variables, using only samples having all four data types available, as described in section 6.3.3. More specifically, the top 50 features (selected and ordered based on NMI values) were plotted along with the identified subgroups. We can see that, among the top 50 features, three data types are represented, biospecimen measurements, DNA methylation and RNA-Seq variables. Patterns of expression can also be compared and appear to be similar within a data type.

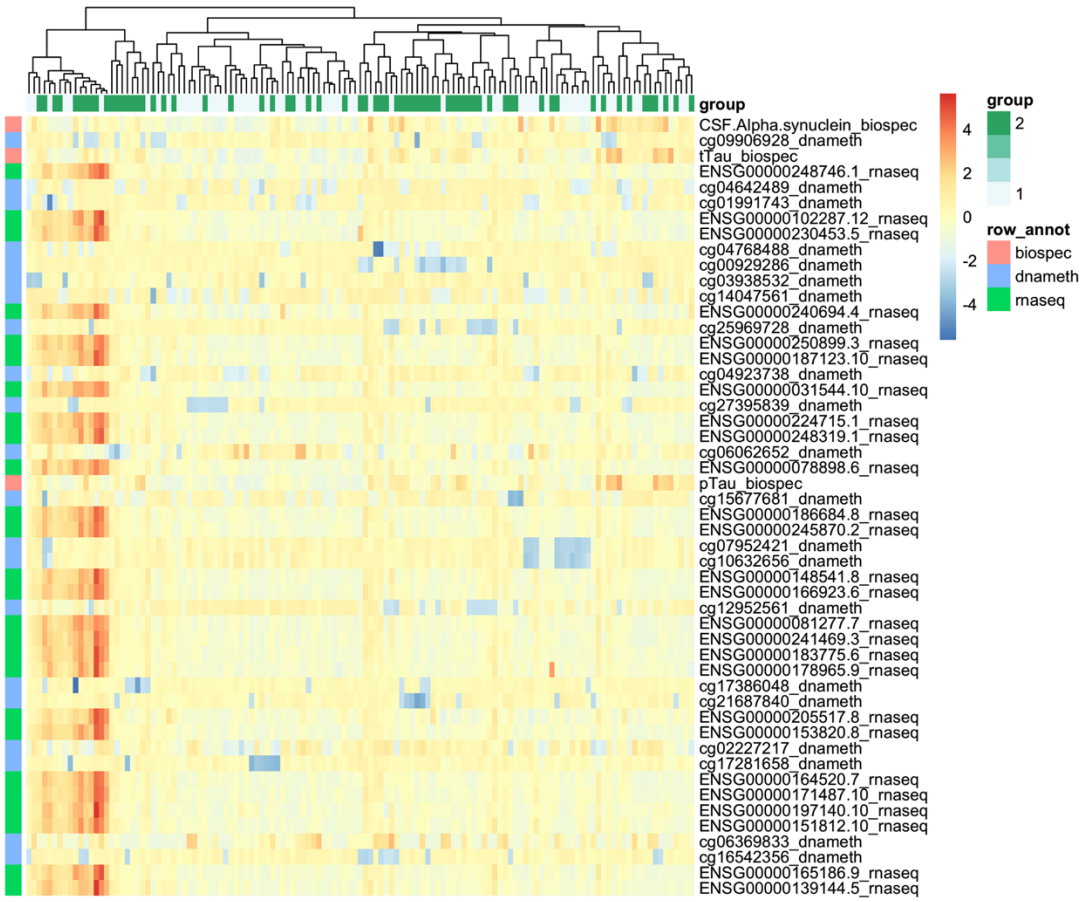


Figure 6.42 - Heatmap of top 50 features from SNF fused network (a row scaling was applied and dendrograms were computed using the 'correlation' metric and the 'complete' method for the columns). Cluster labels, as extracted using SNF output, are reported as 'group'.

6.3.4.3 Comparison between MOFA and SNF results

For a fair comparison between solutions obtained with the two algorithms, MOFA was re-run with the same subset of individuals as SNF. Indeed, the former did not support missing data and some individuals had to be filtered out. The loadings of all variables from MOFA's first factor were compared to NMI values obtained when comparing the best clustering solution to the input data. More specifically, the rankings of these variables were compared for each data type using Spearman's correlation coefficient. Results are presented in Table 6.13.

Correlation coefficients were rather low, highlighting the fact that MOFA and SNF produced different results and thus were highlighting different features of the same dataset.

Table 6.13 - Spearman's correlation results between MOFA and SNF-ranked lists of variables.

MOFA vs SNF	**Spearman's correlation coefficient**
Biospecimen variables	0.2
DNA methylation variables	-0.067
Imaging variables	0.2
RNA-Seq variables	-0.032

6.4 Discussion

6.4.1 Limitations

Due to the heterogeneous nature of the cohort data, there were some limitations to the analyses. For example, most G2019s carriers were represented in the GENPD and REGPD cohorts (see section 6.2.1.1) for which DNA methylation and follow-up RNA-Seq time points were not available. It was thus complex to integrate this data type while at the same time trying to highlight G2019s differences with non-carriers. This strategy might prove successful when RNA-Seq data and DNA methylation data is available for more GENPD and REGPD individuals.

Although MOFA and SNF are designed specifically for multi-omics data they each present different limitations due to their respective designs. Being linear, the MOFA will miss some non-linear relationships present in the data. On the other hand, SNF results will be strongly impacted by the fact that SNF does not accept missing values and will thus be greatly dependent on inherent data imputation and filtering performed beforehand.

Moreover, MOFA and SNF results were quite different (section 6.3.4.3). One possible explanation was that they highlighted different features of a same dataset. However, as SNF input consisted of distance matrices, the choice of distance metric used would have greatly affected the output. It would be of interest to compare the results from MOFA with SNF results obtained using different distance metrics.

6.4.2 Subsequent analyses

Autoencoders, that can be used to produce a reduced set of features from a multi-modal dataset, have been successfully used in multi-omics data integration strategies[178,179]. They might be of interest to study PPMI's data, however, one should carefully consider the interpretability of such results as well as the number of individuals integrated as a low number might lead to poor model performance[180].

6.5 RNA-Seq normalisation strategies

In all MOFA-generated models, most of the variance was explained by RNA-Seq data. Consequently, generated factors relied heavily on the way RNA-Seq data was pre-processed and more specifically, on the normalisation strategy. To assess whether the chosen normalisation, namely VST-transformed, was optimal, we tested two other strategies, namely FPKM (Fragments Per Kilobase Million) and TPM (Transcripts Per Kilobase Million) normalisations, both followed by a log 2 transformation with a 0.25 prior count. We chose to compare the results using biospecimen analysis results, imaging, RNA-Seq and DNA methylation data for PD-cohort individuals.

PCA plots for each one of the strategies (including VST-based), using baseline RNA-Seq data, were created and are presented below.

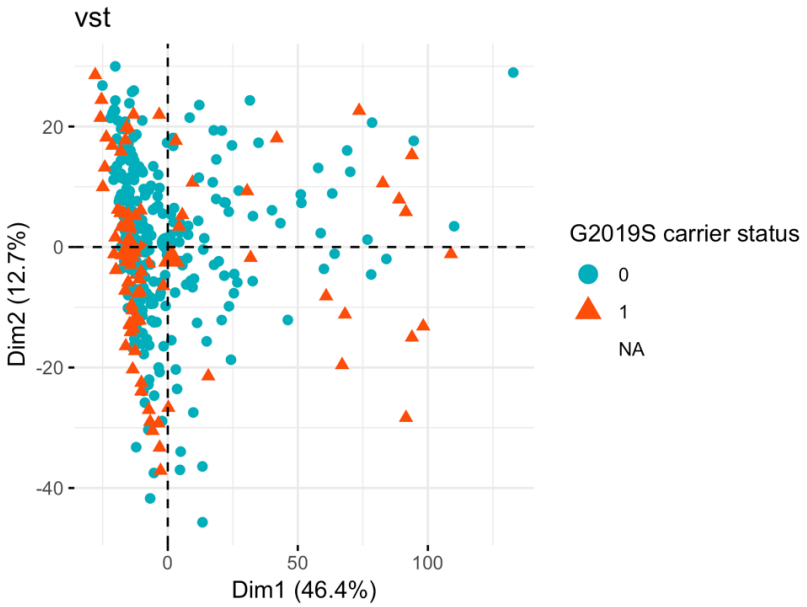


Figure 6.43 - PCA plot of RNA-Seq normalised counts. VST-based normalisation. G2019S carrier status is represented.

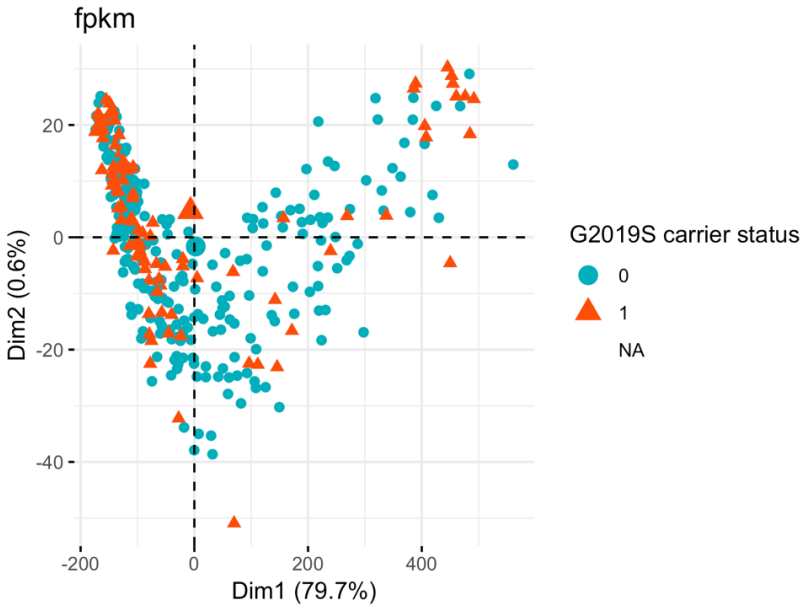


Figure 6.44 - PCA plot of RNA-Seq normalised counts. FPKM-based normalisation. G2019S carrier status is represented.

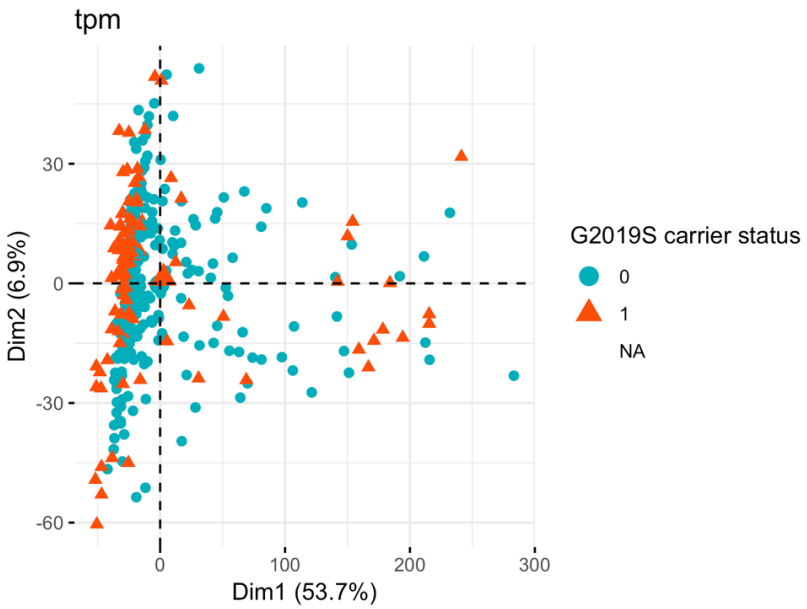


Figure 6.45 - PCA plot of RNA-Seq normalised counts. TPM-based normalisation. G2019S carrier status is represented.

Visually, these plots are quite different and hint that MOFA may produce results models differing greatly. The main model presented in the results section was compared to models using FPKM-based expression values and TPM-based expression values. Factors obtained, as well as variance explained by each data type and factor, are presented in the following figures. VST-based corresponding variance plot is available in Figure 6.16.

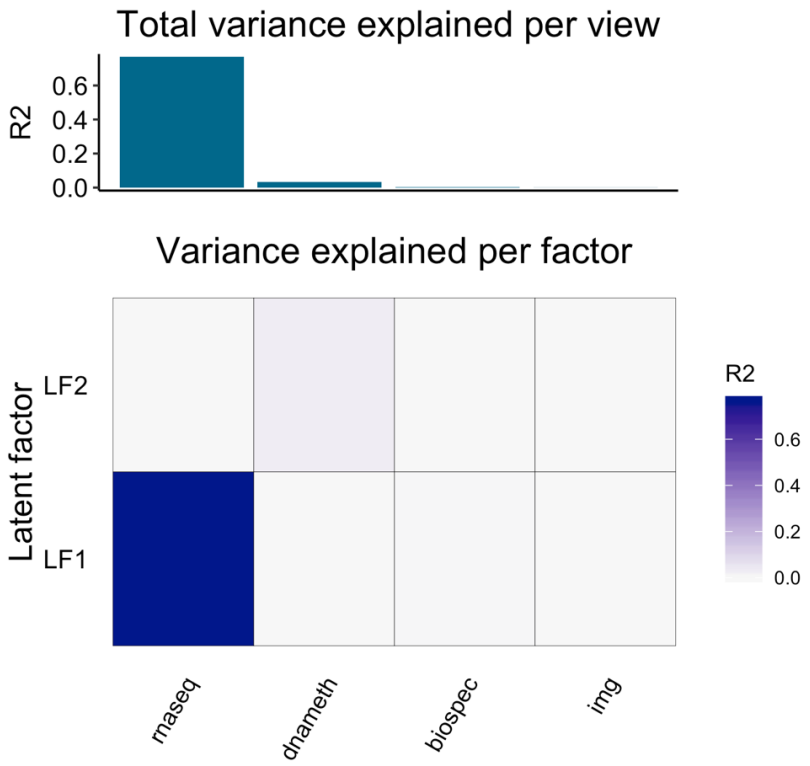


Figure 6.46 - Variance distribution across views and factors for FPKM-based MOFA model.

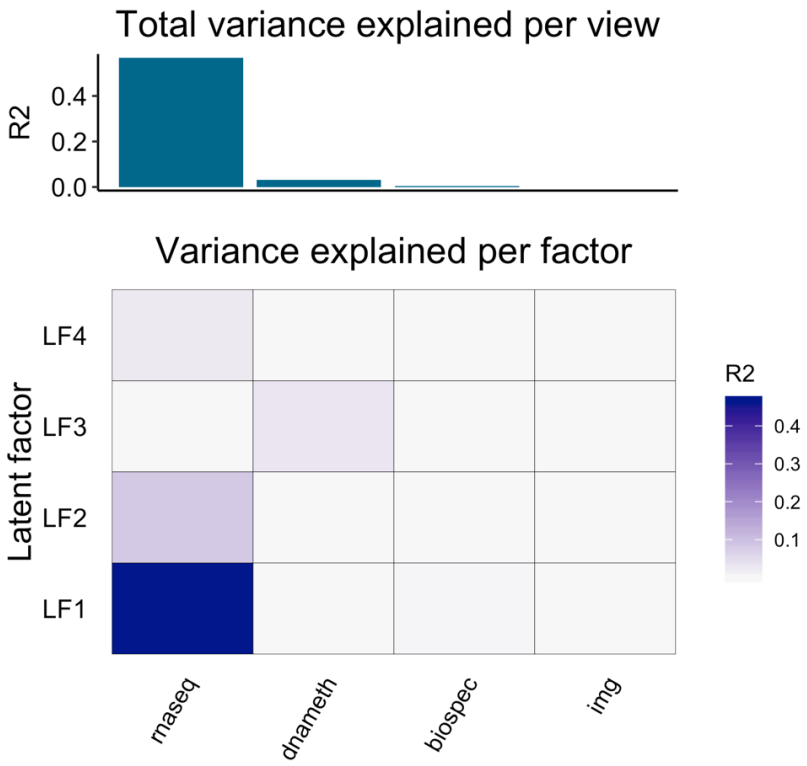


Figure 6.47 - Variance distribution across views and factors for TPM-based MOFA model.

In all results, RNA-Seq data is responsible for most of the explained variance, as expected. DNA methylation explains, in a single factor and usually not shared with other data types, a small proportion of the data variance. Although VST-based and FPKM-based models seem to explain a greater part of variance overall, it is difficult to highlight a 'best' model.

Enrichment analysis was performed, as described in the material and methods section, for each strategy. Results are represented in the next figures in which the terms with an associated FDR smaller than 1% are counted, per factor. Enrichment from VST-based results is represented in Figure 6.18. The FPKM-based model produced no enriched terms using this threshold. Even though more variance was explained for this normalisation, as opposed to TPM-based normalisation for example, this might not have been relevant in terms of biology. The characteristic 'horseshoe' shape[181] seen in Figure 6.44 might be a PCA artefact sometimes observed when many values in the input set are

zeros. This could also explain why not biological relevance could be highlighted using this strategy.

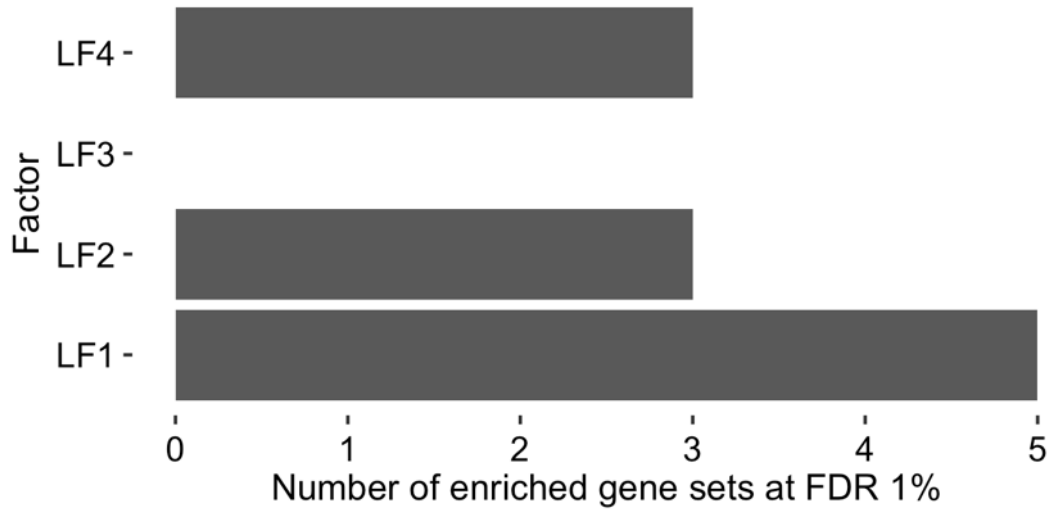


Figure 6.48 - Enriched terms with FDR<1% per factor for TPM-based model.

The number of Reactome enriched terms was much greater when using VST-normalised RNA-Seq counts and suggested that this normalisation strategy might have more power in highlighting biologically relevant features.

Overall, we deemed the VST-based strategy to yield a greater potential in terms of biologically-relevant information and this strategy was used throughout the analyses reported in this chapter.

7. Chapter 7 – Conclusions

This chapter summarises the main findings and conclusions of this thesis. It is divided into four sections. The first and second sections present conclusions from the different chapters and discuss the main limitations of the different projects and analyses. Future directions which could be taken following what was done as part of this thesis project are evoked in a third section. The fourth and last section expose general thoughts about what might be achieved in precision medicine in the near future.

7.1 Conclusions

This thesis focused on the application of different clustering methods, each tailored to a different type of dataset, to highlight novel relevant subgroups of different heterogeneous diseases, namely endotypes. Disease stratification is a trending topic at the moment and is facilitated by the recent surge of available high-volume omics data. Stratification is crucial for the understanding, prevention and treatment of complex heterogenous diseases. Understanding the mechanisms involved in the heterogeneity can help, for example, the research of appropriate drug targets and thus has the potential to improve the quality and outcome of care delivered to patients. Furthermore, following the identification of candidate drug targets, new treatments could be developed and administrated to patients using for example predictive algorithms using information from highlighted endotypes. Moreover, patients at higher risk might be identified early in their disease, and could be monitored more closely. Conversely, patients at higher probability of recovery might be spared unnecessary, harmful or invasive medical procedures which under normal circumstances would have been offered. Using data from EHR, a parallel could be made between characterised heterogeneity and routine measurements available for affected individuals. Also, the current success of treatments

carried out could be compared to the endotypes distribution to determine if a correlation could be established and if a more successful care approach, among those currently existing, could be applied.

Stable, biologically relevant and novel endotypes of AP were found after comparing the results of several clustering algorithms applied to time-series multi-omics data[104]. Different levels of AP severity were found in each of the subgroups. Importantly, aetiology was not found to be associated with subgroups, yet distinct pathways were found to be associated with each of the endotypes, confirming the involvement of discrete processes in the pathogenesis of AP-MODS, and bringing clarity to the heterogeneity of AP. We can unequivocally conclude that molecular subtypes exist in AP, which meet the definition of endotypes. After validation in an independent dataset, the four endotypes were compared to two published endotypes of another disease, ARDS[39], and shown to be overlapping. This suggests that there may be commonality in some molecular mechanisms between different causes of critical illnesses.

I therefore tested the omics profiles of our four AP endotypes in other types of critical illnesses, including sepsis and flu, which, to my knowledge, had not been done before.

Next, complementary stratification strategies were applied to IBD datasets in order to test the hypothesis that CD and UC could be themselves further subdivided into endotypes. Data from a previous study[146] and preliminary analyses carried as part of this thesis suggested that our hypothesis was a promising and worth of more exploration. More specifically, the hypothesis stating that CD was in fact a collection of different subgroups with similar symptoms, was found to be of great interest for its study and has been suggested elsewhere recently[157].

Clustering applied to PD data was then explored. More specifically, I aimed at highlighting how the obtained partitions were related to a given mutation: G2019S. Multi-omics data was used, yet found not to be relevant when trying

to correlate mutation status with potential subgroups. Finding individuals with similar profiles to individuals with the G2019S mutation will be of great interest for the study of PD. However, the number of mutation carrier individuals was too low in my project dataset to draw reliable conclusions.

In conclusion, the analytical approaches and strategies that I have developed in my PhD project, and present here form the basis for a general approach to cohort stratification, and I hope to make a contribution to a better understanding of heterogenous diseases in general.

7.2 Limitations

AP endotypes were highlighted using multi-omics data in chapter 3. Different strategies were employed to generate stratified sets of individuals and compare them fairly and rigorously. Moreover, some overlap was demonstrated between two of the identified AP endotypes and ARDS endotypes identified previously. This finding opens a new avenue to study critical illnesses, what similarities might exist, and how this understanding could be relevant for individuals needing treatment. It would be interesting to add more individuals to this study to further validate and refine the identified clusters. Higher sample numbers would also help strengthen the comparison with ARDS. Although the comparisons were significant, a comparison with a higher number of variables would be a great improvement and would help in understanding this overlap better. Having said that, it is also quite possible that a larger number of individuals analysed, and a larger number of variables may actually uncover greater differences between the various aetiologies of critical illness than it does similarities.

In the data available to me, however, to show that this observed overlap could be seen in other diseases, I used the IGP to compare the 4 AP endotypes in other diseases, for example sepsis, as described in chapter 4. Perhaps the biggest challenge with the type of analyses I have undertaken lies in

heterogeneity in the data. It can be challenging to normalise datasets so that they can be compared with confidence. Indeed, with the rapid advance of data acquisition technologies, data formats and data processing tools change rapidly and are often not directly comparable.

In chapter 5, previously identified subgroups of SNPs relevant to IBD were shown to be promising for patient stratification, especially in CD. Strategies to further characterise them are exposed and discussed. It might be interesting to integrate samples with more measured loci (which could be done directly by genome wide sequencing for example or by imputing the already available data using a reference panel).

Chapter 6 summarises work focusing on PD and more specifically on its stratification relating to the G2019S mutation. This could help in highlighting non-carriers with similar profiles to carriers which could then benefit from a similar treatment approach for example. Stratification using different types of omics data, with the transcriptomics data contributing most to the variation observed, did not permit to highlight correlations with the mutation status as mutation carriers did not cluster together. However, more data and especially more samples with mutation status available, would constitute an invaluable opportunity for the understanding of PD and its relation to the G2019S mutation. Such datasets should be available as part of future releases of the PPMI's project[169].

7.3 Future directions

Many interesting projects could stem from the work presented in this thesis and some of them will be evoked in this section.

For the characterisation of endotypes in AP, in addition to refining them using more samples, other data types could be integrated. For example, metagenomics datasets could be acquired and the relationships between the microbiome and AP subgroups explored.

As mentioned in the previous section, the overlap of omics profiles between different illnesses could be further studied by comparing more datasets. This could be done by comparing our four AP endotypes to other diseases but also to other disease endotypes, similarly to what was done with the two endotypes of ARDS.

The computation of PRS seems to be promising for the study of genetic data and how it relates to disease stratification. This could be further studied by integrating more data from UC/CD but by also reproducing similar analyses in other diseases using the breadth of data available nowadays.

To gain further understanding of PD, without specifically investigating the mutation G2019S, the identified heterogeneity could be studied in more detail and compared with other clinical measurements. This could be reproduced with more data types in order to have a complete picture of this heterogeneity throughout the different data types.

7.4 Thoughts on precision medicine

Precision medicine approaches have the potential to increase our understanding of many diseases, particularly heterogenous conditions as they may be described as homogenous subtypes. This could help improve treatment strategies for affected individuals. Biobanks, collecting different types of data for large number of individuals, such as UK biobank[182] or the 100,000 genome project[183] and data repositories such as the Gene Expression Omnibus[184] are yielding an enormous potential that can be used for these analyses and are crucial in our endeavour to understand disease pathogenesis better.

A crucial step after disease stratification is to translate the findings to actual clinical practice and routine healthcare. Some examples of applications been mentioned in chapter 1 and there are ongoing projects aiming at integrating data in order to personalise medical care. One example is a collaboration

between the NHS, Illumina and Genomics England,[185] which was announced in January 2020 that aims to provide whole genome sequencing as a routine diagnosis for patients affected by rare diseases and some cancers. For eligible individuals, this will directly improve diagnosis and guide treatment decisions.

8. References

1. *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*.; 2012. doi:10.17226/13284

2. Massicotte P, Eddelbuettel D. gtrendsR: Perform and Display Google Trends Queries. 2019. https://cran.r-project.org/package=gtrendsR.

3. Medical Research Council. *The MRC Framework for the Development, Design and Analysis of Stratifed Medicine Research*. Medical Research Council, Swindon, UK; 2018.

4. Mirnezami R, Nicholson J, Darzi A. Preparing for precision medicine. *N Engl J Med*. 2012. doi:10.1056/NEJMp1114866

5. Chen R, Snyder M. Promise of personalized omics to precision medicine. *Wiley Interdiscip Rev Syst Biol Med*. 2013;5(1):73-82. doi:10.1002/wsbm.1198

6. Collins FS, Varmus H. A new initiative on precision medicine. *N Engl J Med*. 2015. doi:10.1056/NEJMp1500523

7. Ashley EA. Towards precision medicine. *Nat Rev Genet*. 2016;17:507-522. https://doi.org/10.1038/nrg.2016.86.

8. Gordon E. Moore. Cramming more components onto integrated circuits. *Electronics*. 1965;38(8):82-84. doi:10.1111/j.1467-9469.2011.00765.x

9. Hayden EC. The $ 1,000 genome. *Nature*. 2014;507:294-295. doi:10.1038/507294a

10. Schork NJ. Personalized medicine: Time for one-person trials. *Nature*. 2015;520(7549):609-611. doi:10.1038/520609a

11. Mole DJ, McClymont KL, Lau S, et al. Discrepancy between the extent of pancreatic necrosis and multiple organ failure score in severe acute pancreatitis. *World J Surg*. 2009;33(11):2427-2432. doi:10.1007/s00268-009-0161-9

12. Tenner S, Sica G, Hughes M, et al. Relationship of necrosis to organ failure in severe acute pancreatitis. *Gastroenterology*. 1997;113(3):899-903. doi:10.1016/S0016-5085(97)70185-9

13. Landsteiner K. Zur Kenntnis der antifermentativen, lytischen und agglutinierenden Wirkungen des Blutserums und der Lymphe. *Zentralblatt für Bakteriol*. 1900;27:357-362. doi:10.1161/01.RES.25.4.500

14. Himsworth HP. Diabetes mellitus: its differentiation into insulin-sensitive and insulin-insensitive types. *Lancet*. 1936;227(5864):127-130. doi:10.1093/ije/dyt203

15. Hugh-Jones P. DIABETES IN JAMAICA. *Lancet*. 1955;269(6896):891-897. doi:10.1016/S0140-6736(55)92530-7

16. Vogel CL, Cobleigh MA, Tripathy D, et al. Efficacy and safety of trastuzumab as a single agent in first-line treatment of HER2-overexpressing metastatic breast cancer. *J Clin Oncol*. 2002;20(3):719-726. doi:10.1200/JCO.20.3.719

17. Smith I, Procter M, Gelber RD, et al. 2-year follow-up of trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer: a randomised controlled trial. *Lancet*. 2007;369:29-36. doi:10.1016/S0140-6736(07)60028-2

18. Davies H, Bignell GR, Cox C, et al. Mutations of the BRAF gene in human cancer. *Nature*. 2002;417:949-954. doi:10.1038/nature00766

19. Sosman JA, Kim KB, Schuchter L, et al. Survival in BRAF V600-mutant

advanced melanoma treated with vemurafenib. *N Engl J Med*. 2012;366(8):707-714. doi:10.1056/NEJMoa1112302

20. Hulsen T, Jamuar SS, Moody AR, et al. From big data to precision medicine. *Front Med*. 2019;6:34. doi:10.3389/fmed.2019.00034

21. Cirillo D, Valencia A. Big data analytics for personalized medicine. *Curr Opin Biotechnol*. 2019;58:161-167. doi:10.1016/j.copbio.2019.03.004

22. Gray M, Meehan J, Ward C, et al. Implantable biosensors and their contribution to the future of precision medicine. *Vet J*. 2018;239:21-29. doi:10.1016/j.tvjl.2018.07.011

23. Rubin RR, Peyrot M. Psychological issues and treatments for people with diabetes. *J Clin Psychol*. 2001;57(4):457-478. doi:10.1002/jclp.1041

24. Li H, Zhu Y, Burnside ES, et al. Quantitative MRI radiomics in the prediction of molecular classifications of breast cancer subtypes in the TCGA/TCIA data set. *npj Breast Cancer*. 2016;2. doi:10.1038/npjbcancer.2016.12

25. Kerem BS, Rommens JM, Buchanan JA, et al. Identification of the cystic fibrosis gene: Genetic analysis. *Science (80-)*. 1989;245(4922):1073-1080. doi:10.1126/science.2570460

26. Green DM. Cystic fibrosis: a model for personalized genetic medicine. *N C Med J*. 2013;74(6):486-487.

27. Cutting GR. Cystic fibrosis genetics: From molecular understanding to clinical application. *Nat Rev Genet*. 2015;16:45-46. doi:10.1038/nrg3849

28. Duarte CW, Willey CD, Zhi D, et al. Expression signature of IFN/STAT1 signaling genes predicts poor survival outcome in glioblastoma multiforme in a subtype-specific manner. *PLoS One*. 2012;7.

doi:10.1371/journal.pone.0029653

29. Li W, Wang R, Yan Z, Bai L, Sun Z. High accordance in prognosis prediction of colorectal cancer across independent datasets by multi-gene module expression profiles. *PLoS One*. 2012;7. doi:10.1371/journal.pone.0033653

30. Ali M, Khan SA, Wennerberg K, Aittokallio T. Global proteomics profiling improves drug sensitivity prediction: Results from a multi-omics, pan-cancer modeling approach. *Bioinformatics*. 2018;34(8):1353-1362. doi:10.1093/bioinformatics/btx766

31. Olivier M, Asmis R, Hawkins GA, Howard TD, Cox LA. The Need for Multi-Omics Biomarker Signatures in Precision Medicine. *Int J Mol Sci*. 2019;20(19):4781. doi:10.3390/ijms20194781

32. Giskeødegård GF, Bertilsson H, Selnæs KM, et al. Spermine and Citrate as Metabolic Biomarkers for Assessing Prostate Cancer Aggressiveness. *PLoS One*. 2013;8. doi:10.1371/journal.pone.0062375

33. Boland MR, Polubriaginof F, Tatonetti NP. Development of A Machine Learning Algorithm to Classify Drugs of Unknown Fetal Effect. *Sci Rep*. 2017;7:12839. doi:10.1038/s41598-017-12943-x

34. Gligorijević V, Malod-Dognin N, Pržulj N. Integrative methods for analyzing big data in precision medicine. *Proteomics*. 2016;16(5):741-758. doi:10.1002/pmic.201500396

35. Feiler T, Gaitskell K, Maughan T, Hordern J. Personalised Medicine: The Promise, the Hype and the Pitfalls. *New Bioeth*. 2017;23(1):1-12. doi:10.1080/20502877.2017.1314895

36. Xiao AY, Tan MLY, Wu LM, et al. Global incidence and mortality of pancreatic diseases: a systematic review, meta-analysis, and meta-regression of population-based cohort studies. *Lancet Gastroenterol*

Hepatol. 2016;1(1):45-55. doi:10.1016/S2468-1253(16)30004-8

37. Forsmark CE, Baillie J. AGA Institute Technical Review on Acute Pancreatitis. *Gastroenterology*. 2007;132(5):2022-2044. doi:10.1053/j.gastro.2007.03.065

38. Mole DJ, Gungabissoon U, Johnston P, et al. Identifying risk factors for progression to critical care admission and death among individuals with acute pancreatitis: A record linkage analysis of Scottish healthcare databases. *BMJ Open*. 2016;6(6):e011474. doi:10.1136/bmjopen-2016-011474

39. Calfee CS, Delucchi K, Parsons PE, Thompson BT, Ware LB, Matthay MA. Subphenotypes in acute respiratory distress syndrome: Latent class analysis of data from two randomised controlled trials. *Lancet Respir Med*. 2014;2(8):611-620. doi:10.1016/S2213-2600(14)70097-9

40. Skouras C, Zheng X, Binnie M, et al. Increased levels of 3-hydroxykynurenine parallel disease severity in human acute pancreatitis. *Sci Rep*. 2016;6:33951. doi:10.1038/srep33951

41. Esteller M. Non-coding RNAs in human disease. *Nat Rev Genet*. 2011;12:861-874. doi:10.1038/nrg3074

42. Crick F. Central dogma of molecular biology. *Nature*. 1970;227(1):561-563. doi:10.1038/227561a0

43. Civelek M, Lusis A. Systems genetics approaches to understand complex traits. *Nat Rev Genet*. 2014;15(1):34-48. doi:10.1038/nrg3575

44. Metzker M. Sequencing technologies - the next generation. *Nat Rev Genet*. 2010;11(1):31-46. doi:10.1038/nrg2626

45. Bentley D, Balasubramanian S, Swerdlow H, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*. 2008;456(7218):53-59. doi:10.1038/nature07517

46. Schwarze K, Buchanan J, Taylor JC, Wordsworth S. Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. *Genet Med*. 2018;20(10):1122-1130. doi:10.1038/gim.2017.247

47. Bird A. Perceptions of epigenetics. *Nature*. 2007;447(7143):396-398. doi:10.1038/nature05913

48. Trapnell C, Roberts A, Goff L, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*. 2012;7(3):562-578. doi:10.1038/nprot.2012.016

49. Maier T, Güell M, Serrano L. Correlation of mRNA and protein in complex biological samples. *FEBS Lett*. 2009;538(24):3966-3973. doi:10.1016/j.febslet.2009.10.036

50. Kim MS, Pinto SM, Getnet D, et al. A draft map of the human proteome. *Nature*. 2014;509(7502):575-581. doi:10.1038/nature13302

51. Tam SW, Pirro J, Hinerfeld D. Depletion and fractionation technologies in plasma proteomic analysis. *Expert Rev Proteomics*. 2004;1(4):411-420. doi:10.1586/14789450.1.4.411

52. Emwas AHM. The strengths and weaknesses of NMR spectroscopy and mass spectrometry with particular focus on metabolomics research. *Methods Mol Biol*. 2015;1277:161-193. doi:10.1007/978-1-4939-2377-9_13

53. Jain A, Murty M, Flynn P. Data clustering: a review. *ACM Comput Surv*. 1999;31(3):264-323.

54. Theodoridis S, Koutroumbas K, Nos K, Bas KM. *Pattern Recognition, Second Edition*.; 2006. doi:10.1007/BF02680460

55. Kira K, Rendell LA. A Practical Approach to Feature Selection. In: *Machine Learning Proceedings 1992*. ; 2014:249-256.

doi:10.1016/b978-1-55860-247-2.50037-1

56. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics*. 2007;23(19):2507-2517.

57. Khalid S, Khalil T, Nasreen S. A survey of feature selection and feature extraction techniques in machine learning. In: *Proceedings of 2014 Science and Information Conference, SAI 2014*. ; 2014. doi:10.1109/SAI.2014.6918213

58. Grabusts P. The Choice of Metrics for Clustering Algorithms. *Environ Technol Resour Proc Int Sci Pract Conf*. 2015;2:70-76. doi:10.17770/etr2011vol2.973

59. Clapham C, Nicholson J. *The Concise Oxford Dictionary of Mathematics*.; 2009. doi:10.1093/acref/9780199235940.001.0001

60. Upton G, Cook I. *A Dictionary of Statistics*.; 2014. doi:10.1093/acref/9780199679188.001.0001

61. Gauthier TD. Detecting trends using Spearman's rank correlation coefficient. *Environ Forensics*. 2001;2(4):359-362. doi:10.1006/enfo.2001.0061

62. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math*. 1987;20:53-65. doi:10.1016/0377-0427(87)90125-7

63. Kaufman L, Rousseeuw PJ. *Finding Groups in Data: An Introduction to Cluster Analysis (Wiley Series in Probability and Statistics)*.; 1990.

64. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*. 1998;95(25):14863-14868.

65. Perou CM, Sørile T, Eisen MB, et al. Molecular portraits of human

breast tumours. *Nature*. 2000;406(6797):747-752. doi:10.1038/35021093

66. MacNaughton-Smith P, Williams WT, Dale MB, Mockett LG. Dissimilarity analysis: A new technique of hierarchical sub-division. *Nature*. 1964;202(4936):1034-1035. doi:10.1038/2021034a0

67. Sokal RR. A statiscal method for evaluating systematic relationships. *Univ Kansas Sci Bull*. 1958;38(2):1409-1438.

68. Ward JH. Hierarchical Grouping to Optimize an Objective Function. *J Am Stat Assoc*. 1963;58(301):236-244. doi:10.1080/01621459.1963.10500845

69. MacQueen J. Some Methods for classification and Analysis of Multivariate Observations. In: *5th Berkeley Symposium on Mathematical Statistics and Probability 1967*. ; 1967:281-297.

70. Divoux A, Tordjman J, Lacasa D, et al. Fibrosis in human adipose tissue: Composition, distribution, and link with lipid metabolism and fat mass loss. *Diabetes*. 2010;59(11):2817-2825.

71. Lefaudeux D, De Meulder B, Loza MJ, et al. U-BIOPRED clinical adult asthma clusters linked to a subset of sputum omics. *J Allergy Clin Immunol*. 2017;139(6):1797-1807. doi:10.1016/j.jaci.2016.08.048

72. Ester M, Kriegel H-P, Sander J, Xu X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Proc 2nd Int Conf Knowl Discov Data Min*. 1996;96(34):226-231.

73. Ankerst M, Breunig MM, Kriegel HP, Sander J. OPTICS: Ordering Points to Identify the Clustering Structure. *SIGMOD Rec (ACM Spec Interes Gr Manag Data)*. 1999;28(2):49-60. doi:10.1145/304181.304187

74. Banfield JD, Raftery AE. Model-Based Gaussian and Non-Gaussian

Clustering. *Biometrics*. 2006;49(3):803. doi:10.2307/2532201

75. Kohonen T. *Self-Organizing Maps*.; 2001.

76. Yeung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL. Model-based clustering and data transformations for gene expression data. *Bioinformatics*. 2001;17(10):977-987. doi:10.1093/bioinformatics/17.10.977

77. Ringnér M. What is principal component analysis? *Nat Biotechnol*. 2008;26(3):303-304. doi:10.1038/nbt0308-303

78. Tamayo P, Slonim D, Mesirov J, et al. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc Natl Acad Sci*. 2002;96(6):2907-2912. doi:10.1073/pnas.96.6.2907

79. Brock G, Pihur V, Datta S, Datta S. clValid : An R Package for Cluster Validation. *J Stat Softw*. 2008;25(4). doi:10.18637/jss.v025.i04

80. Hennig C. Cluster-wise assessment of cluster stability. *Comput Stat Data Anal*. 2007;52(1):258-271. doi:10.1016/j.csda.2006.11.025

81. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res*. 2017;45(Database):353-361. doi:10.1093/nar/gkw1092

82. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: Tool for the unification of biology. *Nat Genet*. 2000;25(1):25-29. doi:10.1038/75556

83. Vazire S. Implications of the Credibility Revolution for Productivity, Creativity, and Progress. *Perspect Psychol Sci*. 2018;13(4):411-417. doi:10.1177/1745691617751884

84. Galar M, Fernández A, Barrenechea E, Bustince H, Herrera F. An overview of ensemble methods for binary classifiers in multi-class

problems: Experimental study on one-vs-one and one-vs-all schemes. *Pattern Recognit*. 2011;44(8):1761-1776. doi:10.1016/j.patcog.2011.01.017

85. Varpa K, Joutsijoki H, Iltanen K, Juhola M. Applying one-vs-one and one-vs-all classifiers in k-nearest neighbour method and support vector machines to an otoneurological multi-class problem. In: *Studies in Health Technology and Informatics*. ; 2011:579-583. doi:10.3233/978-1-60750-806-9-579

86. Adnan MN, Islam MZ. One-vs-all binarization technique in the context of random forest. In: *Computational Intelligence and Machine Learning*. ; 2015:385-390. doi:10.1073/pnas.0407792102

87. Barker M, Rayens W. Partial least squares for discrimination. *J Chemom*. 2003;17(3):166-173. doi:10.1002/cem.785

88. Breiman L. Random Forests. *Mach Learn*. 2001;45(1):5-32.

89. Sun Y V., Hu YJ. Integrative Analysis of Multi-omics Data for Discovery and Functional Studies of Complex Human Diseases. *Adv Genet*. 2016;93:147-190. doi:10.1016/bs.adgen.2015.11.004

90. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*. 2012;28(6):882-883. doi:10.1093/bioinformatics/bts034

91. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007;8(1):118-127. doi:10.1093/biostatistics/kxj037

92. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550. doi:10.1186/s13059-014-0550-8

93. Robinson MD, McCarthy DJ, Smyth GK. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2009;26(1):139-140. doi:10.1093/bioinformatics/btp616

94. Schwanhüusser B, Busse D, Li N, et al. Global quantification of mammalian gene expression control. *Nature*. 2011;473(7347):337-342. doi:10.1038/nature10098

95. Argelaguet R, Velten B, Arnol D, et al. Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol*. 2018;14:e8124. doi:10.15252/msb.20178124

96. Wang B, Mezlini AM, Demir F, et al. Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods*. 2014;11(3):333-337. doi:10.1038/nmeth.2810

97. Falconer SD, Mackay FCT. *Introduction to Quantitative Genetics, 4th Edition*.; 1996.

98. Majumder PP, Ghosh S. Mapping quantitative trait loci in humans: Achievements and limitations. *J Clin Invest*. 2005;115(6):1419-1424. doi:10.1172/JCI24757

99. Dash M, Liu H. Feature selection for classification. *Intell Data Anal*. 1997;1(1-4):131-156. doi:10.3233/IDA-1997-1302

100. Guyon I, Elisseeff A. An Introduction to Variable and Feature Selection. *J Mach Learn Res*. 2011;3(7-8):1157-1182. doi:10.1016/j.aca.2011.07.027

101. Liang Y, Kelemen A. Computational dynamic approaches for temporal omics data with applications to systems medicine. *BioData Min*. 2017;10(1). doi:10.1186/s13040-017-0140-x

102. Grigorov MG. Analysis of time course Omics datasets. *Methods Mol*

Biol. 2011;719:153-172. doi:10.1007/978-1-61779-027-0_7

103. Tarazona S, Balzano-Nogueira L, Conesa A. *Multiomics Data Integration in Time Series Experiments*. Vol 82.; 2018. doi:10.1016/bs.coac.2018.06.005

104. Neyton L, Zheng X, Skouras C, et al. Multiomic definition of generalizable endotypes in human acute pancreatitis. *bioRxiv*. 2019:539569. doi:10.1101/539569

105. Steinberg W, Tenner S. Acute pancreatitis. *N Engl J Med*. 1994;330(17):1198-1210. doi:10.1056/nejm199404283301706

106. Peery AF, Dellon ES, Lund J, et al. Burden of gastrointestinal disease in the United States: 2012 update. *Gastroenterology*. 2012;143(5):1179-1187. doi:10.1053/j.gastro.2012.08.002

107. Kang R, Lotze MT, Zeh HJ, Billiar TR, Tang D. Cell Death and DAMPs in Acute Pancreatitis. *Mol Med*. 2014;20:466-477. doi:10.2119/molmed.2014.00117

108. Mole DJ, Webster SP, Uings I, et al. Kynurenine-3-monooxygenase inhibition prevents multiple organ failure in rodent models of acute pancreatitis. *Nat Med*. 2016;22(2):202-209. doi:10.1038/nm.4020

109. Warndorf MG, Kurtzman JT, Bartel MJ, et al. Early Fluid Resuscitation Reduces Morbidity Among Patients With Acute Pancreatitis. *Clin Gastroenterol Hepatol*. 2011;9(8):705-709. doi:10.1016/j.cgh.2011.03.032

110. Simoes. Predicting Acute Pancreatitis Severity: Comparison of Prognostic Scores. *Gastroenterol Res*. 2011;4(5):216-222. doi:10.4021/gr364w

111. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: A severity of disease classification system. *Crit Care Med*.

1985;13(10):818-829. doi:10.1097/00003246-198510000-00009

112. Imrie CW, Benjamin IS, Ferguson JC, et al. A single-centre double-blind trial of Trasylol therapy in primary acute pancreatitis. *Br J Surg*. 1978;65:337-341. doi:10.1002/bjs.1800650514

113. Puolakkainen P, Valtonen V, Paananen A, Schröder T. C-reactive protein (CRP) and serum phospholipase A2 in the assessment of the severity of acute pancreatitis. *Gut*. 1987;28:764-771. doi:10.1136/gut.28.6.764

114. Banks PA, Bollen TL, Dervenis C, et al. Classification of acute pancreatitis—2012: revision of the Atlanta classification and definitions by international consensus. *Gut*. 2013;62(1):102-111. doi:10.1136/gutjnl-2012-302779

115. Schneider VA, Graves-Lindsay T, Howe K, et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res*. 2017;27(5):849-864. doi:10.1101/gr.213611.116

116. Dobin A, Davis CA, Schlesinger F, et al. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15-21. doi:10.1093/bioinformatics/bts635

117. Liao Y, Smyth GK, Shi W. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*. 2014;30(7):923-930. doi:10.1093/bioinformatics/btt656

118. Tarazona S, Furió-Tarí P, Turrà D, et al. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Res*. 2015;43(21):e140. doi:10.1093/nar/gkv711

119. Lough G, Kyriazakis I, Bergmann S, Lengeling A, Doeschl-Wilson AB. Health trajectories reveal the dynamic contributions of host genetic

resistance and tolerance to infection outcome. *Proc R Soc B Biol Sci*. 2015;282(1819):20152151. doi:10.1098/rspb.2015.2151

120. Giorgino T. Computing and Visualizing Dynamic Time Warping Alignments in R : The dtw Package. *J Stat Softw*. 2009;31(7):1-24. doi:10.18637/jss.v031.i07

121. Strauss T, Von Maltitz MJ. Generalising ward's method for use with manhattan distances. *PLoS One*. 2017;12(1):e0168288. doi:10.1371/journal.pone.0168288

122. Miyamoto S, Abe R, Endo Y, Takeshita JI. Ward method of hierarchical clustering for non-Euclidean similarity measures. In: *Proceedings of the 2015 7th International Conference of Soft Computing and Pattern Recognition, SoCPaR 2015*. ; 2016:60-63. doi:10.1109/SOCPAR.2015.7492784

123. Szekely GJ, Rizzo ML. Hierarchical clustering via joint between-within distances: Extending Ward's minimum variance method. *J Classif*. 2005;22(2):151-183. doi:10.1007/s00357-005-0012-9

124. Hennig C. Dissolution point and isolation robustness: Robustness criteria for general cluster analysis methods. *J Multivar Anal*. 2008;99(6):1154-1176. doi:10.1016/j.jmva.2007.07.002

125. Luo W, Friedman MS, Shedden K, Hankenson KD, Woolf PJ. GAGE: Generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics*. 2009;10(1):161. doi:10.1186/1471-2105-10-161

126. Xia J, Wishart DS. Using metaboanalyst 3.0 for comprehensive metabolomics data analysis. *Curr Protoc Bioinforma*. 2016;55(1):14.10.1-14.10.91. doi:10.1002/cpbi.11

127. Lizio M, Harshbarger J, Shimoji H, et al. Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol*. 2015;16(1).

doi:10.1186/s13059-014-0560-6

128. Andersson R, Gebhard C, Miguel-Escalada I, et al. An atlas of active enhancers across human cell types and tissues. *Nature*. 2014;507:455-461. doi:10.1038/nature12787

129. Chong IG, Jun CH. Performance of some variable selection methods when multicollinearity is present. *Chemom Intell Lab Syst*. 2005;78(1-2):103-112. doi:10.1016/j.chemolab.2004.12.011

130. Bostock M, Ogievetsky V, Heer J. D3 data-driven documents. *IEEE Trans Vis Comput Graph*. 2011;17(12):2301-2309. doi:10.1109/TVCG.2011.185

131. Sweeney TE, Azad TD, Donato M, et al. Unsupervised analysis of transcriptomics in bacterial sepsis across multiple datasets reveals three robust clusters. *Crit Care Med*. 2018;46(6):915-925. doi:10.1097/CCM.0000000000003084

132. Thompson B., Chambers R, Liu K. Acute respiratory distress syndrome. Introduction. *N Engl J Med*. 2017;377(19):1904-1905. doi:10.1016/S0210-5691(06)74495-3

133. Ranieri VM, Rubenfeld GD, Thompson BT, et al. Acute respiratory distress syndrome: The Berlin definition. *JAMA - J Am Med Assoc*. 2012. doi:10.1001/jama.2012.5669

134. Burnham KL, Davenport EE, Radhakrishnan J, et al. Shared and distinct aspects of the sepsis transcriptomic response to fecal peritonitis and pneumonia. *Am J Respir Crit Care Med*. 2017;196(3):328-339. doi:10.1164/rccm.201608-1685OC

135. Davenport EE, Burnham KL, Radhakrishnan J, et al. Genomic landscape of the individual host response and outcomes in sepsis: A prospective cohort study. *Lancet Respir Med*. 2016;4(4):259-271.

doi:10.1016/S2213-2600(16)00046-1

136. Scicluna BP, van Vught LA, Zwinderman AH, et al. Classification of patients with sepsis according to blood genomic endotype: a prospective cohort study. *Lancet Respir Med*. 2017;5(10):816-826. doi:10.1016/S2213-2600(17)30294-1

137. Wong HR, Cvijanovich N, Lin R, et al. Identification of pediatric septic shock subclasses based on genome-wide expression profiling. *BMC Med*. 2009;7:34. doi:10.1186/1741-7015-7-34

138. Sweeney TE, Shidham A, Wong HR, Khatri P. A comprehensive time-course-based multicohort analysis of sepsis and sterile inflammation reveals a robust diagnostic gene set. *Sci Transl Med*. 2015;7(287):287ra271. doi:10.1126/scitranslmed.aaa5993

139. Dunning J, Blankley S, Hoang LT, et al. Progression of whole-blood transcriptional signatures from interferon-induced to neutrophil-associated patterns in severe influenza. *Nat Immunol*. 2018;19(6):625-635. doi:10.1038/s41590-018-0111-5

140. Hofman A, Breteler MMB, Van Duijn CM, et al. The Rotterdam Study: Objectives and design update. *Eur J Epidemiol*. 2007;22(11):819-829. doi:10.1007/s10654-007-9199-x

141. Chawla N V., Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16:321-357.

142. Kuhn M. Building predictive models in R using the caret package. *J Stat Softw*. 2008;28(5):1-26. doi:10.18637/jss.v028.i05

143. Kapp A V., Tibshirani R. Are clusters found in one dataset present in another dataset? *Biostatistics*. 2007;8(1):9-31. doi:10.1093/biostatistics/kxj029

144. Brown LD, Cai TT, Das Gupta A. Interval estimation for a binomial proportion. *Stat Sci*. 2001;16(2):101-117. doi:10.1214/ss/1009213286

145. Agresti A, Coull BA. Approximate is better than “Exact” for interval estimation of binomial proportions. *Am Stat*. 1998;52(2):119-126. doi:10.1080/00031305.1998.10480550

146. Baillie JK, Bretherick A, Haley CS, et al. Shared activity patterns arising at genetic susceptibility loci reveal underlying genomic and cellular architecture of human disease. *PLoS Comput Biol*. 2018;14(3). doi:10.1371/journal.pcbi.1005934

147. Sweeney TE, Wong HR, Khatri P. Robust classification of bacterial and viral infections via integrated host gene expression diagnostics. *Sci Transl Med*. 2016;8(346). doi:10.1126/scitranslmed.aaf7165

148. Han B, Pouget JG, Slowikowski K, et al. A method to decipher pleiotropy by detecting underlying heterogeneity driven by hidden subgroups applied to autoimmune and neuropsychiatric diseases. *Nat Genet*. 2016;48(7):803-810. doi:10.1038/ng.3572

149. Jostins L, Ripke S, Weersma RK, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*. 2012;491(7422):119. doi:10.1038/nature11582

150. Franke A, McGovern DPB, Barrett JC, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn’s disease susceptibility loci. *Nat Genet*. 2010;42(12):1118-1125. doi:10.1038/ng.717

151. Anderson CA, Boucher G, Lees CW, et al. Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat Genet*. 2011;43(3):246-252. doi:10.1038/ng.764

152. Liu JZ, Van Sommeren S, Huang H, et al. Association analyses identify

38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat Genet*. 2015;47(9):979-986. doi:10.1038/ng.3359

153. Goyette P, Boucher G, Mallon D, et al. High-density mapping of the MHC identifies a shared role for HLA-DRB1∗01:03 in inflammatory bowel diseases and heterozygous advantage in ulcerative colitis. *Nat Genet*. 2015;47(2):172-179. doi:10.1038/ng.3176

154. Cleynen I, Boucher G, Jostins L, et al. Inherited determinants of Crohn's disease and ulcerative colitis phenotypes: A genetic association study. *Lancet*. 2016;387(10014):156-167. doi:10.1016/S0140-6736(15)00465-1

155. Hinrichs AS. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res*. 2006;1(34):D590-D598. doi:10.1093/nar/gkj144

156. Khera A V., Chaffin M, Aragam KG, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet*. 2018;50(9):1219. doi:10.1038/s41588-018-0183-z

157. Weiser M, Simon JM, Kochar B, et al. Molecular classification of Crohn's disease reveals two clinically relevant subtypes. *Gut*. 2018;67:36-42. doi:10.1136/gutjnl-2016-312518

158. Warner TT, Schapira AHV, Tatton, et al. Genetic and environmental factors in the cause of Parkinson's disease. *Ann Neurol*. 2003;53(S3):S16-S25. doi:10.1002/ana.10487

159. Klein C, Westenberger A. Genetics of Parkinson' s Disease. *Cold Srping Harb Perspect Med*. 2012;2(1).

160. Elkouzi A. Parkinson's Foundation: Better Lives. Together. Parkinson's Foundation. https://www.parkinson.org/Understanding-

Parkinsons/Causes/Environmental-Factors. Published 2017. Accessed October 22, 2019.

161. Paisán-Ruíz C, Jain S, Evans EW, et al. Cloning of the gene containing mutations that cause PARK8-linked Parkinson's disease. *Neuron*. 2004;44(4):595-600. doi:10.1016/j.neuron.2004.10.023

162. Zimprich A, Biskup S, Leitner P, et al. Mutations in LRRK2 cause autosomal-dominant parkinsonism with pleomorphic pathology. *Neuron*. 2004;44(4):601-607. doi:10.1016/j.neuron.2004.11.005

163. Davis AA, Andruska KM, Benitez BA, Racette BA, Perlmutter JS, Cruchaga C. Variants in GBA, SNCA, and MAPT influence Parkinson disease risk, age at onset, and progression. *Neurobiol Aging*. 2016;37:209.e1-209.e7. doi:10.1016/j.neurobiolaging.2015.09.014

164. Nichols WC, Pankratz N, Hernandez D, et al. Genetic screening for a single common LRRK2 mutation in familial Parkinson's disease. *Lancet*. 2005;365(9457):410-412. doi:10.1016/S0140-6736(05)17828-3

165. Di Fonzo A, Rohé CF, Ferreira J, et al. A frequent LRRK2 gene mutation associated with autosomal dominant Parkinson's disease. *Lancet*. 2005;365(9457):412-415. doi:10.1016/S0140-6736(05)17829-5

166. Gilks WP, Abou-Sleiman PM, Gandhi S, et al. A common LRRK2 mutation in idiopathic Parkinson's disease. *Lancet*. 2005;365(9457):415-416. doi:10.1016/S0140-6736(05)17830-1

167. Do CB, Tung JY, Dorfman E, et al. Web-based genome-wide association study identifies two novel loci and a substantial genetic component for parkinson's disease. *PLoS Genet*. 2011;7(6). doi:10.1371/journal.pgen.1002141

168. Goetz CG, Tilley BC, Shaftman SR, et al. Movement Disorder Society-Sponsored Revision of the Unified Parkinson's Disease Rating Scale

(MDS-UPDRS): Scale presentation and clinimetric testing results. *Mov Disord*. 2008;23(15):2129-2170. doi:10.1002/mds.22340

169. Parkinson's Progression Markers Initiative | About PPMI.

170. Kowarik A, Templ M. Imputation with the R Package VIM. *J Stat Softw*. 2016;74(7). doi:10.18637/jss.v074.i07

171. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43(7):e47. doi:10.1093/nar/gkv007

172. Du P, Zhang X, Huang CC, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*. 2010;11(587). doi:10.1186/1471-2105-11-587

173. Ng AY, Jordan MI, Weiss Y. On Spectral Clustering: Analysis and an Algorithm. In: *Advances in Neural Information Processing Systems*. ; 2001:849-856. doi:10.1.1.19.8100

174. Charrad M, Ghazzali N, Boiteau V, Niknafs A. NbClust : An R Package for Determining the Relevant Number of Clusters in a Data Set. *J Stat Softw*. 2014;61(6). doi:10.18637/jss.v061.i06

175. Fabregat A, Sidiropoulos K, Garapati P, et al. The reactome pathway knowledgebase. *Nucleic Acids Res*. 2016;4(44):481-487. doi:10.1093/nar/gkv1351

176. Frost HR, Li Z, Moore JH. Principal component gene set enrichment (PCGSE). *BioData Min*. 2015;8(25). doi:10.1186/s13040-015-0059-z

177. Yu G, He QY. ReactomePA: An R/Bioconductor package for reactome pathway analysis and visualization. *Mol Biosyst*. 2016;12(2):477-479. doi:10.1039/c5mb00663e

178. Chaudhary K, Poirion OB, Lu L, Garmire LX. Deep learning–based multi-omics integration robustly predicts survival in liver cancer. *Clin Cancer Res*. 2018;24(6):1248-1259. doi:10.1158/1078-0432.CCR-17-0853

179. Zhang L, Lv C, Jin Y, et al. Deep learning-based multi-omics data integration reveals two prognostic subtypes in high-risk neuroblastoma. *Front Genet*. 2018;9(477). doi:10.3389/fgene.2018.00477

180. Mamoshina P, Vieira A, Putin E, Zhavoronkov A. Applications of Deep Learning in Biomedicine. *Mol Pharm*. 2016;13(5):1445-1454. doi:10.1021/acs.molpharmaceut.5b00982

181. Podani J, Miklós I. Resemblance coefficients and the horseshoe effect in principal coordinates analysis. *Ecology*. 2002;83(12):3331-3343. doi:10.1890/0012-9658(2002)083[3331:RCATHE]2.0.CO;2

182. Sudlow C, Gallacher J, Allen N, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Med*. 2015;12(3):e1001779. doi:10.1371/journal.pmed.1001779

183. Torjesen I. Genomes of 100,000 people will be sequenced to create an open access research resource. *BMJ*. 2013;347:f6690. doi:10.1136/bmj.f6690

184. Clough E, Barrett T. The Gene Expression Omnibus database. *Methods Mol Biol*. 2016;1418:93-110. doi:10.1007/978-1-4939-3578-9_5

185. Genomics England and Illumina partner to deliver whole genome sequencing for England's NHS Genomic Medicine Service. https://www.genomicsengland.co.uk/genomics-england-illumina-partner-nhs-genomic-medicine-service/. Accessed January 23, 2020.

Appendices

A. Sample scripts

A.1 time_s_dist.py

```
'''''
3 main methods allowing to generate distances between samples with different time points
input data must present no missing values
2 methods to write the obtained distances in different format (json or abc)

'''

import pandas as pd
from scipy.spatial.distance import hamming
import numpy
from sklearn.decomposition.pca import PCA
from sklearn.metrics import auc
import rpy2
from rpy2.robjects.packages import importr
import rpy2.robjects as ro
from rpy2.robjects import pandas2ri
pandas2ri.activate()
import rpy2.robjects.numpy2ri
rpy2.robjects.numpy2ri.activate()
import scipy.spatial.distance as d


'''''
Dynamic Time Warping function
input: DataFrame with variables as columns and samples as rows, assumes normalisation+extrapolation, time point values should be a column with index='TimePointScale' / sample identifiers should be a column with index='ID'
output: numpy matrix of pairwise distances between samples
'''
def dtw_dist_mat(data_complete_dtw):

    #extract identifier columns
    data_ID = list(set(data_complete_dtw['ID']))

    #create an empty array, will be used as input for dtwclust package in R

    list_ = []

    #for each patient
    for id_ in data_ID:
        #extract part of dataframe related to this identifier
        data_tmp = data_complete_dtw.loc[data_complete_dtw['ID']==id_]
        #get time vector and sort dataframe according to it
        data_tmp = data_tmp.sort_values('TimePointScale')
        #drop time and id columns
        data_tmp=data_tmp.drop('TimePointScale', axis=1)
        data_tmp=data_tmp.drop('ID', axis=1)

        #append the data to a list as required by dtwclust R package
        list_.append(data_tmp.as_matrix())

    #store data into r environment
    ro.globalenv['indexes']=data_ID
```

```
    ro.globalenv['list_']=list_

    #set names for each element of the list
    ro.r('names(list_)<-indexes')

    #load dtwclust library
    ro.r('library(dtwclust)')

    #perform dtw, distance matrix is extracted so method will not change th
e output
    ro.r('res <- dtwclust(list_,type = "hierarchical")')

    #extract distance matrix produced
    ro.r('dm <- attributes(res)$distmat')

    #get it from r envirnoment then format it so it can be used in Python
    dm_dtw = ro.globalenv['dm']
    dm_dtw = ro.r['matrix'](dm_dtw, nrow = len(data_ID))
    dm_dtw = numpy.matrix(dm_dtw)

    return dm_dtw




'''''
Area Under the Curve and PCA function
input: DataFrame with variables as columns and samples as rows, assumes nor
malisation+extrapolation, time point values should be a column with index='
TimePointScale' / sample identifiers should be a column with index='ID'
output: numpy matrix of pairwise distances between samples
'''
def auc_pca(data_,auc_suffix):

    #extract identifier columns
    data_ID = list(set(data_['ID']))
    #create an empty dataframe where the auc values will be added
    df_ = pd.DataFrame(index=range(0,len(data_ID)), columns=data_.columns.v
alues)

    #compute auc for each variable for each patient
    for v in data_:
        #create a vector where auc values will be stored
        col = []
        for id_ in data_ID:

            #extract value vector for this patient for this variable
            df_vals = data_.loc[data_['ID']==id_][v]
            #extract value vector for this patient for time
            df_t = data_.loc[data_['ID']==id_]['TimePointScale']

            #normalise it so that it takes into account the difference in l
enght of each serie
            auc_val = auc(df_t,df_vals)/((max(df_t)-min(df_t)))

                    col.append(auc_val)
               #add this vector as a new column to the auc values DataFrame

               df_[v] = col

          #drop time and identifier columns
          df_ = df_.drop('TimePointScale', axis=1)
          df_ = df_.drop('ID', axis=1)
```

```
107.
108.            #store variables
109.            cols_ = df_.columns.values
110.
111.            #perform pca
112.            pca = PCA(n_components=len(cols_))
113.            pca.fit(df_)
114.            pca_data = pca.transform(df_)
115.            pca_data_comp_1_2 = pca_data[:,0:2]
116.
117.            #extract unique identifiers and add them back to the dataframe
118.            df_['ID']=data_ID
119.
120.            #save AUC values
121.            to_save = df_.copy()
122.            to_save.to_csv(auc_suffix+'auc_values.csv')
123.
124.            #compute each axe's contribution
125.            a = ((pca.explained_variance_/numpy.sum(pca.explained_variance_)
    )*100)[0]
126.            b = ((pca.explained_variance_/numpy.sum(pca.explained_variance_)
    )*100)[1]
127.
128.            #compute Euclidean distances corrected for explained variance fo
    r each axis
129.            dm_pca = d.pdist(pca_data_comp_1_2, lambda u, v: numpy.sqrt((((u
    [0]-v[0])**2)*a)+(((u[1]-v[1])**2)*b)))
130.
131.            #create numpy distance matrix (euclidean)
132.            dm_pca  = d.squareform(dm_pca)
133.            dm_pca = numpy.matrix(dm_pca)
134.
135.            return dm_pca
136.
137.
138.        '''''
139.        PCA and Trajectory function
140.        input: DataFrame with variables as columns and samples as rows, assu
    mes normalisation+extrapolation, time point values should be a column with
    index='TimePointScale' / sample identifiers should be a column with index='
    ID'
141.        output: numpy matrix of pairwise distances between samples
142.        '''
143.        def pca_traj(data_pre_processed):
144.
145.            #isolate new id and time vectors
146.            data_ID_pca = data_pre_processed['ID']
147.            data_TimePointScale_pca = data_pre_processed['TimePointScale']
148.
149.            #drop id and time vectors
150.            data_pre_processed=data_pre_processed.drop('ID', axis=1)
151.            data_pre_processed=data_pre_processed.drop('TimePointScale', axi
    s=1)
152.
153.
154.            #perform pca
155.            pca = PCA(n_components=len(data_pre_processed.columns.values))
156.            pca.fit(data_pre_processed)
157.            pca_data = pca.transform(data_pre_processed)
158.
159.
160.            #isolate first and second components
161.            pca_data_comp_1 = pca_data[:,0]
```

```
            pca_data_comp_2 = pca_data[:,1]

            #add projected values using pca to dataframe
            pca_data_comp_1 = pd.DataFrame(pca_data_comp_1,columns=['PC1']).reset_index(drop=True)
            pca_data_comp_2 = pd.DataFrame(pca_data_comp_2,columns=['PC2']).reset_index(drop=True)
            data_pre_processed = pd.concat([pca_data_comp_1,pca_data_comp_2,data_ID_pca,data_TimePointScale_pca],axis=1)

            var1 = 'PC1'
            var2 = 'PC2'
            #create a dictionary to store the trajectory of each patient through the defined space
            dict_traj = {}
            for id_ in list(set(data_ID_pca)):
                #extract data for each patient
                data_ = data_pre_processed.loc[data_pre_processed['ID']==id_]

                #create an empty trajectory vector
                traj = []

                data_ = data_.reset_index(drop=True)

                i = 0
                #for each time point
                for val in data_['TimePointScale']:
                    if i < len(data_['TimePointScale'])-1:

                        #get value for current time point and for next time point
                        val1 = data_.loc[(data_['TimePointScale'])==val][var1][i]
                        val2 = data_.loc[(data_['TimePointScale'])==val][var2][i]

                        #get the change on x axis and on y axis
                        x_change = data_.loc[(data_['TimePointScale'])==data_['TimePointScale'][i+1]][var1][i+1] - val1
                        y_change = data_.loc[(data_['TimePointScale'])==data_['TimePointScale'][i+1]][var2][i+1] - val2

                        #according to the changes associate value
                        if x_change>=0 and y_change>=0:
                            traj.append(4)
                        else:
                            if x_change<0 and y_change<0:
                                traj.append(2)
                            else:
                                if x_change<0 and y_change>=0:
                                    traj.append(3)
                                else:
                                    traj.append(1)

                    i = i+1
                    #add computed trajectory to dictionary
                    dict_traj[id_] = traj

            #create an empty matrix where the hamming distances between trajectories will be recorded
            matrix = numpy.zeros((len(dict_traj),len(dict_traj)))

```

```
            i = 0
            #compute hamming distances
            for k1,elem_i in dict_traj.items():
                j = 0
                for k2,elem_j in dict_traj.items():
                    matrix[i,j] = hamming(elem_i,elem_j)

                    j = j +1
                i = i+1

            matrix = numpy.matrix(matrix)
            return matrix

        '''''
        Function to create JSON files out of distance matrices
        input: pairwise distance Numpy matrix, array of labels (identifiers)

        output: String following JSON format that can be saved as a JSON file
        '''
        def create_json( datam,labels):
            string_json = ''
            string_json = '{"nodes":['

            #create node element for each id
            for label in labels:

                if label == labels[len(labels)-1]:
                    string_json += '{"id":"'+str(int(label))+'"}'
                else:
                    string_json += '{"id":"'+str(int(label))+'"},'

            string_json += '],"links":['

            #create distance element between each node pair
            for label_i in labels:
                for label_j in labels:

                    if (label_i < label_j):

                        index_i = int(numpy.where(labels==label_i)[0].item(0))
                        index_j = int(numpy.where(labels==label_j)[0].item(0))

                        val=datam.item((index_i,index_j))

                        max_val = datam.max()
                        min_val = datam.min()


                        #scale value so that distances are between 0 and 100

                        val = (val-min_val)/(max_val - min_val)
                        val = val*100


                        if label_i == labels[len(labels)-2] and label_j == labels[len(labels)-1]:
                                string_json += '{"source":'+str(index_i)+',"target":'+str(index_j)+',"value":'+str(val)+'}'
                        else:
```

```
                            string_json += '{"source":'+str(index_i)+',"target":'+str(index_j)+',"value":'+str(val)+'},'

            string_json += ']}'

            return string_json
```

A.2 PGR_script.job

```
#!/bin/bash
#$ -cwd
#$ -l h_vmem=128G

. /etc/profile.d/modules.sh

module add roslin/plink/1.90p
module add R/3.3.2
module add python/2.7.10

data_folder='/exports/eddie/scratch/s1685915/ibd'

###################
#LD pruning for PGR analyses
#uses output from the BB pre-processing
###################
#for CD + Controls
plink \
    --bfile ${data_folder}/results/TEAS.uncleaned.ichip.CD.Control.b37 \
    --indep-pairwise 200 50 0.25 \
    --allow-no-sex \
    --filter-controls \
    --out ${data_folder}/results/TEAS.uncleaned.ichip.PGR.CD.Control.QC

###################
#Filter given LD values
###################
#for CD + Controls
plink \
    --bfile ${data_folder}/results/TEAS.uncleaned.ichip.CD.Control.b37 \
    --extract ${data_folder}/results/TEAS.uncleaned.ichip.PGR.CD.Control.QC.prune.in \
    --allow-no-sex \
    --make-bed \
    --out ${data_folder}/results/TEAS.uncleaned.ichip.PGR.CD.Control.QC

###################
#Tranform OR values
###################
Rscript transform_or.R

###################
#Compute PGRs
###################
plink \
    --bfile ${data_folder}/results/TEAS.uncleaned.ichip.PGR.CD.Control.QC \
    --filter-controls \
    --score ${data_folder}/results/pgr_input_cd_1.txt 1 2 4 center \
```

```
    --extract results/pgr_input_cd_1.txt \
    --out ${data_folder}/results/PGR.Control_1.res

plink \
    --bfile ${data_folder}/results/TEAS.uncleaned.ichip.PGR.CD.Control.QC \
    --filter-cases \
    --score ${data_folder}/results/pgr_input_cd_1.txt 1 2 4 center \
    --extract results/pgr_input_cd_1.txt \
    --out ${data_folder}/results/PGR.CD_1.res

plink \
    --bfile ${data_folder}/results/TEAS.uncleaned.ichip.PGR.CD.Control.QC \
    --filter-controls \
    --score ${data_folder}/results/pgr_input_cd_2.txt 1 2 4 center \
    --extract results/pgr_input_cd_2.txt \
    --out ${data_folder}/results/PGR.Control_2.res

plink \
    --bfile ${data_folder}/results/TEAS.uncleaned.ichip.PGR.CD.Control.QC \
    --filter-cases \
    --score ${data_folder}/results/pgr_input_cd_2.txt 1 2 4 center \
    --extract results/pgr_input_cd_2.txt \
    --out ${data_folder}/results/PGR.CD_2.res
```

A.3 transform_or.R

```
library(data.table)

dat_gp1 <- fread("/exports/eddie/scratch/s1685915/ibd/results/bb_input_cd_1.txt")
dat_gp2 <- fread("/exports/eddie/scratch/s1685915/ibd/results/bb_input_cd_2.txt")

fwrite(dat_gp1[,V4:=log(V4)], "/exports/eddie/scratch/s1685915/ibd/results/pgr_input_cd_1.txt", sep="\t")
fwrite(dat_gp2[,V4:=log(V4)], "/exports/eddie/scratch/s1685915/ibd/results/pgr_input_cd_1.txt", sep="\t")
```