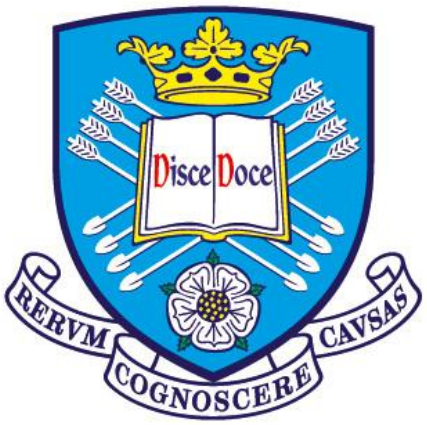


The University Of Sheffield.

Thesis: Unsupervised machine learning of high dimensional data for patient stratification

Sokratis Kariotis

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

September 2022

The University of Sheffield
Faculty of Medicine, Dentistry Health
Department of Infection, Immunity and Cardiovascular Disease
Supervisors: Dr. Dennis Wang, Dr. Allan Lawrie, Dr. Haiping Lu

“I know only one thing: that I know nothing.”
-Socrates

Declaration

This thesis integrates 6 manuscripts written as collaborative work of multiple authors. In two of them I am the primary contributor and listed as first author. Details on my manuscript contributions are noted at the beginning of each chapter. The introduction of this thesis contains parts of the first year literature review written by me. No part of this thesis has been submitted in support of an application for any degree or qualification at the University of Sheffield or any other University or institute of learning.

Sokratis Kariotis

September 2022

Acknowledgements

This dissertation has been the work of almost 4 years from its conception to the writing of the final paper and it included a lot of meetings, conferences and collaborations that would otherwise have been impossible. It was a very fruitful and full experience even if the latter half was completed through COVID years which made everything slower and more difficult for everyone. First of all, I would like to thank my primary supervisor Dr. Dennis Wang who invited me to start and guided me all along the way, mentally through endless meetings and discussions as well as physically offering me the amazing opportunity to study in the very interesting research environment of Singapore. I could not have done this without his help and his care about my wellbeing and academic advancement. I couldn't have asked for a more hands-on and supportive supervisor. I would also like to extend my gratitude to Dr. Allan Lawrie, my second supervisor, for the amazing support and the great amount of time and trust invested in me. He was there to always offer his help and knowledge to anything clinical I ever needed while working with cardiovascular diseases all these years. He was also an integral part of the majority of my research work. Finally, I would also like to thank my third supervisor, Dr. Haiping Lu for his deep expertise and advice on machine learning and computer science subjects.

I would also like to thank all my past and current lab mates for all the discussions, problem solving and the good times we had. The conferences, training, teaching, manuscript submissions were all great and educational experiences only made easier by all the people in my research group.

My sincere thanks to the institute of A*STAR for accepting me for a year's worth of research in their hospitable environment. I had a great year in this amazing institute and country and their hospitality made it very easy for me to live and work abroad. I would especially like to thank my A*STAR group, led by Dr. Dennis Wang. Every person in that group welcomed me and treated me as a member of the group offering any help and support I needed as a PhD student so far away from home and for that I am very grateful.

Most importantly, I would like to acknowledge my family's role not only in this PhD adventure but also anything that I have achieved so far. Simply put I could have not made it this far without them, their trust, help, support and love. My brother George never doubted me and somehow was always sure I'm going to make it. I still remember how nervous he was for me when I was defending my Msc thesis! He was there for anything I wanted to discuss and always had a positive attitude, even when I didn't seem to, a well-known phenomenon during PhD studies. I am very thankful to Dimitris as he was always there for me with solid advice, life views and generally knew how to keep me on track when I was doubting myself or my work. He has done and still doing a lot for me. I also want to thank my father Paschalis for his advice

and experienced opinions that provided me with a level of comfort about my life trajectory. Special thanks to my oldest friend Dimitris whose desire to develop and push forward consistently inspired me to do the same even when I didn't feel like it. The person I can't thank enough is my mother Maria. There is no way I can think of a way to thank her for the million things she has done for me since I was born. She moulded me into what I am today with a tireless, continuous effort that involves everything that had to do with my life. She continuously pushed me academically (sometimes a bit too much, like a traditional Greek mother) and provided me with all the love, emotional and material support I ever needed, always remaining in the background even when I was living, for years, in different countries and continents. She is an unstoppable force of love and support and for that reason I dedicate this PhD to her with love. Definitely couldn't be here without you!

Abstract

The development mechanisms of numerous complex, rare diseases are largely unknown to scientists partly due to their multifaceted heterogeneity. Stratifying patients is becoming a very important objective as we further research that inherent heterogeneity which can be utilised towards personalised medicine. However, considerable difficulties slow down accurate patient stratification mainly represented by outdated clinical criteria, weak associations or simple symptom categories. Fortunately, immense steps have been taken towards multiple omic data generation and utilisation aiming to produce new insights as in exploratory machine learning which showed the potential to identify the source of disease mechanisms from patient subgroups. This work describes the development of a modular clustering toolkit, named Omada, designed to assist researchers in exploring disease heterogeneity without extensive expertise in the machine learning field. Subsequently, it assesses Omada's capabilities and validity by testing the toolkit on multiple data modalities from pulmonary hypertension (PH) patients. I first demonstrate the toolkit's ability to create biologically meaningful subgroups based on whole blood RNA-seq data from H/IPAH patients in the manuscript "*Biological heterogeneity in idiopathic pulmonary arterial hypertension identified through unsupervised transcriptomic profiling of whole blood*". Our work on the manuscript titled "*Diagnostic miRNA signatures for treatable forms of pulmonary hypertension highlight challenges with clinical classification*" aimed to apply the same clustering approach on a PH microRNA dataset as a first step in forming microRNA diagnostic signatures by recognising the potential of microRNA expression in identifying diverse disease sub-populations irrespectively of pre-existing PH classes. The toolkit's effectiveness on metabolite data was also tested. Lastly, a longitudinal clustering approach was explored on activity readouts from wearables on COVID-19 patients as part of our manuscript "*Unsupervised machine learning identifies and associates trajectory patterns of COVID-19 symptoms and physical activity measured via a smart watch*". Two clusters of high and low activity trajectories were generated and associated with symptom classes showing a weak but interesting relationship between the two. In summary, this thesis is examining the potential of patient stratification based on several data types from patients that represent a new, unseen picture of disease mechanisms. The tools presented provide important indications of distinct patient groups and could generate the insights needed for further targeted research and clinical associations that can help towards understanding rare, complex diseases.

Table of contents

List of figures

List of tables

Abbreviations

PAH	Pulmonary arterial hypertension
PH	Pulmonary hypertension
IPAH	Idiopathic pulmonary arterial hypertension
ML	Machine learning
CVD	Cardiovascular diseases
miRNAs	microRNAs
mRNA	messengerRNA
CVDs	cardiovascular diseases
LC-MS	Liquid chromatography - mass spectrometry
NGS	Next Generation Sequencing
RT-qPCR	Reverse transcription quantitative real-time polymerase chain reaction
cDNA	Complementary DNA
CTEPH	Chronic Thromboembolic Pulmonary Hypertension
CTED	Chronic Thromboembolic Pulmonary Vascular Disease

Chapter 1 - Introduction

High-throughput biological experiments are currently generating unprecedented volumes of data which are handled by various types of bioinformatics methods. This new field has to continuously adapt to answer questions based on new data and emerging biological hypotheses. Personalised medicine is one of the forefront concepts that can revolutionise healthcare but relevant research comes with many challenges. This work is focusing on exploring heterogeneity through the power of unsupervised learning as a way to overcome such challenges and make personalised medicine more approachable. It focuses on creating a methodology to assist researchers bypass the complexity of unsupervised learning and extract valuable information from their data. As a complex heterogeneous disease, PH posed as an ideal example to investigate and simultaneously test the aforementioned methodology. The various chapters of this work and their relation to current challenges are shown in **Figure 1**.

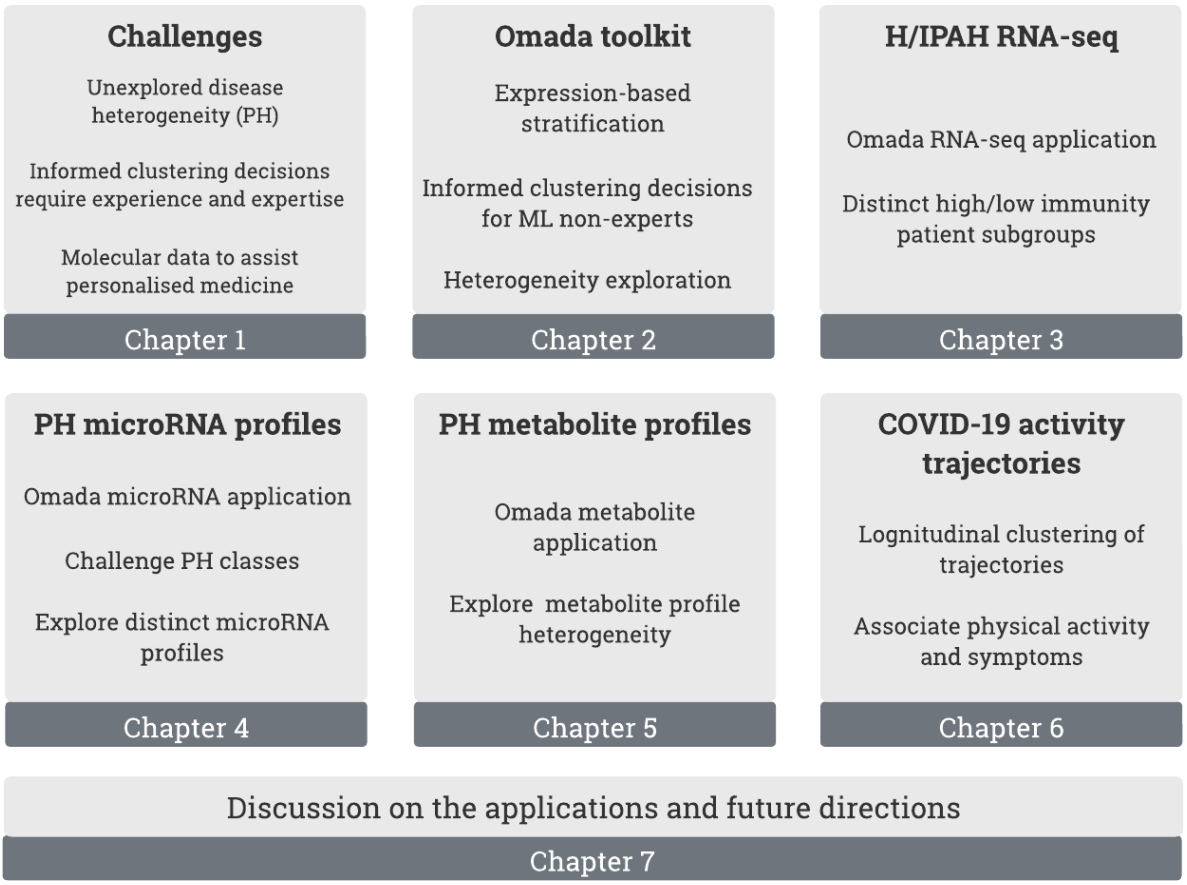


***Figure 1**: Current challenges and the relevant work presented in each chapter*

1.1 Need for patient stratification

Advances in biological and clinical data generation technologies (van Dijk *et al.*, 2018; Ameur, Kloosterman and Hestand, 2019) are enabling researchers to study human physiology at a deeper level (Shashi *et al.*, 2014) by utilising genetic insights and recent methodologies. The higher quality and reduced cost of genetic data increase its granularity and enhance the predictive and discriminative power of existing data-driven models (Pezoulas *et al.*, 2021) generating more accurate models. Previously hidden disease mechanisms could be uncovered and provide answers to important medical questions such as early diagnosis (Choi *et al.*, 2017; Sharma *et al.*, 2021), prognosis (Moura *et al.*, 2016; Westeneng *et al.*, 2018) and development (Pinto *et al.*, 2020). That would allow the allocation of patients in sub-disease cohorts allowing specialised disease treatment plans, the ultimate goal of personalised medicine (Schork, 2019; Peng *et al.*, 2021). This makes patient stratification a prominent goal for future medical research since we keep uncovering the complexity of rare diseases in studies that have discovered subgroups of patients with distinct behaviours (Esteva *et al.*, 2017; Kariotis *et al.*, 2021), phenotypes (Gurovich *et al.*, 2019) and treatment (Hurvitz *et al.*, 2021). However, such distinctions are occasionally based on weak criteria, symptom similarities or even diagnosis of exclusion due to lack of data, as for example in idiopathic pulmonary hypertension (Montani and Simonneau, 2012). The erroneous stratification (or complete lack of it) can occur due to disease heterogeneity which can be caused or affected by multiple sources (Montani and Simonneau, 2012) such as genetic irregularities (Zanwar and Kumar, 2021) or physiological changes. The multimodality of the data needed to answer these questions requires the computational power of machine learning methods to detect signals under enormous amounts of relevant measurements (Schaefer *et al.*, 2020). The latter can be used to explore deeper into molecular mechanisms to uncover the complex structure of diseases that manifests disease heterogeneities.

1.1.1 Disease heterogeneity

A heterogeneous disease is a medical condition which can have multiple etiologies/causes (Cho and Feldman, 2015) or involves different mechanisms that produce similar phenotypes (Manchia *et al.*, 2013). Heterogeneity can be found in several disease types, namely infectious, hereditary (genetic and non-genetic) and physiological diseases and conditions. Its presence can severely increase research complexity and cause serious inaccuracies in phenotype/genotype estimations (Manchia *et al.*, 2013).

Cancer is a well studied disease characterised by significant heterogeneity even within its subtypes. For example, a recent study on prostate tumour heterogeneity (Haffner *et al.*, 2021) discussed both genetic and phenotypic heterogeneity, enhancing previous studies that identified similar results (Haffner *et al.*, 2013; Gundem *et al.*, 2015; Hong *et al.*, 2015). In the subject of breast cancer, (Turashvili and Brogi, 2017) states the inability to classify cancer cases in a clinical context due to intertumor (breast carcinomas from different individuals) and intratumor (presence of heterogeneous cell populations within an individual tumour (Ellsworth *et al.*, 2017)) heterogeneity. Cancer therapeutics and biomarker discovery is also hindered by this heterogeneity (Fisher, Pusztai and Swanton, 2013). Current machine learning methodologies have investigated these distinctions within cancer cases as described in (Laurinavicius *et al.*, 2021; Lee, Park and Kim, 2021) where authors highlight their value in clinical guidance, cancer biology and therapeutics.

Cardiovascular diseases (CVD) entail conditions which affect the heart or blood vessels (Papageorgiou, 2016). Most such diseases are characterised as complex due to the coaction of environmental and genetic factors in their origins and development (Papageorgiou, 2016; Ehret, 2018; Musunuru and Kathiresan, 2019). Research estimates that up to 90% of CVD might be preventable (McGill, McMahan and Gidding, 2008; O'Donnell *et al.*, 2016) highlighting the need for accurate identification of disease subtypes. Metabolomics studies indicated heterogeneity in functional phenotypes in high cardiovascular risk individuals (Mao *et al.*, 2019) while coronary heart disease research showed not only inter-person heterogeneity, driven by risk factors, but also unobserved heterogeneity in individuals with identical risk (Simonetto *et al.*, 2022). This heterogeneity type induced biases during risk factor identification and interpretation of relevant results (Aalen *et al.*, 2015; Balan and Putter, 2020). Risk factors have also shown heterogeneity in subpopulations (Rivera-Andrade and Luna, 2014; Koirala *et al.*, 2021). Current research also showed clinical (Hernandez-Gonzalez *et al.*, 2020) and biological heterogeneity in pulmonary arterial hypertension, a rare complex disease group, identified by variants and genetic expression, respectively.

Infectious diseases compose another heterogeneous category of disorders where the causes consist of various organisms such as bacteria and viruses. As in the previous types, heterogeneity is multifaceted and includes transmission ways or disease progression. More specifically, the ability of hosts to transmit can change the disease dynamics based on multiple factors such as infectiousness, contact rate and infection duration. These mechanisms are described in (VanderWaal and Ezenwa, 2016) where the authors stress the importance of studying pathogen transmission heterogeneity considering the above factors. Another source of heterogeneity was shown to be the causing organism, host susceptibility as well as environmental factors which can influence the risk of infection, as demonstrated for tuberculosis (Trauer, Dodd and Gomes, no date) and malaria (Feachem *et al.*, 2010). Infectious disease control mechanisms are also more complex due to heterogeneity

(Woolhouse *et al.*, 1997). As it became prevalent in recent years, COVID-19 has been the focus of multiple studies whtargeting its inherent heterogeneity. Disease burden (Chen and Assefa, 2021; Vallée, 2022) was the main factor of heterogeneity expression leading to large differences between high and low burden countries with mortality (García-Guerrero and Beltrán-Sánchez, 2021) also contributing. Testing methodologies have also been affected (Berrig, Andreasen and Frost Nielsen, 2022). Very importantly, it was shown that COVID-19 lifestyle (Nikolaidis *et al.*, 2022) and symptoms were highly heterogeneous, ranging from mild to acute across individuals and over time (Rodebaugh *et al.*, 2021).

1.1.2 Diagnostic tests for molecular subtype classification based on molecular heterogeneity

Molecular heterogeneity in cancer is expressed through multiple mechanisms with possible genetic (Burrell *et al.*, 2013), non-genetic and epigenetic origins, including tumour mutational burden and somatic mutations, genomic instability and mutant allele imbalance as well as chromosomal aberrations (Zito Marino *et al.*, 2019). Additionally, tumours can show differences in growth rates, cell surface markers and resistance to therapy (Kreso and Dick, 2014).

In (Jamal-Hanjani *et al.*, 2015; Rich, 2016) authors described this heterogeneity to manifest in two main ways depending on the tumour type. The variations observed between different tumours (different patients, tissues and/or cell types) define the intertumoral heterogeneity, associated with differences in genetic profiles, protein signatures or marker expression, potentially linked to variable treatment responses (Gerdes *et al.*, 2014). The subclonal differentiations within a single tumour are referred to as intratumour heterogeneity which can increase the complexity of cancer prognosis and treatment (Michor and Polyak, 2010).

This heterogeneity introduces a vast amount of complexity in cancer research but can also be utilised to differentiate tumours or patients and create molecular profiles with immense value in prognosis and treatment. To facilitate such studies, machine learning approaches were developed to classify molecular subtypes. In bladder cancer, consensus molecular classification was performed using the results of several classifiers on multiple datasets and cohorts to define six molecular subtypes (Kamoun *et al.*, 2020). On the unsupervised side, (Robertson *et al.*, 2017) used clustering on multiple data types (mRNA, lncRNA, miRNA) to identify 5 expression subtypes with diverse epithelial-mesenchymal transition status, carcinoma-in-situ scores, histologic features and survival. In breast cancer, initial gene expression profiling studies (starting with microarray data) showed distinctions on a transcriptomic level and the potential of additional molecular subtypes (Reis-Filho *et al.*, 2010). In recent years, molecular (Jansen *et al.*, 2005) classifiers and prognostic

multigene classification systems (Loi *et al.*, 2008) have been developed for prediction purposes.

State-of-the-art diagnostic tests for molecular subtype classification, such as Decipher (*Genomic profiling for prostate and bladder cancers*, 2019), are now being tested and validated so they can be used to reveal the underlying biology of tumours. When used alongside clinical information they can greatly assist with personalising treatment and managing disease development. Apalutamide is a treatment for localised prostate cancer and in a recent study (Feng *et al.*, 2021) such a classifier associated molecular subtypes to outcome differences when patients were treated with apalutamide. For another type of prostate cancer, a genomic classifier identified subtypes that were associated with prognosis and could help identify patient candidates for chemohormonal therapy (Hamid *et al.*, 2021). These recent advances highlight the importance of molecular profiling tools due to their ability to harvest crucial heterogeneity information towards more accurate and targeted clinical decisions.

1.1.3 Pulmonary arterial hypertension

Currently, the biomedical field generates enormous amounts of diverse information measuring different aspects of human health. Despite the numerous high-throughput methods (Qiang-long *et al.*, 2014) providing data, rare diseases such as pulmonary hypertension (PH), often require special efforts to tackle relevant problems (limited samples, inconsistent/variable definitions, scattered, unstructured data and inherent heterogeneity) in the race to help diagnose (Vachiéry and Gaine, 2012), treat (Elinoff *et al.*, 2018) or prevent the disease.

As seen in **Figure 2**, an understudied form of PH is pulmonary arterial hypertension (PAH), a subgroup of PH which consists of disorders classified by similar pathological findings, hemodynamic descriptions and disease management approaches (Simonneau *et al.*, 2019). Medically in PAH, the increase in pulmonary artery pressure is driven by progressive pulmonary vascular remodelling. The latter consists of sustained vasoconstriction and dysregulated cell growth (Lan *et al.*, 2018). PAH itself is subdivided in the following categories: Heritable, relating the disease with inherited gene mutations e.g. BPMR2 (Morrell *et al.*, 2019) and recently a larger set of rare mutations (Gräf *et al.*, 2018), drug and toxic induced e.g. SSRI during pregnancy, associated with a second disease such as Connective tissue disease (Kieler *et al.*, 2012), and the less rigidly described idiopathic form, IPAH. The latter category has mostly unknown causes and describes a heterogeneous group of conditions defined by a diagnosis of exclusion (Firth, Mandel and Yuan, 2010). This leads to a heterogeneous population of patients and a difficulty to define how IPAH is structured. Due to the disease complexity and its potential genetic causes (Morrell *et al.*, 2019), unsupervised machine learning based on genomic data, is often utilised to

explore gene expression (the most fundamental level at which the genotype connects to the phenotype) of such diseases, as in (Jiang *et al.*, 2016) where different rare cell types were detected. Thus, gene expression profiles of patients may possess the potential to distinguish subtypes of IPAH.

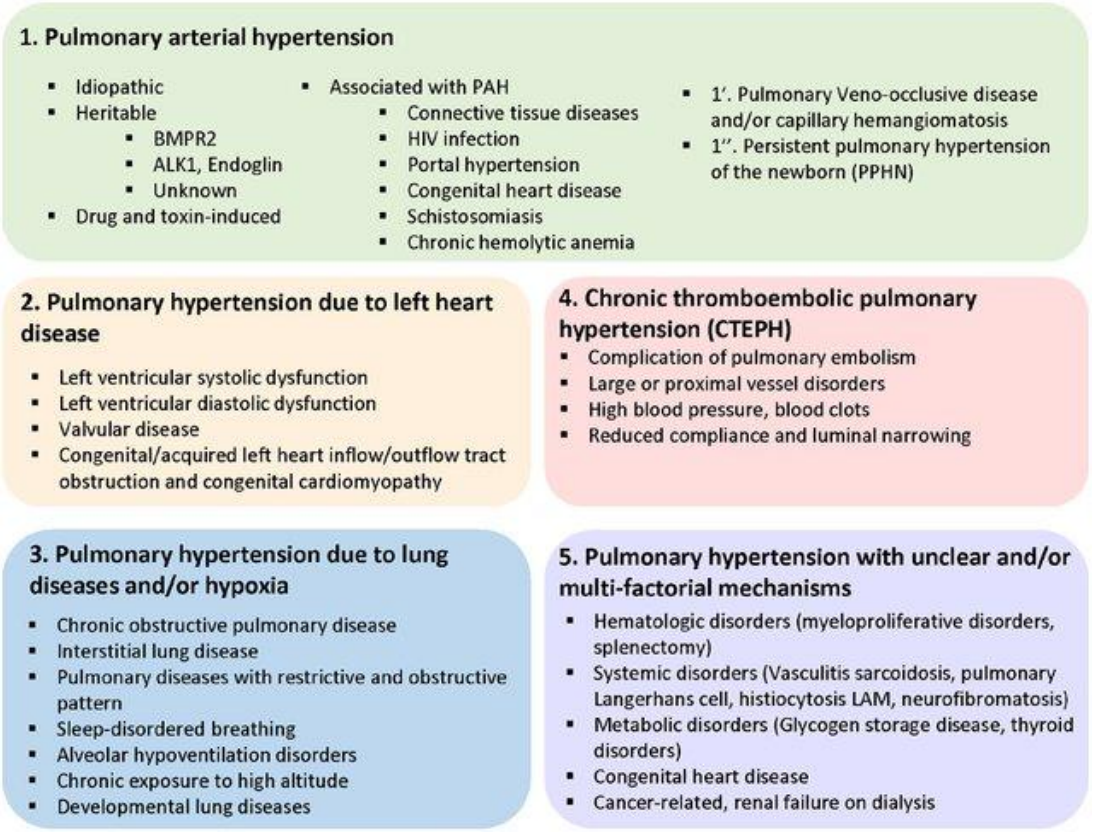


***Figure 2**: Known classification of PH and PAH and the lack of IPAH identified subgroups (Bisserier et al., 2020)*

1.2 Research limitations

However, such rare disease studies are hampered by small patient pools. Therefore, mainstream methods that work well for other diseases are not easily transferable to PAH. Moreover, several data types are prone to biases (Zheng, Chung and Zhao, 2011) further clouding the potential genetic signal underlying patient variations. On a higher level, the molecular complexity allows a specific data type to only provide insights about a single aspect of a certain problematic condition. However, the mechanisms within a health condition context are very often affected by several factors. Research usually focuses on one factor in an effort to reduce complexity or make a specific diagnostic method viable but recently multi-omics studies have started to appear (Chung and Kang, 2019) concerning complex diseases.

As an additional hindering factor, PAH/IPAH sub-structure has not been described yet, which makes the task of assigning patient treatment very hard (Bazan and Fares, 2015) and introduces an element of uncertainty concerning the correct response to a set of symptoms, as demonstrated by (Montani *et al.*, 2010) where different PAH subtypes react very differently to acute vasodilator testing. Although PAH's structure is not entirely hidden (Thenappan *et al.*, 2018), IPAH's heterogeneity may be used to identify patient subgroups that are driven by similar mechanisms thus allowing novel, more precise targeted treatment.

1.3 Molecular data types for clustering

Rapid advances in data generation technologies have fostered the development of large diverse data sets, known as Big Data, to be used in the study and exploration of disease and condition contexts (Prasad *et al.*, 2021). Multiple types of data (Schatz, 2015) such as genomic sequencing, proteomics, non-coding RNA, epigenetics, physical activity and metabolomics can be applied to answer different research questions either individually or in combinations to identify patterns and explore sample differences (Bayat, 2002). Aside from exploration, the development of cutting edge methodologies aims to understand these enormous datasets and find ways to reduce costs that relate to data computation and management or effective and economical treatment and general currently costly or even unfeasible healthcare applications.

As made apparent, these complex molecular datasets have the potential to answer critical questions in healthcare but they don't come without problems (Marx, 2013). Biological data always represent one aspect of a very complicated system that cannot be captured in its entirety in a single experiment. More often than not these datasets only contain a part of the hidden biological signal making its detection even harder and in some cases the challenging (Davidson, Overton and Buneman, 1995; Gligorijević and Pržulj, 2015) integration of data types necessary. Furthermore, the high-dimensionality of the data, despite the opportunities it provides (Quackenbush, 2007), poses an obstacle to most methods used to decryptify noisy biological data as a large amount of useless or even misleading information can drive methodologies (Clarke *et al.*, 2008). Researchers need to exclude as much of this data as possible in order to extract meaningful insights from the context they explore. Additionally, the appropriate methods need to be utilised for maximising the amount of extracted vital information as well as the interpretability of any results. The generation of methods and their application to the most common types of molecular data are described below.

1.3.1 RNA sequencing data

The quantitative analysis of the expression of genes is the most direct way to study genome regulation. Gene expression was initially measured through microarrays (Lamot *et al.*, 2015) with a few thousands of genes screened at a time which resulted in large expression datasets. However, microarray technology incurred high experiment costs and sensitivity, low specificity probes and manufacturer technical biases impacting the actual expression estimation (Jaksik *et al.*, 2015). Transitioning from arrays to deep sequencing, RNA-sequencing became the main tool for transcriptome profiling (Wang, Gerstein and Snyder, 2009) providing more precise and affordable measurements illuminating the complexity of disease (Costa *et al.*, 2013).

Contrary to previous methodologies, high-throughput next-generation sequencing (NGS) is directly determining the transcript sequence, by sequencing complementary DNA (cDNA) (Wang, Gerstein and Snyder, 2009), instead of targeting specific genomic regions and limiting their outputs. As demonstrated in **Figure 3,** RNA sequencing starts from total RNA extraction, followed by isolating and filtering certain RNA types. The remaining material is then converted to cDNA which is used to construct a sequencing library to be amplified by polymerase chain reaction (PRC). It should be noted that multiple protocols for library creation exist and they detect specific transcripts that might be of interest in individual studies. A detailed table of such protocols can be found in (Kukurba and Montgomery, 2015).

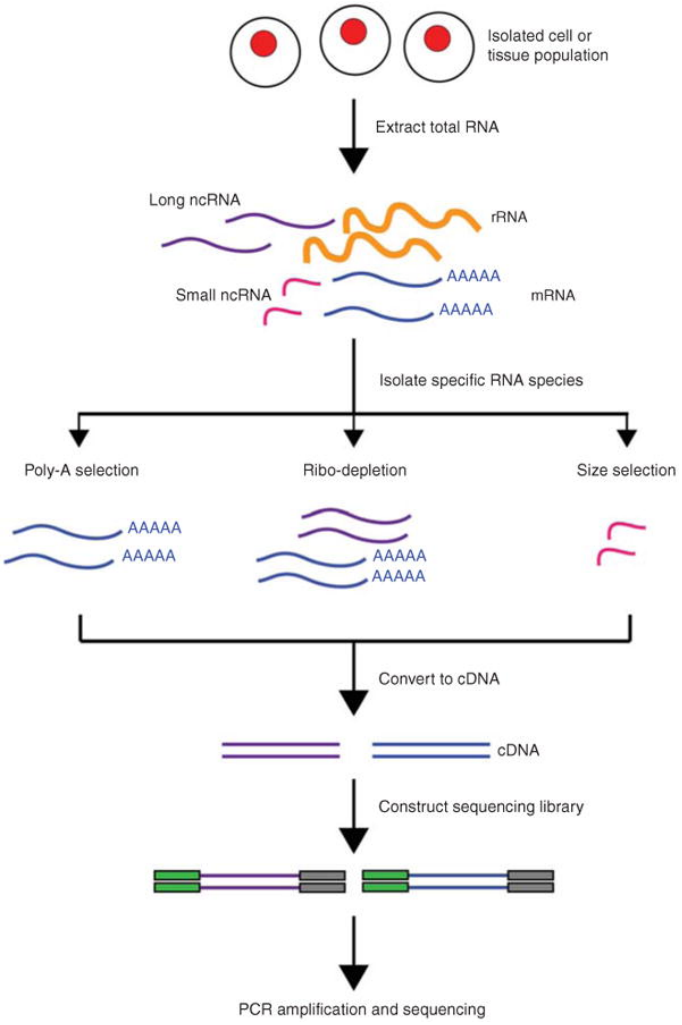


Figure 3: *"Overview of RNA-Seq. First, RNA is extracted from the biological material of choice (e.g., cells, tissues). Second, subsets of RNA molecules are isolated using a specific protocol, such as the poly-A selection protocol to enrich for polyadenylated transcripts or a ribo-depletion protocol to remove ribosomal RNAs. Next, the RNA is converted to complementary DNA (cDNA) by reverse transcription and sequencing adaptors are ligated to the ends of the cDNA fragments. Following amplification by PCR, the RNA-Seq library is ready for sequencing.", taken from (Kukurba and Montgomery, 2015).*

The amplified RNA-seq libraries (reads) created by such experiments need a few additional processing steps to be used in downstream analysis which requires a quantified representation of the expression landscape. Initially, sequencing reads need to be aligned to the reference genome of the organism of interest (Howe *et al.*, 2013). Such genomes are pre-assembled and updated to be as accurate as possible in order for all further analyses to be based on a realistic representation of the nucleic acid sequence. Reads are then assembled into transcripts either using reference transcript annotations (identified functional elements of the genome sequence) or in cases where a reference is not available, i.e. when assembling small bacteria genomes, de novo sequence assembly methods (Steinegger, Mirdita and Söding, 2019). Finally, the desired expression of the included gene is estimated by counting

the number of reads that aligned to each annotated region. For traditional sequencing that aims to estimate only the function of protein coding genes, only reads that align to exons and full length transcripts are considered. In other cases, like microRNAs, different non-coding regions are used.

Genomic sequencing data are being analysed in the field of bioinformatics for diverse research purposes (Corchete *et al.*, 2020). Such data provide measurements about DNA as well as sequence variation across individuals used in genotype-phenotype association studies in PAH (Lane *et al.*, 2000; Gräf *et al.*, 2018). Lately, imaging data (e.g. magnetic resonance imaging) utilised machine learning methods to predict outcomes of PAH patients (Dawes *et al.*, 2017). Current research is focusing on high-throughput RNA sequencing for disease profiling (Lightbody *et al.*, 2019) and tools assessing its quality (Planet *et al.*, 2012; Thrash, Arick and Peterson, 2018). Its product, relative expression of genes, is widely used to uncover gene behaviour in contexts of interest and was recently used in induced PAH in mice where they marked CREB as the master transcription factor in the pathogenesis of PAH (Xiao, Xie and Lian, 2018). Furthermore, these interactions can form gene/pathway networks, e.g. based on co-expression (Pita-Juárez *et al.*, 2018), revealing a higher-level organisation of a cell. Naturally, RNA-seq data from different tissues and organs show a different regulatory picture which allows the study of specific contexts within the human genome. However, in some cases tissue samples are too invasive to acquire and researchers have to compromise. For example, the lack of lung tissue biopsy studies use whole blood RNA measurements as a surrogate, as demonstrated in (Blankley *et al.*, 2014).

1.3.2 microRNA data

As mentioned, multiple forms of RNA sequencing can be performed to generate data to help answer different questions. A special form, microRNAs(miRNAs), are found in various tissues and blood plasma, are composed of a short sequence (averaging 22 nucleotides) and contribute to gene regulation by binding to messenger RNA (mRNA) repressing protein production or causing post-translational silencing (Cannell, Kong and Bushell, 2008; O'Brien *et al.*, 2018). miRNAs have only recently started to be explored because their small length requires a generation method with the extreme precision only offered by current next-generation sequencing technologies. Generating miRNAs is part of the high-throughput next-generation sequencing and requires specific libraries (Lu, Meyers and Green, 2007) that target and amplify miRNA regions as well as annotation databases (Kozomara and Griffiths-Jones, 2014).

Reverse transcription quantitative real-time polymerase chain reaction (RT-qPCR) is one of the prominent technologies through which we can quantitate mrRNAs (Adams, 2020).

Quantitative PCR, whether involving a reverse transcription step or not, is routinely used in molecular biology labs and has revolutionised the way in which research is carried out due to its relatively simple pipeline (**Figure 4**). Its advantages over standard PCR include the ability to visualise which reactions have worked in real time and without the need for an agarose gel. It also allows truly quantitative analysis. One of the most common uses of qPCR is determining the copy number of a DNA sequence of interest. Using absolute quantitation, the user is able to determine the target copy numbers in reference to a standard curve of defined concentration in a far more accurate way than ever before. RT-qPCR, on the other hand, allows the investigation of gene expression changes upon treatment of model systems with inhibitors, stimulants, small interfering RNAs (siRNAs) or knockout models, etc. This technique is also routinely used to detect changes in expression both prior to (as quality control) and after (confirmation of change) RNA-Seq experiments.

RT-qPCR combines two workflows, reverse transcription PCR (RT-PCR) and quantitative real time PCR (qPCR, **Figure 4B**) to allow measuring of RNA levels using cDNA in a qPCR to rapidly detect expression changes. More specifically, during the RT-PCR step RNA from the sample is isolated to then be used to generate cDNA with reverse transcriptase (**Figure 4A**). Subsequently, PCR is used to amplify specific regions. In qPCR (**Figure 4B**), DNA is isolated and amplified instead, with fluorescent probes used to quantitate the PCR product. This quantification method is often used to detect pathogen presence and measure copy number of certain DNA sequences. As a combination (**Figure 4C**), RT-qPCR isolates RNA followed by the generation of cDNA before the aforementioned qPCR quantification facilitating rapid quantification of changes on expression.

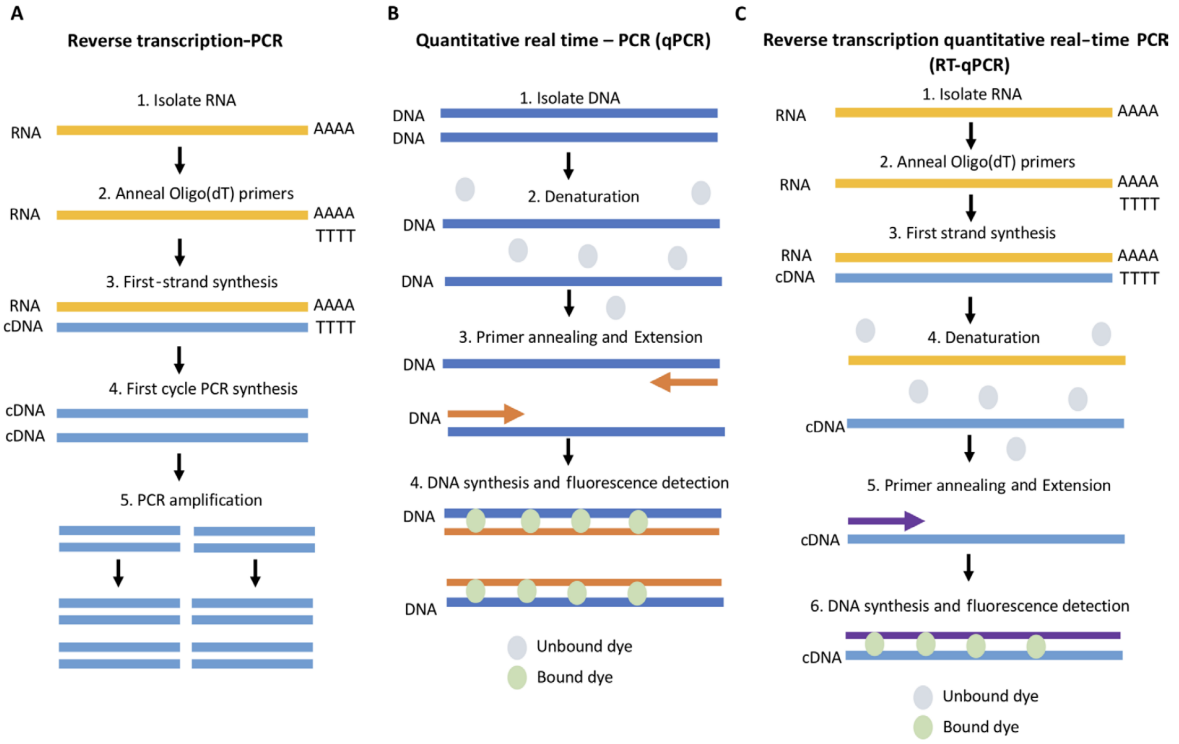


***Figure 4**: "(A) RT-PCR workflow. RNA is isolated and cDNA is generated via reverse transcription (RT); PCR is then carried out to amplify areas of interest. (B) qPCR schematic. DNA is isolated and amplified; amplification is quantitated using a probe which fluoresces upon intercalation with double-stranded DNA. (C) RTqPCR procedure. RNA is isolated and cDNA generated before commencing a qPCR procedure.", taken from (Adams, 2020).*

Since research started looking into non-coding regions and their importance in gene regulation (Qureshi *et al.*, 2014; Statello *et al.*, 2021) sequencing protocols focused on that species of small RNA (miRNAs). With gene regulation being a very important function, miRNAs have been found to play multiple integral roles including cell development, differentiation, division etc (Cimmino *et al.*, 2005; Xiao and Rajewsky, 2009). That being the case, the bioinformatics field has been creating tools to improve the quality of miRNAs (Potla, Ali and Kapoor, 2021) as well as methods and pipelines to take advantage of the information hidden in these accurately identified small genome regions. Their clinical application in infectious diseases (Drury, O'Connor and Pollard, 2017) and their potential on a variety of human disease groups (Li and Kowdley, 2012) has been discussed in recent years. More specifically, in the context of cardiovascular diseases (CVDs) (Li *et al.*, 2017) miRNAs were found to be highly expressed and synergistically work to play an important role in heart tissues. Specific miRNAs were shown to be involved in cardiac hypertrophy by directly affecting cell cycle related genes (Carè *et al.*, 2007). In PAH a set of miRNAs was noted as potential biomarkers (Errington *et al.*, 2021) with (Santos-Ferreira *et al.*, 2020) highlighting the potential for further analysis and the need to translate miRNA research to treatment implementations.

1.3.3 Metabolomic data

Metabolomics describe the study of metabolites, small molecules found in cells, biological fluids or tissues. Metabolites, being substrates and products of metabolism, can directly represent or measure biochemical activity and reflect a molecular phenotype (Gieger *et al.*, 2008).

A variety of, often complementary, methods have been developed to extract, detect and quantify the activity of metabolites (Roessner and Beckles, 2009) due to the enormous diversity of existing chemical structures and abundance variations. In most cases the metabolites are divided into subsets and preparation and analytical methods are applied individually depending on the subsets characteristics (functional groups, structural similarity etc) as demonstrated in **Figure 5**. This method heterogeneity can create problems in integrating metabolome results and as a consequence multiple methods have been developed to report (Sumner *et al.*, 2007) and test (Martin *et al.*, 2015) compatibility.

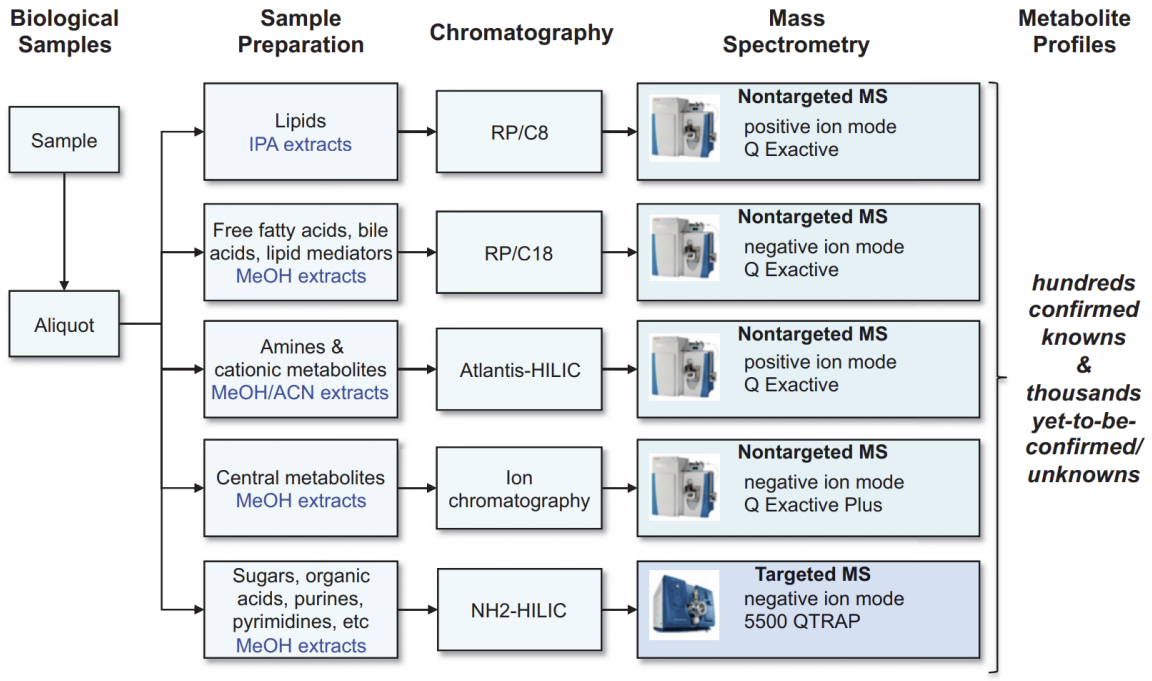


***Figure 5**: "Illustration of the liquid chromatography–mass spectrometry (LC-MS)-based metabolomics platform used at the Broad Institute of MIT and Harvard. A comprehensive metabolomics platform that uses targeted and nontargeted LC-MS methods to measure lipids, metabolites of intermediate polarity such as free fatty acids and bile acids, and polar metabolites. IPA, isopropanol; RP, reversed phase; HILIC, hydrophilic interaction liquid chromatography." as seen in (Clish, 2015)*

Many methodologies and tools have been developed to make metabolite analysis more efficient and assist in biological interpretation (Lamichhane *et al.*, 2018). As a result, metabolites have become increasingly popular in researching metabolism related conditions and diseases. More specifically, a metabolome profiling method showed potential novel biomarkers pointing to inborn metabolism error in patients (Coene *et al.*, 2018). Similarly, microbiome sequencing and metabolome data were utilised by a machine learning model to identify metabolic profiles (Yin *et al.*, 2020). In another integration recent study, COVID-19 specific genome-scale metabolic models were analysed and showed differences in cholesterol metabolism regulation and metabolic pathways related to host response and are potential antiviral targets (Režen *et al.*, 2022). More metabolic pathways of Natural Product Metabolism were explored in Medicinal Plants when metabolic and NGS data were integrated (Scossa *et al.*, 2018). Finally, the current and future utility in precision medicine was described in (Clish, 2015).

1.3.4 Physical activity data

Initially, physical activity was recorded through questionnaires (InterAct Consortium *et al.*, 2012) but since the development of wearable technologies physical activity measurements have been of great study interest (Füzéki, Engeroff and Banzer, 2017). It has been shown that exercise affects multiple systems in the human body at a molecular level (Hawley *et al.*, 2014) but the majority of them are still unexplored (Neufer *et al.*, 2015). Wearable monitors are now able to provide a number of measurements or physiological markers such as heart-rate, energy burned, step count and ventilation (Freedson *et al.*, 2012). These data have been complemented by environment variables (such as education, ethnicity etc) and used to investigate the effects of activity (or inactivity) in various populations (Atkin *et al.*, 2017). Regression models have been used with physical activity data like energy expenditure to assess mortality from various sources (Mok *et al.*, 2019). Accelerometer measurements were used to classify several rehabilitation categories (Skovbjerg, Honoré and Mechlenburg, 2022) and wearable data were used to passively assess COVID through machine learning (Sarwar and Agu, 2021).

1.4 Machine learning, a field of exploration and prediction

Machine learning (ML) is an area of computer science aiming to discover patterns within volumes of data. This field creates methods which require the use of algorithms and statistics to understand patterns and either differentiate data groups or predict where new data points belong. The main characteristic of such methods is their ability to learn from input data and in principle refine their results as more data become available. The explosion of ML popularity and utility is based on three trends

(Fradkov, 2020). First, the ever increasing amount of data generated in practically every field where variables can be measured allowed ample inputs for ML algorithms to train and improve. Secondly, as technology improves the cost of (parallel) computing is being reduced while available memory dramatically increases the limits and computational power of statistics. Finally, the above improvements allowed the development of new ML algorithms, such as deep learning and neural networks, which are harnessing the computational benefits and big data information to tackle the prediction and stratification problems from new angles.

1.4.1 Unsupervised learning / clustering

For exploratory projects, such as discovering underlying disease branches and differentiating features, unsupervised methods are best suited, as demonstrated in (Kallenberg *et al.*, 2016) where they are able to extract a feature hierarchy related to mammographic risks. Commonly known as clustering, grouping objects based on their inherent similarities (Rokach and Maimon, 2005), has been used extensively in many fields. In image recognition, clustering is used to discover the distribution of objects in a feature space (Kheradpisheh, Ganjtabesh and Masquelier, 2016). Robotics use clustering towards object detection (Zapf *et al.*, 2018) while Recommender Systems build user groups and improve recommendation quality (Das *et al.*, 2014). In medicine, it has been used in a wide variety of subjects, from medical diagnosis (Wu, Duan and Du, 2015) to analysing drug-drug interaction networks (Udrescu *et al.*, 2016). The advantage that sets clustering methods apart from other machine learning approaches is their ability to function without any prior knowledge about the objects' class. In terms of disease biology this means no prior literature of clinical classifications.

The variety of clustering approaches, as depicted in **Figure 6**, often makes the identification of the appropriate algorithm a difficult task. Algorithm efficiency can be affected by several factors, such as the type, distribution and volume of data, the expected nature of similarities between samples and each algorithm's specific strengths and weaknesses.

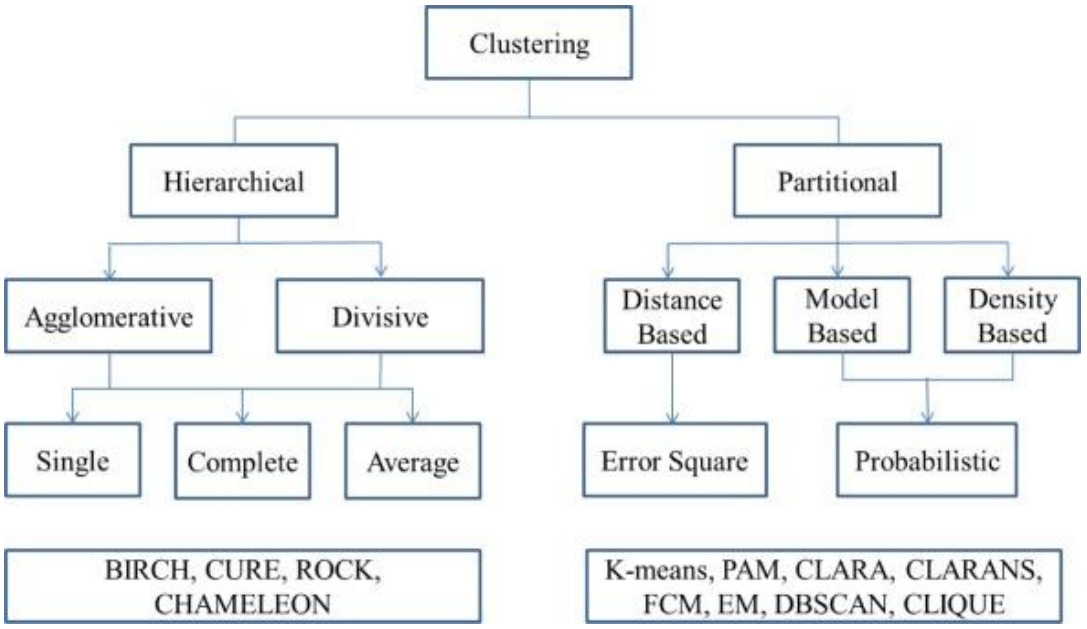


***Figure 6**: Taxonomy of clustering approaches (Saxena et al., 2017)*

Certain clustering approaches (that encapsulate more specific methods) are preferred in medical research and bioinformatics due to the complex nature of medical data. Hierarchical clustering connects samples to form clusters based on their distance, while density-based clustering defines clusters as areas of higher density compared to the remainder of the data set. Probability distribution-based clustering is based on distribution models where clusters can be defined as objects that most likely belong to the same distribution. Applications of the above include building regulatory networks (Fiers *et al.*, 2018), discovering disease subtypes (Keenan *et al.*, 2018), inferring gene function/coexpression (Chen *et al.*, 2008) and reducing dimensionality (Lee *et al.*, 2018). Currently, spectral clustering is a popular technique that performs feature reduction based on the first eigenvectors, before clustering (Jia *et al.*, 2014). Due to its implementation simplicity and promising performance this technique has been used in bioinformatics, e.g. to cluster cells utilising single-cell RNA-sequencing (Park and Zhao, 2018) as well as different fields, e.g. for speech recognition (Li *et al.*, 2018).

1.4.2 Clustering analysis challenges

As noted, RNA-seq potentially possesses the power to discriminate between patients and healthy individuals by revealing gene activity. This potential can be harvested by clustering gene expression, a method potentially capable of deciphering the features that characterise a health condition or a clinical phenotype (Burgel, Paillasseur and Roche, 2014). Clustering can be very effective in PAH, as it does not require prior knowledge of the branches in which each patient belongs, allowing independent

clusters of patients and controls to be formed within this context of heterogeneity. Such clusters are further analysed either using independent clinical data or differentially expressed genes. These analyses enable the association of such features with a specific cluster of samples as well as cross-cluster comparisons. Ideally, a set of significant associations can point to biologically meaningful phenotypes.

However, clustering work can be complex even before the application of an unsupervised approach (Hennig *et al.*, 2015). The inherent grouping of the dataset has to be assessed before any algorithm is applied. Sample sets with no grouping potential (i.e. healthy volunteer samples with no underlying condition of interest) do not offer an interesting dataset for clustering as they will most likely base the sample partitioning on variables of no interest. In most cases the validity of clusters in a dataset is tested after the clustering analysis and can be also referred to as assessing the number of clusters. Most methods have to assume the number of clusters before any calculation therefore this initial choice greatly affects any outputs. Since in most real datasets there isn't a priori knowledge about existing clusters an estimation is needed by the user. A number of indexes (such as the silhouette score (Shahapure and Nicholas, 2020)) have been created to measure how far apart the clusters are and how compactly their samples form indicating the number of clusters. However, each index only provides an indication of quality from a different perspective and one must be careful not to present misleading pictures of cluster quality. Another important matter is the dimensionality of a dataset (the number of samples and features) as it provides the statistical power in algorithm formula calculations. Large dimensional imbalances can cause skewed calculations and obscure otherwise detectable clusters i.e. too many patients of a specific condition may cause the algorithm to ignore a much smaller set of very different patients due to their low count. The total size of the dataset can also be a challenge. Very large datasets, known as Big Data, can cause computational problems and even disable certain algorithms. It is well observed that Big Data need special handling in clustering projects (Zhou and Wang, 2016; Saeed, Al Aghbari and Alsharidah, 2020) such as data preprocessing and feature selection (Devi, Gayathri Devi and Sabrigiriraj, 2018), the latter being very important for gene expression data where we deal with thousands of genes. Selecting the most effective clustering methodology is also a very important challenge for this type of analysis as results may greatly vary. Certain data types and research questions best fit the formulas of certain algorithms (Hennig *et al.*, 2015).

1.5 Hypothesis and aims

The central hypothesis of this thesis is that a semi-automated machine learning toolkit applied to high-dimensional readouts from patients can provide a robust and effective way of partitioning diseases. I aimed to create and test this approach on a variety of disease contexts and data profiled from patients with the following aims:

A. **Create a modular toolkit that helps users without extensive machine learning experience to perform an in-depth unsupervised analysis of RNA sequencing data (Chapter 2).** Most current research that aims to identify heterogeneity within expression datasets only utilises a small variety of methodologies due to the specialisation and experience needed to apply more complex machine learning methods. For that reason I have focused on creating a set of easily expandable tools assisting with decisions that need to be addressed during this type of work. A large number of clustering parameters and relevant metrics are integrated into these tools in order to allow non-ML experts to generate gene expression based subgroups and justify their decisions along the way supported by machine learning theory.

B. **Apply the toolkit on mRNA data from patient blood to detect subgroups of pulmonary arterial hypertension (Chapter 3).** The Omada toolkit is providing educated estimates on integral machine learning questions based on well studied algorithms and formulas. However, I needed to test its effectiveness in real disease data to validate its ability towards the identification of meaningful patient subgroups, a central question in current research especially in the context of disease treatment and diagnosis. Therefore, I aimed to apply each tool on a H/IPAH cohort and biologically separate this heterogeneous group of patients while providing valuable insights towards patient group characterisation.

C. **Test the toolkit on other molecular data types collected from pulmonary hypertension patients (Chapter 4 & 5).** The application and validation of Omada on RNA sequencing datasets led to the question whether other omics data can be utilised by this pipeline. This part of my work aimed to test whether the toolkit is applicable to other popular molecular data such as metabolomics and miRNAs and whether it can produce interesting subgroups and biological insights.

D. **Test longitudinal clustering on the activity data of COVID-19 patients (Chapter 6).** Partition of patients according to longitudinal criteria is a very different approach to traditional clustering on data from a single timepoint as we have to consider the relationship between measurements across timepoints. Using data generated by wearable technology, I aimed to identify physical activity subgroups in COVID-19 patients based on multiple activity over-time measurements utilising Fretchet distances, additional trajectory related distances and transformations.

Chapter 2 - Omada: Robust clustering of transcriptomes through multiple testing

2.1 Background

The main goal of this research was to explore the utility of unsupervised learning on gene expression data and potentially provide valuable tools that help with partitioning of samples in a biologically meaningful way. Driven by the need to avoid the application of default clustering algorithms in modern research, which in most cases will provide suboptimal results, I created a set of tools, named Omada, that make critical decisions during any clustering analysis. Each step in such an analysis needs to be supported by established machine learning theory and be able to justify any decisions it makes. Each tool is built independently to allow the addition of methods to increase its future utility. The work presented in the article describing the tools of Omada aims to assist with one of the following clustering analyses problems:

- *Help with initial unsupervised learning dataset feasibility*
- *Identify the most robust clustering algorithm per dataset*
- *Estimate the set of genes that provide the most stable clustering*
- *Estimate the most probable number of clusters*
- *Provide clustering with estimated optimised parameters*
- *Meta-analysis with cluster gene signatures*

The tools are solely based on the expression of genes and agnostic to any clinical information therefore only affected by the genetic profile of each patient. In this way all identified categories of patients only stem from gene expression patterns. This toolkit, Omada, is published as a package in Bioconductor and can be found at 10.18129/B9.bioc.omada.

2.2 Contribution

For this publication I was the first author from the conception of the study to the implementation and writing of the manuscript. More specifically, I was the main contributor in writing the manuscript, implementing all unsupervised learning work as well as the majority of the analyses of the results in both main and supplementary documents. I also created all the relevant code published along with the paper in bioconductor. Sections not generated by me are in italics and brackets in text.

2.3 Manuscript 1

Omada: Robust clustering of transcriptomes through multiple testing

Sokratis Kariotis[1,2,4], Tan Pei Fang[4], Haiping Lu , Chris Rhodes[5], Martin Wilkins[5], Allan Lawrie[5], Dennis Wang[1,4,5]#

[1]Department of Neuroscience, University of Sheffield, Sheffield UK,

[2]Department of Infection, Immunity & Cardiovascular Disease, University of Sheffield, Sheffield, UK,

[3]Department of Computer Science, University of Sheffield, Sheffield UK.

[4]Singapore Institute for Clinical Sciences, A*STAR Research Entities, Singapore

[5]National Heart and Lung Institute, Imperial College London, London, UK

Corresponding authors: Sokratis Kariotis, Dennis Wang

Short-title: Machine learning tools for automated transcriptome clustering analysis

Abstract

Cohort studies increasingly collect biosamples for molecular profiling and are observing molecular heterogeneity. High throughput RNA sequencing is providing large datasets capable of reflecting disease mechanisms. Clustering approaches have produced a number of tools to help dissect complex heterogeneous datasets, however, selecting the appropriate method and parameters to perform exploratory clustering analysis of transcriptomic data requires deep understanding of machine learning and extensive computational experimentation. Tools that assist with such decisions without prior field knowledge are nonexistent. To address this we have developed a suite of tools to automate these processes and make robust unsupervised clustering of transcriptomic data more accessible through automated machine learning based functions. The efficiency of each tool was tested with five datasets characterised by different expression signal strengths to capture a wide spectrum of RNA expression datasets. Our toolkit's decisions reflected the real

number of stable partitions in datasets where the subgroups are discernible. Within datasets with less clear biological distinctions, our tools either formed stable subgroups with different expression profiles and robust clinical associations or revealed signs of problematic data such as biased measurements.

Introduction

The rapid development of next-generation sequencing boosted the quantitative analysis of gene expression in a variety of human tissues and organs[1,2] generating valuable resources[3] for downstream investigative analysis. In recent years, such analyses aim to elucidate disease mechanisms[4] and construct genomic profiles[5] to explain diagnosis[6], prognosis and treatment patterns. However, transcriptomic profiles can be heterogeneous due to several causes pertaining to technical biases that produce batch effects[7], cellular diversity[8], disease heterogeneity[9] as well as differences between individuals and populations[10,11]. In turn, this heterogeneity hinders traditional research efforts[12] aiming to define structures especially under complex diseases which led to the utilisation of the field of machine learning towards this data demanding goal. More specifically, unsupervised machine learning i.e. clustering, in the form of transcriptomic profiling based on sequencing data[13–15], has been explored in terms of symptomatic heterogeneity in complex diseases revealing differences in molecular states[16–18] and phenotypes described by the gene expression of diseased tissue. However, deep medical and molecular knowledge is required to identify solvable problems and interpret results within the context of various diseases and conditions. Simultaneously, specialised knowledge and experience is needed to create functional, efficient and insightful models which generate reproducible solutions. Despite the inherent power of these models, most times a default model is not sufficiently tuned to the specific dataset thus unable to extract essential information. Many models have been compared, tested and found to work on different data and research questions[19,20] highlighting that no single model is always optimal without tuning (or optimising) on the specific dataset at hand, especially with state-of-the-art methodologies[21,22].

Machine learning (ML) is currently being used in many forms and combinations[23], for different types of projects within diverse fields of biomedical research[24–26]. Supervised and unsupervised methods are being developed to address specific questions and/or data problems as the pace of new data generation increases rapidly. Big data has made the importance of tailored methodologies essential for specialised datasets[27,28], as speed and accuracy pose an even greater obstacle, especially when handling sizable medical data. The impact of machine learning in biomedical sciences has risen considerably with the multitude of methodologies

leading to previously unfeasible computations[29,30]. Unsupervised learning proved to be an invaluable tool towards exploring heterogeneity in complex diseases since its functioning without any prior knowledge or assumptions of sample labels. Due to this diversity, there is a need for methods that support non-expert users to utilise the characteristics of various methodologies in their unsupervised work. One of the most important aspects of sample partitioning is the stability of the generated groups as unstable clusters, usually imply the lack of signal which should be present and drive the clusters. Signals can take many forms, for example the level of gene activity in RNA sequencing datasets. Cluster instability can be caused inherently by the data points or by the type and application quality of a clustering technique.

With the above obstacles in mind, we introduce Omada, a toolkit with multiple functions based on cluster stability and machine learning formulas to provide assistance to both experienced and inexperienced users during the steps from dataset assessment to the formation of the subgroups. Each function's results are based on machine learning theory and multiple metrics to ensure that a wealth of methods will be considered, in the current version and in the future, during the decision and clustering process.

Methods

This toolkit consists of a pipeline that takes in a gene expression matrix to identify transcriptomic subgroups of samples (Figure 1 and Supplementary Figure 1). Starting from a matrix of gene expression values (e.g. transcripts per million from RNAsequencing), the most suitable clustering method is chosen, followed by selecting the transcript features for clustering and determining the number of clusters and memberships.

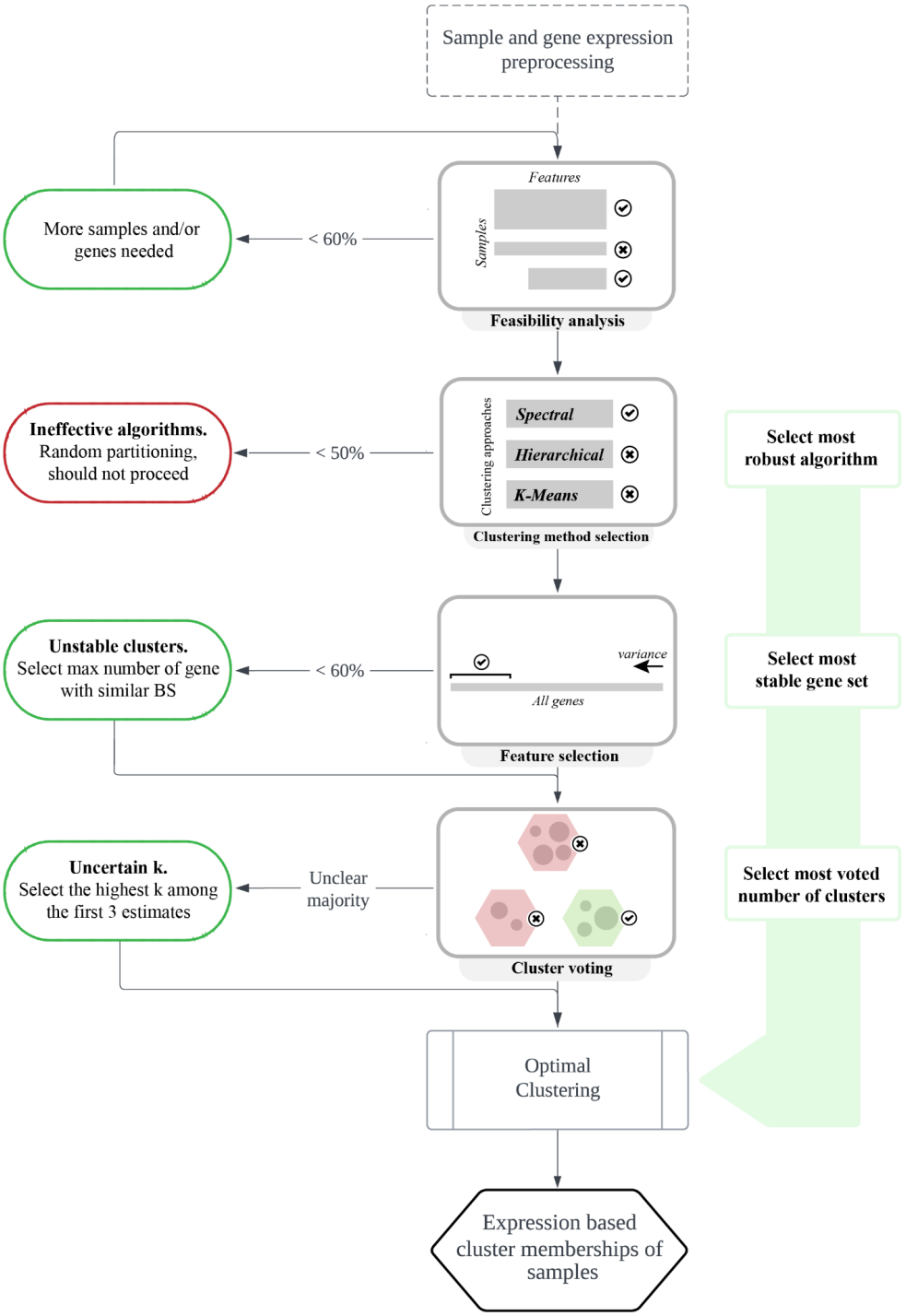


Figure 1: *An overview of steps for discovering gene expression subgroups using Omada's clustering tools. First, processing, quality control and feasibility analysis are ensuring the input data are suitable for clustering. Then, choose the most robust clustering methodology that provides the most consistent partitions. Next, determine the genes that are useful for discriminating samples and provide the most stable clusters. Finally, determine the number of subgroups that potentially exist in the cohort by selecting the number of clusters (k) supported by the majority of internal machine learning indexes. The end result, after the final optimised clustering, consists of the assignment of a cluster to each sample driven solely by its expression profile.*

Sample and gene expression preprocessing

A preprocessing step is recommended by the user before the application of these tools on any dataset to heighten the chances of any underlying important signal to be discovered. Data biases and format can often drive clustering attempts to focus on discriminating data points based solely, or mostly, on known information producing no new insights irrespectively of the method used[21,31]. To address this, it is recommended to attempt to remove/normalise any data points that might be introducing strong biases to allow the novel signal to be detected. Furthermore, numerical data may need to be normalised in order to account for potential misdirecting quantities (i.e. outliers) or specifically transformed to satisfy an algorithm's input criteria. Data points or samples have to be filtered based on field knowledge to allow the data to answer specific scientific questions. Expression data should go through proper quality control depending on the manner of collection to identify outliers and remove unreliable datapoints. For microarrays it's important to assess sample, hybridization and overall signal qualities along with signal comparability and potential biases. Array correlations through PCA and correlation plots should also be considered[32]. RNA sequencing experiments also produce data that need to be controlled for potential trimming of adapter sequences, low quality reads, uncalled based and contaminants by using a plethora of available tools[33,34]. Additionally, qPCR generated data should be checked for abnormal amplification, positive and negative control samples, control on PCR replicate variation and determine reference gene expression stability and deviating sample normalisation factors[35]. As for the number of genes, it is advised for larger genesets (>1000 genes) to filter down to the most variable ones before the application of any function as genes that do not vary across samples do not contribute towards identifying heterogeneity. Moreover, large genesets require increased computational power and extended runtime without adding any real value due to the large number of non useful genes. Lastly, it is important to note that technical artefacts, such as sampling location or machine specifications, may drive clustering causing the formation of very distinct clusters which can solely be attributed to relevant biases. It is very important for those cases to be identified and extracted insights should be disregarded as they do not reflect real signals or data trends.

Determining clustering potential

At the start of each study, we assess the suitability of the input dataset for clustering to ensure general dataset attributes do not influence the process (Figure 1). The

number of samples and features -i.e. genes-, as well as the balance of the two dimensions, directly affects the capabilities of clustering methods to handle the dataset. An inadequate number of samples does not provide enough training power[36], while an overabundance of samples might clutter the provided information and confuse most methodologies[37]. Similarly, too few features can lead to weak clustering criteria and too many features might lead a methodology away from the features that can really differentiate between clusters of samples. Therefore, to estimate the feasibility of a clustering procedure on a specifically sized dataset we rely on measurable metrics of cluster quality, such as stability. Clusters of high stability denote both a partitionable dataset as well as a dataset-suitable methodology[38]. The feasibility score of any dataset is a function of both dimensions as well as the number of classes requested. As such, if too many or a single class is requested of a relatively small dataset the calculation will reflect low feasibility due to insufficient samples and/or features to form the desired classes.

Simulating datasets

To assess the quality of the dataset to be used, our toolkit includes two functions for simulating datasets of different dimensionalities for stability assessment. We use those to understand the relation between the number of samples, genes and cluster sizes. The first function we use to simulate datasets allows for tuning the number of samples (n), genes (m) and clusters (c). Each cluster contains $\frac{n}{c}$ samples drawn from a normal distribution with a different mean and standard deviation. Each mean is drawn from a sequence of c evenly spaced integers that belong to the range $[5,\ c\ *\ 10]$. Each standard deviation is similarly drawn from a range of $[1,\ c\ *\ 2]$. To estimate the difference between distributions, we calculate the two sided Kolmogorov's D statistic between each pair of distributions representing the generated classes and plot the empirical cumulative distribution function (EDCF).

Subsequently, we calculate the stability of each *k* (number of clusters for a particular clustering run) using the clusterboot function in R package fpc v2.2-3. The number of clusters *k* to be considered belong to $k \in [number\ of\ classes\ -\ 2,\ number\ of\ classes\ +\ 2]$, with a minimum of *k* = 2. The maximum and average stabilities over all *k* are reported, providing a stability-based quality score that provides an insight in deciding whether a prospect dataset is suitable for a clustering study.

To assess the clustering feasibility of an existing dataset this tool kit also provides a similar function which generates a simulated dataset based on an input dataset and the user's estimation of the number of classes. The number of samples and genes equal those of the input dataset and its mean (m_{input}) and standard deviation (sd_{input})

affect those of each generated class within the dataset. Specifically, if $n \in (1, 2, 3, \ldots)$ is the number of classes, each class mean (m_{class}) equals $m_{input} * 10 * n$ and each class standard deviation (sd_{class}) equals $sd_{input} * 2 * n$.

Intra-method Clustering Agreement

Unsupervised learning offers a multitude of methods to be applied on specific types of data due to their nature (e.g. numeric, binary) or underlying signal to be detected. Most studies employ widely-used methods (e.g. hierarchical clustering) without exercising any kind of selection method that would point towards the most effective methodology. Selecting an appropriate approach requires extensive machine learning and data analysis knowledge coupled with tuning and testing of multiple different algorithms. To enable non-machine learning expert users to utilise the vast capabilities of this field and avoid default limited efficiency methodologies we present a clustering selection tool that offers an intelligent selection method with unbiased results through parameter randomization. The nature of this selection method allows any number of well established unsupervised methods to be considered.

To address the lack of class labels and thus a performance measure in unsupervised models, we compare how consistently different approaches partition our data when one or more parameters change. As high consistency we define the high agreement score calculated between different variations of a clustering algorithm. When two different clustering runs agree on the partitioning of the samples they also show robustness since they do not randomly assign samples to subgroups but rather are driven by the underlying structure of the data.

We implemented a tool (Figure 1) to calculate an average agreement score per clustering approach by comparing a number of runs within each of the three clustering approaches (hierarchical[39], k-means[40], spectral clustering[41]) using multiple parameters (kernels, measures, algorithms) specifically based on the data set provided. The number of comparisons (*c*), between runs of the same approach, is an additional overarching parameter of this tool and contributes to the agreement score. For each comparison, the parameters of the two runs are drawn randomly from a predefined set (Table 1) selected randomly with replacement while not allowing the same parameters to be used within one comparison. In the interest of performance and computational time we suggest three comparisons to be used. Depending on *c*, we generate variations of the base clustering algorithms (package kernlab v0.9-29),

along with the various distance measures and clustering categories they belong to. Within each pair of clustering runs the agreement is calculated using the adjusted Rand Index (package fossil v0.3.7), the corrected-for-chance version of the original Rand index[42], which is based on the number of times any pair of points is partitioned in the same subgroup throughout different clusterings runs. To calculate the agreement within each clustering algorithm (spectral, k-means, hierarchical) we are considering pairs of runs using the same algorithm but different parameters. For those pairs the agreement is averaged across clustering runs and *k* number of clusters tested. The algorithm that presents the highest intra-method agreement over a logical range of clusters ($k \in [2,x]$) is noted as the most appropriate clustering of the samples based on a detected signal. A logical range of *k* is considered a set of successive *k*'s (where $k \geqslant 2$) that is most probable to exist within our data, often determined by prior knowledge of the data, previous studies or domain expertise. This selection procedure is mainly affected by the type and size of the data leading similar datasets to opt for the same method due to the specific mathematical formulas within each algorithm.

Spectral clustering algorithm[41]

Given a set of points $S = \{s_1 \dots, s_n\}$ in R^l that we want to cluster into ***k*** subsets:

1. *Form the affinity matrix* $A \in R^{n \times n}$ *defined by* $A_{ij} = \exp(-\|s_i - s_j\|^2 / 2\sigma^2)$ *if* $i \neq j$, *and* $A_{ii} = 0$
2. *Define D to be the diagonal matrix whose (i, i)-element is the sum of A's i-th row, and construct the matrix* $L = D^{-l/2} A D^{-l/2}$
3. *Find* $x_1, x_2, \dots, x_k$, *the k largest eigenvectors of L (chosen to be orthogonal to each other in the case of repeated eigenvalues), and form the matrix* $X = [x_1 x_2 \dots x_k] \in {}^{n \times k}$ *by stacking the eigenvectors in columns*
4. *Form the matrix Y from X by renormalizing each of X's rows to have unit length (i.e.* $Y_{ij} = X_{ij} / (\Sigma_j X^2_{ij})^{1/2}$)
5. *Treating each row of Y as a point in* R^k, *cluster them into k clusters via K-means or any other algorithm (that attempts to minimize distortion)*
6. *Finally, assign the original point* s_i *to cluster j if and only if row i of the matrix Y was assigned to cluster j*

Hierarchical clustering algorithm (average linkage)

Given a set of points S = {s_1 ... ,s_n} that we want to cluster into ***k*** subsets:

1. *Initialise with n clusters, each containing one data point (s_i)*
2. *Compute the between-cluster distance D(r, s) as the between-object distance of the two data points in clusters r and s respectively, r, s =1, 2, ..., n. Let the square matrix D = (D(r, s)). Various distances can be used (euclidean, manhattan, canberra, minkowski, maximum).*
3. *Find the most similar pair of clusters r and s, such that D(r, s) is minimum among all pairwise distances*
4. *Merge r and s to a new cluster t and compute the between-cluster distance D(t, k) for any existing cluster k ≠ r, s. Once the distances are obtained, delete the rows and columns corresponding to the old cluster r and s in the D matrix, as r and s do not exist anymore. Then add a new row and column in D corresponding to cluster t.*
5. *Repeat Step 3 a total of n − 1 times until there is only one cluster left.*
6. *Decide on a point to cut the cluster tree created above so as to obtain the desirable number of clusters (k)*

K-means[40]

K: kernel matrix, k: number of clusters, w: weights for each point, tmax: optional maximum number of iterations, $\{\pi_c^{(0)}\}_{c=1}^k$: optional initial clusters

1. *If no initial clustering is given, initialize the k clusters $\pi_1^{(0)}, \dots, \pi_k^{(0)}$ (i.e. randomly). Then, set t = 0*
2. *For each a_i and every cluster c, compute*

$$d(a_i, m_c) = K_{ii} - \frac{2\sum_{\alpha_j \in \pi_c^{(t)}} w_j K_{ij}}{\sum_{\alpha_j \in \pi_c^{(t)}} w_j} + \frac{\sum_{\alpha_j, \alpha_l \in \pi_c^{(t)}} w_j w_l K_{jl}}{\left(\sum_{\alpha_j \in \pi_c^{(t)}} w_j\right)^2}$$

3. *Find $c^*(a_i) = \text{argmin}_c d(a_i, m_c)$, resolving ties arbitrarily. Compute the updated clusters as*

$$\pi_c^{(t+1)} = \{a: c^*(a_i) = c\}$$

4. *If not converged or $t_{max} > t$, set $t = t + 1$ and go to Step 2; Otherwise, stop and output final clusters*

$$\{\pi_c^{(t+1)}\}_{c=1}^k$$

Table 1 *| The clustering algorithms, their approach category and the various distance measures tested*

Clustering algorithms	Category	Distance measures/kernels	Additional parameters
K-means	Partitioning	Hartigan-Wong, Lloyd, Forgy, MacQueen	-
Hierarchical	Hierarchical	Euclidean, Manhattan, Minkowski, Canberra	Average, complete, median (linkage)
Spectral	Graph Theory	Rbfdot, Polydot, Tanhdot, Laplacedot, Vanilladot, Anovadot, Splinedot	-

Feature set subsampling

While gene expression data provide measures on the thousands of transcripts in the transcriptome, not all of them may provide discriminative information on the samples and may not be useful for clustering. Moreover, most clustering algorithms are heavily affected by a large number of features both computationally due to input size and in performance due to misdirecting data noise[43]. A common strategy to select interesting and potentially useful RNA features is to measure their variance across samples and select the ones with the highest scores instead of those that are either housekeeping or do not differentiate in our context. In this tool, we exclude RNA features that remain stable across samples and are therefore unable to offer any discriminatory power to our unsupervised machine learning models. Furthermore, the exhaustive feature selection procedure incrementally considers all the genes in the feature set and takes into account the stability of all generated test clusters and number of cluster ranges. This step does not require any deep knowledge or filtering decisions by the user.

Based on this observation our sample selection step, which is a part of the tool for bootstrap resampling of features presented in Figures 1C and 2A, first ranks features in a descending order of variance (var() function from the Stats R package) across samples, generating a list of the most variable features. Subsequently, multiple datasets of all samples and subsets of features are generated. All subsets draw a different number of features from the top of the variance list with replacement. The

first dataset uses a relatively small number of features (*n*), depending on the total number of features (*N*) and the granularity of the result desired. The following datasets re-draw from the initial list increasing the number of features by *n*, ending up with $\frac{N}{n}$ datasets.

Stability-based assessment of feature sets

To assess the suitability of each resampled feature set for our clustering, we measure the average stability of the clusters they generate per run when a clustering method is applied over a range of *k*'s (Figure 2B). First, the clustering range, where the stability of each dataset will be calculated, is selected. For each dataset we generate the bootstrap stability for every *k* within range. To calculate each bootstrap stability score the data is randomly sampled with replacement and clustered internally using a spectral approach. We then compute the Jaccard similarities between the original clusters and the most similar clusters in the resampled data. The above procedure results in a stability score for each *k* and each dataset. We then calculate the final stability of each dataset by averaging the stability over *k*. The genes that comprise the dataset with the highest stability are the ones that compose the most appropriate set for the downstream analysis.

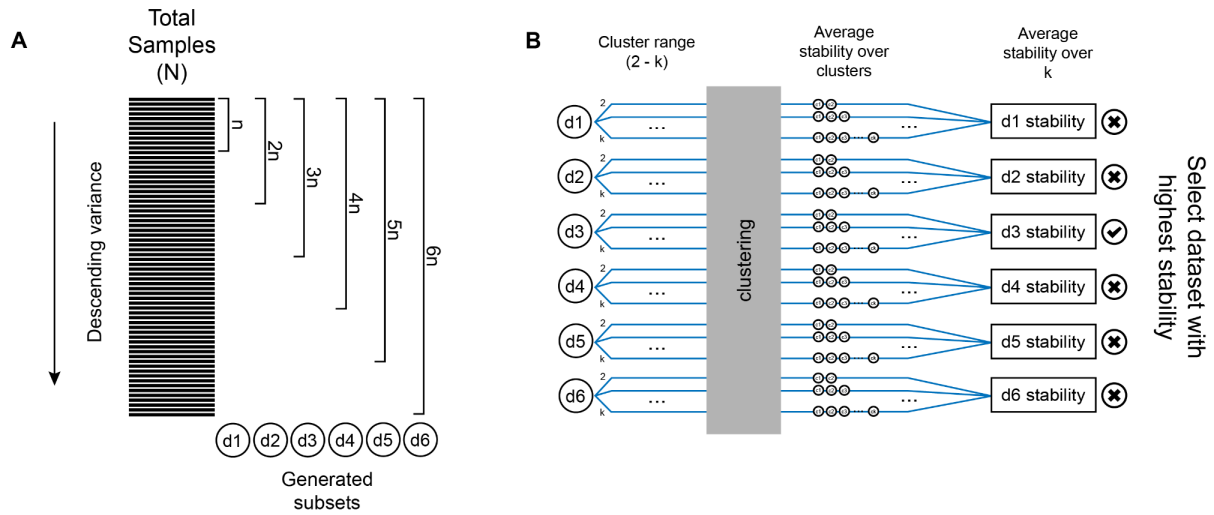


***Figure 2**: Sample selection overview. (A) Ranking of samples based on their variance across features and the subsequent generation of datasets of increasing size. (B) Calculation of the stability score of each generated dataset. Initially, we select a cluster range to run our clustering method for each dataset. After the clustering procedure, we calculate and average the stability over the generated clusters. Finally, we average the stabilities over k per dataset and determine a final stability score for each dataset. The features of the dataset with the highest stability are the ones that compose the most appropriate set for the downstream pipeline*

Choosing k number of clusters

Most clustering methods require the number of *k* clusters to be defined as a parameter before the application of the algorithm on the data. The lack of a concrete way to determine the real number of clusters in a dataset led many studies to base their estimation on field/prior knowledge or various estimation methods such as the Silhouette score[44]. However, each method favours different aspects of the generated clusters (i.e. how compact clusters are and how far apart cluster centres are) and therefore suits specific datasets and may introduce bias towards the selection of *k*. To encompass these different angles in one methodology, avoid the risk of selecting an ineffective index and present a more general solution, this tool uses an ensemble learning approach (Figure 1) where multiple internal cluster indexes contribute to the decision making[45]. This approach prevents any bias from specific metrics and frees the user from making decisions on any specific metric and assumptions on the optimal number of clusters.

Initially, the value of the 15 indexes is calculated for each *k* within a cluster range of $[2, x]$, where *x* is a logical upper limit of the number of clusters realistic for our dataset. The means over *k* are calculated per index and the optimal *k* is estimated by majority voting of the 14 means that evaluate the compactness and/or the distance between different subgroups. The selection of indexes can be found in Table 2. It is important to note that the most important aspect of determining *k* is minimum loss of information which directs us to overestimate and not underestimate *k*[43] while interpreting the voting results. Furthermore, cases that present only a single *k* as the optimal number of clusters should be treated with caution in case they are a result of a biased dataset.

***Table 2**: The list of 15 internal indexes used to estimate the optimal number of clusters (k).*

Internal index	Ideal	Formula	Source
Calinski-Harabasz	max	$\frac{(\frac{between\ cluster\ variation}{k-1})}{(\frac{within\ cluster\ variation}{n-k})}$	46
Dunn	max	$\frac{min(intercluster\ distance)}{max(\ distance\ between\ all\ pairs)}$	47
Pbm	max	$\frac{1}{k} * \frac{E^T}{E^W} * D_B$	48
Tau	max	$\frac{concordant\ pairs - discordant\ pairs}{\sqrt{N_B * N_W \frac{N_T(N_T-1)}{2}}}$	49
Gamma	max	$\frac{concordant\ pairs - discordant\ pairs}{concordant\ pairs + discordant\ pairs}$	50
C index	min	$\frac{S_W - S_{MIN}}{S_{MAX} - S_{MIN}}$	51
Davies–Bouldin	min	$\frac{1}{n} * \sum_{i=1}^{n} max(\frac{clusters\ scatter\ difference}{cluster\ seperation})$	52
Mcclain rao	min	$\frac{N_B}{N_W} * \frac{S_W}{S_B}$	53
sd_dis	min	$a\ *\ (avg\ scattering\ for\ clusters)\ +$ $total\ separation\ between\ clusters$	54
Ray–Turi	min	$\frac{1}{n} * \frac{within-cluster\ dispersion}{min\ of\ the\ sq.\ distances\ between\ all\ the\ cluster\ barycenters}$	55
g_plus	min	$\frac{2 * s^-}{N_T(N_T - 1)}$	56
Silhouette	max	$average\ distance\ between\ clusters$	44
s_dbw	min	$mean\ dispersion\ of\ clusters\ +\ between\ cluster\ density$	57
Compactness	max	$Intra\ -\ Cluster\ distance$	58
Connectivity	max	$The\ extent\ by\ which\ the\ items\ are\ placed\ in\ the$ $same\ cluster\ as\ their\ nearest\ neighbours\ in\ the\ data\ space$	-

All indexes are using different formulas to score a partitioning, measuring one or both of the following concepts: a) how compact each cluster is and b) how well the clusters separate. For each index we present which value is preferred (min or max) and its source.

For the formulas: k = number of clusters, n = number of data points, E^T = sum of the distances of all the points to the barycenter G of the entire dataset, E^W = sum of the distances of the points of each cluster to their barycenter, N_B = pairs constituted of points which do not belong to the same cluster, N_B = pairs constituted of points which belong to the same cluster, $N_T = N_W + N_B$, S_W = sum of the N_W distances between all the pairs of points inside each cluster, S_{MIN} = sum of the N_W smallest distances between all the pairs of points in the entire data set, S_{MAX} = sum of the N_W largest distances between all the pairs of points in the entire data set, S_B = sum of the between-cluster distances, α = weight equal to the value of average scattering of clusters obtained for the partition with the greatest number of clusters.

Optimal parameter tuning

Previous steps have selected the optimal method, number of features and clusters. To perform the optimal clustering we automate the selection of parameters for each method so that manual tuning is not required. Towards that goal we utilise cluster stabilities to decide on the parameters (which depend on the specific algorithm i.e. kernels in k-means and spectral clustering, linkage method in hierarchical clustering) selected by this toolkit. All available parameters (Table 1) participate in the selection procedure where we measure the average bootstrap stability of the clusters (clusterboot function in R package fpc v2.2-3) using the previously determined optimal *k* and feature set for each parameter. The parameter that produces the highest stability is used for the optimal clustering run.

Test datasets

Five datasets were used to validate different capabilities of the Omada package. First, two datasets were simulated by Omada's functions. Function feasibilityAnalysisDataBased() was used to generate a multi-class dataset with 359 samples and 300 genes based on the contents and dimensions of the original RNA-seq data[18] and composed of five groups of samples drawn from five different distributions with means (5,16,27,38,50) and sd (1,3,5,7,10), representing the five classes. Function feasibilityAnalysis() simulated a single-class dataset of 100 samples and 100 genes drawn from a single distribution. For the multissue Pan-cancer dataset we downloaded RNAseq expression data for 2244 samples and 253 genes representing three types of cancers: breast (n=1084), lung (n=566) and colon/rectal (n=594) downloaded through cbioportal[59] from TCGA PanCancer Atlas[60]. The mRNA expression was in the form of z-scores relative to normal samples where

we applied an extra step of arcsine normalisation. After filtering for tissue-specific genes[61] for the three cancer types we retained 243 genes. Next, we utilised a PAH dataset (25,955 genes) generated from 359 patient samples with idiopathic and heritable pulmonary arterial hypertension (IPAH/HPAH). The transcriptomic data can be found in the EGA (the European Genome-phenome Archive) database under accession code EGAS0000100553265[62] (restricted access) and all pre-processing details and parameters used can be found in [18]. Finally, we used an RNA dataset from the whole blood of 238 mothers during midgestation (26-28 weeks of pregnancy). Read counts were extracted from GEO (accession number GSE182409[63]) and were then read into R and converted into TPM using the *convertCounts* function available in the *DGEobj.utils package*. For the purpose of clustering, we mapped the TPM dataset to the list of 24,070 genes used in the PAH dataset described in a previous section.

Results

Omada was applied to five diverse gene expression datasets to demonstrate its utility in guiding cluster analysis and identifying plausible subgroups of samples. Two datasets were simulated by our tools. The simulated dataset with multiple distinct classes was used to determine Omada's ability to accurately estimate *k* with reasonable stability when we know the existence of sample classes. In contrast, samples in the single-class simulated dataset were drawn from a single class and used to demonstrate the toolkit's ability to point towards the lack of sample subgroups by indicating inconclusive low scores throughout the analysis. A multi-tissue pancan dataset was introduced to assess Omada's capability to generate signal-based clusters that closely follow the tissue-specific patient sample distributions. In addition, to determine whether Omada can identify distinct heterogeneous subgroups from data without any prior classification information but potential present heterogeneity, we used a whole blood RNA-seq dataset from patients with pulmonary arterial hypertension (PAH)[18]. Lastly, implementation of the toolkit on a whole blood expression dataset (GUSTO) was included to demonstrate a case with potential technical biases and no known subgroups since it is composed of healthy participants.

For the above, we measured its consistency on algorithm, feature and number of clusters (*k*) selection and the stability of the generated clusters for a particular *k* ($stability_k$) and the average across *k's* ($stability_{avg}$). It's important to note that the value of this validation is derived from the fact that unstable clusters should not be

interpreted as this instability comes from problematic data or an incorrect approach. On the other hand, it's worth noting that cluster stability only provides a way to assess the potential underlying data structure and further information is required to fully validate the clusters[38] , ideally using biological criteria.

Identifying known clusters

Multi-class dataset: Five distinct simulated expression classes representing heterogeneity

Omada's basic function is to help identify samples that come from different sources and group together samples that come from the same source. Towards that end, we simulated a dataset with five sets of expression profiles with 50 samples each and 120 genes sourced from five unique distributions of expression data that represent heterogeneity within our samples. The means and standard deviations of each class are presented in Figure 3A, depicting the expression differences. Additionally, the empirical Cumulative Distribution Functions (ECDFs) of the five simulated classes (Figure 3B) as well as the high average Kolmogorov-Smirnov distances (D_{avg}=88.3%, supplementary Table 1) show distinct differences between the distributions in respect to the expression in the simulated RNA-seq dataset. To demonstrate the effect of different sample and gene numbers, multiple datasets were simulated with an increasing number of samples and genes (Figure 3C). The calculated cluster stabilities, where each value represents the stability over a range of *k* and a specific number of samples and features, show five or less samples per class provide highly unstable and unreliable clusters. The minimum acceptable stability threshold of 60% was achieved with at least 20 samples and a reliable stability of 75% was achieved using 1000 samples.

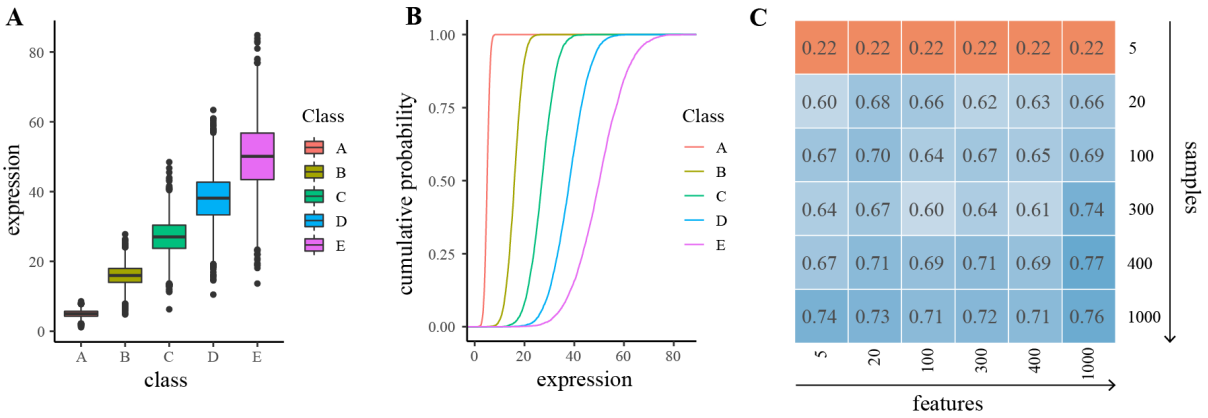


***Figure 3**: A) Expression boxplots for the five clusters showing the means and standard deviations B) The cumulative probability (as calculated from the empirical cumulative distribution function) for the five clusters calculated by a two sided Kolmogorov-Smirnov Test C) Average over-k stabilities for simulated datasets of increasing sample and gene numbers. A small number of samples consistently provides extremely unstable clusters (orange) while increasing both numbers consistently produces datasets that pass the stability threshold of 0.6 (blue).*

To test the ability of the clustering tools to produce stable clusters in various contexts we first apply them in sequence on strategically simulated data. The data are composed of distinct classes (based on class mean m_{class} and standard deviation sd_{class}) and due to that strong signal our tools are expected to determine an accurate *k* with reasonable stability, scoring above 60%. To allow for a more direct comparison, we used a multi-class simulated dataset (see *Test datasets* in *Methods*) based on the original RNA-seq data[18]. When considering ranges of *k* we are using [2, 6] clusters to observe a broader range of results for comparison reasons. First, the clustering feasibility tool showed that the highest stability was 78% (Supplementary Table 4) providing a strong indication of stability across our clusters. Since we selected a limited range of $k \in [2, 6]$ where the stability should remain high, the averaged -over every tested - stability ($stability_{avg}$) of 72% indicates a dataset of adequate size and class definition to proceed to clustering analysis. It should be noted that when large ranges of *k* are selected the average stability will naturally decrease as the calculations will take into account *k*'s much larger or smaller than the actual number of classes in the data. In such cases the user can review the individual *k* stabilities generated as part of this tool to conclude whether those values are satisfying i.e a minimum of 60%. Next, we calculated the partitioning agreement of three clustering algorithms and spectral clustering showed the highest average score of 56% (Figure 5A and Supplementary Table 4). Partitioning agreement scores should be interpreted across algorithms applied on the same dataset rather than as absolute values keeping in mind that a score below 50% represents a random partitioning and subsequently a non-robust clustering. In the subsequent feature selection step, the highest average stability was registered when using all 300 features ($stability_{avg}$ = 78%, Supplementary Table 4), not discarding any feature as they all demonstrated

very similar variance due to the nature of the simulated data. Finally, 8 out of 15 internal metrics voted five clusters as the optimal number during the k estimation step (Figure 5B) providing a confident estimation above 50%.

Single-class dataset: Homogeneously simulated dataset

To demonstrate Omada's ability to identify datasets without any present clusters where all patients belong to one class, we used the single-class simulated dataset (see *Test datasets* in *Methods*). All potential *k* of two or higher achieved low scores with average and maximum stabilities of 45% and 55%, respectively (supplementary Table 2). It is recommended to avoid clustering analysis on such low score datasets and instead opt for scores of at least 60%. Ideally, stabilities of 80-90% are considered very strong[64], however the potential of several signals in transcriptomic data and the exploration across multiple *k* generally decreases the output stability to an acceptable threshold of 60-70%. Next, Figure 4A shows the overall low partitional consistencies (averaged over all tested *k*) for all algorithms with spectral average partition agreement of 52%, kmeans average partition agreement of 3% and hierarchical average partition agreement of 26%. With the best performing algorithm showing an agreement of around 50% we can assume that the tested algorithms are randomly assigning memberships, therefore we cannot achieve a robust model with the current data. When using spectral clustering to select the most appropriate set of genes, the cluster stability rapidly dropped below 50% when using more than 20 genes (Figure 4B) indicating that the algorithm got worse in assigning memberships as we considered more simulated genes. Finally, the ensemble voting step showed the majority of the votes supporting five clusters (Figure 4C), a significant variation from the single simulated class of this dataset. In such unexpected outputs, one should examine the generated metric scores. In this case, the vast majority of metric scores are worse when we are testing single-class instead of multi-class simulations (supplementary Table 3) inferring lower cluster quality, i.e lower compactness and smaller distance between clusters. Additionally, worse scores (decided according to Table 2) infer higher uncertainty during the voting process.

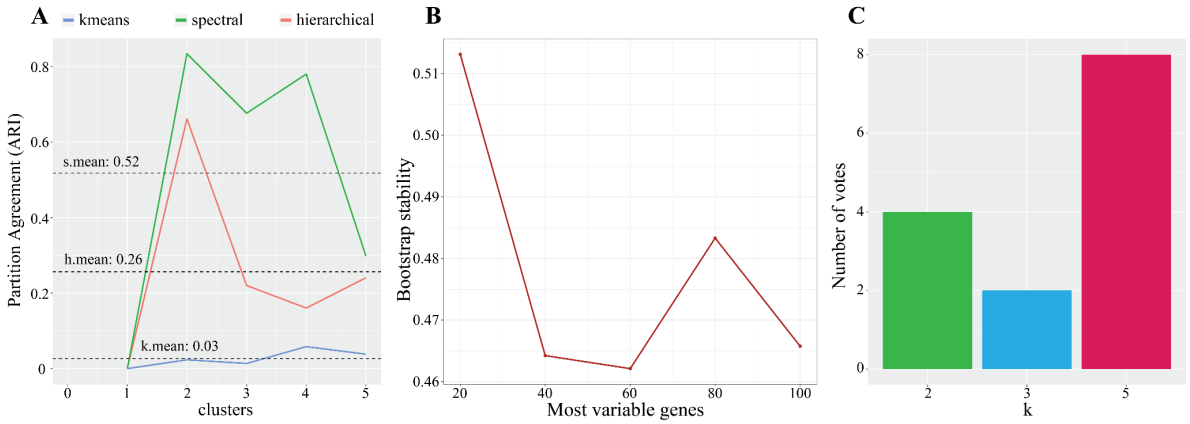


Figure 4: Performance criteria for single-class simulated dataset. The results demonstrate low scores for the majority of steps. A) shows the average partition agreement of all three algorithms below the 52% mark indicating very unstable clustering runs overall. In B) the stability of every possible subsets of genes does not surpass 51.3% underlying overall unstable clusters. C) shows five clusters as the first estimate (voted by 8 metrics), significantly different from the one class this dataset contains.

Discriminating cancer types from pan cancer tissue expression data

An integral capability of Omada is to accurately stratify patients according to any biologically relevant signal present in expression data and detect differences stemming from genes, pathways, tissues etc. Real multi-tissue samples are often the focus of exploratory studies as they present cell-type differences but still unknown factors that may discriminate them. Using expression data from multiple cancer types (Pan-cancer dataset as described in *Test datasets* in *Methods*), we expect our tools to identify clusters that are consistent with the samples' tissues of origin. Due to the different types of tumours we explored the potential cluster range of [2, 5] for each pipeline step. The clustering feasibility of the dataset (2244 samples, 243 genes) presented an average stability of 88% and maximum stability of 100% (Supplementary Table 5) providing confidence for the downstream analysis. Spectral clustering showed the highest consistency (partition agreement$_{avg}$ = 63% closely resembling the simulated multiclass dataset, Figure 5A) and was therefore deemed as the most robust. In this example hierarchical clustering showed high instability, as shown in Figure 5A, demonstrating the importance of selecting the appropriate algorithm to create a robust model. According to our selection tool, all 243 genes produced the most stable set of clusters with a stability of 96% (Supplementary Table 5) which coupled with the high algorithm robustness indicated a model that most likely detects a signal in the data. Additionally, a very important observation is that all genes were deemed important to produce nearly perfectly stable clusters agreeing with the filtering of genes based on the cancer type annotations we performed prior to this clustering analysis. The ensemble voting tool estimated our dataset to contain three clusters of samples with the support of 57% of the metrics (Figure 5B). When

comparing these results with the simulated five-class dataset, both achieved higher certainty on the five clusters (>50%, Figure 5B) reflecting the rigid differences between the clusters when dealing with cancer tissues and simulated classes. In the case of the pancan partitioning, the breast, lung and colon/rectal samples almost perfectly grouped in their respective clusters (Figure 5C).

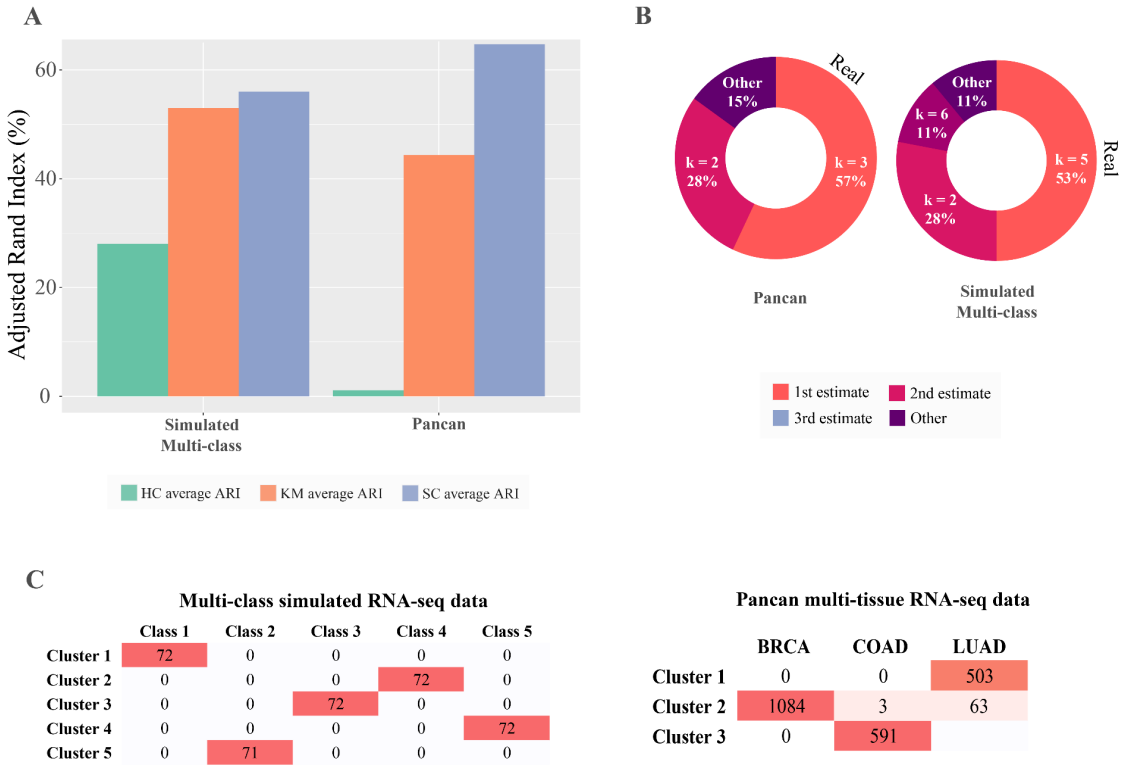


Figure 5: Performance criteria for two heterogeneous datasets, simulated multi-class and Pancan dataset. The multi-class dataset contains artificial samples from five distinct clusters and the Pancan dataset is composed of three different cancer types presenting biological heterogeneity. A) The agreement between the predicted and true clusters (Adjusted Rand Index) from three different clustering algorithms (HC: hierarchical, KM: K-means, SC: spectral clustering) applied to the two datasets. B) Shows the real number of clusters for the dataset (black text) and the three most likely number of clusters k, with estimates of their percent probability. C) The contingency tables of the combinations between generated clusters (1st estimates) and real classes in the datasets. Darker red colour intensity denotes higher frequency.

RNAseq data from diseased tissue with unknown heterogeneity

It is important for Omada to be able to robustly identify patient subgroups when heterogeneity for the cohort has not been previously characterised. We applied our tools on such a dataset (PAH dataset as described in *Test datasets* in *Methods*) to assess whether they can still produce stable clusters that differ in terms of

expression profiles and other phenotypic measures. The feasibility for this dataset's simulation showed an average stability of 61% and a maximum stability of 74% both acceptable to proceed with the clustering analysis (Supplementary Table 7). A notable 13% difference between average and maximum stability provides a positive indication that a specific *k* might prove significantly more stable downstream. The spectral clustering technique recorded the highest partitional consistency (partition $agreement_{avg}$ = 86% and partition $agreement_{max}$ = 96%, Supplementary Table 7) when we examined each algorithm's partition agreement for up to ten clusters. The bootstrapping subset selection tool estimated the 300 most variable genes as the most stable clustering parameter with a maximum stability of 73% (Figure 6A) showing an impressive reduction from the initial gene set (25,955) and ensuring the removal of a lot of data noise. According to the ensemble voting tool two clusters were voted by 71% of the internal metrics followed by *k* = 3 (14%) and *k* = 5 (7%). Despite the strong indication of two clusters, *k* = 5 was selected to prevent loss of information occuring when smaller embedded clusters are disregarded. As shown by the downstream analysis, fully presented in [18], selecting the higher *k*, even as a second estimate, allowed us to detect strong expression profiles. After considering cluster sizes the three predominant subgroups showed significant differences in expression, immunity and survival profiles as well as risk category distributions (Figure 6B, C).

RNA-seq data from healthy whole blood tissue

Next we tested how Omada would discriminate samples from healthy individuals from a single tissue type. Generally, in studies based on a dataset with no discernible heterogeneity to be explored - i.e a dataset without patients of dissimilar outcomes or controls - clustering algorithms may not be robust and may generate variable results. Useful partitionings might still be formed, such as unforeseen disease subgroups, but these observations must be validated. Towards that end we used the GUSTO RNA dataset of 238 mothers, as seen in *Test datasets* in *Methods*. During determining clustering potential our simulated dataset showed $stability_{avg}$ = 56% and $stability_{max}$ = 59% (Supplementary Table 8), a similar low-stability score as in the simulated single-class (45% and 55%). We examined a k-range of [2, 5] where spectral and k-means clustering showed very similar internal average partitional agreements of 61% and 60% and very high maximum agreements of 93% and 88% (Supplementary Table 8), respectively. The extremely high agreement scores should be interpreted with caution as they might not reflect a very strong signal but an underlying bias that partitions samples in similar groups repeatedly, over-powering the parameter changes. The 50 most variable genes were estimated to produce the most stable clustering with maximum stability = 71% (Figure 6A). Similarly to the agreement scores, a small number of genes driving the most stable clusters (starting from

24,070 genes) might indicate either a strong expression signal or a pre-existing bias. When estimating the number of clusters, two (46%) and three (40%) clusters were voted by the majority showing a general consensus. Considering all the above strong indications, we need to assess the dataset and the resulting subgroups for potential biases before relying on the cluster memberships. Towards that end and utilising clinical data, the association results show the dataset might be biased based on technical batches with sequencer machine and flowID presenting significant differences between clusters (1.39e-03 and 2.55e-06, respectively) with hospital location coming close to significance with p-value = 0.072 (Supplementary Table 9). Additional statistical tests and regression analysis with maternal and foetal physiological and clinical phenotypes did not show any association with the clusters. The expression profiles of the two clusters show visible differences as do the t-SNE and PCA analyses (Figure 6E) with the first principal component of the latter explaining 79% of the variance in the GUSTO dataset.

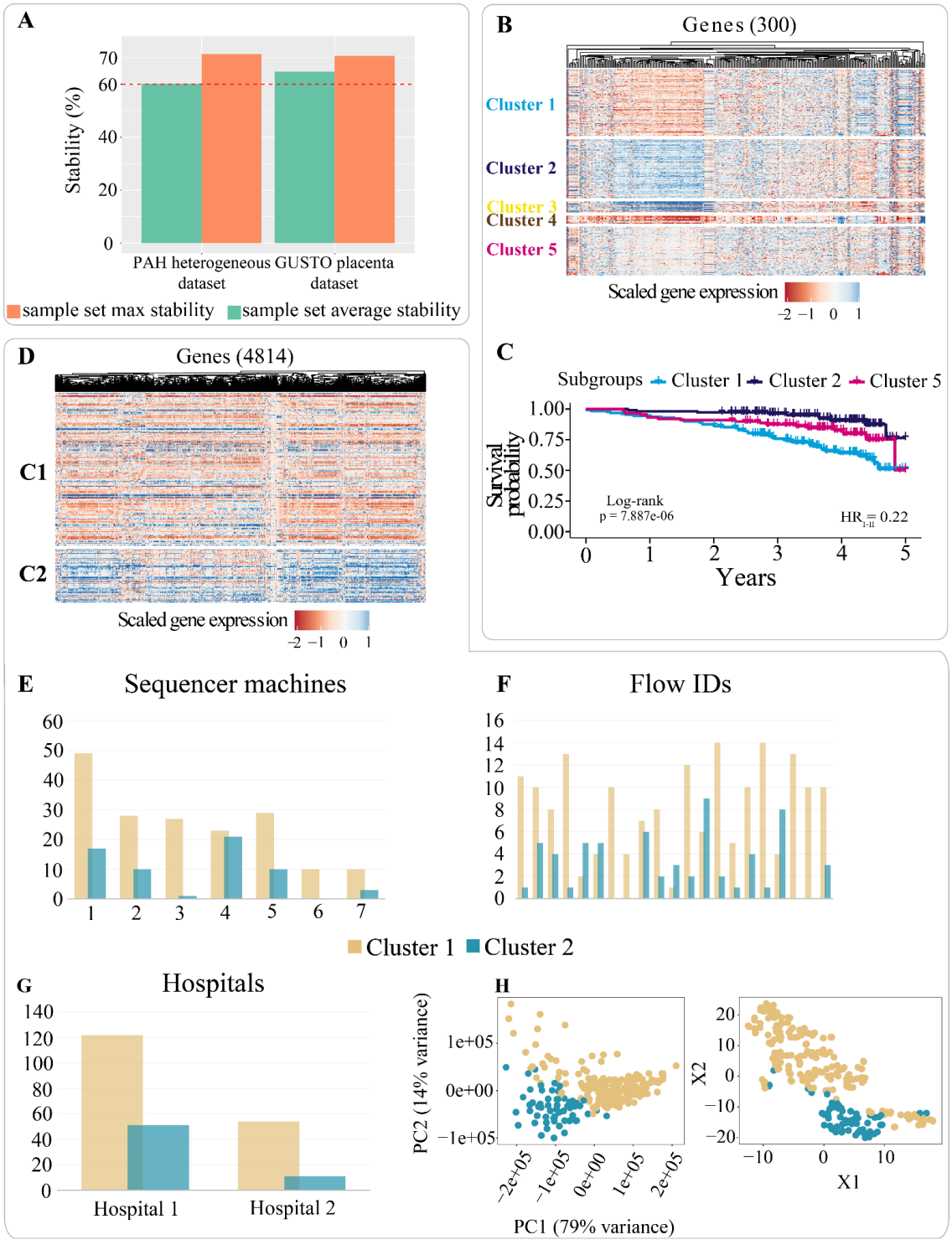


Figure 6: *Performance criteria for PAH and GUSTO datasets which have no known subgroups. Panel A) depicts the average and max sample set stabilities (percentage) for both datasets. The red dashed line represents the threshold of a stable clustering (60%). The PAH RNA-seq dataset contains expression of IPAH patients with panel B) showing the gene expression heatmap and C) survival profiles for discovered subgroups. The GUSTO dataset contains expression from healthy maternal whole blood with panel D) showing the gene expression heatmap. The following panels show the distribution of cluster members across E) sequencer machines(chi-square p-value 1.39e-03), F) flow IDs (chi-square p-value 2.55e-06) and G) hospitals where the data were collected (chi-square p-value 0.072). H) t-SNE and PCA plots of the expression profiles with labelling of the two discovered subgroups.*

Discussion

Our toolkit is designed to answer multiple questions arising during transcriptomic exploratory studies that target to uncover heterogeneity and subtypes within any condition that might be driven by expression changes. With the plurality of unsupervised methods available and their specialised nature, the selection of the most appropriate approach is a multi-factor problem. A lot of technical decisions are required in the procedure starting with a dataset and completed with a meaningful set of subgroups. To assist with this problem, our toolkit initially assesses the potential of a target dataset and provides estimates of the most appropriate method, gene set and number of subgroups finally outputting a partition based on optimised parameters unburdening the user of specialised decision making (Figure 1). All individual tools are computed internally and therefore do not require prior deep knowledge of machine learning by the user. All results, intermediate and final, are observable and each step is justified by multiple measures and indices representing widely used machine learning techniques.

Applying unsupervised learning on expression datasets is often not a straightforward task as it contains the element of uncertainty mainly introduced by the lack of knowledge on the data points. No methods or metrics can give a definitive answer to the main clustering questions, as presented in previous sections, therefore each tool has to be used with caution, i.e. determining the dataset clustering potential is an indication rather than a clear sign that partitioning the dataset will yield informational subgroups. Additionally, clustering can often contain non-deterministic steps allowing for each function to behave slightly differently between similar runs. To reduce the uncertainty and provide a reliable set of tools, this toolkit has been applied on various gene expression datasets where its efficiency has been demonstrated. However, it is important to note that despite the use of multiple methodological approaches within this toolkit the inherent exploratory characteristics of clustering do not allow for clusters of definite value, instead they are meant to be dealt with scientific caution and biological validation. Aside from the actual memberships, the functions in this package can also reveal useful information about the input data. The GUSTO RNA-seq dataset showed that biases can be discovered by applying simple tests, such as PCA or tSNE, in conjunction with the cluster members. It is also possible for Omada to hint towards the existence of a single class, and therefore no heterogeneity, by consistently revealing low partition agreement and stability scores across multiple functions, as demonstrated in our single-class dataset example. Furthermore, Omada can help in selecting a small group of genes with potential partitioning capabilities as the feature selection step is expected to greatly reduce the number of genes which in most cases count to thousands. This toolkit is currently based on specific clustering techniques and metrics but its modular nature

allows its extension to accommodate different data types that come in the form of continuous numeric values such as microRNA, metabolite or single cell RNA datasets. Furthermore, the structure of this toolkit allows for additional approaches to be added in the future to the pool of clustering algorithms to be tested keeping up with the current state of the art techniques.

Code availability

Code will be available on github and as a bioconductor software package (Omada) at 10.18129/B9.bioc.omada.

Data availability

The expression datasets used in this work can be accessed through the following sources: The two simulated, by Omada, datasets (single and multi-class) can be accessed and downloaded at https://github.com/BioSok/OmadaSimulatedDatasets. The Pan cancer tissue expression data can be accessed through (https://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq). The transcriptomic data used in this study can be accessed through the EGA (the European Genome-phenome Archive) database under accession code EGAS00001005532. In compliance with the Ethics under which these data and samples have been collected, the transcriptomic data are available through restricted access for approved researchers who agree to the conditions of use, i.e. keeping it secure and only using it for approved purposes. To apply for access please contact cohortcoordination@medschl.cam.ac.uk. You will receive an application form within 30 days. The 'UK National PAH Cohort Study Data Access Committee' will review requests within 3 months of receipt of the completed application form and if approved, provide details for access to the RNAseq data stored at the EGA. All requesters must agree to the data access conditions found in EGA. The data used to generate statistics, plots and figures are accessible through our interactive portal found in https://sheffield-university.shinyapps.io/ipah-rnaseq-app/. The GUSTO expression dataset is available in NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under the accession numbers GSE182409 (Corresponding Reviewer token number: qjolmmeudnofnsv).

Acknowledgements

The UK National Cohort of Idiopathic and Heritable PAH is supported by grants from the British Heart Foundation (SP/12/12/29836 & SP/18/10/33975) and the UK Medical Research Council (MR/K020919/1). Additional samples from the Sheffield Teaching Hospitals Observational Study of Pulmonary Hypertension, Cardiovascular and other Respiratory Diseases were supported by British Heart Foundation (PG/11/116/29288). S.K. is supported by a Donald Heath Ph.D. Studentship award and A*STAR Research Attachment Programme (ARAP) award.

The GUSTO study group includes Allan Sheppard, Amutha Chinnadurai, Anne Eng Neo Goh, Anne Rifkin-Graboi, Anqi Qiu, Arijit Biswas, Bee Wah Lee, Birit Froukje Philipp Broekman, Boon Long Quah, Chai Kiat Chng, Cheryl Shufen Ngo, Choon Looi Bong, Christiani Jeyakumar Henry, Daniel Yam Thiam Goh, Doris Ngiuk Lan Loh, Fabian Kok Peng Yap, George Seow Heong Yeo, Helen Yu Chen, Hugo P. S. van Bever, Iliana Magiati, Inez Bik Yun Wong, Ivy Yee-Man Lau, Jeevesh Kapur, Jenny L. Richmond, Jerry Kok Yen Chan, Joanna Dawn Holbrook, Johan G. Eriksson, Joshua J. Gooley, Keith M. Godfrey, Kenneth Yung Chiang Kwek, Kok Hian Tan, Krishnamoorthy Naiduvaje, Leher Singh, Lin Lin Su, Lourdes Mary Daniel, Lynette Pei-Chi Shek, Marielle V. Fortier, Mark Hanson, Mary Foong-Fong Chong, Mary Rauff, Mei Chien Chua, Michael J. Meaney, Mya Thway Tint, Neerja Karnani, Ngee Lek, Oon Hoe Teoh, P. C. Wong, Peter David Gluckman, Pratibha Keshav Agarwal, Rob Martinus van Dam, Salome A. Rebello, Seang Mei Saw, Shang Chee Chong, Shirong Cai, Shu-E Soh, Sok Bee Lim, Stephen Chin-Ying Hsu, Victor Samuel Rajadurai, Walter Stunkel, Wee Meng Han, Wei Wei Pang, Yap Seng Chong, Yin Bun Cheung, Yiong Huak Chan and Yung Seng Lee.

Author contributions

SK and DW conceived the tools. SK undertook computational work and drafted the work with DW. AL, CR, TPF and MW participated in the data acquisition of the work. All authors revised it critically for important intellectual content; and gave final approval of the version submitted for publication.

References (Manuscript 1)

1. Yu, N. Y.-L. *et al.* Complementing tissue characterization by integrating transcriptome profiling from the Human Protein Atlas and from the FANTOM5 consortium. *Nucleic Acids Res.* **43**, 6787–6798 (2015).

2. Keen, J. C. & Moore, H. M. The Genotype-Tissue Expression (GTEx) Project: Linking Clinical Data with Molecular Analysis to Advance Personalized Medicine. *J Pers Med* **5**, 22–29 (2015).

3. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).

4. Wang, L. *et al.* RNA sequencing-based longitudinal transcriptomic profiling gives novel insights into the disease mechanism of generalized pustular psoriasis. *BMC Med. Genomics* **11**, 52 (2018).

5. Neff, R. A. *et al.* Molecular subtyping of Alzheimer's disease using RNA sequencing data reveals novel mechanisms and targets. *Sci Adv* **7**, (2021).

6. Saeidian, A. H., Youssefian, L., Vahidnezhad, H. & Uitto, J. Research Techniques Made Simple: Whole-Transcriptome Sequencing by RNA-Seq for Diagnosis of Monogenic Disorders. *J. Invest. Dermatol.* **140**, 1117–1126.e1 (2020).

7. Tran, H. T. N. *et al.* A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.* **21**, 12 (2020).

8. Xing, Q. R. *et al.* Unraveling Heterogeneity in Transcriptome and Its Regulation Through Single-Cell Multi-Omics Technologies. *Front. Genet.* **11**, 662 (2020).

9. Firth, A. L., Mandel, J. & Yuan, J. X.-J. Idiopathic pulmonary arterial hypertension. *Dis. Model. Mech.* **3**, 268–273 (2010).

10. Koirala, B. *et al.* Heterogeneity of Cardiovascular Disease Risk Factors Among Asian Immigrants: Insights From the 2010 to 2018 National Health Interview Survey. *J. Am. Heart Assoc.* **10**, e020408 (2021).

11. Rivera-Andrade, A. & Luna, M. A. Trends and heterogeneity of cardiovascular disease and risk factors across Latin American and Caribbean countries. *Prog. Cardiovasc. Dis.* **57**, 276–285 (2014).

12. Manchia, M., Cullis, J., Turecki, G., Rouleau, G. A. & Uher, R. The impact of phenotypic and genetic heterogeneity on results of genome wide association studies of complex diseases. *PLoS One* (2013).

13. Vidman, L., Källberg, D. & Rydén, P. Cluster analysis on high dimensional RNA-seq data with applications to cancer research - An evaluation study. *PLoS One* **14**, e0219102 (2019).

14. Ren, Z., Wang, W. & Li, J. Identifying molecular subtypes in human colon cancer using gene expression and DNA methylation microarray data. *Int. J. Oncol.* **48**, 690–702 (2016).

15. Sotiriou, C., Neo, S. Y. & McShane, L. M. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proceedings of the* (2003).

16. Lapointe, J. *et al.* Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 811–816 (2004).

17. Wu, F. *et al.* Single-cell profiling of tumor heterogeneity and the microenvironment in advanced non-small cell lung cancer. *Nat. Commun.* **12**, 2540 (2021).

18. Kariotis, S. *et al.* Biological heterogeneity in idiopathic pulmonary arterial hypertension identified through unsupervised transcriptomic profiling of whole blood. *Nat. Commun.* **12**, 7104 (2021).

19. Xu, D. & Tian, Y. A Comprehensive Survey of Clustering Algorithms. *Annals of Data Science* vol. 2 165–193 Preprint at https://doi.org/10.1007/s40745-015-0040-1 (2015).

20. Reddy, C. K. & Vinzamuri, B. A Survey of Partitional and Hierarchical Clustering Algorithms. *Data Clustering* 87–110 Preprint at https://doi.org/10.1201/9781315373515-4 (2018).

21. Jamail, I. & Moussa, A. Current State-of-the-Art of Clustering Methods for Gene Expression Data with RNA-Seq. in *Applications of Pattern Recognition* (IntechOpen, 2020).

22. Ezugwu, A. E. *et al.* A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Eng. Appl. Artif. Intell.* **110**, 104743 (2022).

23. Cao, Y., Geddes, T. A., Yang, J. Y. H. & Yang, P. Ensemble deep learning in bioinformatics. *Nature Machine Intelligence* **2**, 500–508 (2020).

24. Park, C., Took, C. C. & Seong, J.-K. Machine learning in biomedical engineering. *Biomed Eng Lett* **8**, 1–3 (2018).

25. Choy, G. *et al.* Current Applications and Future Impact of Machine Learning in Radiology. *Radiology* **288**, 318–328 (2018).

26. Stafford, I. S. *et al.* A systematic review of the applications of artificial intelligence and machine learning in autoimmune diseases. *NPJ Digit Med* **3**, 30 (2020).

27. Hulsen, T. *et al.* From Big Data to Precision Medicine. *Front. Med.* **6**, 34 (2019).

28. Wang, X., Williams, C., Liu, Z. H. & Croghan, J. Big data management challenges in health research—a literature review. *Brief. Bioinform.* **20**, 156–167 (2019).

29. Nayyar, A., Gadhavi, L. & Zaman, N. Machine learning in healthcare: review, opportunities and challenges. *Machine Learning and the Internet of Medical Things in Healthcare* 23–45 Preprint at https://doi.org/10.1016/b978-0-12-821229-5.00011-2 (2021).

30. Gaba, D. & Mittal, N. 2. Implementation and classification of machine learning algorithms in healthcare informatics: approaches, challenges, and future scope. *Computational Intelligence for Machine Learning and Healthcare Informatics* 21–34 Preprint at https://doi.org/10.1515/9783110648195-002 (2020).

31. Wang, C., Gao, X. & Liu, J. Impact of data preprocessing on cell-type clustering based on single-cell RNA-seq data. *BMC Bioinformatics* **21**, 440 (2020).

32. Eijssen, L. M. T. *et al.* User-friendly solutions for microarray quality control and pre-processing on ArrayAnalysis.org. *Nucleic Acids Res.* **41**, W71–6 (2013).

33. Andrews, S. & Others. FastQC: a quality control tool for high throughput sequence data. Preprint at (2010).

34. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* vol. 30 2114–2120 Preprint at https://doi.org/10.1093/bioinformatics/btu170 (2014).

35. D'haene & Hellemans. The importance of quality control during qPCR data analysis. *Int. Drug Discov.*

36. Baccarella, A., Williams, C. R., Parrish, J. Z. & Kim, C. C. Empirical assessment of the impact of sample number and read depth on RNA-Seq analysis workflow performance. *BMC Bioinformatics* **19**, 423 (2018).

37. Wang, J., Yue, S., Yu, X. & Wang, Y. An efficient data reduction method and its application to cluster analysis. *Neurocomputing* **238**, 234–244 (2017).

38. Hennig, C. Cluster-wise assessment of cluster stability. *Comput. Stat. Data Anal.* **52**, 258–271 (2007).

39. Hartigan, J. A. *Clustering Algorithms*. (John Wiley & Sons, Inc., 1975).

40. Dhillon, I. S. *A Unified View of Kernel K-means, Spectral Clustering and Graph Cuts*. (Computer Science Department, University of Texas at Austin, 2004).

41. Ng, A. Y., Jordan, M. I. & Weiss, Y. On Spectral Clustering: Analysis and an algorithm. in *Advances in Neural Information Processing Systems 14* (eds. Dietterich, T. G., Becker, S. & Ghahramani, Z.) 849–856 (MIT Press, 2002).

42. Rand, W. M. Objective Criteria for the Evaluation of Clustering Methods. *J. Am. Stat. Assoc.* **66**, 846–850 (1971).

43. Rodriguez, M. Z. *et al.* Clustering algorithms: A comparative approach. *PLoS One* **14**, e0210236 (2019).

44. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* vol. 20 53–65 Preprint at https://doi.org/10.1016/0377-0427(87)90125-7 (1987).

45. Polikar, R. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine* **6**, 21–45 (2006).

46. Calinski, T. & Harabasz, J. A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods* vol. 3 1–27 Preprint at https://doi.org/10.1080/03610927408827101 (1974).

47. Dunn†, J. C. Well-Separated Clusters and Optimal Fuzzy Partitions. *Journal of Cybernetics* vol. 4 95–104 Preprint at https://doi.org/10.1080/01969727408546059 (1974).

48. Pakhira, M. K., Bandyopadhyay, S. & Maulik, U. Validity index for crisp and fuzzy clusters. *Pattern Recognition* vol. 37 487–501 Preprint at https://doi.org/10.1016/j.patcog.2003.06.005 (2004).

49. Kendall, M. G. A New Measure of Rank Correlation. *Biometrika* vol. 30 81 Preprint at https://doi.org/10.2307/2332226 (1938).

50. Baker, F. B. & Hubert, L. J. Measuring the Power of Hierarchical Cluster Analysis. *Journal of the American Statistical Association* vol. 70 31–38 Preprint at https://doi.org/10.1080/01621459.1975.10480256 (1975).

51. Hubert, L. J. & Levin, J. R. A general statistical framework for assessing categorical clustering in free recall. *Psychological Bulletin* vol. 83 1072–1080 Preprint at https://doi.org/10.1037//0033-2909.83.6.1072 (1976).

52. Davies, D. L. & Bouldin, D. W. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence* vol. PAMI-1 224–227 Preprint at https://doi.org/10.1109/tpami.1979.4766909 (1979).

53. McClain, J. O. & Rao, V. R. CLUSTISZ: A program to test for the quality of clustering of a set of objects. *J. Mark. Res.* **12**, 456–460 (1975).

54. Halkidi, M., Batistakis, Y. & Vazirgiannis, M. On Clustering Validation Techniques. *J. Intell. Inf. Syst.* **17**, 107–145 (2001).

55. Ray, S. & Turi, R. H. Determination of number of clusters in k-means clustering and application in colour image segmentation. in *Proceedings of the 4th international conference on advances in pattern recognition and digital techniques* 137–143 (Citeseer, 1999).

56. Rohlf, F. J. Methods of Comparing Classifications. *Annual Review of Ecology and Systematics* vol. 5 101–113 Preprint at https://doi.org/10.1146/annurev.es.05.110174.000533 (1974).

57. Halkidi, M. & Vazirgiannis, M. Clustering validity assessment: finding the optimal partitioning of a data set. *Proceedings 2001 IEEE International Conference on Data Mining* Preprint at https://doi.org/10.1109/icdm.2001.989517.

58. Song, Y. Class compactness for data clustering. in *2010 IEEE International Conference on Information Reuse & Integration* 86–91 (IEEE, 2010).

59. cBioPortal for Cancer Genomics. https://www.cbioportal.org/datasets.

60. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics* vol. 45 1113–1120 Preprint at https://doi.org/10.1038/ng.2764 (2013).

61. Kim, P. *et al.* TissGDB: tissue-specific gene database in cancer. *Nucleic Acids Res.* **46**, D1031–D1038 (2018).

62. Kariotis, S. & Jammeh, E. *BioSok/spectral_clustering_of_IPAH: v1.0.1*. (2021). doi:10.5281/zenodo.5549872.

63. Pan, H. *et al.* Integrative Multi-Omics database (iMOMdb) of Asian Pregnant Women. *Hum. Mol. Genet.* (2022) doi:10.1093/hmg/ddac079.

64. Tibshirani, R. & Walther, G. Cluster Validation by Prediction Strength. *J. Comput. Graph. Stat.* **14**, 511–528 (2005).

2.4 Supplementary information from Manuscript 1

Omada: Robust clustering of transcriptomes through multiple testing

Kariotis et al

Supplementary methods

To determine the significance of the difference between simulated distributions, used during the dataset simulation, the Kolmogorov's D statistics are shown in Supplementary Table 1.

Supplementary results

All results generated from our tools application on the available datasets can be found in Supplementary Table 2 (simulated single-class dataset), Supplementary Table 4 (simulated multi-class dataset), Supplementary Table 5 (pancan dataset), Supplementary Table 7 (I/PAH dataset) and Supplementary Table 8 (GUSTO dataset). Additionally, the confusion matrices for simulated multi-class and pancan datasets, for which we know the actual labels, are presented in Supplementary Table 6.

Supplementary tables

Supplementary Table 1 *| Kolmogorov's D statistic for a simulated dataset containing 5 clusters (A, B, C, D, E). Each distance D has a p-value of less than 2.2e-16*

AvsB	AvsC	AvsD	AvsE	BvsC	BvsD	BvsE	CvsD	CvsE	DvsE
0.99	0.99	1	1	0.84	0.97	0.97	0.65	0.88	0.54

Supplementary Table 2 *| Simulated single-class dataset results, composed of 100 samples and 100 genes drawn from a single distribution*

Tool	Parameters	Results
Dataset clustering feasibility	100 samples, 100 features	max stability = 0.55 average stability = 0.45
Clustering method selection	max clusters = 6 comparisons = 3	Spectral PA = **0.52** Kmeans PA = 0.03 Hierarchical PA = 0.26
Sample set selection	min(k) = 2 max(k) = 6 feature step = 20	Optimal features = 20 max stability = 0.51 average stability = 0.47
K estimation	min(k) = 2 max(k) = 6 method = spectral	Optimal k = 5

**max clusters: the maximum number of clusters to be tested starting from 2, range [2, max clusters]*
**min(k): minimum number of clusters to be tested*
**max(k): maximum number of clusters to be tested*
**feature step: the number of features by which the generated datasets grow. Also, the smallest dataset to be tested*

Supplementary Table 3 *| The scores of all internal indexes used to decide on the ensemble voting of the number of clusters for the multi and single class simulated datasets. The scores for the most voted k are presented for each dataset along with the ideal score (min/max) for each index*

	Multi Class (k=5)	One Class (k=6)	Ideal
Calinski-Harabasz	608.5824	5.325228	max
Dunn	0.672101	0.373477	max
Pbm	740.0956	0.709396	max
Tau	0.50501	0.192052	max
Gamma	0.896494	0.335348	max
C index	0.01731	0.324603	min
Davies–Bouldin	0.993194	2.897971	min
Mcclain rao	0.303852	0.907469	min
sd_dis	0.233813	0.479863	min
Ray–Turi	0.301727	2.20407	min
g_plus	0.016422	0.108995	min
Silhouette	0.479414	0.050476	max
s_dbw	0	0	min
Compactness	0	145.1075	max
Connectivity	7.512624	6.277666	max

***Supplementary Table 4** | Simulated multi-class dataset results, where distribution is represented by around 120 samples and 3 clusters are used as a default parameter*

Tool	Parameters	Results
Dataset clustering feasibility	359 samples, 300 features	max stability = 0.78 average stability = 0.72
Clustering method selection	max clusters = 6 comparisons = 3	Spectral PA = **0.56** Kmeans PA = 0.53 Hierarchical PA = 0.28
Sample set selection	min(k) = 2 max(k) = 6 feature step = 25	Optimal features = 300 max stability = 0.84 average stability = 0.78
K estimation	min(k) = 2 max(k) = 6 method = spectral	Optimal k = 5

**max clusters: the maximum number of clusters to be tested starting from 2, range [2, max clusters]*
**min(k): minimum number of clusters to be tested*
**max(k): maximum number of clusters to be tested*
**feature step: the number of features by which the generated datasets grow. Also, the smallest dataset to be tested*

***Supplementary Table 5** | Pancan multi-tissue RNA-seq dataset B results using 5 tumour classes, filtering genes with expression variance less than 5*

Tool	Parameters	Results
Dataset clustering feasibility	801 samples, 20531 features	max stability = 0.87 average stability = 0.70
Clustering method selection	max clusters = 5 comparisons = 3	Spectral PA = **0.60** Kmeans PA = 0.46 Hierarchical PA = 0.12
Sample set selection	min(k) = 2 max(k) = 6 feature step = 50	Optimal features = 350 max stability = 0.85 average stability = 0.85
K estimation	min(k) = 2 max(k) = 6 method = spectral	Optimal k = 3

**max clusters: the maximum number of clusters to be tested starting from 2, range [2, max clusters]*
**min(k): minimum number of clusters to be tested*
**max(k): maximum number of clusters to be tested*
**feature step: the number of features by which the generated datasets grow. Also, the smallest dataset to be tested*

Supplementary Table 6 *| The confusion matrix of cluster estimations and actual classes of samples for simulated multi-class and pancan dataset*

	Simulated multi-class dataset				
	Class 1	**Class 2**	**Class 3**	**Class 4**	**Class 5**
Cluster 1	72	0	0	0	0
Cluster 2	0	0	0	72	0
Cluster 3	0	0	72	0	0
Cluster 4	0	0	0	0	72
Cluster 5	0	71	0	0	0
	Pancan multi-tissue dataset				
	BRCA	**COAD**	**KIRC**	**LUAD**	**PRAD**
Cluster 1	152	0	0	0	0
Cluster 2	0	73	145	0	136
Cluster 3	53	1	1	2	0
Cluster 4	0	4	0	139	0
Cluster 5	95	0	0	0	0

Supplementary Table 7 *| RNA-seq iPAH/HPAH dataset results as shown in (Kariotis et al., 2021)*

Tool	**Parameters**	**Results**
Dataset clustering feasibility	359 samples, 25955 features	max stability = 0.74 average stability = 0.61
Clustering method selection	max clusters = 10 comparisons = 3	Spectral PA = **0.86** Kmeans PA = 0.17 Hierarchical PA = 0.57
Sample set selection	min(k) = 2 max(k) = 10 feature step = 50	Optimal features = 300 max stability = 0.73 Average stability = 0.61
K estimation	min(k) = 2 max(k) = 10 method = spectral	Optimal k = 5

**max clusters: the maximum number of clusters to be tested starting from 2, range [2, max clusters]*
**min(k): minimum number of clusters to be tested*
**max(k): maximum number of clusters to be tested*
**feature step: the number of features by which the generated datasets grow. Also, the smallest dataset to be tested*

Supplementary Table 8 *| Gestational diabetes dataset (GUSTO) results*

Step	Parameters	Results
Dataset clustering feasibility	238 samples, 24,070 features	max stability = 0.59 average stability = 0.56
Clustering method selection	max clusters = 5 comparisons = 3	Spectral PA = **0.61** Kmeans PA = 0.60 Hierarchical PA = 0.12
Sample set selection	min(k) = 2 max(k) = 6 feature step = 50	Optimal features = 50 max stability = 0.71 average stability = 0.65
K estimation	min(k) = 2 max(k) = 6 method = spectral	Optimal k = 2

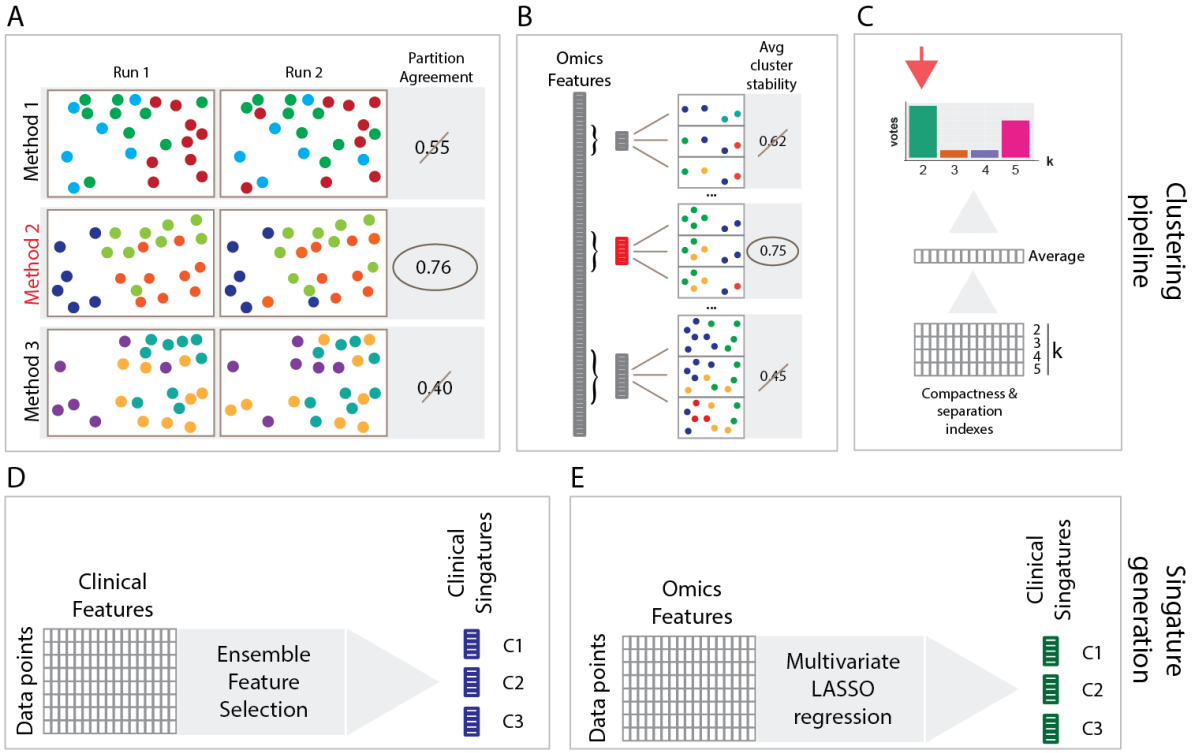
**max clusters: the maximum number of clusters to be tested starting from 2, range [2, max clusters]*
**min(k): minimum number of clusters to be tested*
**max(k): maximum number of clusters to be tested*
**feature step: the number of features by which the generated datasets grow. Also, the smallest dataset to be tested*

Supplementary Table 9 *| P-values of chi-square statistical and generalised linear regression analysis results for clinical variables and maternal phenotypes for the GUSTO dataset*

	Method	p-value
Hospital (2 locations)	chi-square	0.072
Time before/after 11am	chi-square	0.351
Sequenced location (2 locations)	chi-square	0.468
Sequenced machine (7 machines)	chi-square	1.39e-03
Sequenced Flow (22 flows)	chi-square	2.55e-06
Ethnicity (Chinese/Malay/Indian)	chi-square	0.15
Full/pre term	chi-square	0.53
Mother GDM	chi-square	1
Infant sex	chi-square	0.119
Pre-eclampsia	chi-square	0.153
ppBMI (under/normal/overweight)	chi-square	0.235
bookingBMI, WHO class(under/normal/overweight)	chi-square	0.235

Total GWG IOM (ppBMI) (Inadequate/Normal/Excessive)	chi-square	0.577
Total GWG IOM (bookBMI) (Inadequate/Normal/Excessive)	chi-square	0.063
Rate of GWG IOM (ppBMI) (Inadequate/Normal/Excessive)	chi-square	0.311
Rate of GWG IOM (bookBMI) (Inadequate/Normal/Excessive)	chi-square	0.425
Average maternal age	GLM	0.299
Average gestational weeks	GLM	0.604
Average ppBMI	GLM	0.379
Average bookingpBMI	GLM	0.539
Average total GWG (ppweight)	GLM	0.655
Average total GWG (bookingweight)	GLM	0.258
Average rate of GWG	GLM	0.294
Average EPDS	GLM	0.602
Average STAI state	GLM	0.25
Average STAI trait	GLM	0.195
Average fasting glucose	GLM	0.696
Average 2hr post glucose	GLM	0.762

Supplementary figures



***Supplementary Figure 1**: A) Clustering method selection based on the highest partition agreement between multiple differently parameterized runs B) Selection of the most cluster stable omics feature subset C) Estimating the optimal number of clusters for a dataset based on majority voting of several compactness and separation internal machine learning indexes D) Generating signature groups of clinical features/variables that drive each cluster by a multistep ensemble feature selection process utilising several machine learning classifiers E) Generating signature groups of omics features based on directional coefficients of a LASSO regression model*

Chapter 3 - Biological heterogeneity in idiopathic pulmonary arterial hypertension identified through unsupervised transcriptomic profiling of whole blood

3.1 Background

This piece of research focused on understanding the heterogeneity of idiopathic and heritable pulmonary arterial hypertension (IPAH) patients using RNA profiles observed in their blood. Our Omada toolkit was applied to a cohort of IPAH patients profiled by whole blood RNA-seq to identify what is differentiating them in a way that is agnostic to clinical characteristics. The work presented in this article attempts to answer four questions:

- *Can clustering identify stable patient subgroups for idiopathic PAH?*
- *How many distinct RNA-based subgroups of IPAH patients?*
- *Are any subgroups significantly associated with clinical outcomes?*
- *What is the relationship between gene and clinical signatures for RNA subgroups?*

The first part of this study is composed of the application of Omada to explore and partition the PAH patients while generating robust and reproducible results supported by reliable machine learning algorithms and metrics. We showed a robust methodology which we validated through supervised learning and interpreted biologically. The second part generated gene and clinical signatures with supervised approaches (classification and regression) which we combined and validated with additional expression datasets. Finally, to show the subgroup distinctions, we performed extensive biological downstream analyses which included the subjects of survival, disease risk, immune expression and differential cell composition as well as gene-clinical variable correlations.

3.2 Contribution

For this publication I was the first co-author from the conception of the study to the implementation and writing of the manuscript and supplementary. More specifically, I was the main contributor in writing the manuscript and supplementary, implementing the code and all unsupervised learning work as well as the analyses of the results in both main and supplementary documents. Sections/figures not generated by me are in italics and brackets in text specifying the author.

3.3 Manuscript 2

Biological heterogeneity in idiopathic pulmonary arterial hypertension identified through unsupervised transcriptomic profiling of whole blood

Sokratis Kariotis[1,2], Emmanuel Jammeh[1,2], Emilia M Swietlik[3], Josephine A. Pickworth[2], Christopher J Rhodes[4], Pablo Otero[4], John Wharton[4], James Iremonger[2], Mark J. Dunning[1], Divya Pandya[3], Thomas S Mascarenhas[1], Niamh Errington[1,2], A. A. Roger Thompson[2,5], Casey E. Romanoski[6], Franz Rischard[6], Joe G.N. Garcia[6], Jason X.-J. Yuan[7], Tae-Hwi Schwantes An[8], Ankit A. Desai[8], Gerry Coghlan[9], Jim Lordan[10], Prof Paul A. Corris[10,] Luke S Howard[4], Robin Condliffe[2,5], Prof David G. Kiely[2,5,11], Colin Church[12], Joanna Pepke-Zaba[13], Mark Toshner[3,13], Stephen Wort[4], Stefan Gräf [3], Prof Nicholas W Morrell[3], Prof Martin R Wilkins[4], Prof Allan Lawrie[2,11]*, Dennis Wang[1,14,15]*

On behalf of the NIHR BioResource – Rare Diseases (BRIDGE) PAH Consortium and the UK National PAH Cohort Study Consortium; *these authors jointly supervised this work.

[1] Department of Neuroscience, University of Sheffield, Sheffield UK, [2] Department of Infection, Immunity & Cardiovascular Disease, University of Sheffield, Sheffield, UK, [3] Department of Medicine, University of Cambridge, Cambridge, UK, [4] National Heart and Lung Institute, Imperial College London, London, UK, [5] Sheffield Pulmonary Vascular Disease Unit, Royal Hallamshire Hospital, Sheffield, UK, [6] Department of Cellular and Molecular Medicine, University of Arizona, Tucson, AZ, USA, [7] Department of Medicine, University of California, San Diego, La Jolla, CA, USA. [8] Department of Medicine, Indiana University, Indianapolis, IN, USA, [9] Royal Free Hospital, University College London, London, UK, [10] Newcastle University, Newcastle, UK, [11] Insigneo institute for in silico medicine, Sheffield, UK, [12] University of Glasgow, Glasgow, UK, [13] Royal Papworth Hospital, Cambridge, UK, [14] Department of Computer Science, University of Sheffield, Sheffield UK.,[15] Singapore Institute for Clinical Sciences, Singapore, Singapore.

Corresponding authors: a.lawrie@sheffield.ac.uk & dennis.wang@sheffield.ac.uk

Abstract

Idiopathic pulmonary arterial hypertension (IPAH) is a rare but fatal disease diagnosed by right heart catheterisation and the exclusion of other forms of pulmonary arterial hypertension, producing a heterogeneous population with varied treatment response. Here we show unsupervised machine learning identification of three major patient subgroups that account for 92% of the cohort, each with unique whole blood transcriptomic and clinical feature signatures. These subgroups are associated with poor, moderate, and good prognosis. The poor prognosis subgroup is associated with upregulation of the ALAS2 and downregulation of several immunoglobulin genes, while the good prognosis subgroup is defined by upregulation of the bone morphogenetic protein signalling regulator NOG, and the C/C variant of HLA-DPA1/DPB1 (independently associated with survival). These findings independently validated provide evidence for the existence of 3 major subgroups (endophenotypes) within the IPAH classification, could improve risk stratification and provide molecular insights into the pathogenesis of IPAH.

Introduction

Pulmonary arterial hypertension (PAH) is a rare but devastating disease characterised by sustained pulmonary vasoconstriction and progressive pulmonary vascular remodelling. This leads to an increase in pulmonary vascular resistance and pulmonary artery pressure, resulting in right heart failure and death[1]. The cause of idiopathic PAH (IPAH) remains unknown and diagnosis is derived from the exclusion of other forms of PAH, resulting in a heterogeneous group of patients who have significant differences in survival and treatment response across clinical cohort and registry studies[2,3,4,5].

The pathobiology of PAH involves the complex interaction of resident vascular cells, including endothelial cells, arterial smooth muscle cells and fibroblasts, with infiltrating inflammatory cells, and has been shown to be regulated by an ever growing number of molecular and genetic mechanisms[6,7,8]. We have identified both rare mutations[9] and common variants[10] in heritable and idiopathic PAH (H/IPAH) that have provided further insight into the genetic underpinning of PAH. Additional proteomic[11], metabolomic[12] and transcriptomic[13] studies have described diagnostic and prognostic biomarkers that add to our increasing understanding of the molecular mechanisms that regulate disease in this cohort. In Rhodes et al. we compared clinically defined H/IPAH cases to healthy controls and defined an imperfect

diagnostic signature for H/IPAH; however, we have not previously examined the molecular heterogeneity that exists within H/IPAH cases. Deep RNA profiling of blood samples have provided accessible biomarkers to detect rare diseases[14] and defined molecular mechanisms behind myocardial infarction[15]. We therefore investigated whether transcriptomic profiling of whole blood can provide more granular molecular 'endophenotypes' of H/IPAH to stratify patients better than is currently permissible with the standard clinical classification. Furthermore, we hypothesised that these transcriptome-defined subgroups would provide additional insights into biological pathways driving disease, and potential drug targets offering a route to precision medicine approaches for H/IPAH.

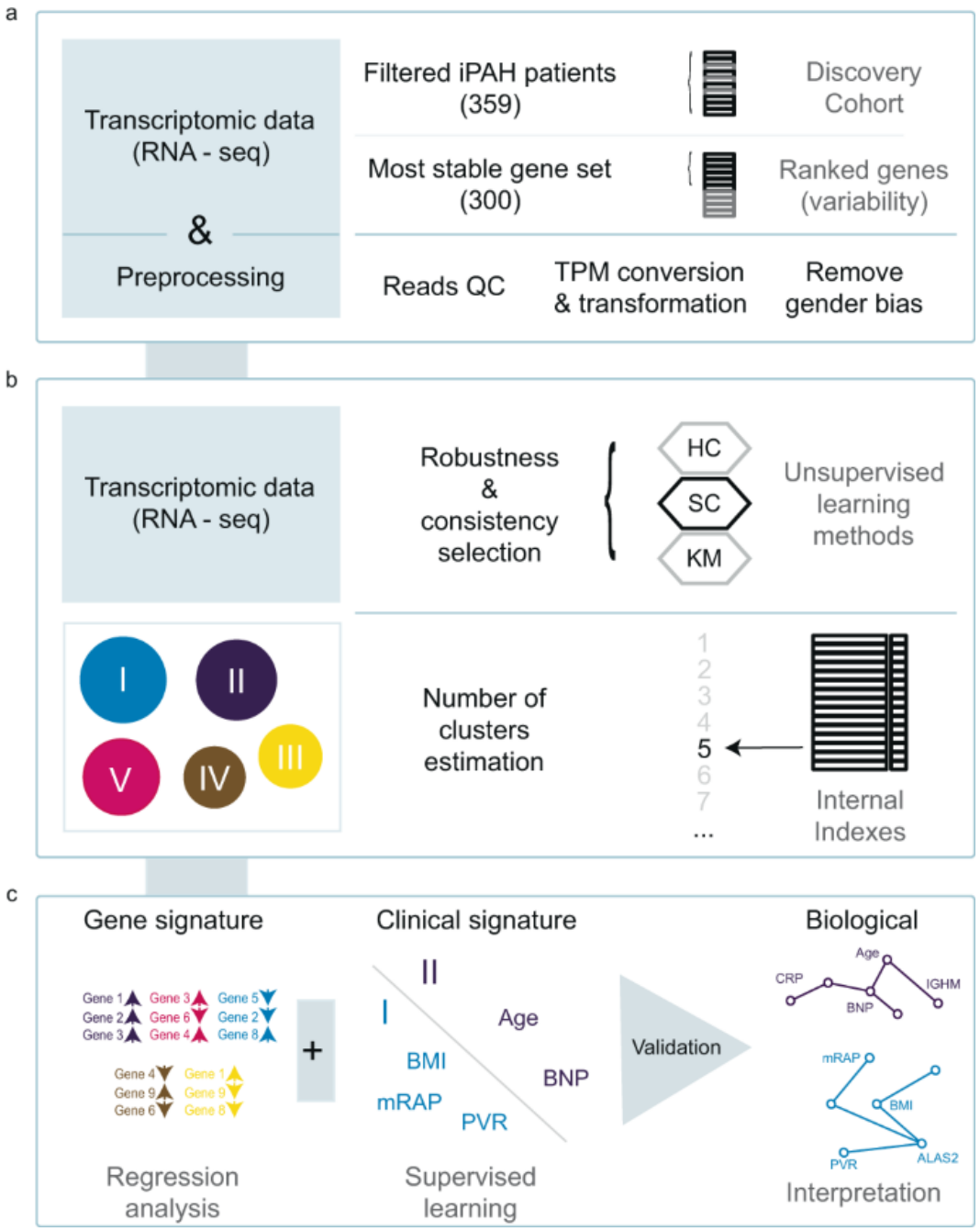
In this study, assessment of transcriptome patterns in whole blood was conducted using unsupervised machine learning agnostic to the clinical definitions and descriptors of H/IPAH. We describe the unbiased partitioning of patients into multiple distinct transcriptomic subgroups that associate with different survival properties, each with predictive clinical and genetic features. Specifically, we highlight the potential role of immunity and immune genes in discrimination of PAH endophenotypes associated with differential patient outcomes. These data further highlight the concept that inflammation is an important mediator of PAH pathogenesis[16,17,18,19,20,21,22] and the discovery of distinct immune subgroups from blood cytokine profiles of patients with PAH[16,17,18]. Finally, we identify a specific panel of clinical features that describe each transcriptomic subgroup and replicate these subgroups in a validation cohort who did not undergo full transcriptomic profiling using their clinical phenotype data. The gene expression profile of key cluster associated genes was subsequently confirmed, and the correlation with key clinical variables validated in both internal and external validation cohorts, thereby validating our approach, and providing an alternative method to define these endophenotypes without the need for transcriptomic data.

Results

Unsupervised cluster analysis of whole-blood transcriptomes reveals five distinct subgroups of H/IPAH

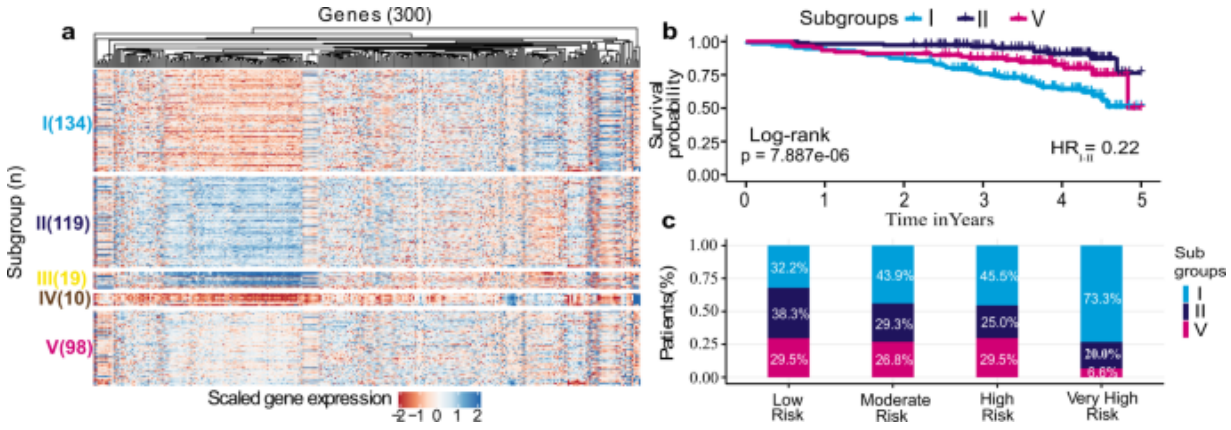
Whole blood samples from patients with H/IPAH (n = 359) were processed for RNA-sequencing as previously described[13]. Samples from 359 patients and 21 samples collected from a second time point underwent RNAseq data processing to reduce noise, and gene filtering to remove gender bias as sex chromosomes produced the highest variation in gene expression during clustering (**Supplementary Fig. 1**). Sample collection site did not produce any discernible effect on clustering (**Supplementary Figs. 2** and **3**). Simultaneously, the 300 genes that produced the most stable expression dataset were utilised to identify unique subgroups of gene expression profiles and describe the biological and clinical descriptors of these subgroups (**Fig. 1**). A clustering algorithm for selection and majority voting of multiple internal validation indexes (**Supplementary Data 1**) allowed us to identify as statistically optimal five distinct and stable subgroups of patients' profiles (**Fig. 2a**) while retaining the maximum heterogeneity information found in our dataset. The largest of the patient clusters identified was subgroup I (n = 129), which had poorer survival (53%, five-year median survival from sampling; **Fig. 2b**). The second largest, subgroup II (n = 112), demonstrated the best survival (78%, 5 years from sampling; Fig. 2b). Subgroup V (n = 89) demonstrated a mixed gene expression pattern and average survival outcome compared to subgroups I and II (**Fig. 2a, b**). Subgroups III (n = 19) and IV (n = 10) also demonstrated distinct gene expression patterns, with subgroup III most similar to subgroup II, and subgroup IV similar to subgroup I both in terms of gene expression level and survival outcomes. Due to the small size of subgroups III and IV (making statistical significance unattainable), we focused further characterisation of genetic and clinical correlates for subgroups I, II and V. The 33 HPAH patients in our PAH cohort showed an equal distribution (~10%) among the subgroups of our initial clustering (**Supplementary Table 1**), indicating that the inclusion of HPAH, or the small number of mis-classified patients, did not drive the partitioning procedure. An additional clustering pipeline exclusively utilising 313 samples from patients with IPAH (i.e. excluding those with HPAH, or re-classified PH) also showed five subgroups (Supplementary Fig. 4), where there were also a group of patients with poorer survival (clusters B and E, n = 149), a group with good survival (A and C, n = 109) and a group with moderate survival (D, n = 55).

Fig. 1: Overview of IPAH subgroup identification methodology.



a *A cohort of 359 IPAH patients and a set of 300 genes are selected for clustering based on RNA data quality and variability of expression across samples.* ***b*** *Spectral clustering of patients using expression values (TPM) was benchmarked against hierarchical clustering (HC) and k-means clustering (KM), and the optimal number of IPAH subgroups was selected based on internal indexes.* ***c*** *Associated gene expression and clinical features were identified and validated in independent cohorts.*

Fig. 2: Gene expression profiles, survival and risk categories that demonstrate five distinct subgroups.



***a** The expression heatmap for the five discovered subgroups showing distinct expression profiles. **b** Kaplan–Meier survival curves for the three predominant subgroups demonstrating the difference in survival profiles (from RNA sampling) for a span of 5 years along with two-sided log-rank test p values. **c** The percentage of predominant subgroups I, II and V patients across REVEAL risk categories. High- and very-high-risk populations mostly consist of subgroup I patients (45.5% and 73.3%, respectively), while the low-risk population is mostly composed of subgroup II (38.3%) and V (29.5%) patients. Fisher's exact test showed a statistically significant difference (two-sided p value = 0.024) between subgroups I and II for low- and very-high-risk categories.*

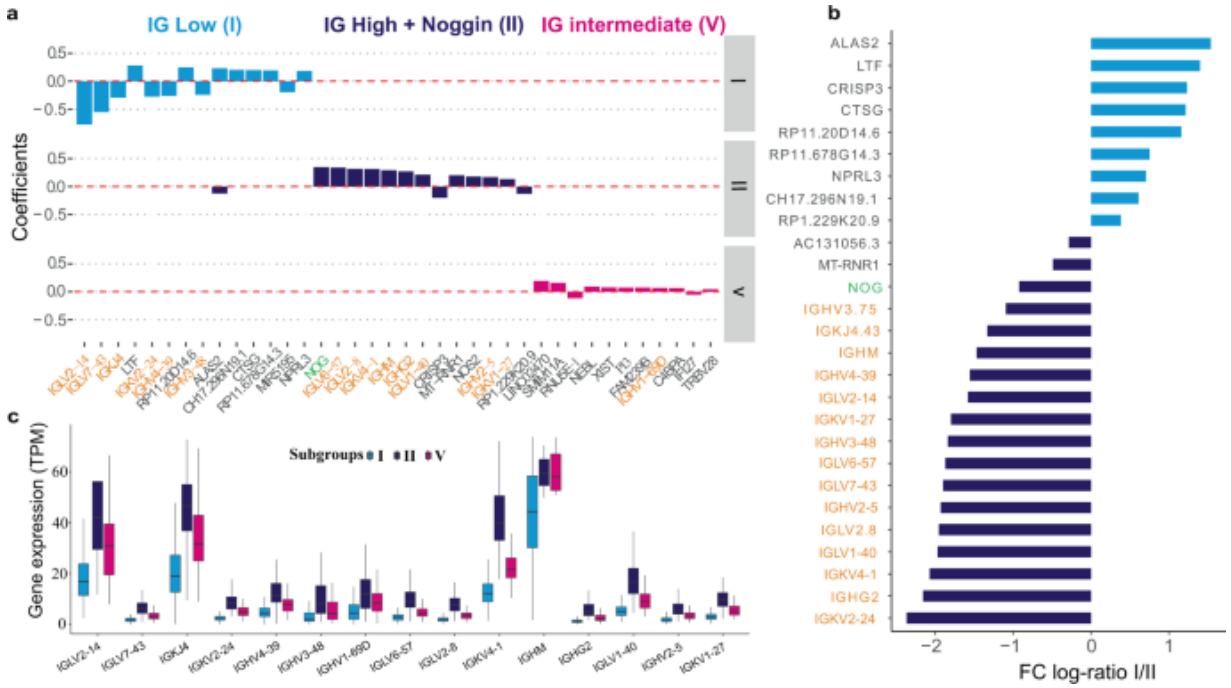
In order to determine whether the survival differences between the three main (largest) transcriptomic subgroups were also associated with disease severity in the surviving patients, we calculated the REVEAL 2.0 risk score[4] across all risk levels: low (n = 146), moderate (n = 41), high (n = 44) and very high (n = 15). Subgroup I which had the worst survival also had both the highest percentage of patients in high-risk categories (medium 43.9%, high 45.5% and very high 73.3%) and a lower percentage (32.2%) in the low-risk category (**Fig. 2c**). In contrast, subgroup II which had the best survival contained the largest group of low risk patients (38.3%), a proportion significantly different to subgroup I (z-test p = 0.01422, **Fig. 2c**). The distribution of subgroup V was uniform across the risk groups, except for a small proportion of very-high-risk patients (6.6%). Age and sex were also included as covariates with the subgroups in a Cox regression model. Age above 52 years (median) was significantly associated with poor survival (HR = 2.29) while gender showed no relationship with overall survival. Even with these covariates, subgroup I was still significantly associated with survival and was the biggest risk factor (HR = 3.83) for poor outcome (**Supplementary Fig. 5**). Within each subgroup, a small number of patients had a second time-point sample collected on average after 463 days. Patients with these longitudinal samples (n = 19) were found to either remain within their subgroup or transition from either subgroup I (poor prognosis) or II (good prognosis) to the moderate prognosis subgroup V (**Supplementary Fig. 6a**). Interestingly, no patient transitioned from subgroup II (best survival) directly to subgroup I (worst survival) or

vice versa over time, 9 patients changed through the moderate subgroup, while 12 stayed in the same subgroup. Additionally, no functional class changes observed with almost all samples belonging to functional class III. When including transcriptomes from healthy volunteers in our cluster analysis, the highest proportion of healthy volunteers (39.1%) grouped with subgroup II patients (better prognosis) (**Supplementary Fig. 6b**). To further investigate the defining characteristics of the three largest subgroups, we interrogated both their gene expression profiles and clinical features to define their endophenotype.

Relative expression of immunoglobulins define RNA-based subgroups of IPAH

We next interrogated the three largest RNA-based subgroups using a multivariate penalised regression to identify the relationship between gene expression profiles and each of the three subgroups. The most parsimonious model revealed 57 genes with measurable association to the subgroups. ALAS2 (erythroid ALA-synthase), a catalysing haeme biosynthesis enzyme, appeared in the signatures for both subgroups I and II, and was the most differentially expressed gene (>2-fold) between the two subgroups. Several immunoglobulin light chain genes (IGKV and IGLV) were key markers for the subgroups, and these were found to be either downregulated in subgroup I (poor prognosis) or upregulated in subgroup II (good prognosis; **Fig. 3a**). Other than immunoglobulins, Noggin, a bone morphogenetic protein 4 antagonist, and inhibitor of hypoxia-induced proliferation[23], was the gene with the highest positive regression coefficient for subgroup II, underlining its association with good prognosis. BMP antagonist Noggin and immunoglobulin genes associated with the good prognostic subgroup II were all downregulated by more than twofold in subgroup I (**Fig. 3b**), fitting with contemporary understanding of perturbed BMP and inflammatory signalling in PAH pathogenesis[16,21,24]. Across the three major subgroups, the relative expression level of immunoglobulins ranged from low, intermediate and high for subgroups I, V and II, respectively (**Fig. 3a, c**), while Noggin showed significantly higher expression in subgroup II (**Supplementary Fig. 7**).

Fig. 3: Genes associated with subgroups I (low survival), II (high survival) and V (intermediate survival).



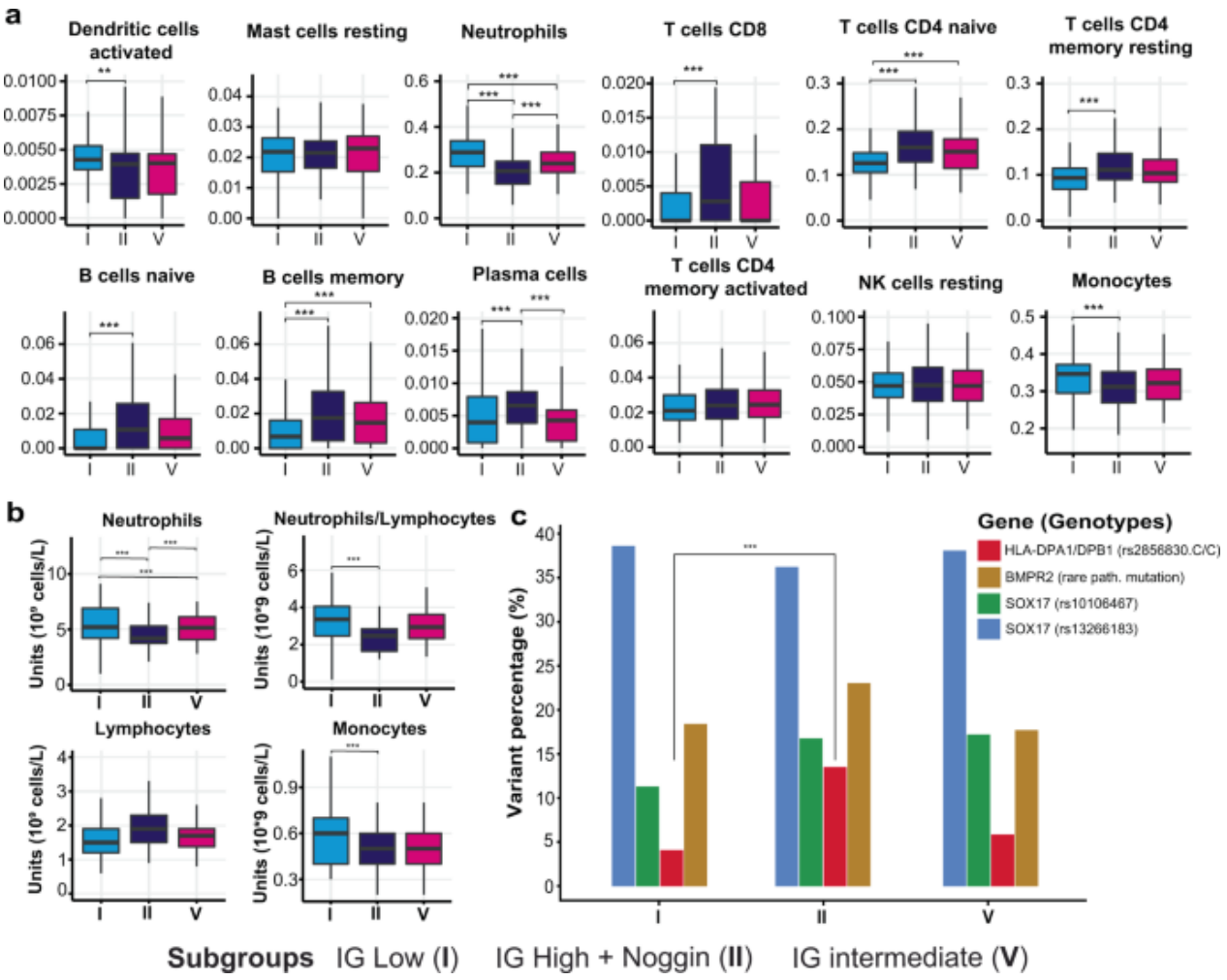
a *Genes with the highest 5% of LASSO coefficients across subgroups I, II and V.* ***b*** *Average expression fold change (log2 scaled) of the signature genes between subgroups I and II, with significance notations. Genes over-expressed in subgroup I are denoted by light blue bars while genes primarily expressed in subgroup II are represented by dark blue bars.* ***c*** *Expression level of immunoglobulin genes selected by LASSO across the three predominant subgroups with medians shown. Subgroups I (n = 134), V (n = 98) and II (n = 119) can be defined as having low, intermediate and high immunoglobulin characteristics. Vertical centre line represents the median, top and bottom bounds of the box represent the first and third quartile, while the tips of the whiskers represent min and max values.*

Differential immune cell composition between IPAH subgroups

To ascertain whether the large expression differences in immunoglobulin genes associated with subgroups I and II also corresponded to different levels of immune activity, we deconvoluted the RNA profiles to estimate the proportions of immune cell types in each sample. Significant differences ($p < 0.01$) in the proportion of lymphocytes and neutrophils were observed between samples in subgroup I and II (**Fig. 4a** and **Supplementary Fig. 8**). In particular, CD4/CD8 T cells and memory B cells were significantly more abundant in subgroup II where we observed upregulation of immunoglobulins. The lower proportion of lymphocytes (B cells and T cells) and higher proportion of neutrophils in the poor prognosis subgroup I was found to be statistically significant (**Supplementary Table 2**) and validated by clinical whole-blood cell counts (**Fig. 4b**). A higher neutrophil–lymphocyte ratio is known to

be an indicator of poor overall survival[25]. The differences observed in CD4 T cells and memory B cells may be due to changes in MHC class II antigen presentation genes, such as HLA-DP. We have previously identified the HLA-DPA1/DPB1 rs2856830 genotype to be strongly associated with survival in a large IPAH GWAS study, with the C/C homozygous genotype conferring increased survival compared with the T/T genotype, despite similar baseline disease severity[10]. Consistent with this genotype association with prognosis, we found that there was a significantly higher proportion of patients (p = 0.009) with the C/C genotype in subgroup II (good survival) compared with subgroup I (poor survival). This difference in variant frequencies between subgroups was not seen in known genetic risk factors for H/IPAH[9], including BMPR2 and SOX17 (**Fig. 4c**, **Supplementary Fig. 9** and **Supplementary Table 3**).

Fig. 4: Immunity cell composition across PAH transcriptomic subgroups.



a *CIBERSORT estimation of relative cell abundance in patients of subgroups I (n = 129), II (n = 112) and V (n = 89) using two-sided test and Bonferroni adjusted mean difference significance notation, p-values in Supplementary Table 7 .Vertical centre line represents the median, top and bottom bounds of the box represent the first and third quartile, while the tips of the whiskers represent min and max values.* ***b*** *Whole-blood cell counts across subgroups I*

(n = 129), II (n = 112) and V (n = 89) using two-sided test and Bonferroni adjusted mean difference significance notation. p-values in Supplementary Table 7. Vertical centre line represents the median, top and bottom bounds of the box represent the first and third quartile, while the tips of the whiskers represent min and max values. ***c*** *Proportion of patients in each subgroup with DNA variants in HLA-DPA1/DPB1 (rs2856830), SOX17 (rs10106467 and rs13266183, homozygous and heterozygous), BMPR2 (rare pathogenic variant). Notably, pI-II (HLA-DPA1/DPB1) = 0.009. Generated using a two-sample test for equality of proportions with continuity correction. *P value ≤ 0.05, **p value ≤ 0.01, ***p value ≤ 0.001. [The whole blood deconvolution was performed by Pablo Otero]*

Common clinical characteristics across RNA subgroups

Patients in this cohort were diagnosed at a median age of 45 years (IQR = 35–59 years) and sampled at a median age of 52 years (42–64) with an average of 5.3 years' time between diagnosis and sampling. As shown in Table 1, patients in subgroup I were significantly older (p value < 0.01) at 57 [45–70] years than the other subgroups. Consistent with the incidence rate of IPAH in the UK population[3], patients in the cohort were predominantly females (70%). Patients in the subgroups were also predominantly females with 62%, 73% and 70% in subgroups I, II and V, respectively. Across the whole cohort, 16.4% of patients presented positive pulmonary vasodilator response, 44.4% were in Functional Class (FC) III at sampling date with 6-minute walk distance (6MWD) of 387 m and a mean N-terminal (NT)-proBNP of 222.5 [78.9–1162.8] ng/ml. When the cohort was stratified, subgroup I had the highest proportion of FC III (50.4%), whereas subgroup II had the highest proportion of patients for FC I and II (16.5% and 41.3%, respectively, p value = 0.013). The lowest 6MWD (median = 327 m, p < 0.01) and the highest N-terminal (NT)-proBNP was (median = 345.0 ng/ml, p = 0.055) were observed in patients from subgroup I (poorest survival group). Diagnostic RHC across the cohort showed mean pulmonary arterial pressure (mPAP) was 54 (46–61) mmHg, pulmonary arterial wedge pressure (PAWP) was 10 (7–12) mmHg and CO was 3.8 (3.0–4.9) l/min at diagnosis. The cohort at the time of sampling, 143 (40.2%) of the patients were FC II and 158 (44.4) FC III with a median 6MWD of 387 m, pulmonary vascular resistance (PVR) was 8.9 Wood units and an NT-proBNP 222.5 ng/ml suggestive of a slight improvement of disease phenotype in response to vasodilator therapy. The full demographics table can be found in **Supplementary Data 2**.

Table 1 Major clinical characteristics of the three main RNA subgroups in the discovery cohort (n = 359) at the time of sampling.

	Low-risk subgroup II (high immunoglobulin)	Intermediate-risk subgroup V (intermediate immunoglobulin)	High-risk subgroup I (low immunoglobulin)	All patients
n	112	89	129	359
Age, years	46 [37–56]	52 [41–62]	57 [45–70]	52 [42–64]
Age at diagnosis, years	41 [31–51]	46 [37–55]	52 [42–67]	47 [35–59]
Gender:Female	82 (73%)	69 (78%)	80 (62%)	253 (70%)
Vasoresponse	10 (21.7%)	6 (13.6%)	6 (16.2%)	23 (16.4%)
Treatments				
Phosphodiesterase 5 Inhibitors (PDE5i)	12 (15.4%)	16 (21.9%)	22 (21.8%)	53 (19.4%)
Endothelin receptor antagonist (ERA)	6 (7.69%)	13 (17.8%)	8 (7.92%)	33 (12.1%)
PDE5i & ERA combination	42 (53.8%)	30 (41.1%)	53 (52.5%)	134 (49.1%)
Prostacyclin therapy	3 (3.85%)	1 (1.37%)	3 (2.97%)	7 (2.56%)
Calcium channel blockers	15 (19.2%)	13 (17.8%)	14 (13.9%)	45 (16.5%)
WHO functional class				
I	18 (16.5%)	10 (11.2%)	6 (4.7%)	35 (9.8%)
II	45 (41.3%)	36 (40.4%)	44 (34.1%)	143 (40.2%)
III	43 (39.4%)	40 (44.9%)	65 (50.4%)	158 (44.4%)
IV	3 (2.8%)	3 (3.4%)	14 (10.9%)	20 (5.6%)
6-minute walking distance, m	397 [338–500]	420 [367–464]	327 [183–390]	387 [300–449]
NT-proBNP, ng/l	131.7 [54.5–362.0]	185.5 [76.3–463.5]	345.0 [91.0–1556.1]	222.5 [78.9–1162.8]
Forced expiratory volume [% predicted]	92 [82–101]	84 [72–98]	78 [66–98]	85 [68–100]
Forced vital capacity [% predicted]	101 (20)	99 (24)	93 (29)	97 (24)
Transfer factor of lung for carbon monoxide [% predicted]	93 [87–106]	97 [92–101]	88 [67–96]	94 [87–103]
Diagnostic Right Heart Catheter Study				
Mean pulmonary artery pressure, mmHg	47 [39–60]	52 [37–65]	56 [41–65]	51 [39–63]
Mean right atrial pressure, mmHg	8 [4–10]	8 [4–11]	11 [6–14]	9 [4–12]
Mean pulmonary arterial wedge pressure, mmHg	10 [7–12]	10 [8–13]	12 [10–14]	11 [8–13]
Cardiac Index, l/min/m^2	2.3 [1.6–2.8]	2.2 [1.7–2.4]	1.9 [1.5–2.5]	2.2 [1.6–2.5]
Pulmonary vascular resistance, Wood Units	8.1 [5.7–14.1]	15.0 [5.9–16.1]	8.4 [5.9–13.2]	8.9 [5.7–15.0]

Intervals describe first and third quartiles. Parentheses describe standard deviation (SD).

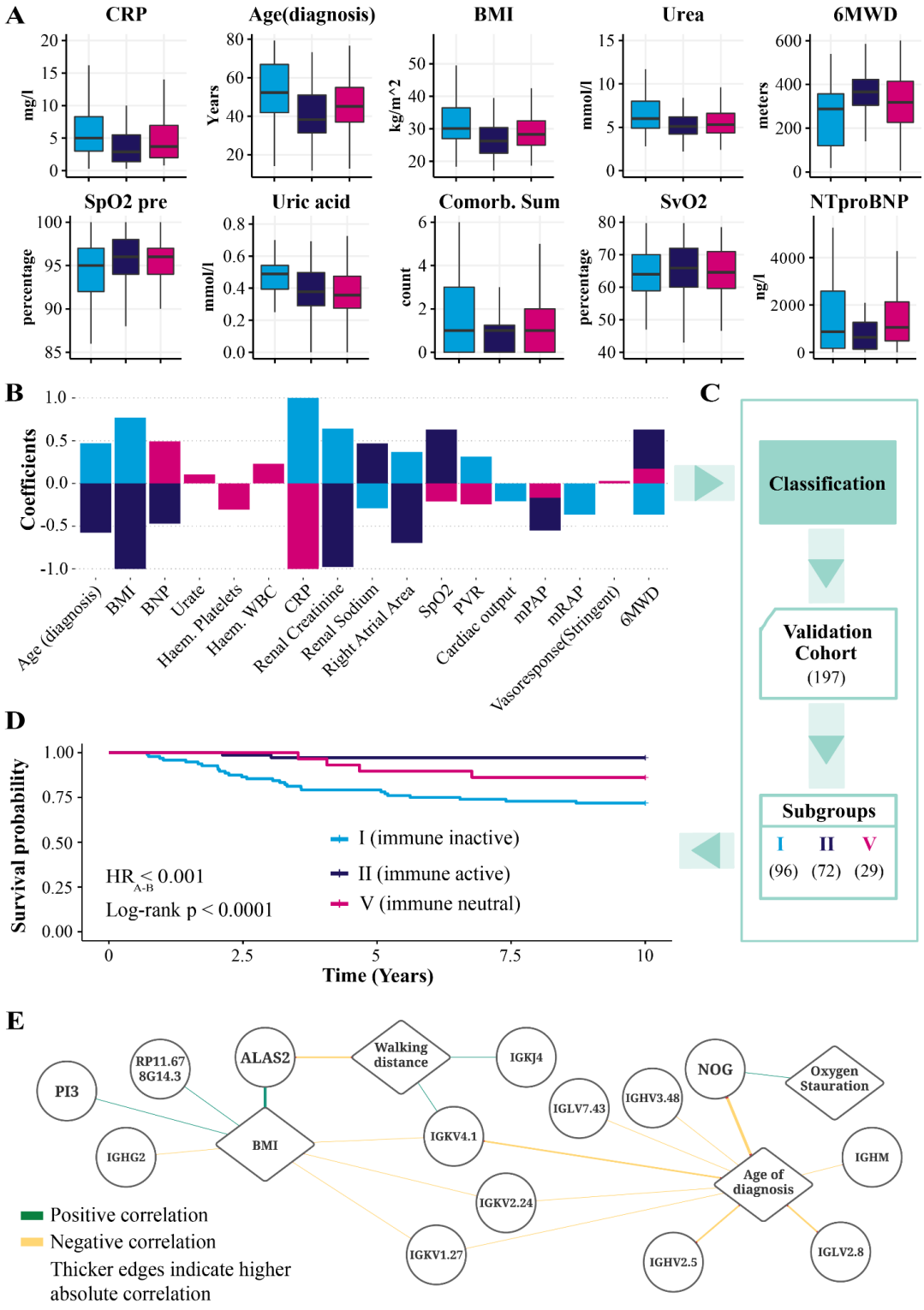
[Initial statistics generated by Emilia M Swietlik]

Clinical signatures describe RNA-based subgroups

Identification of specific clinical characteristics associated with each transcriptome-derived subgroup could explain how the gene expression patterns manifest into differences in patient outcome. We therefore used supervised machine learning with feature selection to identify the most important clinical features to describe the subgroups. The full list of clinical features used by the multivariate classifiers are described in Supplementary section 'Clinical features identification: Supervised learning' and in table format in **Supplementary Data 2**. Each clinical feature was assessed individually in a univariate model (**Fig. 5a** and **Supplementary Fig. 10**) and in combination with other features (multivariate model). Ensemble feature selection was used to identify reliable sets of clinical features that describe signatures for the RNA subgroups. The most important features in the signature for

subgroup I irrespective of feature selection method were C-reactive protein (CRP), creatinine, age of diagnosis, body mass index (BMI) and 6MWD. For subgroup II, the important features were CRP, creatinine, age of diagnosis, BMI and 6MWD, oxygen saturation (pre-6MWD) and right atrial area (RAA) (by echocardiography). CRP, 6MWD, urate, pulmonary vascular resistance (PVR), white blood cell count (WBC) and positive acute vasodilator challenge (at diagnostic right heart catheter) characterised subgroup IV.

Fig. 5: Clinical variables descriptive of RNA subgroups and used for classification of new patients.



a *Comparison of clinical variables deemed most important from our univariate feature selection model across subgroups I (n = 129), II (n = 112) and V (n = 89). Vertical centre line represents the median, top and bottom bounds of the box represent the first and third quartile, while the tips of the whiskers represent min and max values.* ***b*** *Clinical variables selected by ensemble feature selection from models predictive of each subgroup. Coefficients shown for each variable are from the most predictive support vector machine classifiers. [Source data generated by Emmanuel Jammeh]* ***c*** *Selected clinical features are*

used to classify 197 IPAH patients from an independent validation cohort. ***d*** *Kaplan–Meier survival curves per predicted subgroup in the validation cohort confirming the difference in survival outcomes between subgroups along with log-rank test p values.* ***e*** *Gene and clinical variable correlation network. Diamond nodes represent clinical variables drawn from the clinical signatures. Round nodes represent genes drawn from the gene signature generated by our LASSO model. Edges denoted Spearman rank correlation and have been thresholded to 0.25 and two-tailed test p value < 1.11 × 10−5. Additional P values found in Supplementary Table 8.*

CRP and 6MWD were the only clinical features present in signatures for subgroup I, II and V. Higher CRP was a marker for subgroup I, whereas lower levels indicated subgroups II and V. In contrast, 6MWD was negatively associated with subgroup I and positively with subgroups II and V. CRP showed a 37.19% increase in subgroup I compared to the average for subgroups II and V, 20.75% reduction in subgroup V compared to the average for subgroup I and II and 47.86% reduction in subgroup II compared to the average for subgroups I and V. 6MWD was 29.05% lower in subgroup I compared to the average for II and V, and increased by 7.63% in subgroup V compared to the average for II and I and 16.97% increase in subgroup II compared to the average for I and V. Five clinical features were present in signatures for subgroup I and II but had opposite coefficients (**Fig. 5b**). Higher age of diagnosis, BMI, RAA and creatinine are associated in subgroup I, whereas lower levels of those three features are associated with subgroup II. Subgroup I has 17.8% higher average age compared to the average for II and V and 21.2% lower in subgroup II compared to the average for I and V. BMI was 13.1% higher in subgroup I compared to the average for II and V, and 12.9% lower in subgroup II compared to the average for I and V. Additionally, creatinine was higher by 12.8% in subgroup I compared to the average for II and V and lower by 14.4% in subgroup II compared to the average for I and V. RAA was higher by 6.8% in subgroup I and lower by 6.3% in subgroup II. In contrast, there was a 27.6% reduction of renal sodium in subgroup I compared to the average for II and V, and 26.7% increase in subgroup II compared to the average for I and V.

Validation of clinical signatures on an independent cohort

To validate the relationship between clinical and gene features in the RNA subgroups, we used the clinical feature signatures of the subgroups to classify patients in an independent cohort of H/IPAH patients (n = 197) where whole-blood RNA profiling was not performed (**Fig. 5c**). Similar to the discovery cohort, patients were diagnosed at a median age of 52 years (IQR = 39–67) and 67% were female. In all, 17.7% of the patients showed positive pulmonary vasodilator response and the majority were categorised in Functional Class III (66%) with a 6MWD of 295 m (170–396) and

NT-proBNP of 796 ng/pl (128–1092). Their mPAP was 51 mmHg (42–57) and PAWP was 9 mmHg (6–11). The clinical features associated with RNA subgroups from the discovery cohort were used to classify this validation cohort. Our supervised approach identified three subgroups similar to our discovery cohort subgroups I, II and V (Table 2). These subgroups also displayed differences in their 10-year survival outcome from diagnosis (**Fig. 5d**). Those characterised as subgroup I based on their clinical features (corresponding to the low Noggin and immunoglobulin expression subgroups from RNAseq) (n = 96) demonstrated the lowest survival of 71% from the time of diagnosis. Subgroup V (corresponding to the immune neutral, intermediate RNAseq subgroup) (n = 31) also had an intermediate survival of 86%, while patients in Subgroup II (corresponding to the best surviving subgroup with upregulated Noggin and immunoglobulin genes) showed a very high survival rate of 97.2% (n = 96). These results provide key validation of the existence of endophenotypes for the three major subgroups of patients within the H/IPAH clinical classification group, and that these new subgroups can be identified using routinely collected clinical features associated with RNA dysregulation.

Table 2 Major clinical characteristics of the three subgroups within the validation cohort (n = 197) at time of diagnosis.

	Subgroup I	Subgroup II	Subgroup V	All patients
n	96	70	31	197
Age, years	65 [55–74]	40 [29–49]	45 [30–63]	54 [39–67]
Gender: female	56 (58%)	58 (83%)	18 (58%)	132 (67%)
Vasoresponse	7 (16.3%)	8 (21.6%)	2 (12.5%)	17 (17.7%)
Treatments				
Phosphodiesterase 5 inhibitors (PDE5i)	23 (29.5%)	10 (18.5%)	8 (29.6%)	41 (25.8%)
Endothelin receptor antagonist (ERA)	3 (3.85%)	5 (9.26%)	1 (3.70%)	9 (5.66%)
PDE5i & ERA combination	48 (61.5%)	29 (53.7%)	17 (63.0%)	94 (59.1%)
Prostacyclin agonist	1 (1.28%)	2 (3.70%)	1 (3.70%)	4 (2.52%)
Calcium channel blockers	3 (3.85%)	8 (14.8%)	0 (0.00%)	11 (6.92%)
WHO functional class				
I	4 (4.2%)	11 (15.7%)	2 (6.5%)	17 (8.6%)
II	22 (22.9%)	26 (37.1%)	15 (48.4%)	63 (32.0%)
III	60 (62.5%)	32 (45.7%)	12 (38.7%)	104 (52.8%)
IV	10 (10.4%)	1 (1.4%)	2 (6.5%)	13 (6.6%)
6-minute walking distance, m	306 (152)	419 (123)	409 (120)	360 (148)
NT-proBNP, ng/l	492 [196; 1327]	188 [90.0; 400]	266 [128; 499]	303 [128; 1092]
Forced expiratory volume [% predicted]	84 (21)	90 (19)	90 (17)	87 (20)
Forced vital capacity [% predicted]	94 (21)	98 (20)	99 (15)	96 (20)
Transfer factor of lung for carbon monoxide [% predicted]	92 (15)	98 (17)	96 (11)	95 (15)
Diagnostic Right Heart Catheter Study				
Mean pulmonary artery pressure, mmHg	50 [43; 57]	48 [42; 58]	49 [41; 57]	49 [42; 57]
Mean right atrial pressure, mmHg	9 [7; 12]	7 [5; 10]	6 [3; 7]	8 [5; 12]
Mean pulmonary Arterial wedge pressure, mmHg	10 (3)	8 (4)	8 (4)	9 (4)
Cardiac Index, l/min/m	2.0 [1.7; 2.5]	2.1 [1.7; 2.5]	2.0 [1.8; 2.4]	2.0 [1.7; 2.5]
Pulmonary vascular resistance, Wood Units	11 [7; 14]	11 [9; 15]	12 [10; 14]	11 [8; 15]

Intervals describe first and third quartiles. Parentheses describe standard deviation (SD). [Initial statistics generated by Emilia M Swietlik]

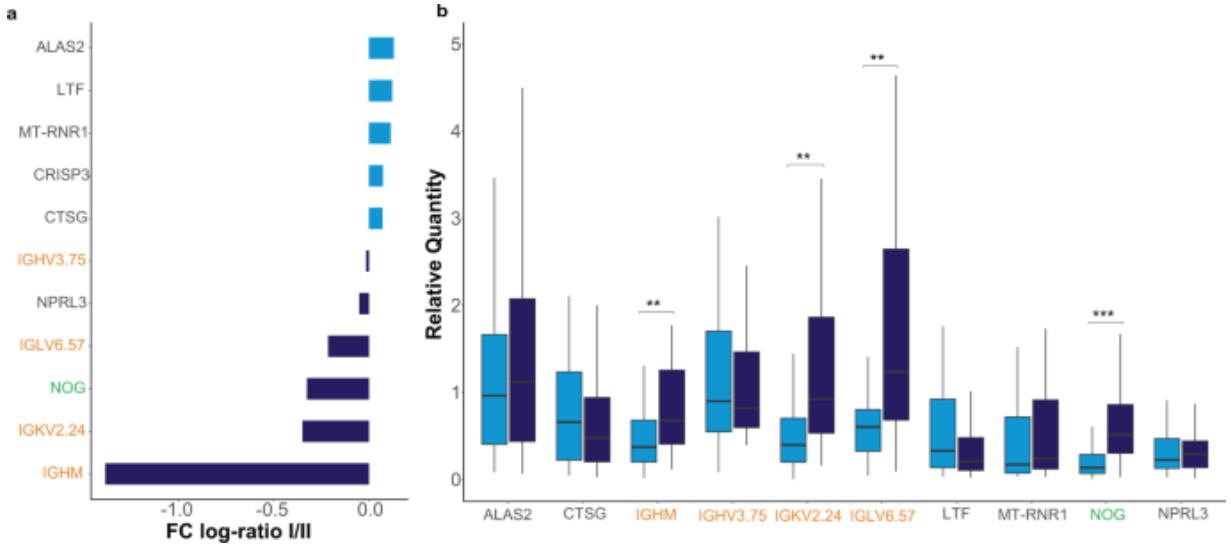
Clinical signatures are associated with subgroup-specific genes

We assessed the relationship between gene and clinical features of the subgroups by measuring the correlation between the most predictive features in both signatures. Immunoglobulins IGHV2.5, IGKV4.1, IGLV2.8 and IGHM (Spearman rho = −0.354, -0.342, −0.334, −0.297, respectively, p value <1.11 × 10−5) are negatively correlated with age of diagnosis (**Fig. 5e**). Indeed, we observed lower expression of immunoglobulin in poor prognosis subgroup I where there were older patients. Noggin was negatively correlated with age of diagnosis (rho = −0.443) but positively correlated with oxygen saturation (rho = 0.275). Interestingly, ALAS2 correlated most strongly with BMI (rho = 0.382) but showed an inverse correlation with 6MWD (rho = −0.323). This is consistent with our observations in the poor prognostic subgroup I where patients with higher expression of ALAS2 also had higher BMI and shorter walk distances. Genes negatively correlated with BMI included immunoglobulins (IGKV4.1, IGKV2.24 and IGKV1.27).

Gene expression in clinical-feature defined subgroups

Although the RNAseq whole transcriptome was not measured in this internal validation cohort, we compared gene expression differences between subgroups in this cohort using TaqMan PCR for 17 of the 27 genes (GAPDH used as the endogenous control gene) previously associated with the subgroups and/or clinical variable correlations. Nine of the 11 genes we measured demonstrated a fold change between subgroup I and II in the same direction as the discovery cohort (**Fig. 6a**). Differences in expression of key genes (IGHM, IGKV2.24, IGLV6.57 and NOG) were significant ($p < 0.01$) between subgroups I and II (**Fig. 6b** and **Supplementary Table 4**).

Fig. 6: Genes of interest with data based on our qPCR results of 91 patients (I = 53, II = 38) of the validation cohort.



a *Mean expression fold change (log2 scaled) of the signature genes between validation subgroup I (immune inactive) and II (immune active). The fold ratio was generated based on negative delta Ct values (vs GAPDH). Genes over-expressed in subgroup I are denoted by light blue bars while genes primarily expressed in subgroup II are represented by dark blue bars.* ***b*** *The relative quantity (RQ) of each gene of interest relative to GAPDH using a two-sided t-test with medians and significant differences shown with pI-II (IGHM) = 8.256 × 10−3, pI-II (IGKV2.24) = 2.373 × 10−3, pI-II (IGLV6.57) = 5.908 × 10−3 and pI-II (NOG) = 1.233733 × 10−4. **p < 0.01, ***p < 0.001. Vertical centre line represents the median, top and bottom bounds of the box represent the first and third quartile, while the tips of the whiskers represent min and max values.*

Clinical signatures are associated with subgroup-specific genes

The correlations between gene and clinical features observed in the discovery cohort were also examined in our validation cohort of 91 subjects, and also in an external cohort of 32 subjects with RNA collected from PBMCs[26]. We found that 64 of the 90 (71%) correlations measured in these two independent cohorts were consistent with our discovery cohort (**Supplementary Table 5**).

Discussion

In this study we describe a machine learning approach to identify transcriptome associated subgroups or endophenotypes of patients with heritable or idiopathic PAH. We defined five distinct clinical subgroups based on clinical presentation, severity and survival. The three largest subgroups displayed significantly different clinical characteristics, severity and survival outcomes suggesting that a molecular classification for PAH may be possible. We also identified patients that progressed through these subgroups over time with treatment and disease progression, the majority of which remaining within their subgroup with only a few transitioning to and from the intermediate subgroup V. The dysregulation of immunoglobulin genes, NOG and ALAS2, were most predictive of the subgroups with the best and worst prognosis, suggesting that these genes are key in determining patient outcome, and may therefore represent future drug targets but also a tool to identify patients responsive to current treatments. Estimates of cell counts in whole blood revealed elevated levels of lymphocytes, in particular T cells, and lower levels of inflammatory markers in the better prognosis subgroup. We further generated classifiers based on associated clinical features of these new RNA subgroups and used it to identify subgroups that differed in survival outcome in an independent cohort.

The most striking difference between the best and worst surviving subgroups was in immunoglobulin transcription. The upregulation of transcripts coding for the variable domain of immunoglobulin light chains (IGLV and IGKV genes) that participate in antigen recognition were markers of subgroup II, while their downregulation were markers of subgroup I. Differential levels of IGVL and IGKV gene transcripts, as seen in subgroup I, may control self-reactivity of human antibodies[27], and the reduction in the diversity of light chains has been associated with several autoimmune diseases, including systemic lupus erythematosis (SLE), type 1 diabetes, and myasthenia gravis[28, 29]. The association between autoimmunity and PAH has long been discussed. There are known associations with autoimmune diseases in other forms of PAH such

as systemic sclerosis, SLE, Sjogren's, etc., and the dysregulation of immune cells including T cells, B cells[30] and natural killer cells[31] are well described in IPAH, further validating that our unbiased approach has identified important subgroups. While we detected significant differences in lymphocyte, neutrophil and CRP levels in the blood samples of subgroup I patients, deeper genomic characterisation of T cell receptor and B cell receptors may be needed to understand the role of adaptive immunity on PAH progression.

Beyond the differences in immunoglobulin genes, the expression patterns that defined each subgroup also highlighted haeme biosynthesis through ALAS2 was a marker for subgroup I and correlated with greater disease severity. Previous gene expression studies across multiple forms of PH, including IPAH, showed significantly increased expression of ALAS2 in both systemic sclerosis-associated PAH (SSc-PAH) and IPAH[32]. In that study, in IPAH patients increased ALAS2 levels also demonstrated strong correlation with right atrial pressure, pulmonary vascular resistance, pulmonary artery saturation and cardiac index[32]. These data, and our own observations (**Fig. 3**), are suggestive of a role for ALAS2, iron[33] and hepcidin[34, 35] in pulmonary vascular remodelling and PH. Subgroup II with better prognosis can be partially defined by the downregulation of ALAS2 and increased expression of NOG, a BMP antagonist with high-affinity binding to BMP4 (ref. 36) which has been shown to inhibit hypoxia-induced proliferation of PASMC[23], and previously associated with BMI in PAH[37] has been proposed as a potential therapeutic target[38]. The role of Noggin in the low-risk group is particularly interesting given the proposed role of both Gremlin and Noggin in the mechanism of action for Sotatercept in the treatment of PAH[39].

Previous studies have identified clinical features collected during the diagnosis of PAH that also have prognostic utility. The clinical features identified here share many commonalities with those previously included in widely used risk scores (e.g. REVEAL, ERS) assessment for PAH, including, for example, 6-MWD, WHO functional class, and NT-proBNP[4, 40, 41]. This provided further validation that the transcriptomic profile associated with these subgroups provide insight into the biology of disease, and perhaps future drug targets. In addition to biomarkers such as CRP which is known to be elevated in PAH and CTEPH and shown to be predictive of outcome and sensitive to therapies[42] and NT-proBNP with high levels highly prognostic of right ventricular failure[43], age of diagnosis, BMI and renal function were also identified. Renal function has previously been associated with outcome in PAH, although likely because of cardiac function[44]. The age of diagnosis is often discussed as a consequence of genetics[45], or occurrence of co-morbidities; however, in our study the age of diagnosis was most strongly associated with the immunoglobulin light chain genes and Noggin. Carriers of BMPR2 mutations often present with PAH at a younger age and have a worse survival[46] so the association with Noggin is interesting in the context of perturbed BMP signalling. However, the patients with BMPR2 mutation did

not cluster within one subgroup perhaps fitting with the concept that it is dysfunctional TGFβ/BMP signalling rather than the precise mutation that is important.

There is a well-described sex-paradox in PAH[47] with a 4:1 female to male prevalence but the worse survival in male patients[48, 49]. During our initial analyses of the RNAseq data, we identified subclusters exclusively defined by sex genes. To mitigate against any gender bias, we excluded sex-chromosome-associated genes in our preprocessing steps of the analysis pipeline. Although we cannot reject the possibilities of the aforementioned genes contributing towards PAH or resilience, we believe that their removal ensures that the clustering algorithm captures heterogeneity independent of sex-associated expression variation. However, the interactions between gender and other autosomal genes in the context of PAH require further study.

The application of unsupervised learning from molecular profiles of IPAH is a powerful approach for revealing subgroups within a heterogeneous population that has not been defined clinically. Most studies employ widely used clustering algorithms without exploring their data suitability. By contrast, in this study we determine spectral clustering as the most consistent method in detecting differences and subsequently partitioning RNA-sequencing samples using robust performance criteria. Furthermore, previous studies have focused on clustering all PAH cases using a small set of immune markers, and captured immune phenotypes overlooked by the broad clinical classifications[50]. We used a much larger set of features, i.e., the whole transcriptome and clustered cases lacking causal pathologies, and also found immune phenotypes that differentiated the subgroups. While we controlled for confounding factors that affect clustering, such as gender-associated genes (**Supplementary Fig. 1**), there may yet be other hidden factors, such as viral infections related to age and gender that could influence patterns observed from whole blood[51]. The large degree of validation of the subgroups using both transcriptomic and clinical features to define them provides strong evidence that these endophenotypes are reproducible and may be useful to risk stratify or biologically classify subgroups of IPAH patients. However, further transcriptomic studies profiling patients at multiple timepoints are required to fully understand the dynamics of the immune components we identified, the frequency of acute infections, and the impact on PAH phenotype.

Transcriptomic profiling of the blood samples coupled with clinical data from IPAH patients provides an insight into endophenotypes that may describe this heterogeneous disease based on RNA expression. The use of additional ‘omic' biomarkers to provide further molecular profiles (e.g. DNA, protein, metabolites) as stable biomarkers for stratifying patients could further improve our algorithmic predictions of patient outcomes and reveal endophenotypes to be targeted

therapeutically. Furthermore, these data hold promise that these molecular endophenotypes may be tractable to existing therapies, may offer an alternative approach to tailor, and assess individual treatment response, in PAH as well as offering insights into disease pathogenesis that can be targeted by therapeutics as a precision medicine approach[52] in PAH and potentially other diseases to drive molecular clinical classification suited to the future precision medicine era in healthcare.

Methods

Study design

The Cohort study of idiopathic and heritable PAH is an observational, prospective and longitudinal study of patients with idiopathic and heritable PAH (clinicaltrials.gov NCT019072950). The Sheffield Teaching Hospitals Observational Study of Pulmonary Hypertension, Cardiovascular and other Respiratory Disease (UK REC Ref 18/YH/0441) is a longitudinal study of patients with suspected pulmonary hypertension or an associated cardiovascular or respiratory condition. Follow-up information is collected as a part of routine clinical care every 6 months. The study allows recruitment of both incident and prevalent cases. Patients consented to the study agreed to have blood taken for next-generation sequencing and other omics studies. Healthy adult controls were recruited for comparison studies. The subsequent whole-blood sample collection process is described in ref. 13. *[written by other co-authors]*

Ethics

All UK samples were obtained following informed consent into the UK National Cohort Study of Idiopathic and Heritable Pulmonary Arterial Hypertension (clinicaltrials.gov NCT01907295; UK REC Ref. 13/EE/0203) and/or the Sheffield Teaching Hospitals Observational Study of Pulmonary Hypertension, Cardiovascular and other Respiratory Disease (UK REC Ref 18/YH/0441). Data were obtained from samples collected at the University of Arizona Pulmonary Hypertension clinic between 2012 and 2015 following institutional guidelines and following informed consent. *[written by other co-authors]*

Participants

Patients diagnosed with H/IPAH, PVOD or PCH, relatives of index cases and unrelated healthy controls were recruited at nine UK centres and followed up by a median of 7.9 years. In total, 358 patients (Supplementary Fig. 11) of which 96.7% were further verified to be H/IPAH, 13 relatives, and 21 healthy controls recruited to the H/IPAH Cohort study were analysed. Both prevalent and incident cases were allowed. Prevalent cases were defined as diagnosed earlier than 6 months before the study initiation. Patients in Cohort study were followed longitudinally as part of their clinical PAH care. All cases were diagnosed between March 1994 and November 2016, and diagnostic classification was made according to international guidelines[53]. Patients with PAH associated with anorexigen exposure were considered as IPAH, whereas HPAH was defined by the presence of a positive family history of PAH. Clinical, functional and haemodynamic characteristics at the time of PAH diagnosis were prospectively entered into the database. The date of diagnosis corresponded to that of confirmatory right heart catheterisation.

Following diagnosis, subsequent treatments and follow-ups were at the discretion of the treating physician, according to the contemporary guidelines. In most centres, patients were seen every 3–6 months with an assessment of functional status and exercise capacity. Right heart catheterisation was repeated when considered necessary by the responsible clinician. Study visits were performed every 6 months. Healthy controls had been sampled only once and had clinical information recorded from the time of sampling. *[written by other co-authors]*

Clinical data capture, processing and quality control

Pseudonymised results of routinely performed clinical tests reported in either clinical case notes or electronic medical records (EMR) were stored in web-based OpenClinica (OC) data capture system (Community edition). Twenty electronic Clinical Case Report Forms (eCRFs) distributed across seven events (Diagnostic, Continuous data, Follow-up, Epidemiology questionnaire, Suspension, Relatives, Unrelated healthy control) were constructed to accommodate routinely available clinical information. Details regarding data verification procedures were previously described in detail[54].

Information about participants' status was collected every 6 months (via National Health System Digital Spine portal or an equivalent local system). Current analysis

was performed on the census performed on 31 January 2020. Two risk assessment strategies were applied to the data. Reveal risk score[4] and abbreviated ERS risk scores[55] were calculated in all patients who had the necessary minimum phenotypic information available. Patients who died or were transplanted were suspended on the day of the event, patients who withdrew from the study were censored on the date of the last visit, the reason for withdrawal was recorded. *[written by other co-authors]*

Missingness assessment and imputation

Missingness rates, patterns and causes were assessed per individual, variable and centre and visualised with vim package v5.1.1R (**Supplementary Fig. 12**). Multiple imputation by the chain equations method was used to impute missing data (mice v3.8.0 package R)[56]. The imputation model included all variables that were necessary in the analysis model, including cumulative baseline hazard function and variables that predicted both the incomplete variable and if the incomplete variable was missing like the centre and whether the case was incident or prevalent. Quality of predictors was assessed using outflux–influx plot. Numerical data were imputed with predictive mean matching (pmm), factors with two levels were imputed using logistic regression, factors with more than two levels with multinomial logit model and ordered factors with more than two levels with the ordered logit model. Transformed variables (BMI, ratios, score sums) were imputed as just another variable as well as passively with good concordance. The visiting sequence was set to 'monotone' to speed up convergence. The number of iterations was set to 20. Following the rule of thumb proposed by White et al.[57] that the number of imputations should be at least equal to the percentage of incomplete cases, the procedure was performed at m = 50. The convergence of the algorithm was checked, and the means and standard deviations of imputed values were plotted over 20 iterations. The streams of numerical and factor variables intermingled and showed no trends at later iterations. Factors influencing the accuracy of the imputation include the variability in time between diagnosis and sampling, higher missingness in clinical data for prevalent cases (diagnosed sometimes many years ago), and differences in measurement error between centres which followed different protocols for clinical data collection. *[written by Emilia M Swietlik and Divya Pandya]*

RNA data preprocessing

A number of preprocessing steps were required to prepare the raw sequencing data for unsupervised machine learning. High-throughput sequencing generated raw pair-end counts of 205,259 transcripts across 508 samples that belong to GenCode Release 28 (GRCh38.p12). Consequently, Salmon (https://combine-lab.github.io/salmon/) was used to estimate the relative abundance of the transcripts (TPM, units of transcripts per million) which were then mapped to genes (n = 60,144) using the tximport R package. Only genes with more than two reads (in a transcript level) in at least 95% of control and patient samples were considered and 11 additional male genes were removed (n = 25,955). Hyperbolic arcsine transformation (package base v3.6.0) was applied to the final RNAseq TPM matrix. Further information on quality control of samples and genes can be found in the Supplementary Methods. The RNA-sequencing and clinical data of healthy controls were not used in the main pipeline of this study. A secondary clustering with all patient and healthy samples was implemented to demonstrate the lack of pure patient and healthy subgroups within our cohort (**Supplementary Fig. 6b**). Principal component analysis of expression profiles from samples with a second replicate clustered together according to the first four principal components (**Supplementary Fig. 13**).

Spectral clustering: gene expression subgroup identification

We performed cluster analysis to partition IPAH patients to distinct RNA-based groups. The spectral clustering model (package kernlab v0.9-29) was selected as the most suitable unsupervised learning algorithm based on the highest partitional consistency when comparing multiple dissimilar algorithms (**Supplementary Table 6**). For the spectral clustering method, data points (i.e. patients) are embedded and partitioned in a low-dimensional space in the form of a similarity graph, rather than being characterised by more than 25,000 gene dimensions. High partitional consistency was defined as the high adjusted Rand Index (package fossil v0.3.7) and low standard deviation calculated between different variations of each clustering algorithm (k-means, spectral, hierarchical clustering), as described in Clustering algorithm selection. For the selection of the most appropriate clustering algorithm we utilised 25,955 genes across 359 IPAH patient samples (discovery cohort) after further filtering for repeated same-visit samples and non-H/IPAH diagnosis. We compared three fundamentally different methods (hierarchical, k-means and spectral) and use partitioning consistency to determine which method picks up an underlying signal from our data type (RNA-sequencing). As shown in **Supplementary Fig. 14**,

spectral clustering showed the highest consistency (Adjusted Rand Index) in detecting differences and subsequently partitioning patients in similar clusters independently of the kernel. Notable is the difference in intra-agreement of spectral (~75%) and k-means (−13%) clustering, which highlights the importance of the extra step of mapping data in a low-dimensional space (as a similarity graph) in spectral clustering. To run the main spectral clustering partitioning we first selected the most relevant gene set by ranking all genes based on the variability of their expression across patient samples using the stats v3.6.0R package (Supplementary section 'Feature selection of genes'). Subsequently, several candidate gene sets of increasing size were drawn from the top ranking gene list and the one that generated subgroups of highest stability, according to package fpc v2.2-3, was selected (Supplementary section 'Highest stability gene set'; **Supplementary Fig. 15**). This resampling bootstrap approach determined that the most stable gene set was composed from the 300 most variable genes. As observed in **Supplementary Fig. 15**, the average stability is expected to peak towards the smaller sets of genes because in most biological cases the relevant genes tend to be fewer than hundreds and usually the ones that show the most variance across patients/samples. However, variability across all patients does not always mean partitioning power (i.e. clearer cluster separation). Some distinctive features (in our case genes) might have received a lower variance ranking since they could effectively discriminate only between smaller groups of patients but score less overall variance across the entire cohort. Therefore they would only be included in larger genesets, as described in the geneset selection method step. For that reason, we might observe smaller peaks following the highest one, as in **Supplementary Fig. 15**. Another reason for this observation can be that the stability score consists of an averaging of the stabilities across different numbers of clusters (k) starting from two and usually reaching above five (in our case up to six). As the clustering algorithm takes into account higher ks, lower variance genes might contribute to the finer partitions thus increasing the stability. Due to the stability averaging over k, this effect becomes less prominent but still can be reflected in the formation of late lower peaks.

For the secondary clustering run with only IPAH, 1700 variable genes were selected as the most stable gene set for clustering.

The number of IPAH subgroups was estimated through ensemble learning[58] utilising 15 internal indexes calculated using the package diceR v0.6.0 (Supplementary sections 'Optimal number of subgroups k' and 'Internal Index Voting'). A representation of patient flow across k can be found at Supplementary Fig. 16. The Radial Basis function kernel was used as the similarity measure with five target subgroups, identified as the optimal number of subgroups by an ensemble learning method. We elected to investigate k = 5 subgroups, since in clustering contexts it is safer to overestimate than underestimate the number of subgroups to prevent loss of information. However, k = 3 subgroups were voted from the vast majority of methods and we expect them to be the main subgroups. Further information on the selection

of clustering algorithms and parameters can be found in the Supplementary Methods.

Analysis of subgroup differences

Survival analysis was performed (R package survival 3.1-7) on the main (**Supplementary Fig. 17**) and validation cohorts to identify the survival differences between subgroups. Kaplan–Meier survival curves from diagnosis and sampling were calculated for the main patient cohort (per spectral subgroup) as well as the validation cohort (per predicted subgroup). Subsequently, two multivariate Cox models were fitted and Hazard ratios calculated on the main cohort once adjusting for gender and once adjusting for the composite clinical signature discovered by supervised machine learning. Gene signatures for each subgroup were identified using LASSO regression models with cross-validation (package glmnet 3.0-1). The variables with the 5% highest coefficients for each class were highlighted (**Supplementary Fig. 18**), and the full list of non-zero coefficients for each class can be found in Supplementary Data 3. The pathfinder R package was used to highlight enriched gene pathways between subgroups and differential expression analysis using the DESeq2 package[59] was performed on genes associated with the subgroups (**Supplementary Fig. 19**). Gene expression differences across subgroups are presented in **Supplementary Fig. 20**. All statistical tests between subgroups were two-sided and Bonferonni adjusted for multiple testing. *[regression analysis was performed by Thomas S Mascarenhas]*

Identifying clinical signatures of subgroups

The dataset was initially cleaned and filtered on 119 features that were identified by a domain expert from the original 887 features that described the dataset. Subsequently, any feature that had more than 5% missing data was dropped, and categorical features numerically encoded.

All ML tasks were carried out using Scikit-learn[60] ML framework version 0.23.2 in a Python 3 environment. As machine learning classifiers, we used Logistic Regression (LR), support vector machines (SVM), Random Forest (RF) and k-nearest neighbour (kNN). RF is a powerful ensemble learning technique especially for high-dimensional classification tasks. Further details about classifier training and feature selection can be found in the Supplementary Methods.

Classification of new patients using signatures

Each clinical signature was used to develop a classification model trained on the discovery cohort to classify new patients into the RNA-based subgroups. Classification models were built using SVM[61], RF[62], LR[63] and KNN[64]. The candidate signature that obtained the best performance was selected. This process was repeated for all signature sizes, s = 1 to s = 20, for subgroups I, II and V. A final signature for each subgroup was selected based on a compromise between the fewest number of features (s = 1 to s = 20) and classification performance. Final selected signatures for each of the subgroups were pooled to create a composite signature, which was then used in a multi-class classification model. The model was trained on the discovery dataset to discriminate between subgroups I, II and V, used to predict subgroup membership of an unseen validation dataset. The predicted subgroup membership was then used to calculate survival of predicted subgroups. Survival of the predicted subgroups was compared to known survival of subgroups in the discovery dataset for validation purposes. *[written by Emmanuel Jammeh]*

qPCR on validation cohort

Frozen Tempus tubes collected from patients in the validation cohort, collected under the UK National Cohort study, were obtained; RNA was extracted using Maxwell® 16 LEV simplyRNA Blood Kit (Cat.# AS1310) as described in the manufacturer's instructions on the Maxwell® 16 Instrument (Cat.# AS2000). Extracted RNA was transcribed using the High-Capacity-RNA-to-cDNA kit (Thermo Fisher Cat.# 437406) following the manufacturer's instructions. Resultant cDNA was analysed using custom TaqMan array cards (Thermo Fisher Cat.#4342249) with Fast Advanced Mastermix (Thermo Fisher Cat.# 4444964); damples were run 8 to a card across 25 cards with 24 primer probes (Thermo Fisher) per sample (18S-Hs99999901_s1, ACTB-Hs00357333_g1, ALAS2-Hs01085701_m1, BMPR2-Hs00176148_m1, C4BPA-Hs00426339_m1, CRISP3-Hs00195988_m1, CTSG-Hs00175195_m1, GAPDH-Hs02786624_g1, HPRT1-Hs02800695_m1, IFI27-Hs01086373_g1, IGHM-Hs00941538_g1, IGHV3-75-Hs03832008_sH, IGKV2-24-Hs06671746_g1, IGLV6-57-Hs01696637_s1, LINC00221-Hs01382601_m1, LTF-Hs00914334_m1, MT-RNR1-Hs02596859_g1, NEBL-Hs01067284_m1, NOG-Hs00271352_s1,

NOS2-Hs01075529_m1, NPRL3-Hs00429221_m1, PI3-Hs00160066_m1, SMIM11A;SMIM11B-Hs00938773_m1, XIST-Hs01079824_m1). These assays were performed in duplicate using the Applied Biosystems 7900HT Fast real-time PCR system with the TaqMan Low Density Array card block following calibration using the TaqMan Low Density Array Calibration Kit (Thermo Fisher Cat.# 10341465). Ct values were determined with Automatic thresholding in the SDS2.4 software. GAPDH-Hs02786624_g1 was used as a control. Relative quantity was calculated using the ΔΔCt method. *[written by Josephine A. Pickworth*

External cohort validation

An external validation cohort of patients with Group 1 PAH prospectively recruited at the University of Arizona Pulmonary Hypertension clinic between 2012 and 2015 following institutional guidelines and informed consent was used. The cohort comprised 84 subjects with Group 1 PAH of whom 32 were diagnosed with idiopathic PAH. For each subject, demographics and clinical variables were collected[26]. PBMCs were stored in RNAlater as previously described. In total, approximately 3600 million clusters with paired-end 75 bp reads (~35M cluster per sample) were generated from PBMC-derived RNA. *[correlation data provided by Ankit A. Desai]*

Clinical variable and gene correlations

We calculated correlations between the clinical and gene signatures we generated in previous steps of this study. For discovery and validation cohorts we used the rcorr function of R package Hmisc (version 4.5-0). For the external validation we used the values found in ref. 26.

Study approval

Study approval for the use of sample and data were obtained from the UK National PAH Cohort Study Data Access Committee (clinicaltrials.gov NCT01907295; UK REC Ref 13/EE/0203), and the Sheffield Teaching Hospitals Observational Study of Pulmonary Hypertension, Cardiovascular and other Respiratory Diseases Scientific Advisory Board (UK REC Ref 18/YH/0441).

Data availability

The transcriptomic and clinical data used in this study have been deposited in the EGA (the European Genome-phenome Archive) database under accession code EGAS00001005532. In compliance with the Ethics under which these data and samples have been collected, the transcriptomic data are available through restricted access for approved researchers who agree to the conditions of use, i.e. keeping it secure and only using it for approved purposes. To apply for access please contact cohortcoordination@medschl.cam.ac.uk. You will receive an application form within 30 days. The 'UK National PAH Cohort Study Data Access Committee' will review requests within 3 months of receipt of the completed application form and if approved, provide details for access to the RNAseq data stored at the EGA. All requesters must agree to the data access conditions found in EGA. The data used to generate statistics, plots and figures are accessible through our interactive portal found in https://sheffield-university.shinyapps.io/ipah-rnaseq-app/. Source data are provided with this paper.

Code availability

Additionally, the code used to generate the results of this study is publicly available at https://zenodo.org/badge/latestdoi/299615578 (ref. 66). [*Feature selection related code was written by Emmanuel Jammeh*]

Acknowledgements

The UK National Cohort of Idiopathic and Heritable PAH is supported by grants from the British Heart Foundation (SP/12/12/29836 & SP/18/10/33975) and the UK Medical Research Council (MR/K020919/1). Additional samples from the Sheffield Teaching Hospitals Observational Study of Pulmonary Hypertension, Cardiovascular and other Respiratory Diseases were supported by British Heart Foundation (PG/11/116/29288). We gratefully acknowledge financial support from the UK Department of Health via the NIHR comprehensive Biomedical Research Centre award to Imperial College Healthcare NHS Trust, Cambridge Biomedical Research Centre, and Guy's and St Thomas' NHS Foundation Trust in partnership with King's College London and King's College Hospital NHS Foundation Trust and the NIHR Imperial Clinical Research Facility. Sheffield NIHR Clinical Research Facility award to Sheffield Teaching Hospitals Foundation NHS Trust. S.K. is supported by a Donald Heath Ph.D. Studentship award; C.J.R. is supported by a British Heart Foundation

Intermediate Basic Science Research fellowship (FS/15/59/31839). N.E. is supported by an EPSRC Centre for Doctoral Training; A.A.R.T. is supported by a British Heart Foundation Intermediate Clinical Research fellowship (FS/18/13/33281); N.W.M. is a British Heart Foundation Professor and NIHR Senior Investigator. A.L. is supported by a BHF Senior Basic Science Research fellowship (FS/18/52/33808). E.J. is supported by the Academy of Medical Sciences Springboard (ref: SBF004/1052). M.R.W. is in receipt of a British Heart Foundation Centre for Research Excellence award (RE/18/4/34215). M.J.D. and D.W. are supported by the NIHR Sheffield Biomedical Research Centre. We thank and thank all the patients and their families who contributed to this research, the UK Pulmonary Hypertension Association for their support, NIHR BioResource volunteers for their participation, and gratefully acknowledge NIHR BioResource centres, NHS Trusts and staff for their contribution. We thank the National Institute for Health Research, NHS Blood and Transplant, and Health Data Research UK as part of the Digital Innovation Hub Programme. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

Author information

Contributions

SK, EJ and EMS contributed equally. All authors made substantial contributions to the conception or design and data acquisition of the work. S.K., E.J., E.M.S., J.A.P., C.J.R., P.O., J.W., J.I., M.J.D., D.P., T.S.M., N.E., A.A.R.T., C.E.R., F.R., J.G.N.G., J.X.-J.Y., T.-H.S., A.A.D., G.C., J.L., P.A.C., L.S.H., R.C., D.G.K., C.C., J.P.-Z., M.T., S.W., S.G., N.W.M., M.R.W., A.L. and D.W. performed the analysis and/or interpretation of data. S.K., E.J., E.M.S., A.L. and D.W. drafted the work and all authors revised it critically for important intellectual content; and gave final approval of the version submitted for publication; and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References (Manuscript 2)

1. Galiè, N. et al. Risk stratification and medical therapy of pulmonary arterial hypertension. Eur. Respir. J. 53, 1801889 (2019).
2. Hurdman, J. et al. ASPIRE registry: assessing the spectrum of pulmonary hypertension identified at a REferral centre. Eur. Respir. J. 39, 945–955 (2012).
3. Ling, Y. et al. Changing demographics, epidemiology, and survival of incident pulmonary arterial hypertension: results from the pulmonary hypertension registry of the United Kingdom and Ireland. Am. J. Respir. Crit. Care Med. 186, 790–796 (2012).
4. Benza, R. L. et al. Predicting survival in patients with pulmonary arterial hypertension: the REVEAL Risk Score Calculator 2.0 and comparison with ESC/ERS-based risk assessment strategies. Chest 156, 323–337 (2019).
5. Bergemann, R. et al. High levels of healthcare utilization prior to diagnosis in idiopathic pulmonary arterial hypertension support the feasibility of an early diagnosis algorithm: the SPHInX project. Pulm. Circ. 8, 2045894018798613 (2018).
6. Thompson, A. A. R. & Lawrie, A. Targeting vascular remodeling to treat pulmonary arterial hypertension. Trends Mol. Med. 23, 31–45 (2017).
7. Schermuly, R. T., Ghofrani, H. A., Wilkins, M. R. & Grimminger, F. Mechanisms of disease: pulmonary arterial hypertension. Nat. Rev. Cardiol. 8, 443–455 (2011).
8. Southgate, L., Machado, R. D., Gräf, S. & Morrell, N. W. Molecular genetic framework underlying pulmonary arterial hypertension. Nat. Rev. Cardiol. 17, 85–95 (2020).
9. Gräf, S. et al. Identification of rare sequence variation underlying heritable pulmonary arterial hypertension. Nat. Commun. 9, 1416 (2018).
10. Rhodes, C. J. et al. Genetic determinants of risk in pulmonary arterial hypertension: international genome-wide association studies and meta-analysis. Lancet Respir. Med. 7, 227–238 (2019).

11. Rhodes, C. J. et al. Plasma proteome analysis in patients with pulmonary arterial hypertension: an observational cohort study. Lancet Respir. Med. 5, 717–726 (2017).
12. Rhodes, C. J. et al. Plasma metabolomics implicates modified transfer RNAs and altered bioenergetics in the outcomes of pulmonary arterial hypertension. Circulation 135, 460–475 (2017).
13. Rhodes, C. J. et al. Whole blood RNA profiles associated with pulmonary arterial hypertension and clinical outcome. Am. J. Respir. Crit. Care Med. https://doi.org/10.1164/rccm.202003-0510OC (2020).
14. Frésard, L. et al. Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. Nat. Med. 25, 911–919 (2019).
15. Vanhaverbeke, M., Veltman, D., Janssens, S. & Sinnaeve, P. R. Peripheral blood RNAs and left ventricular dysfunction after myocardial infarction: towards translation into clinical practice. J. Cardiovasc. Transl. Res. https://doi.org/10.1007/s12265-020-10048-x (2020).
16. Pickworth, J. et al. Differential IL-1 signaling induced by BMPR2 deficiency drives pulmonary vascular remodeling. Pulm. Circ. 7, 768–776 (2017).
17. Farkas, D. et al. Toll-like receptor 3 is a therapeutic target for pulmonary hypertension. Am. J. Respir. Crit. Care Med. 199, 199–210 (2019).
18. Hameed, A. G. et al. Inhibition of tumor necrosis factor-related apoptosis-inducing ligand (TRAIL) reverses experimental pulmonary hypertension. J. Exp. Med. 209, 1919–1935 (2012).
19. Frid, M. G. et al. Immunoglobulin-driven complement activation regulates proinflammatory remodeling in pulmonary hypertension. Am. J. Respir. Crit. Care Med. 201, 224–239 (2020).
20. Mamazhakypov, A., Viswanathan, G., Lawrie, A., Schermuly, R. T. & Rajagopal, S. The role of chemokines and chemokine receptors in pulmonary arterial hypertension. Br. J. Pharmacol. https://doi.org/10.1111/bph.14826 (2019).

21. Hurst, L. A. et al. TNFα drives pulmonary arterial hypertension by suppressing the BMP type-II receptor and altering NOTCH signalling. Nat. Commun. 8, 14079 (2017).
22. Steiner, M. K. et al. Interleukin-6 overexpression induces pulmonary hypertension. Circ. Res. 104, 236–244, 28p following 244 (2009).
23. Yang, K. et al. Noggin inhibits hypoxia-induced proliferation by targeting store-operated calcium entry and transient receptor potential cation channels. Am. J. Physiol. Cell Physiol. 308, C869–C878 (2015).
24. Bell, R. D. et al. TNF induces obliterative pulmonary vascular disease in a novel model of connective tissue disease associated pulmonary arterial hypertension (CTD-PAH). Arthritis Rheumatol. https://doi.org/10.1002/art.41309 (2020).
25. Foris, V. et al. Neutrophil-to-lymphocyte ratio as a prognostic parameter in pulmonary arterial hypertension. 4.3 Pulmonary Circulation and Pulmonary Vascular Diseases. https://doi.org/10.1183/13993003.congress-2016.pa2479 (2016).
26. Romanoski, C. E. et al. Transcriptomic profiles in pulmonary arterial hypertension associate with disease severity and identify novel candidate genes. Pulm. Circ. 10, 2045894020968531 (2020).
27. Collins, A. M. & Watson, C. T. Immunoglobulin light chain gene rearrangements, receptor editing and the development of a self-tolerant antibody repertoire. Front. Immunol. 9, 2249 (2018).
28. Panigrahi, A. K. et al. RS rearrangement frequency as a marker of receptor editing in lupus and type 1 diabetes. J. Exp. Med. 205, 2985–2994 (2008).
29. Vander Heiden, J. A. et al. Dysregulation of B cell repertoire formation in myasthenia gravis patients revealed through deep sequencing. J. Immunol. 198, 1460–1473 (2017).
30. Nicolls, M. R., Taraseviciene-Stewart, L., Rai, P. R., Badesch, D. B. & Voelkel, N. F. Autoimmunity and pulmonary hypertension: a perspective. Eur. Respir. J. 26, 1110–1118 (2005).
31. Ormiston, M. L. et al. Impaired natural killer cell phenotype and function in idiopathic and heritable pulmonary arterial hypertension. Circulation 126, 1099–1109 (2012).

32. Cheadle, C. et al. Erythroid-specific transcriptional changes in PBMCs from pulmonary hypertension patients. PLoS ONE 7, e34951 (2012)
33. Rhodes, C. J. et al. Iron deficiency in pulmonary arterial hypertension: a potential therapeutic target. Eur. Respir. J. 38, 1453–1460 (2011).
34. Rhodes, C. J. et al. Iron deficiency and raised hepcidin in idiopathic pulmonary arterial hypertension: clinical prevalence, outcomes, and mechanistic insights. J. Am. Coll. Cardiol. 58, 300–309 (2011).
35. Ramakrishnan, L. et al. The Hepcidin/Ferroportin axis modulates proliferation of pulmonary artery smooth muscle cells. Sci. Rep. 8, 12972 (2018).
36. Mehler, M. F., Mabie, P. C., Zhang, D. & Kessler, J. A. Bone morphogenetic proteins in the nervous system. Trends Neurosci 20, 309–317 (1997).
37. Al-Khafaji, K. H. A., Al-Dujaili, M. N. & Al-Dujaili, A. N. G. Assessment of noggin level in pulmonary arterial hypertension patients. Curr. Iss. Pharm. Med. Sci. 31, 122–130 (2018).
38. Boucherat, O. & Bonnet, S. NOGGIN: a new therapeutic target for PH? Focus on “Noggin inhibits hypoxia-induced proliferation by targeting store-operated calcium entry and transient receptor potential cation channels. Am. J. Physiol. Cell Physiol. 308, C867–C868 (2015).
39. Humbert, M. et al. Sotatercept for the treatment of pulmonary arterial hyprtension. N. Engl. J. Med. 384, 1204–1215 (2021).
40. Benza, R. L. et al. The REVEAL registry risk score calculator in patients newly diagnosed with pulmonary arterial hypertension. Chest 141, 354–362 (2012).
41. Benza, R. L. et al. Prognostic implications of serial risk score assessments in patients with pulmonary arterial hypertension: A Registry to Evaluate Early and Long-Term Pulmonary Arterial Hypertension Disease Management (REVEAL) analysis. J. Heart Lung Transpl. 34, 356–361 (2015).

42. Quarck, R., Nawrot, T., Meyns, B. & Delcroix, M. C-reactive protein: a new predictor of adverse outcome in pulmonary arterial hypertension. J. Am. Coll. Cardiol. 53, 1211–1218 (2009).
43. Nickel, N. et al. The prognostic impact of follow-up assessments in patients with idiopathic pulmonary arterial hypertension. Eur. Respir. J. 39, 589–596 (2012).
44. Kaiser, R., Seiler, S., Held, M., Bals, R. & Wilkens, H. Prognostic impact of renal function in precapillary pulmonary hypertension. J. Intern. Med. 275, 116–126 (2014).
45. Phillips, J. A. et al. Synergistic heterozygosity for TGFβ1 SNPs and BMPR2 mutations modulates the age at diagnosis and penetrance of familial pulmonary arterial hypertension. Genet. Med. 10, 359–365 (2008).
46. Evans, J. D. W. et al. BMPR2 mutations and survival in pulmonary arterial hypertension: an individual participant data meta-analysis. Lancet Respir Med. 4, 129–137 (2016).
47. Austin, E. D. Gender, sex hormones, and pulmonary arterial hypertension. Adv. Pulm. Hypertens. 10, 160–166 (2011).
48. Foderaro, A. & Ventetuolo, C. E. Pulmonary arterial hypertension and the sex hormone paradox. Curr. Hypertens. Rep. 18, 84 (2016).
49. Hester, J., Ventetuolo, C. & Lahm, T. Sex, gender, and sex hormones in pulmonary hypertension and right ventricular failure. Compr. Physiol 10, 125–170 (2019).
50. Sweatt, A. J. et al. Discovery of distinct immune phenotypes using machine learning in pulmonary arterial hypertension. Circ. Res. 124, 904–919 (2019).
51. Bongen, E. et al. Sex differences in the blood transcriptome identify robust changes in immune cell proportions with aging and influenza infection. Cell Rep 29, 1961–1973.e4 (2019).
52. Morrell, N. W. et al. Genetics and genomics of pulmonary arterial hypertension. Eur. Respir. J. 53, 1801899 (2019).
53. Galiè, N. et al. 2015 ESC/ERS Guidelines for the diagnosis and treatment of pulmonary hypertension. The Joint Task Force for the Diagnosis and Treatment of Pulmonary

Hypertension of the European Society of Cardiology (ESC) and the European Respiratory Society (ERS). Eur. Respir. J. 46, 903–975, 1855–1856 (2015).

54. Swietlik, E. M. et al. Bayesian inference associates rare KDR variants with specific phenotypes in pulmonary arterial hypertension. Circ Genom Precis Med. 14, e003155 https://doi.org/10.1161/CIRCGEN.120.003155 (2021).
55. Hoeper, M. M. et al. Mortality in pulmonary arterial hypertension: prediction by the 2015 European pulmonary hypertension guidelines risk stratification model. Eur. Respir. J. 50, 1700740 (2017).
56. van Buuren, S. & Groothuis-Oudshoorn, K. mice: multivariate imputation by chained equations in R. J. Stat. Softw. 45, 1–67 (2011).
57. White, I. R., Royston, P. & Wood, A. M. Multiple imputation using chained equations: issues and guidance for practice. Stat. Med. 30, 377–399 (2011).
58. Polikar, R. Ensemble based systems in decision making. IEEE Circuits Syst. Mag. 6, 21–45 (2006).
59. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550 (2014).
60. Pedregosa, F. et al. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830 (2011).
61. Choi, E., Schuetz, A., Stewart, W. F. & Sun, J. Using recurrent neural network models for early detection of heart failure onset. J. Am. Med. Inform. Assoc. 24, 361–370 (2017).
62. Breiman, L. Random Forests. Mach. Learn. 45, 5–32 (2001).
63. Hosmer, D. W., Jr., Lemeshow, S. & Cook, E. D. Applied Logistic Regression, Second Edition: Book and Solutions Manual Set (Wiley-Interscience, 2001).
64. Peterson, L. K-nearest neighbor. Scholarpedia J. 4, 1883 (2009).
65. Kariotis, S., Wang, D. & Lawrie, A. PAH sequencing study—EGA European Genome-Phenome Archive. https://ega-archive.org/studies/EGAS00001005532 (2021).
66. Kariotis, S. & Jammeh, E. BioSok/spectral_clustering_of_IPAH: v1.0.1. https://doi.org/10.5281/zenodo.5549872 (2021).

3.4 Supplementary information from Manuscript 2

Heterogeneity of Idiopathic pulmonary arterial hypertension revealed through unsupervised transcriptomic profiling of whole blood

Kariotis et al

Supplementary methods

Cohort population and study design

This expression based study utilizes the UK national H/IPAH cohort including all patients for which idiopathic and heritable PAH was identified in one of the National Centres for Pulmonary Hypertension within the UK, Golden Jubilee National Hospital (n=32), Imperial College Healthcare NHS Trust (n=119), Newcastle Pct (n=31), Papworth Hospital NHS Foundation Trust (n=68), Royal Brompton And Harefield NHS Foundation Trust (n=33), Royal Free Hampstead NHS Trust (n=20) and Sheffield Teaching Hospitals NHS Foundation Trust (n=63). Full written, informed consent with the local ethical committee was required for clinical data as well as blood sampling with the intent of next generation sequencing. The subsequent whole blood sample analysis is described in [1] . Complete information at **Supplementary Data 2**. [*written by other co-authors*]

Missingness

Clinical data for RNA sequenced patients was assessed for missingness, and patterns within the missing data, prior to analysis. Diagnosis and cohort visit 1 data were analysed separately, using the R packages VIM and Naniar. When looking at clinically relevant variables, overall missingness was 21.35% and 48.64% across the diagnostic and visit 1 datasets, respectively. When focusing on diagnostic variables, the highest percentage of missing data were present in the following variables: BNP (75.77%), left atrial size (73.54%), Troponin (72.42%), NT-proBNP (71.31%), right atrial area (71.03%). High levels of missing data for BNP and NTproBNP are due to centres carrying out only one of these tests to assess BNP levels. For the other variables, echocardiography is not required for diagnosis and the clinical blood tests chosen are centre dependent. When focusing on cohort visit 1, the highest percentages of

missingness can be seen in Troponin (96.94%), jugular venous pressure (89.42%), left atrial size (88.86%), total lung capacity (86.91%) and SvO2 (86.91%). When looking at variables with the lowest levels of missingness, age (diagnosis) (0.28%), history of syncope (0.28%) , ankle swelling (0.56%) and ascites (0.56%), and weight (2.22%), rank lowest amongst diagnostic data. In comparison, for visit 1, the following variables have the lowest percentages of missingness: functional class (0.84%), ankle swelling (1.11%) and ascites (1.11%), renal sodium (5.01%) and renal urea (5.01%). These variables are recorded routinely when joining the cohort study (for diagnostic variables) or routinely checked as part of clinical examinations. When analysing missingness per centre, diagnostic data had lower levels of missingness than visit 1. This is unsurprising, as the diagnostic event is more comprehensive than subsequent visits. For the diagnostic dataset, Glasgow had the highest rate of missingness at 36.20% compared to Imperial and Hammersmith which had 2 the lowest percentage missingness at 19.10%. For cohort visit 1, Glasgow again showed the highest rate of missingness at 71.96%, while Papworth had the lowest at 39.10%. Imperial and Hammersmith accounted for 33.43% of patients in the study, while the other centres contributed between 5.57% and 18.10% of participants, thus inflating missingness statistics. Centre specific differences also exist, with Sheffield using the shuttle test to assess exercise performance in the place of 6MWD. Specific values for missingness calculated for classifier variables only can be seen in **Supplementary Figure 12**, with NT-proBNP exhibiting the highest missingness and age at the time of diagnosis with the lowest missingness, for both diagnosis and cohort visit 1 datasets. [*written by Emilia M Swietlik*]

Ethnicity

H/IPAH Cohort study collected and coded the self-reported ethnicity information as per The Office of National Statistics: White: A – British, B – Irish, C – Any other White background, Mixed: D – White and Black Caribbean, E – White and Black African, F – White and Asian, G – Any other mixed background, Asian or Asian British: H – Indian, J – Pakistani, K - Bangladeshi, L – Any other Asian background, Black or Black British: M – Caribbean, N – African, P – Any other Black background, Other Ethnic Groups: R – Chinese, S – Any other ethnic group, Z – Not stated. [*written by other co-authors*]

Sample and gene selection preprocessing

The initial gene expression dataset consisted of 508 samples, both patients and controls. A number of samples had to be filtered out to ensure a high quality, unskewed input sample set for all subsequent clustering runs. Initially, the first occurrences of three samples were removed from the RNA-seq matrix as they were

repeated samples from the same visit of the corresponding patient/control with identical gene expression values. Moreover, 11 patients were excluded as their diagnosis of PAH was not of the idiopathic form and an additional eleven as they were diagnosed with a different form of PH (Pulmonary veno-occlusive disease). Finally, 10 relatives of IPAH patients were excluded due to potentially sharing underlying genetic characteristics with the corresponding IPAH samples. As part of the preprocessing for the clustering two gene filtering steps were implemented. Firstly, only genes that have more than two reads (in a transcript level) in at least 95% of control and patient samples were considered. This step reduces the number of genes from 60,144 (that occurred after the transcript transformation) to 25,966. Additionally, 11 male genes were removed as they were driving all following clusterings forming subgroups composed entirely by male or female samples. After selecting for genes and samples a patient gene expression dataset of 385 samples and 25955 genes was generated. The dataset was used to determine the most fitted clustering algorithm in terms of robustness and partition consistency, estimate the optimal number of subgroups and as input for the gene filtering step. A large number of clinical variables, 121 from the original clinical file and an additional 760 from OpenClinica study database, were uniformly measured at the different medical and research centers that provided samples as described in the Cohort / Study design section. The date related clinical entries were updated according to the census date of 30.01.2019. Specifically, the date of sampling (defined as the first visit to the corresponding medical center when the blood sample was taken) was composed of additional data on patient samples and healthy volunteers from Imperial and Sheffield local databases. They were mapped to the sample ids to be used either by the clustering algorithms or the downstream biological analyses. Samples with multiple trials were flagged and technical repeats were excluded from the sample pool. Samples with multiple visits (where blood was collected on different dates) were dated to retain the chronological relation between them as useful longitudinal information for the biological analyses.

Feature selection of genes

Next generation sequencing methods measure the expression of thousands of genes providing a huge number of dimensions (>20,000) per sample. However, diseases are usually regulated by smaller groups of genes rather than thousands, as reviewed for PAH in (Ma and Chung 2017). Therefore, the majority of genes are expected not to contribute to PAH development. Additionally, gene filtering helps in reducing the computational burden of most clustering algorithms also affecting their performance by removing misdirecting noise, as shown in (Rodriguez et al. 2019). Investigating IPAH subgroups concerns the structure underneath the idiopathic form of the disease and is characterised by the complete lack of sample labels. Standard feature selection methods, used for RNAsequencing data, can not be used in this case as

they require labels, e.g. in (Rodriguez et al. 2019; Wenric and Shemirani 2018) where they use disease and control labelling. Therefore, to investigate disease subgroups we ranked all genes based on the variability of their expression across patient samples, as expression variance indicates interesting gene behaviour in our disease context. Each gene was scored according to its variability across the 385 patients using the var() function from the Stats R package. Subsequently, all 25,955 genes were ranked based on that score. To generate the candidate gene sets for the determination of the most stable one, multiple subsets of the top ranked genes were extracted, each time increasing the size by 100.

Highest stability gene set

In our case of feature selection we needed to select one of the gene sets to base the clustering on. Since there are no known IPAH-related genes in the literature we moved forward with the gene sets (one per pipeline) that generated the subgroups of highest stability. We used the established clusterboot function from the fpc R package. The function assesses the clusterwise stability by resampling the data in a bootstrap approach. Then it computes the Jaccard similarities of the original subgroups to the most similar subgroups in the resampled data. Spectral clustering (kernlab R package), 50 resampling runs, k between 2 and 6 and a seed of 28588 for reproducibility purposes was used. We generated multiple gene sets starting from the top 100 ranked genes and increasing the size of the gene set by 100 until we reach the total number of 25,955 genes.

Clustering algorithm selection

There is a wealth of methods in the unsupervised learning field that are appropriate for certain data types. Most studies employ widely-used methods (e.g. hierarchical clustering) without utilising any kind of selection method that would point towards a certain effective methodology. In this study we aimed to examine a group of diverse algorithms that cover different clustering approaches. Since we lacked labels, and thus a performance measure, we compared the partitioning consistency of the different approaches on the expression data. As good consistency we defined high agreement and low standard deviation calculated between different variations of a clustering algorithm. When two different clustering runs agree on the partitioning of the samples they show robustness since they do not randomly assign samples to subgroups but rather are driven by the underlying structure of the data. (**Supplementary Table 6**) presents the various algorithms used, along with various distance measures and clustering categories they belong to. The preprocessed RNA-seq dataset and k2,10 were used for the determination of the algorithm

agreements. Within each pair of clustering runs the agreement was calculated using the adjusted Rand Index, the corrected for-chance version of the original Rand index (Rand 1971), which is based on the number of times any pair of 4 points is partitioned in the same subgroup throughout different clusterings. To calculate the intra-agreement of each clustering algorithm (spectral, k-means, hierarchical) we considered only pairs in which both runs were based on the same algorithm. For those pairs the agreement was averaged across clustering runs and ks. In a similar way, the standard deviation per algorithm was calculated.

Optimal number of subgroups k

Estimating the actual number of subgroups is a key decision for every clustering algorithm and is usually based on field knowledge, specific study decisions or statistical indexes. Due to the lack of prior knowledge about the structure of IPAH we are unable to set the number of subgroups based on literature or biology. Since this is an exploratory study (which does not include a predetermined number of subgroups) we are utilizing various indexes depending on the questions we are addressing.

For the case of the IPAH subgroups, we are using ensemble learning, the process where multiple indexes (or experts) participate in the selection of the optimal value of a machine learning decision [2] . Specifically, since we can not use any class related labels, we estimate the optimal number of subgroups using voting among 15 internal indexes that evaluate the compactness and/or the distance between different subgroups (see **Supplementary methods: Internal Index Voting**).

Determining the optimal number of clusters (k) is an inherently difficult task in unsupervised machine learning as it is always an educated estimation, since we do know the actual number of categories within our data. Indeed some of the 14 used indexes are bound to not work on our data type (RNA-seq) and that is why we used an ensemble/voting method to estimate k (**supplementary section Internal index voting**) since we cannot base our estimation on any one index. The voting result (**Supplementary Data 1**) showed the clear majority of indexes to favour up to 5 clusters, with a preference to 2 clusters. The most important aspect in selecting the number of clusters in a data set is retaining as much information as possible, therefore selecting the highest supported k minimizes information lost. Following that notion, we retained the highest voted k = 5, where we discovered 3 distinct adequate sized clusters and 2 small clusters that, despite their interesting gene expression profiles (**Figure 2A**), were unable to show any statistical significance in follow-up work due to their small size. In **Supplementary Figure 16**, we demonstrate the flow of patients between clusterings along with the cluster sizes and the proportion/count of transferring patients across k. The colored nodes represent our subgroups I, II,III, IV and V. According to the clustering tree, the 3 main subgroups I, II

and V remain clustered together when k = 2 (in a 341 sized cluster) and k = 3 (in a 295 sized cluster). This indicates that for k < 5 we are missing the information that separates these 3 distinct subgroups. The two smaller subgroups (III and IV) mostly originate from a group of patients (circled in green) that dissociates early on from main subgroups I, II and V implying that these samples show some differences even when less subgroups are requested. The remaining samples that end up in subgroup III have a common parent with subgroup II.

Gene signatures of subgroups

A number of biological analyses are used to explore IPAH sub-structure. For the patients of each subgroup survival (Kaplan-Meier curves), response to vasodilators (IPAH treatment), gender and functional class (Fisher's exact test) are calculated and compared while measuring significance. Additionally, the difference in the age of diagnosis (one-way anova test) and a number of known PAH-related genes are examined across subgroups. We perform a driver gene discovery analysis (LASSO regression model) which can indicate the most influential genes whose literature-generated annotations are investigated. The results of the patient 5 clustering are used to draw genetic differences between the two groups. IPAH subgroups are subjected to differential expression analysis and subsequently to pathway analysis in order to interpret the genes' involvement in terms of functionality. The p values of each gene when considering the fold change between subgroups I and II were calculated using a Welch Two Sample t-test on the raw values and presented in the **Supplementary Table 4**. The absolute lower cut off values of fold change (log2 scaled) is 0.28. [*Initial regression performed by Thomas S Mascarenhas*]

Internal index voting

To implement ensemble learning, we used the majority voting rule among 15 internal validation indexes, selecting as the optimal number of subgroups (k) the one voted by the most indexes. The various indexes[6–17] , (McClain and Rao 1975) and (Ray and Turi 1999) results can be found in Supplementary Data 1. All of the above are based on variations of the same idea, to score a partitioning on how compact each subgroup is and how well the subgroups separate. No index can select the real number of subgroups with perfect accuracy, therefore the "voting of experts" method can provide a safer alternative to using any one index. The preprocessed RNAseq dataset after the appropriate gene filtering step was used for this work.

Clinical variable associations / classification

Survival analysis using a Kaplan-Meier estimate [18] was undertaken to compare the time until death between patient subgroups, a measure able to overcome issues such as subjects withdrawing from the study or not experiencing death during the course of the study's observations. In this cohort withdrawal etiologies include withdrawal by clinician, transplant, leaving the country, and loss to follow up. As the clinical data for the patients did not include a cause of death, it was decided to limit the duration of the analysis to minimise the inclusion of other causes of death. A duration of 10 years from diagnosis was selected as it is a period during which almost 80% of IPAH deaths occur if conventional therapy is used (Kang et al. 2014), while also being sufficient time to allow for useful statistical analysis. Patients who did not die during this period were considered to be alive for the analysis and had their survival time set to 10 years. The Kaplan-Meier model was created using the survival R package to compare survival time between the subgroups, and subsequently plotted using survminer R package (ggsurvplot function). Cox Regression was undertaken to show any statistically significant survival differences between the subgroups. A Cox model was selected as it has been shown to be a more flexible alternative to parametric methods, and does not require the distribution of survival times to be stated (Bradburn et al. 2003). It was noted that patient survival could be affected by factors other than their subgroup membership. Therefore, survival analysis was repeated using a multivariate Cox regression which included the patients' age at diagnosis, sex, and New York Heart Association (NYHA) functional class. This method allows for adjustment to the impact of these other factors, and shows an estimate of their respective strength of effect.

A frequency table was created for vasoresponders, gender and each functional class within each patient subgroup. Pairwise comparisons were made between the subgroups using Fisher's exact test. This test was performed using the rcompanion R package and a Bonferroni correction was used as multiple comparisons were made.

A comparison of the age at IPAH diagnosis was made between the patient subgroups using a one-way anova test. This test assumes that the observations within each subgroup are normally distributed, and that the data are homoscedastic with equal standard deviation between the subgroups. The normality of the data within each subgroup was confirmed visually by producing a histogram for the age of diagnosis for each subgroup, as well as by plotting a histogram of the residuals for the anova model to ensure that these followed an approximately 6 normal distribution, and creating a Q-Q plot of the residual values. The homoscedasticity was assessed by plotting the residual values against the fitted value, and by using the car R package to perform a Levene's test.

A regression model was used for feature selection of genes whose expression most significantly drove subgroup membership. The RNA-seq counts were split into a training and testing set with a ratio of 70:30. The glmnet R package uses penalised maximum likelihood in order to fit a regression. The model family was set to “multinomial” as the model was to be used to predict a nominal dependent variable with multiple categories, subgroup membership, given gene expression data. As an additional parameter, the type.multinomial was set to “grouped”, meaning that multinomial coefficients for a variable were included or excluded together. This also demonstrated an increase in model prediction accuracy during initial testing of models. Ridge, elastic-net, and lasso models were created using the training set, their parameters optimised by cross-validation (cv.glmnet function), and then used to predict the subgroup membership of samples in the test data-set. The models run on the entire data-set to produce coefficients for each gene relating to each subgroup. The elastic-net regression model was preferred based on its ability to select strongly correlated variables in or out together, a useful feature when dealing with genes which may be correlated due to sharing a biological pathway (Zou and Hastie 2005). From the regression results heat maps were produced showing the coefficients for genes in each subgroup. Genes in the top 5% for largest coefficients were selected for further investigation to identify common pathways or functions. It was decided to investigate genes with both positive and negative coefficients, as genes with decreased gene expression which drove subgroup membership are still of biological interest.

The pathfinder R package was used to demonstrate enriched pathways between subgroups. Genes with an absolute LFC in the highest 10% for their subgroup, and with an adjusted p value ≤ 0.05, were inputted to the package. [*Initial regression analysis and statistics tests performed by Thomas S Mascarenhas*]

Identifying clinical signatures of RNA subgroups

The following clinical variables were used during the clinical signature pipeline: age_diagnosis, sex, diagnosis, drug_exposure, bmi, functional_class, 6 minute walking distance, oxygen saturation (pre), oxygen saturation (post), mRAP, mPAP, mPAWP, cardiac output, SvO2, vasoresponse (lenient), vasoresponse (stringent), FEV1, FVC, TLC, KCO, right Atrial Area, Right Ventricle, Tricuspid Apse, emphysema Category, Fibrosis Category, Thromboembolic disease, NT-proBNP, BNP, Urate, inflammation CRP, haem. HB, haem. WBC, haem. Platelets, Renal Sodium, Renal Potassium, Renal Urea, Renal Creatinine, Metabolic Syndrome, Comorbidity HHT, Comorbidity epistaxis, Comorbidity bleed, Comorbidity AVM, Comorbidity Cirrhosis, Comorbidity Hepatitis, Comorbidity PPH, Comorbidity DM1, Comorbidity DM2, Comorbidity Hypothyroidism, Comorbidity SLE, Comorbidity SS, Comorbidity Ankylosing Spondylitis, Comorbidity Sjogren, Comorbidity UCTD, Comorbidity

Necrotising Vasculopathies, Comorbidity Overlap Syndrome, Comorbidity Polymyalgia Rheumatica, Comorbidity COPD, Comorbidity Asthma, Comorbidity OSA, Comorbidity CAD, Comorbidity CVA, Comorbidity PAD, Comorbidity HTN, Comorbidity Arrhythmia, Comorbidity Hyperlipid, Comorbidity PE, Comorbidity Heterotaxy, Comorbidity Asplenia, Comorbidity CKD, Comorbidity CA, PVR, Sum of Comorbidities, Any Comorbidity.

Ensemble feature selection[19] based on recursive feature elimination (RFE)[20,21] and a linear SVM[22] as the estimator, was used to ensure robust identification of the smallest set of clinical features (signature), from the clinical features, that best describe each subgroup. RFE feature ranking was based on absolute weights of features from the SVM, which quantifies the contribution or importance of each feature towards the multivariate construction of a hyperplane separating the subgroups. The regularisation parameter of the SVMs was set to C=1. The discovery dataset used for feature selection was resampled without replacement into 500 subsamples (90% of samples), for each subgroup over signature sizes (s) ranging between s=1 to 20. Each 7 resampled dataset was further divided into bootstrap samples (k), with k=50. Feature values of each bootstrap sample were normalised to improve feature selection performance. RFE-SVM[20] was used to rank features for each bootstrap sample, and an aggregate rank was calculated for each feature using all k bootstrap rankings. This process was repeated over all resampled datasets, resulting in 500 candidate signatures for each signature size, s. Each candidate signature was then used to develop a classification model, which was then trained on the discovery dataset to discriminate between a given subgroup from all other subgroups. Classification models were built using support vector machines (SVM)[23], random forest (RF)[24], logistic regression (LR)[25], and knearest neighbour (KNN)[26] .

LR was implemented using sklearn.linear_model.LogisticRegression using l2 penalty and default values used for all other parameters. SVM was implemented using sklearn.svm.LinearSVC with regularisation parameter C set to 1, and default values used for all other parameters. RF was implemented using sklearn.ensemble.RandomForestClassifier with default values used for all other parameters. kNN was implemented using sklearn.neighbors.KNeighborsClassifier with a number of neighbours (n_neighbors) set to 5 and default values used for all other parameters.

For feature selection tasks, we used ensemble feature selection based on recursive feature elimination (RFE) technique. RFE is a backward feature elimination technique that iteratively prunes the least informative feature(s) from a training dataset. A RFE based on a linear SVM starts by using all features to train an SVM model and ranks all features according to importance. The least ranked feature is removed from the training dataset and the SVM model refitted. This is iteratively done until only the required number of features remain. All features are also ranked according to importance.

Ensemble feature selection aggregates several feature rankings into a single consensus feature ranking to ensure robustness of the feature selection process and of selected features. Feature importance measures used for feature ranking are based on the hyperplane weight vector of a linear support vector machine (SVM). The weight vector quantifies the contribution of each feature to the construction of the hyperplane, and is used for ranking features according to importance.

Ensemble feature selection[56] based on recursive feature elimination (RFE)[57,58] and a linear SVM[59] as the estimator, was used to ensure robust identification of the smallest set of clinical features (signature) that best describe each subgroup. RFE feature ranking was based on absolute weights of features from the SVM, which quantifies the contribution or importance of each feature towards the multivariate construction of a hyperplane separating the subgroups. The regularisation parameter of the SVMs was set to C=1. The discovery dataset used for feature selection was resampled without replacement into 500 subsamples (90% of samples), for each subgroup over signature sizes (s) ranging between s=1 to 20. Each resampled dataset was further divided into bootstrap samples (k), with k=50. Feature values of each bootstrap sample were normalised to improve feature selection performance. RFE-SVM[57] was used to rank features for each bootstrap sample, and an aggregate rank was calculated for each feature using all k bootstrap rankings. The set of feature rankings R, aggregated over all bootstrap samples, is calculated as

$$R \sim \left(\sum_{i=1}^{k} \quad r \quad {}_{i}^{1} \sum_{i=1}^{k} \quad r \quad {}_{i}^{2} \cdots \sum_{i=1}^{k} \quad r \quad {}_{i}^{N-1} \sum_{i=1}^{k} \quad r \quad {}_{i}^{N}\right) (1)$$

where k is the number of bootstrap samples, N is the total number of features in the dataset, and r_i^n is the rank of feature n in bootstrap sample i. This process was repeated over all resampled datasets, resulting in 500 candidate signatures for each signature size, s. The candidate signature that obtained the best performance was selected. This process was repeated for all signature sizes, s=1 to s=20, for subgroups I, II and v. A final signature for each subgroup was selected based on a compromise between the fewest number of features (s=1 to s=20) and classification performance. Final selected signatures for each of the subgroups were pooled to create a composite signature, which was then used to develop a multi-class classification model. The model was trained on the discovery dataset to discriminate between subgroups I, II and V, used to predict subgroup membership of an unseen validation dataset. The predicted subgroup membership was then used to calculate survival of predicted subgroups. Survival of the 8 predicted subgroups was compared to known survival of subgroups in the discovery dataset for validation purposes. [*This whole section was written by Emmanuel Jammeh*]

Differential expression analysis

Differential expression analysis was performed between patients' subgroups. The raw un-normalised counts which were the output of the Salmon quantification were used. The DESeqDataSetFromTximport function was used to create an input data-set for DESeq2 which included the raw count data, and the subgroup membership for each sample. Rows with fewer than a total of 10 counts were excluded in order to decrease computing time. Utilizing the apeglm R package the log fold change (LFC) shrinkage was performed on the results in order to reduce noise from genes with low counts, while retaining genes with large fold changes. This is an alternative to introducing filtering thresholds or pseudocounts which have disadvantages such as resulting in the loss of genes with true expression differences. This method was used to create pairwise comparisons of gene LFC between the control subgroup and each patient subgroup. Using the LFC data, genes with a log2fold change of greater than +/-1.5 were selected, and ranked by their p-value. The LFC data used ensembl92 gene IDs, so the biomaRt R package was used to add HUGO Gene Nomenclature Committee (HGNC) gene names and a brief description to allow for easier identification of genes of interest. [*Performed by Thomas S Mascarenhas*]

Secondary clustering analysis with healthy volunteers

All RNA-seq samples (508) are utilised along with the genes that provide the most information when attempting to predict whether each sample belongs to a patient or a healthy volunteer. We used the patient and control labels as ground truth. Utilizing the labels, we ranked the genes according to the amount of information they contribute in distinguishing the two classes. To determine the amount of information each gene contributed towards separating IPAH and healthy samples we used the Information Gain Criterion from the Biocomb R package. The 25,955 genes were scored and ranked. To generate the candidate gene sets for the determination of the most stable one, multiple subsets of the top ranked genes were extracted, each time increasing the size by 50. While investigating the differences between disease and healthy samples we can utilise the only ground truth we have established, the partitioning of the samples into patient and control groups. This knowledge enables the selection of the number of subgroups (k) to be based on the average subgroup purity of each k and compare them to select the k with the highest average purity.

Additional clustering pipeline of IPAH patients

To estimate the impact the 33 HPAH patients had on our main clustering pipeline, we examined their distribution across subgroups and ran an additional clustering pipeline including exclusively the 313 IPAH samples. As in the main pipeline, we utilised spectral clustering with the rbfdot kernel, the 300 most variable genes and identical preprocessing (section **Sample and gene selection preprocessing**).

Supplementary Tables

Supplementary Table 1 *| Distribution of HPAH samples across the 5 subgroups and their proportion in each subgroup.*

I	II	III	IV	V
12(9.3%)	6(5.35%)	3(15.7%)	1(10%)	11(12.3%)

Supplementary Table 2 *| Bonferroni adjusted p-values for various white blood cell counts.*

Subgroup Pairs	*Lymphocytes*	*Eosinophils*	*Monocytes*	*Neutrophils*	*Neutrophils / Lymphocytes*
I - II	0.13	0.18	0.0076	7.2e-12	0.0061
I - III	0.14	1	1	8.0e-04	0.1100
I - IV	1	1	0.0390	8.7e-03	1
I - V	1	1	0.5900	1.3e-03	1
II - III	0.47	1	1	1	1
II - IV	1	0.87	1	3.2e-10	1
II - V	1	1	1	4.4e-04	0.23
III - IV	1	1	1	2.7e-06	1
III - V	0.16	1	1	1.3e-01	0.0670
IV - V	1	1	1	6.4e-08	1

Supplementary Table 3 *| The distribution of BMPR2 mutations across all patient subgroups (n=357).*

	I	II	III	IV	V
BMPR2	22 (20.5%)	23 (26.1%)	4 (26.6%)	4 (50%)	18 (26.4%)
Not BMPR2	107	88	15	8	68

Supplementary Table 4 *| All fold changes p values of the gene signature for subgroups calculated using a two-sided t-test.*

Gene	P value	Gene	P value
ALAS2	1.60E-12	MIR5195	2.36E-13
LTF	5.40E-06	IGKJ4.43	1.92E-17
CRISP3	4.99E-07	IGHM	8.50E-16
CTSG	1.53E-05	IGHV4.39	2.69E-12
RP11.20D14.6	1.73E-08	IGLV2.14	6.82E-22
RP11.678G14.3	4.53E-09	IGKV1.27	2.42E-10
NPRL3	0.006121994	IGHV3.48	2.55E-17
CH17.296N19.1	6.31E-05	IGLV6.57	2.01E-11
RP1.229K20.9	0.002453699	IGLV7.43	5.59E-18
AC131056.3	0.08654707	IGHV2.5	4.21E-08
MT.RNR1	0.4689423	MIR5195	2.36E-13
NOG	1.05E-11	IGKJ4.43	1.92E-17

Supplementary Table 5 *| Correlations between discovery, validation and external validation cohorts. In the discovery set, Spearman correlations were calculated between RNA-req TPM values, the validation set between negative delta Cts values with GAPDH used as the endogenous control gene. Green cells denote agreement in the directionality of gene-clinical variable correlations between the validation sets and the discovery set. Red cells denote disagreement/opposite correlation between the two data sets. Notation of (**) denotes a p-value of less than 0.01, (***) denotes a p-value of less than 0.001 (using asymptotic approximated p-value by using the t distribution , while no stars denote non-significant p-values.*

Genes	Discovery [n= 359]			Validation [n =91]			External Validation [n =32]		
	Age	BMI	SixMWD	Age	BMI	SixMWD	Age	BMI	SixMWD
ALAS2	0.2***	0.38***	-0.32***	-0.006	0.06	-0.1	-0.08	0.31	-0.35
C4BPA	0.01	0.02	-0.07	0.14	-0.11	-0.13	0.02	-0.19	0.25
CRISP3	0.16*	0.14	-0.16*	0.11	0.016	-0.22	-0.07	0.14	0.05
CTSG	0.19***	0.14*	-0.12*	0.07	0.06	-0.17	0.06	0.19	-0.12
IFI27	0.19**	0.16*	-0.22***	0.24*	0.004	-0.28*	0.1	0.1	-0.32
IGHM	-0.29***	-0.18***	0.2***	-0.42***	-0.21*	0.18	-0.15	0.33	-0.02
IGHV3.48	-0.27***	-0.08	0.14*	-0.31	0.34	-0.5*	-0.11	0.22	0.12
IGKV2.24	-0.25***	-0.26***	0.24***	-0.45***	-0.17	0.11	-0.07	0.14	0.15
IGLV6.57	-0.21***	-0.16*	0.15*	-0.5***	-0.08	0.16	-0.31	0.32	0.04
LTF	0.18**	0.19***	-0.13*	0.08	0.12	-0.23*	0.01	0.08	0.01
NEBL	0.01	0.05	0.01	-0.09	-0.01	-0.23	0.03	0.12	-0.16
NOG	-0.44***	-0.19***	0.2***	-0.58***	-0.2*	0.18	-0.13	0.02	0.14
NPRL3	0.05	0.18*	-0.13*	-0.002	-0.09	-0.1	0.03	-0.06	-0.41
PI3	0.15**	0.25***	-0.17**	0.1	0.04	-0.21	-0.13	-0.02	0.31
SMIM11A	-0.03	-0.2***	0.12*	-0.24*	-0.13	0.06	0.1	-0.04	0.18

Supplementary Table 6 *| The clustering algorithms, their approach category and the various distance measures tested.*

Clustering algorithms	Category	Distance measures
K-means	Partitioning	rbfdot, polydot, tanhdot, laplacedot
Hierarchical	Hierarchical	euclidean, manhattan, minkowski, canberra
Spectral	Graph Theory	rbfdot, polydot, tanhdot, laplacedot

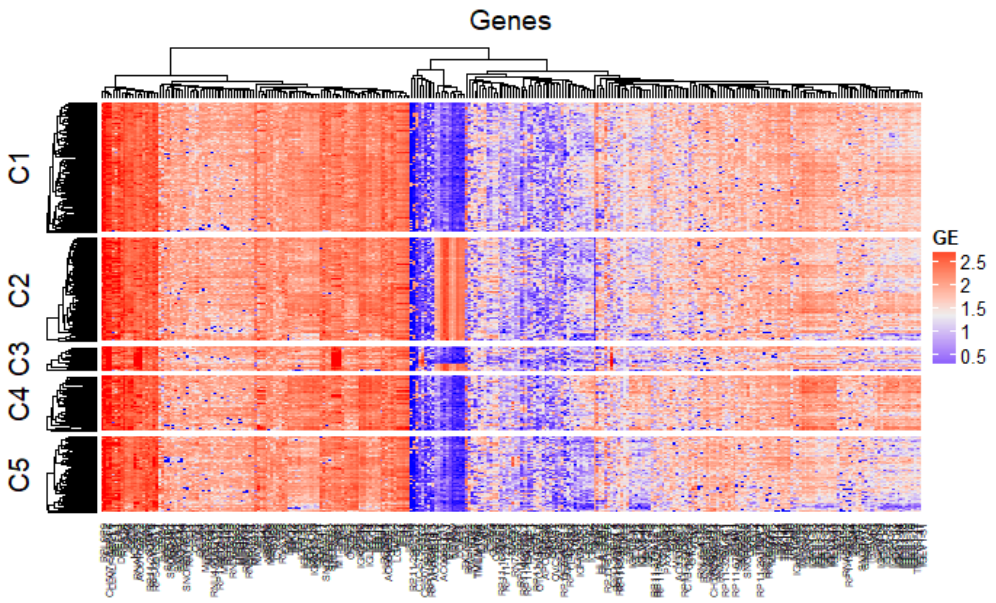
Supplementary Table 7 *| P-values relative to main Figure 4*

Relative CIBERSORT abundance between subgroups	**P value**
pI-II(Dendritic cells activated)	0.011
pI-II(Neutrophils)	$4.4 \times 10{-}11$
pI-V(Neutrophils)	$2.0 \times 10{-}3$
pII-V(Neutrophils)	$1.7 \times 10{-}3$
pI-II(T cells CD8)	$4.8 \times 10{-}5$
pI-II(T cells CD4 naive)	$1.9 \times 10{-}8$
pI-V(T cells CD4 naive)	$3.8 \times 10{-}3$
pI-II(T cells CD4 memory resting)	$2.3 \times 10{-}5$
pI-II(B cells naive)	$2 \times 10{-}5$
pI-II(B cells memory)	$2.5 \times 10{-}6$
pI-V(B cells memory)	$3.9 \times 10{-}3$
pI-II(Plasma cells)	$6.4 \times 10{-}4$
pII-V(Plasma cells)	$6.5 \times 10{-}5$
pI-II(Monocytes)	0.0053
Whole blood cell counts across subgroups	**P value**
pI-II (Neutrophils)	*$7.2 \times 10{-}12$*
pI-V (Neutrophils)	*$8.0 \times 10{-}4$*
pII-V (Neutrophils)	*$4.4 \times 10{-}4$*
pI-II (Neutrophils/Lymph.)	*0.0061*
pI-II (monocytes)	*0.0076*

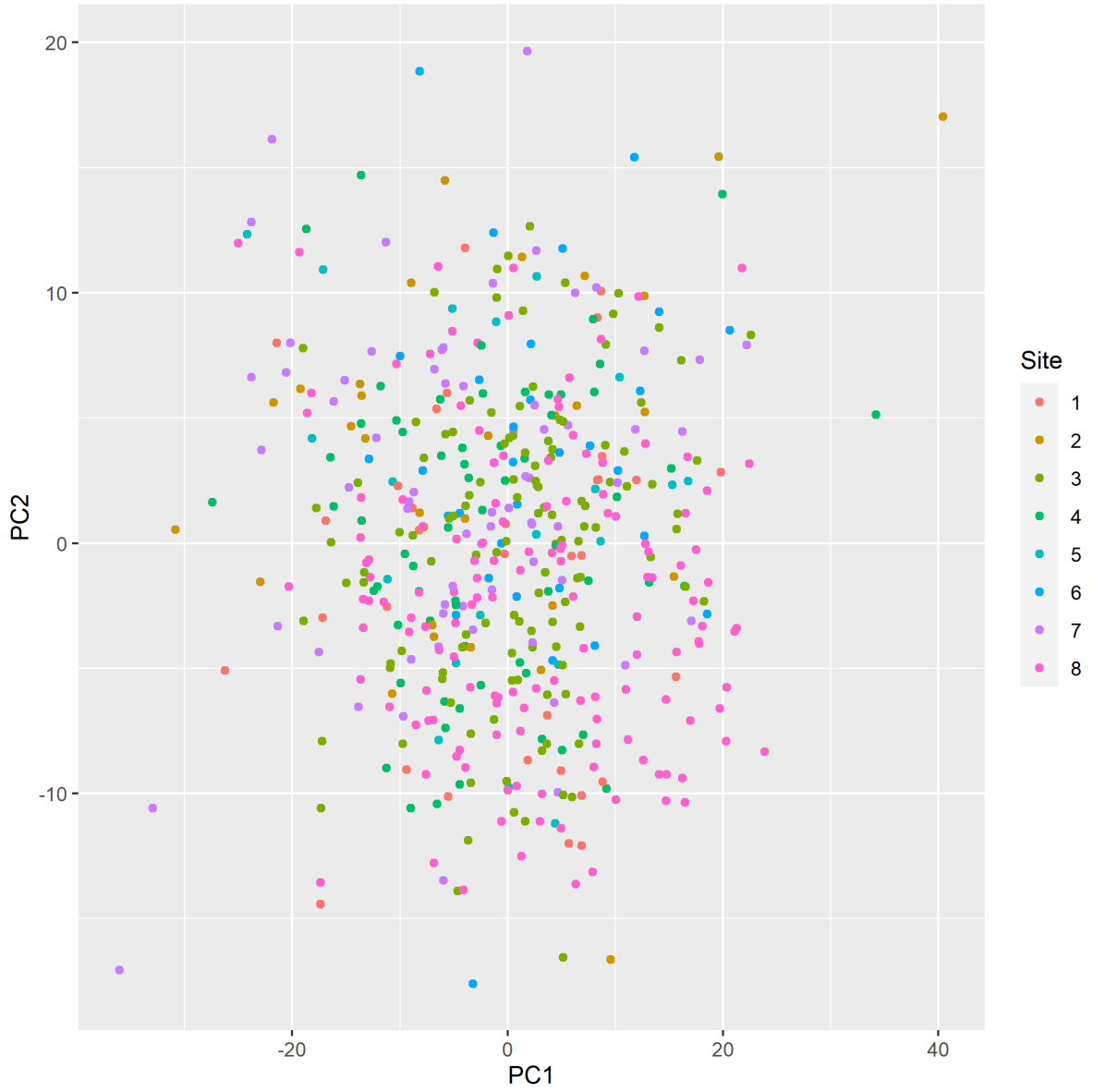
Supplementary Table 8 *| P-values relative to main Figure 5*

Correlation P values between gene and clinical variables	P value
BMI-ALAS2	$1.27 \times 10{-}11$
BMI-PI3	$3.17 \times 10{-}6$
BMI-IGHG2	$4.13 \times 10{-}6$
BMI-RP11.678G14.3	$8.22 \times 10{-}6$
BMI-IGKV1.27	$9.32 \times 10{-}6$
BMI-IGKV2.24	$3.09 \times 10{-}6$
BMI-IGKV4.1	$9.55 \times 10{-}7$
6MWD-IGKV4.1	$2.83 \times 10{-}6$
6MWD-IGKJ4	$2.08 \times 10{-}6$
6MWD-ALAS2	$7.52 \times 10{-}10$
AoD-IGHV2.5	$3.72 \times 10{-}10$
AoD-IGLV2.8	$1.06 \times 10{-}9$
AoD-IGHM	$6.2 \times 10{-}8$
AoD-NOG	$3.18 \times 10{-}17$
AoD-IGHV3.48	$7.7 \times 10{-}7$
AoD-IGLV7.43	$1.04 \times 10{-}6$
AoD-IGKV4.1	$6.35 \times 10{-}10$
AoD-IGKV2.24	$4.19 \times 10{-}6$
AoD-IGKV1.27	$3.93 \times 10{-}7$
OxygenSat-NOG	$1.11 \times 10{-}6$

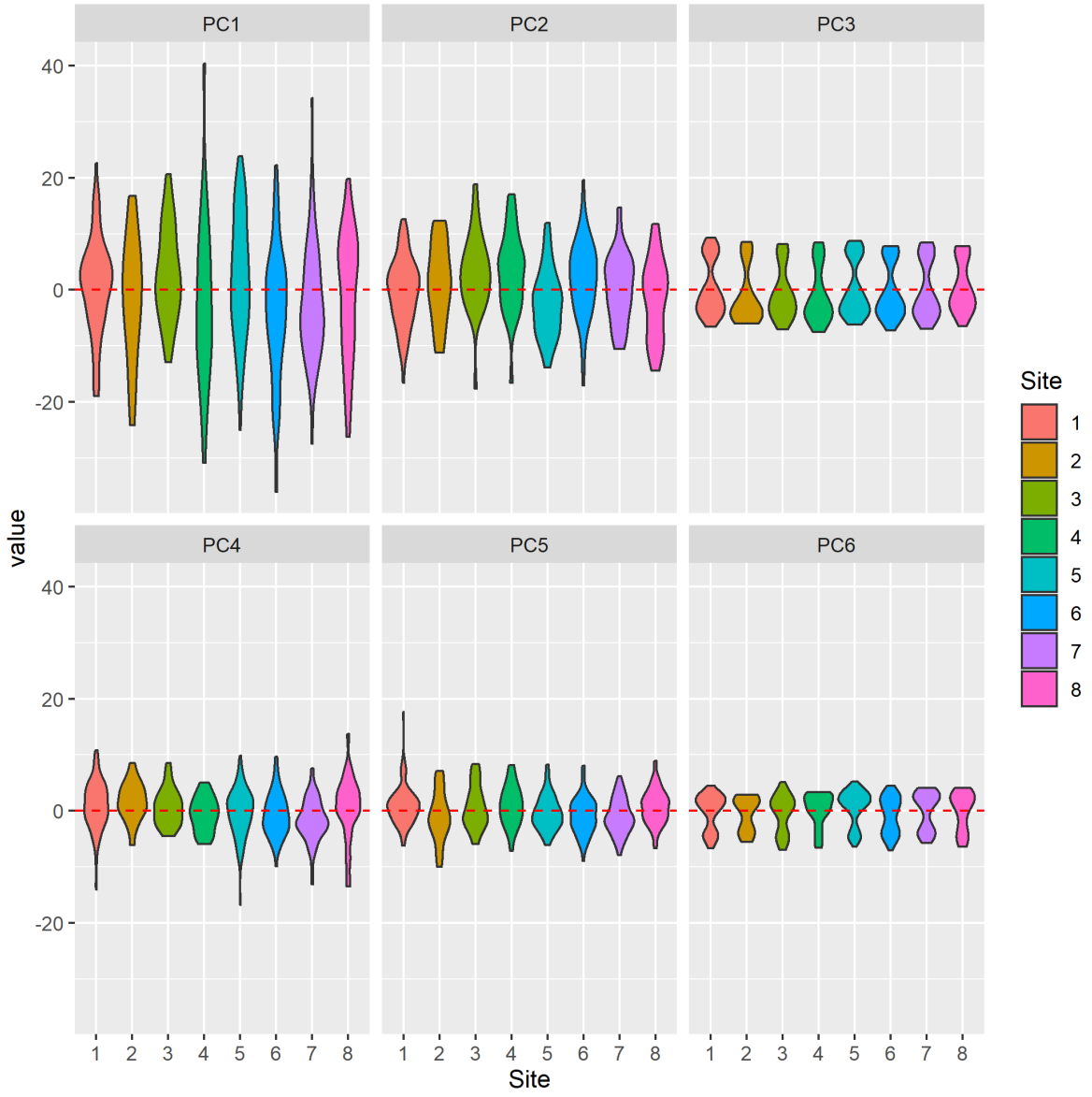
Supplementary Figures



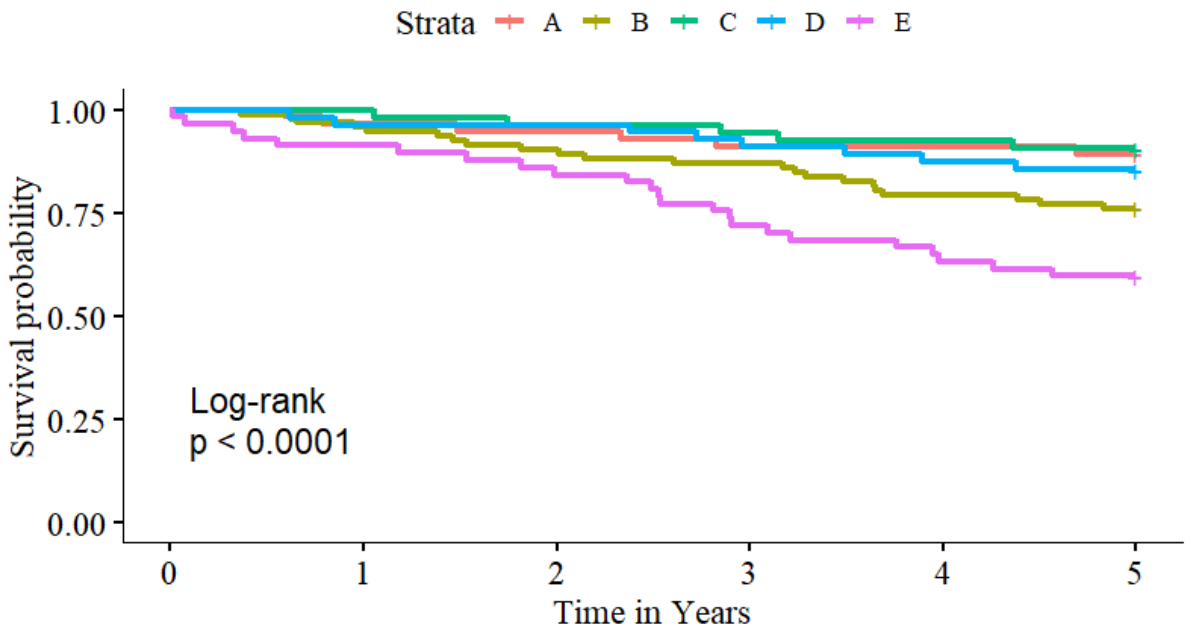
***Supplementary Figure 1**: Heatmap of gene expression after clustering with 11 male genes included. Separation of subgroups consisted solely of males (C2 and C3) or females (C1, C4, C5), thus obstructing the capturing of the disease signal. 11 male specific genes (PRKY, TTTY15, AC006032.1, RPS4Y1, EIF1AY, KDM5D, TXLNG2P, USP9Y, ZFY, DDX3Y, UTY) were observed to drive the initial clustering of genomic expression profiles.*



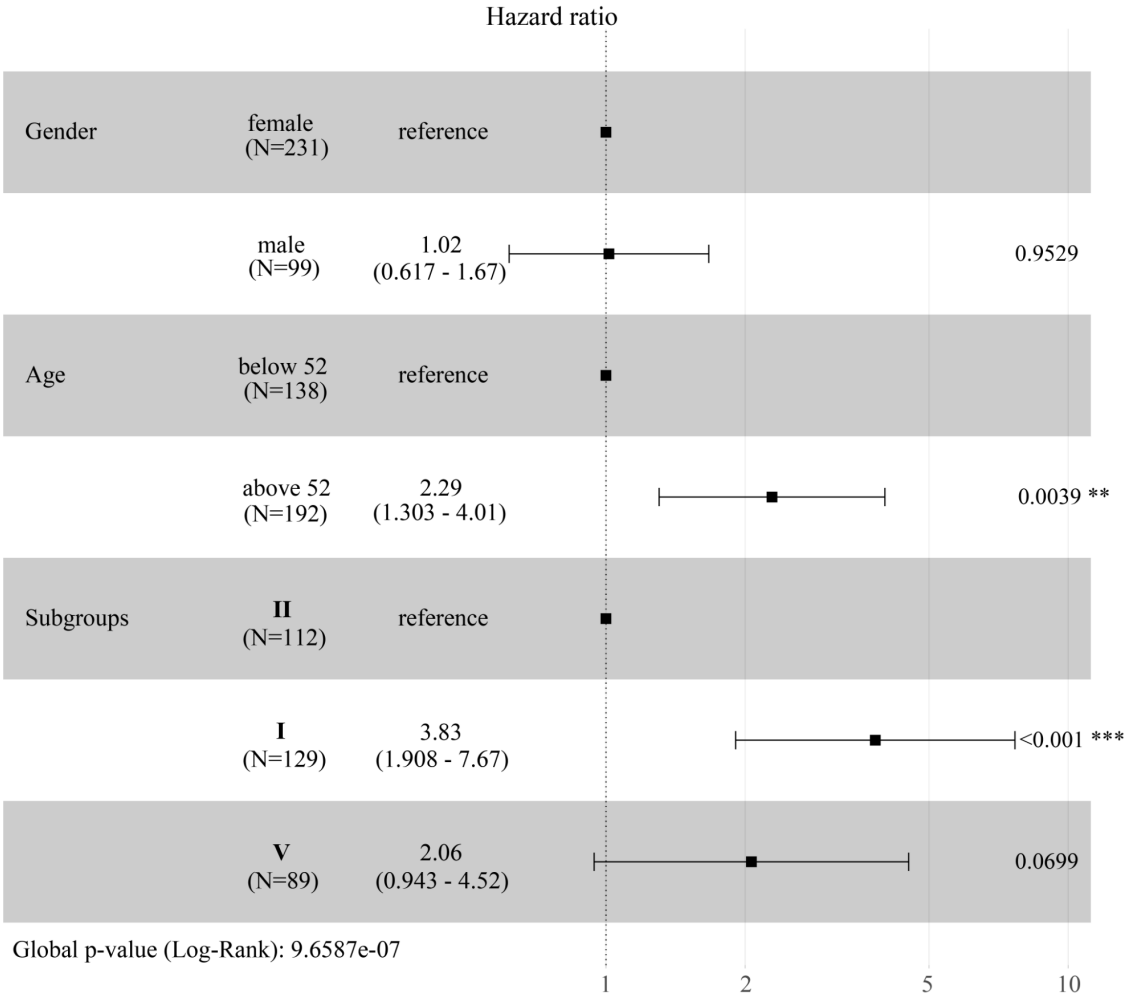
Supplementary Figure 2: *Plot of the first two principal components of the RNA-seq data derived from the 10,000 most variable genes according to IQR in our dataset. Each dot represents a distinct sample in the dataset coloured according to the institute that provided that sample. No discernible effect is seen due to the sample collection site. [Generated by Mark Dunning]*



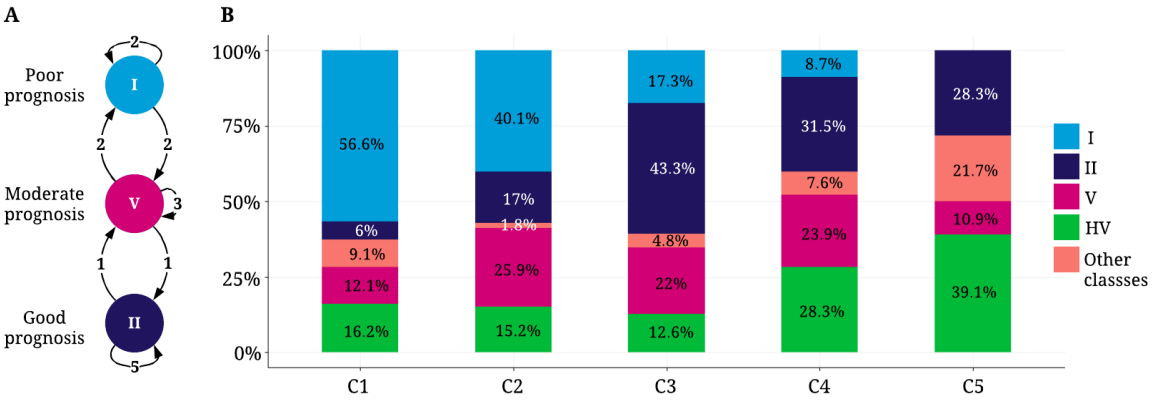
Supplementary Figure 3: *Boxplots showing the distribution of the first eight principal components of the RNA-seq dataset (n = 10,000 most variable genes) grouped according to the Site that provided the sample. No discernible effect is seen due to sample collection site. Red line represents the median, top and bottom bounds of the box represent min and max values. [Generated by Mark Dunning]*



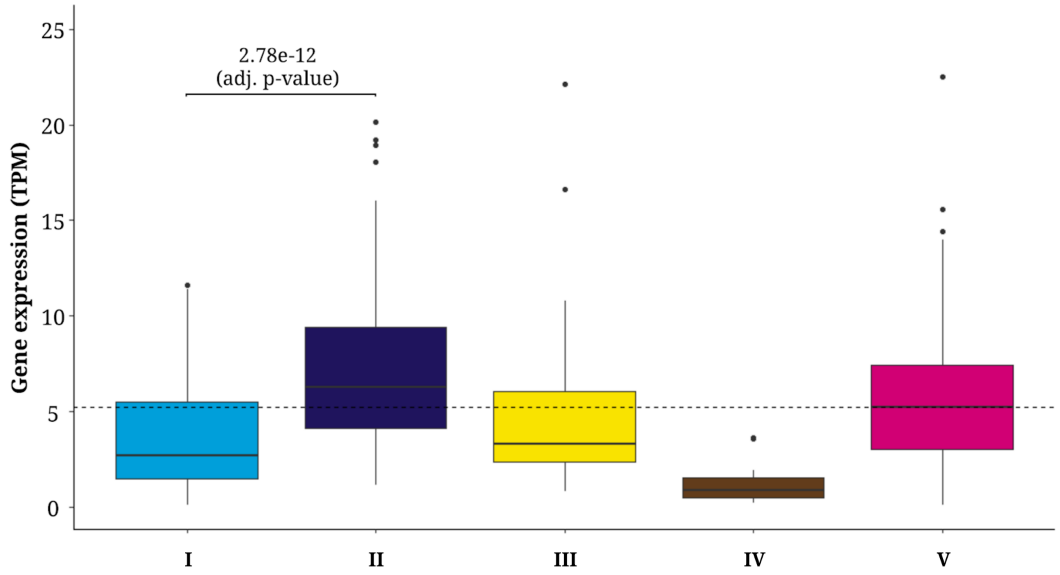
Supplementary Figure 4: *Survival of patients in clusters(A, B, C, D, E) created by clustering only IPAH samples with two-sided log rank test p-value.*



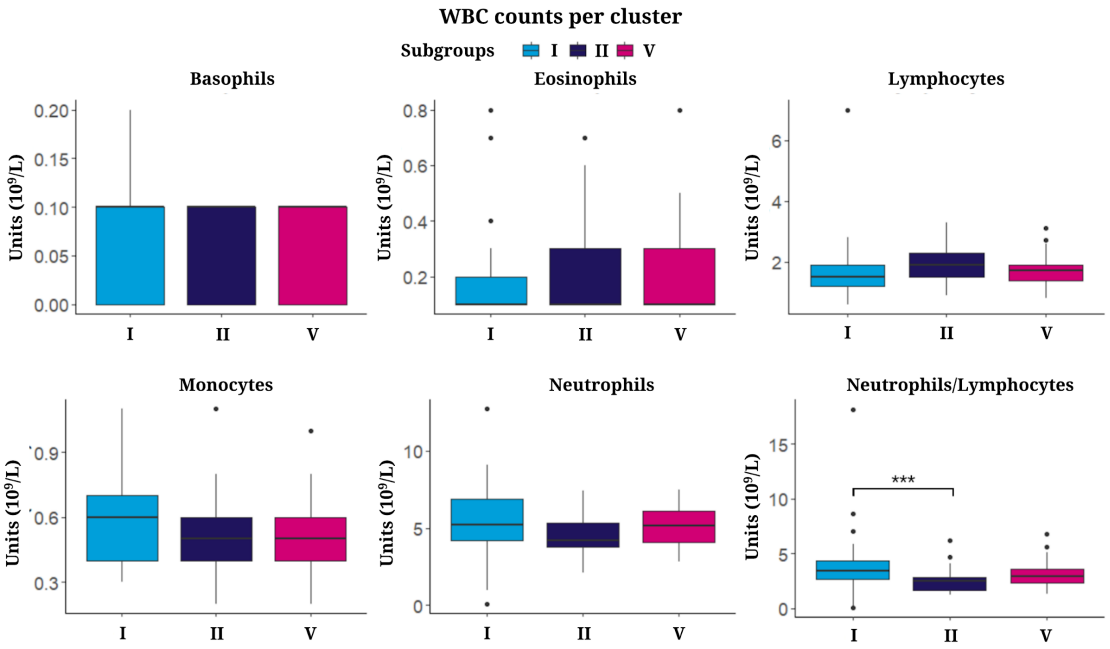
Supplementary Figure 5: *Hazard Ratio of discovery cohort clustering adjusted for gender and age category of patients. Notation of (**) denotes a two-sided log rank test p-value of less than 0.01, (***) denotes a p-value of less than 0.001, while no stars denote non-significant p-values. Data are presented as median values and error bars as 95% confidence intervals. Gender did not reveal any relationship with survival while an age over 52 was significantly associated with poor survival (HR=2.29). The most significant association with poor survival was found for patients classified in subgroup I.*



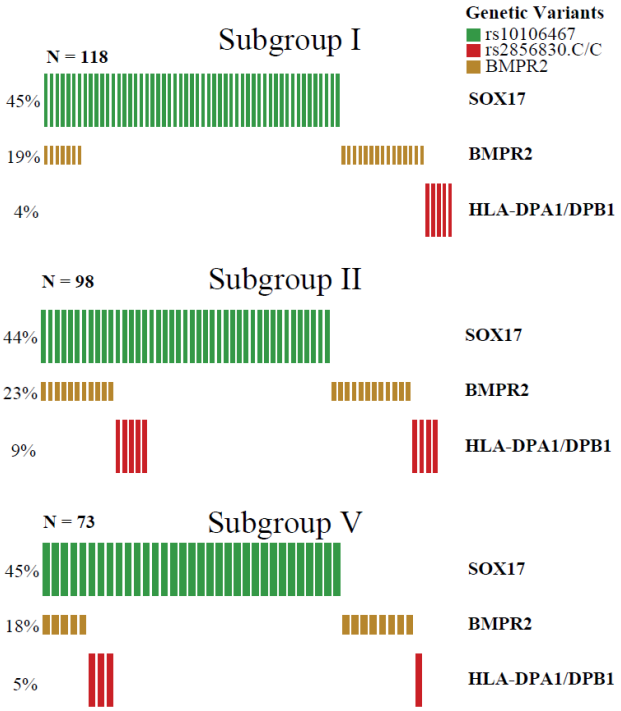
Supplementary Figure 6 *| Step-wise progression of IPAH. From healthy individuals to subgroup II, then V, then I. (A) longitudinal samples change across subgroups with different prognoses. (B) The five subgroups (C1-C5) from our secondary clustering pipeline which included patients and healthy volunteers. Each subgroup contains subgroups from our original clustering (I, II, V and the other classes -III and IV-) and healthy volunteers (HV).*



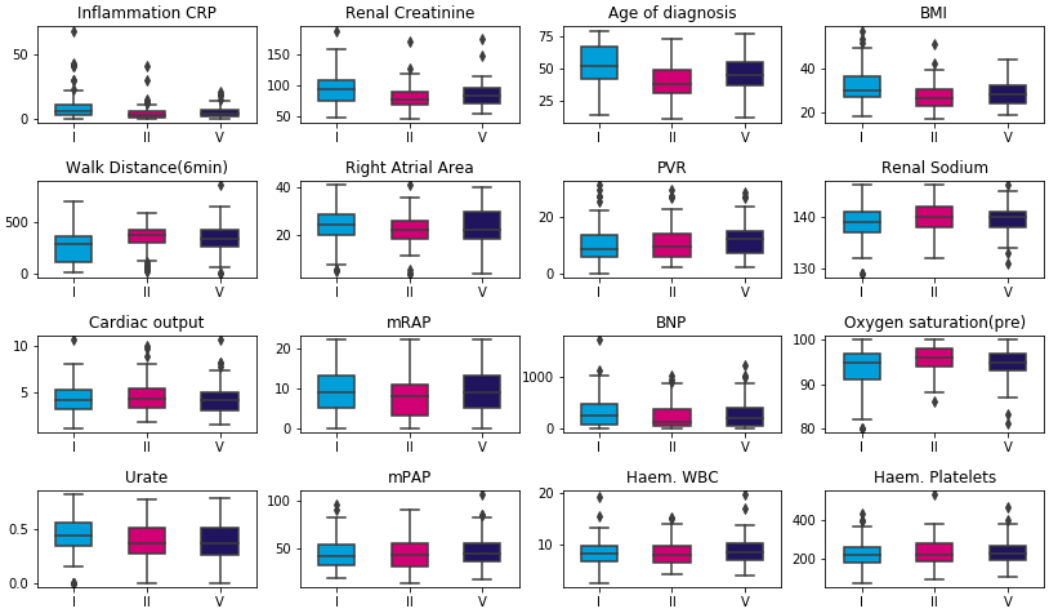
Supplementary Figure 7*: Noggin presents significantly higher expression in the high survival/low risk subgroup II. Furthermore, it is shown to be upregulated in the same subgroup based on our LASSO regression analysis. Boxplots were calculated for I (n=134), II (n=119), III (n=19), IV (n=10) and V(n=98). Vertical line represents the median, top and bottom bounds of the box represent the first and third quartile, while the tips of the whiskers represent min and max values.*



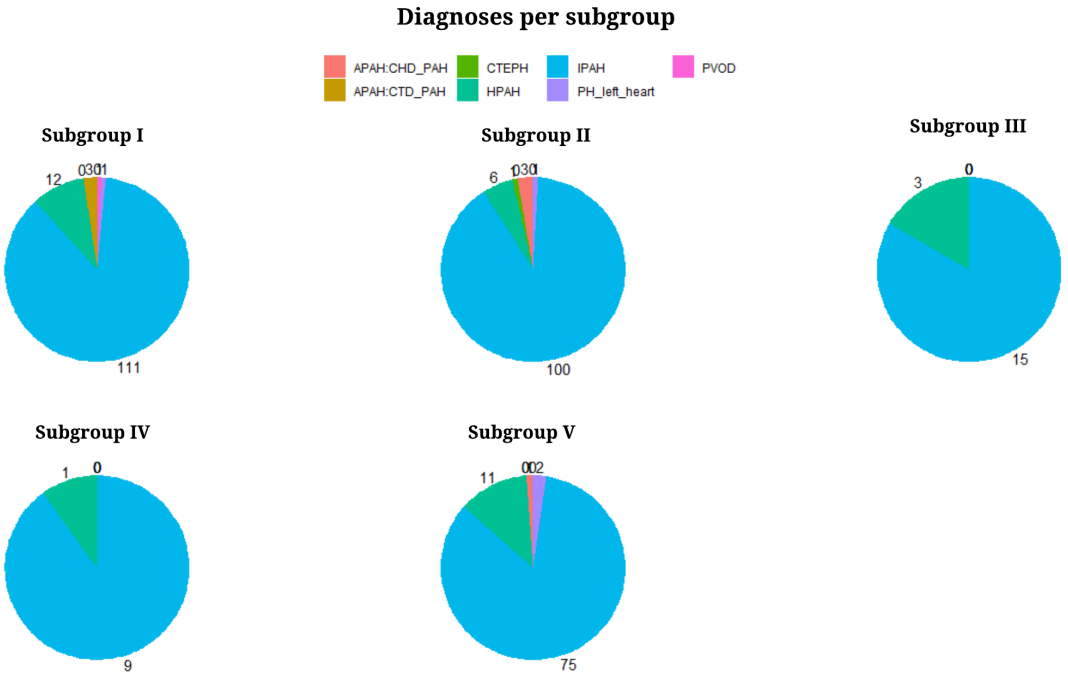
Supplementary Figure 8: *Quantity of various types of white blood cells across subgroups I (n=41), II (n=43) and V(n=28). Vertical line represents the median, top and bottom bounds of the box represent the first and third quartile, while the tips of the whiskers represent min and max values.*



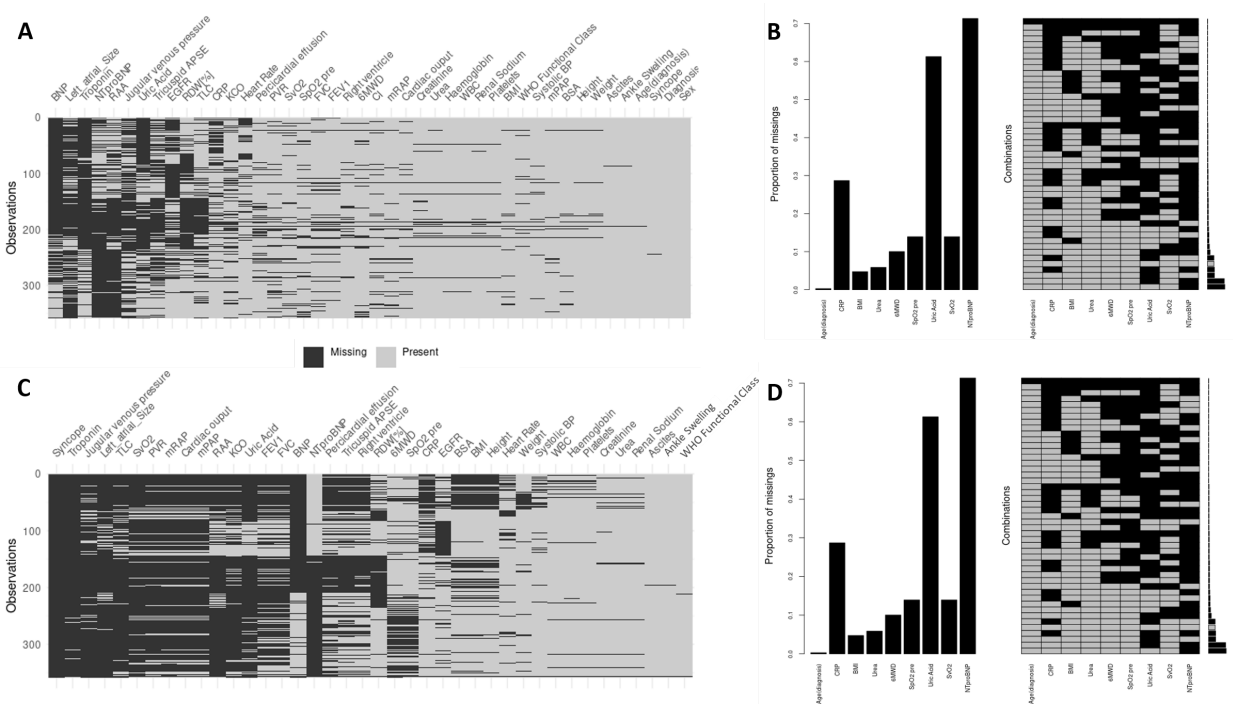
Supplementary Figure 9: *Oncoprint of variants previously associated with PAH across RNA subgroups. Presence of any pathogenic BMPR2 variant is labeled for each patient while the presence of specific SNPs for SOX17 and HLA-DPA1/DPB1 are labeled.*



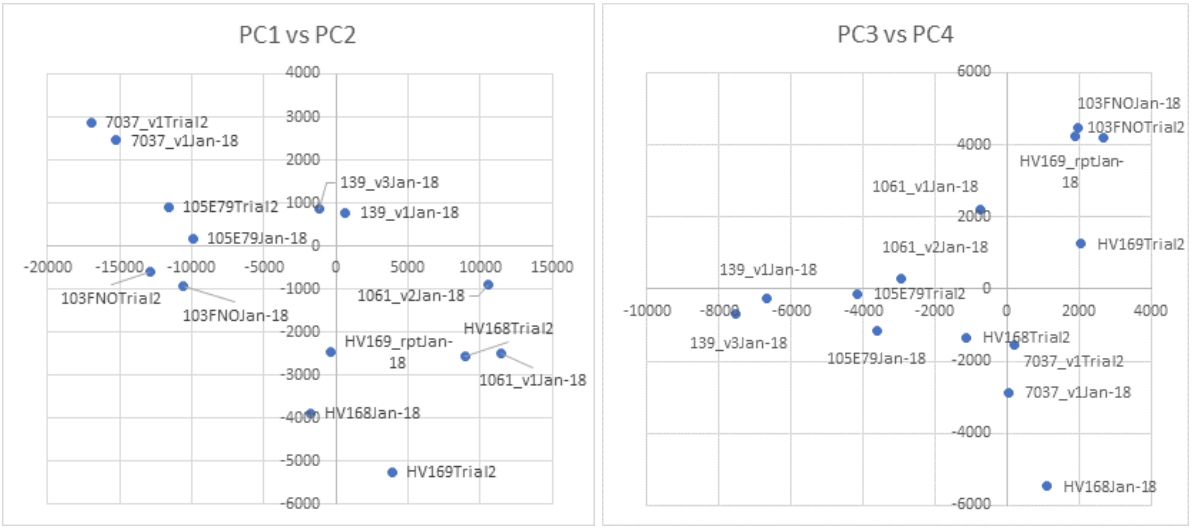
***Supplementary Figure 10**: Comparative measurements of all clinical variables from feature selection across RNA-based subgroups for subgroups I(n=129), II(n=112) and V(n=89). Vertical line represents the median, top and bottom bounds of the box represent the first and third quartile, while the tips of the whiskers represent min and max values.*



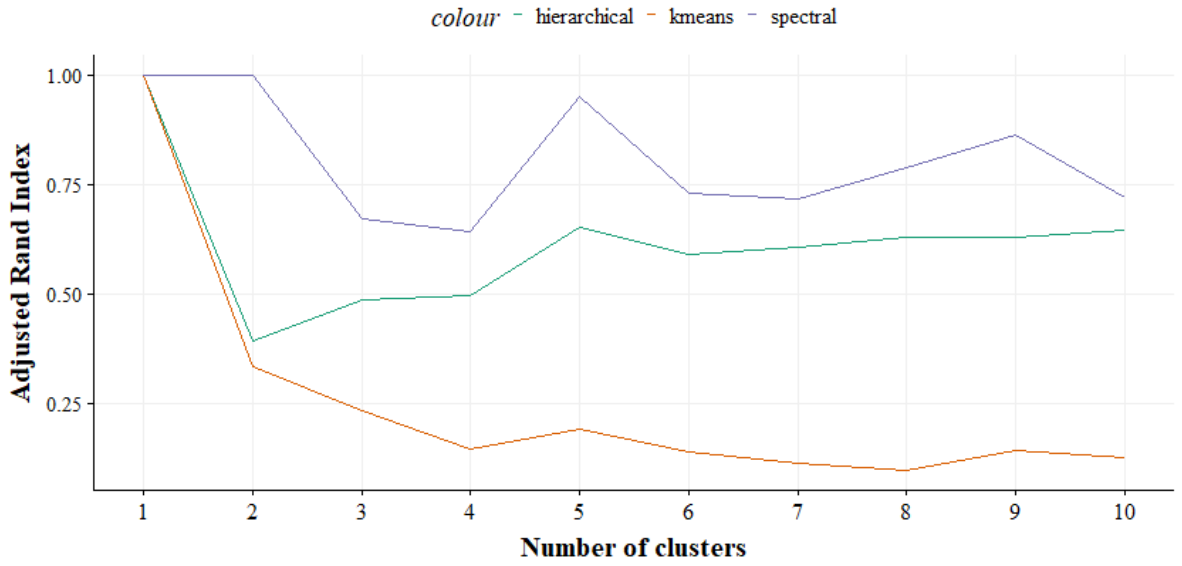
***Supplementary Figure 11:** Revised diagnoses for each subgroup*



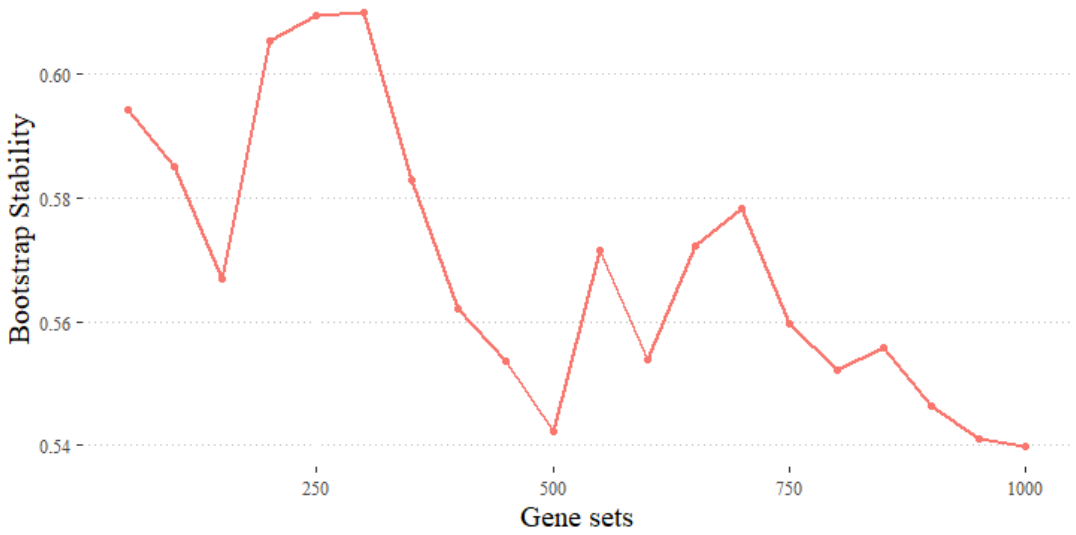
***Supplementary Figure 12**: A) Heatmap showing missingness across important clinical variables for the diagnostic dataset. B) Barchart showing the proportion of missing data and chart showing the combinations of missing data for the classifier variables from the diagnostic dataset. C) Heatmap showing missingness across important clinical variables for the cohort visit 1 dataset. D) Barchart showing the proportion of missing data and chart showing the combinations of missing data for the classifier variables from the cohort visit 1 dataset. [Generated by Emilia M Swietlik and Divya Pandya]*



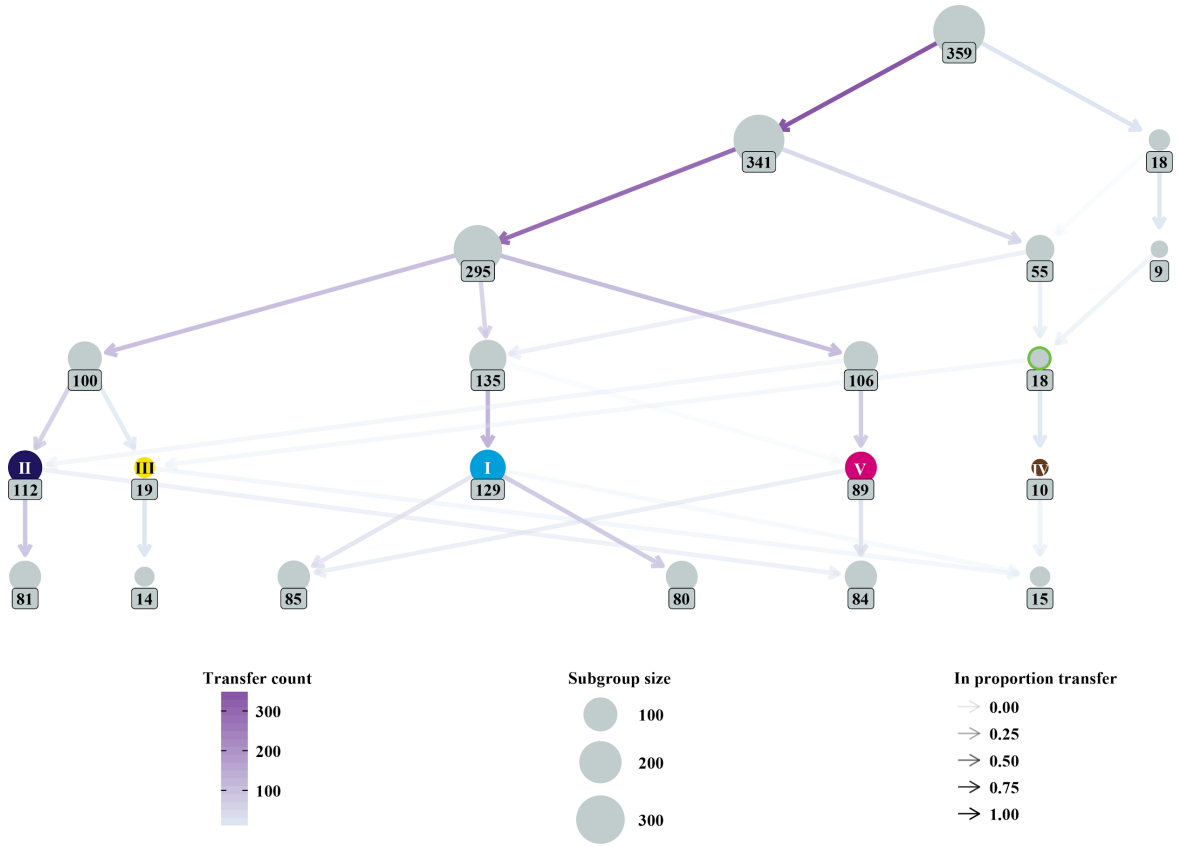
***Supplementary Figure 13:** Principal components analysis of expression profiles from samples with a second replicate that was RNA sequenced (labeled as "...Trial2"). Both replicates are clustered together according to the first four principal components. [Generated by Thomas S Mascarenhas]*



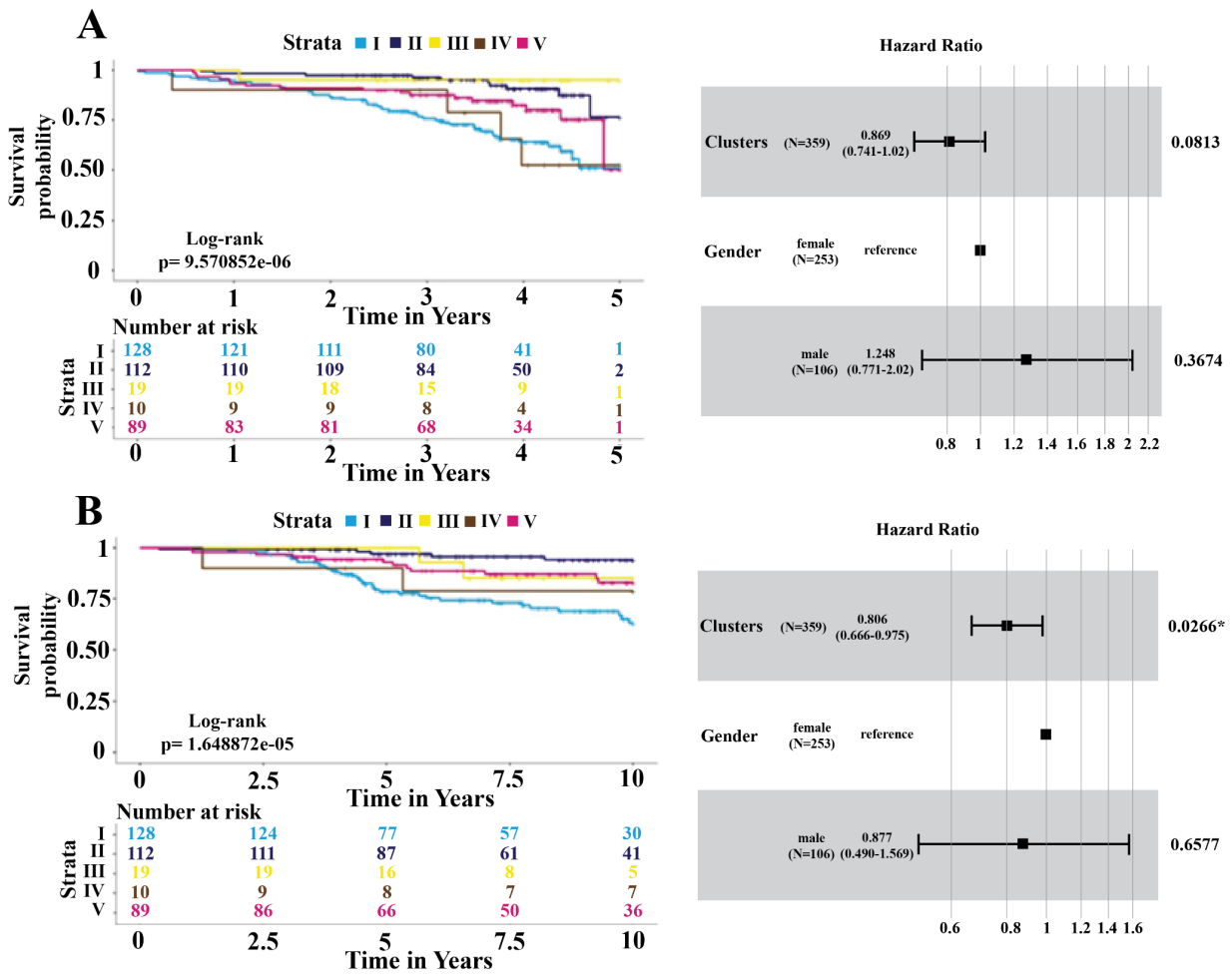
Supplementary Figure 14: *The average adjusted rand index (ARI) of three clustering methods: spectral (blue line), hierarchical (green line) and k-means (red line) clustering. For each method 3 different distance measures/kernels were used and their ARI was averaged for each method.*



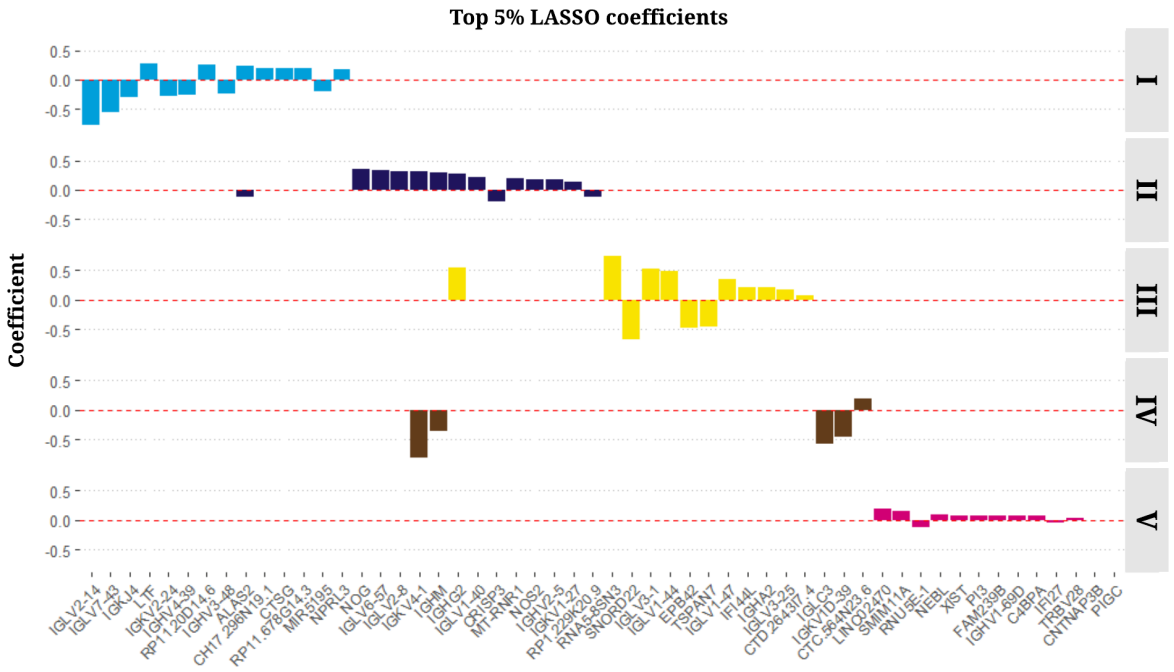
Supplementary Figure 15*: The bootstrap resampling describes stability of clustering as a function of increasing size (by 50) of gene sets. The highest stability is observed for the gene set that contains the top 300 variable genes across patients*



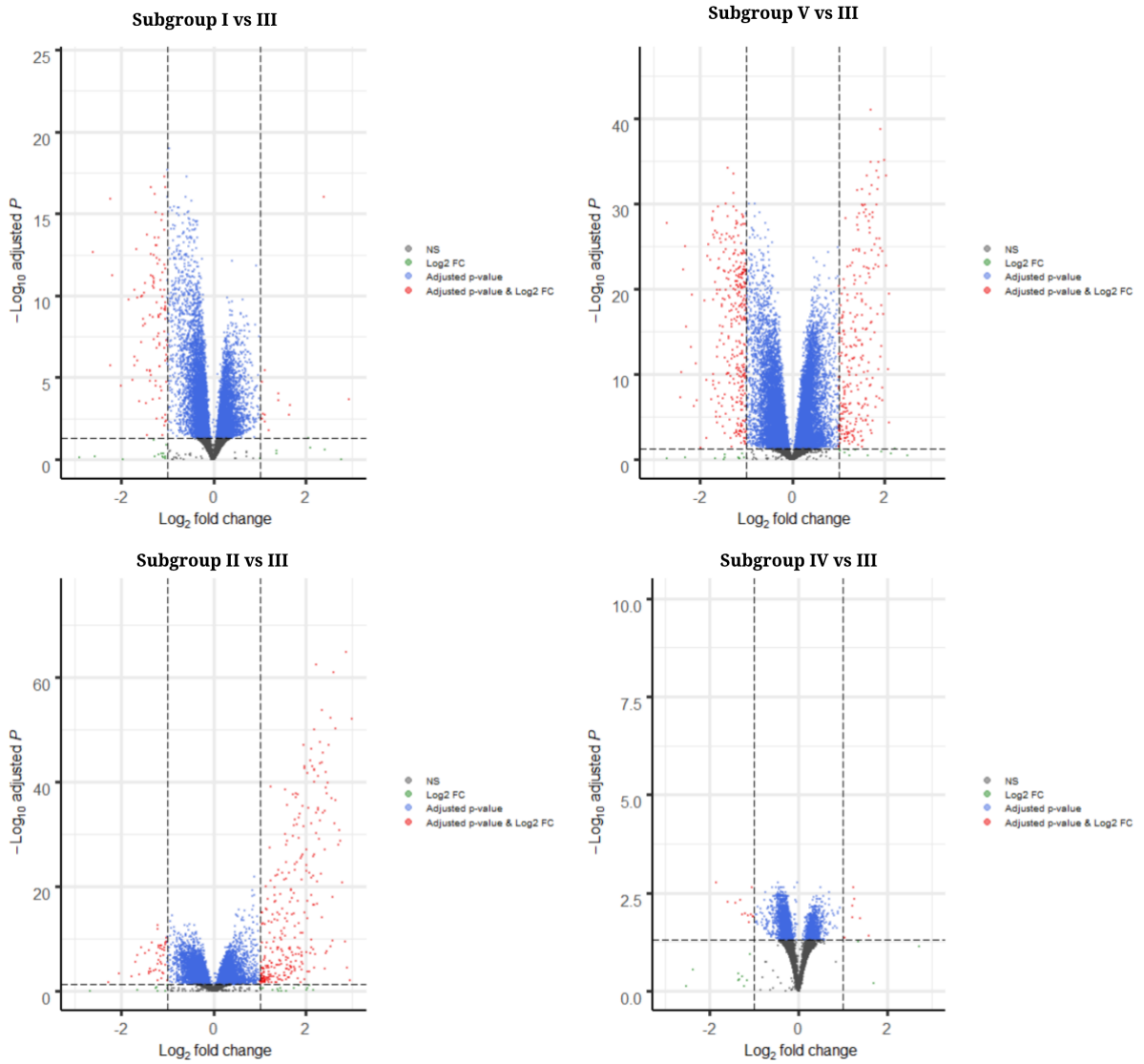
Supplementary Figure 16: *Clustree visualisation of the 5 subgroups (colored) discovered by our spectral clustering methodology. Edges represent the transfer of patients between clusterings of different k. Their opaqueness indicates the amount of patients that transferred.*



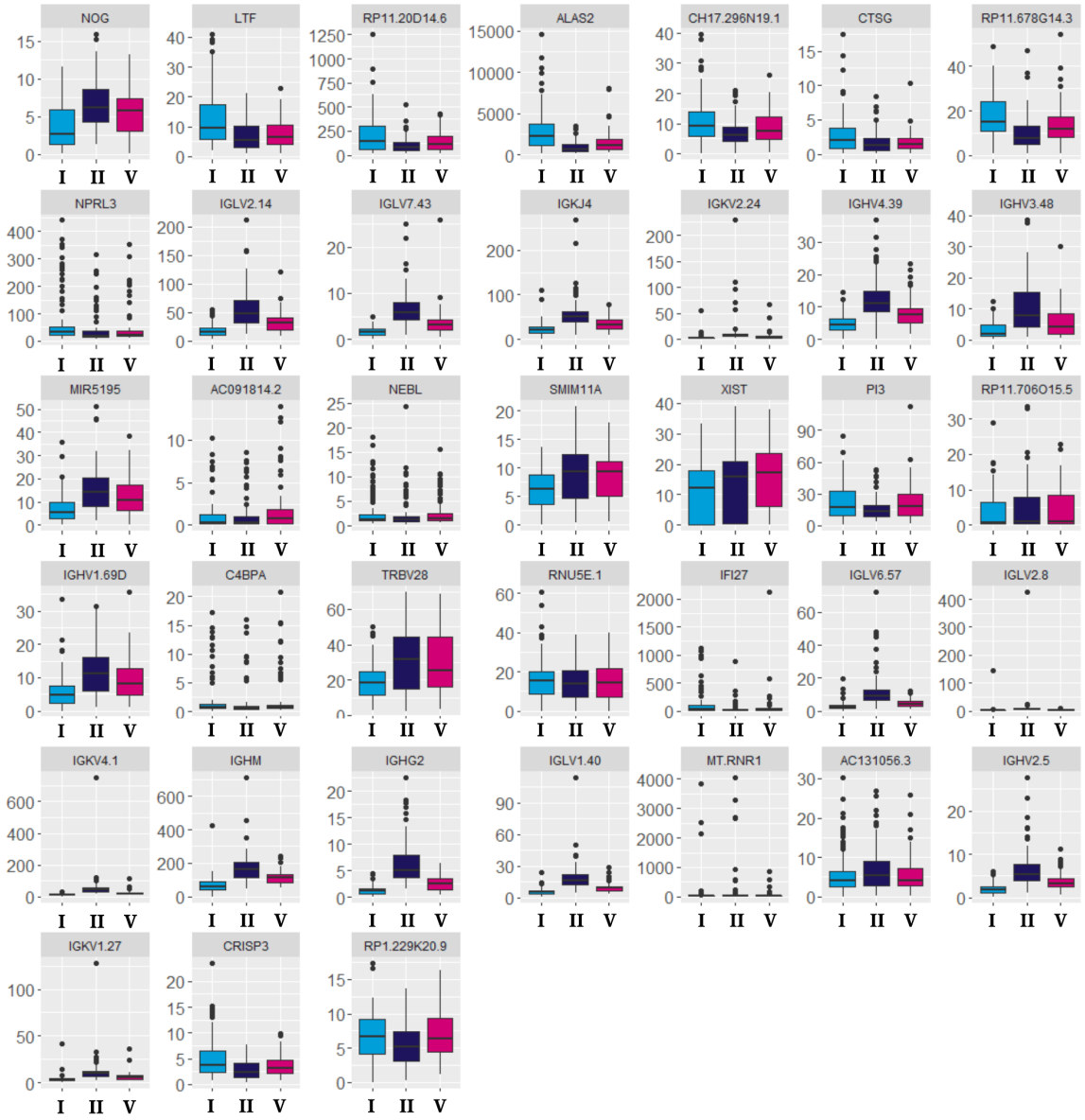
Supplementary Figure 17: *Survival curves and gender hazard ratios from Cox Proportional Hazards model using data (A) from sampling and (B) time of diagnosis. Forest plot data are presented as median values and error bars as 95% confidence intervals.*



***Supplementary Figure 18**: Genes with the top 5% of Lasso coefficients showing gene signatures for each subgroup.*



Supplementary Figure 19: *Volcano plots of differential expression between pairs of subgroups. Dots coloured in red are genes with > 2 fold change and bonferroni log adjusted two-sided t-test p<0.01.*



Supplementary Figure 20: *Gene signature expression (TPM) across subgroups I(n=129), II(n=112) and V(n=89). Vertical line represents the median, top and bottom bounds of the box represent the first and third quartile, while the tips of the whiskers represent min and max values.*

References (Manuscript 2 supplementary)

1. Rhodes, C. J. et al. Whole Blood RNA Profiles Associated with Pulmonary Arterial Hypertension and Clinical Outcome. Am. J. Respir. Crit. Care Med. (2020) doi:10.1164/rccm.202003-0510OC.
2. Polikar, R. Ensemble based systems in decision making. IEEE Circuits and Systems Magazine vol. 6 21–45 (2006).
3. Sweatt, A. J. et al. Discovery of Distinct Immune Phenotypes Using Machine Learning in Pulmonary Arterial Hypertension. Circ. Res. 124, 904–919 (2019).
4. Kherbeck, N. et al. The role of inflammation and autoimmunity in the pathophysiology of pulmonary arterial hypertension. Clin. Rev. Allergy Immunol. 44, 31–38 (2013).
5. Hemnes, A. R. & Humbert, M. Pathobiology of pulmonary arterial hypertension: understanding the roads less travelled. Eur. Respir. Rev. 26, (2017).
6. Calinski, T. & Harabasz, J. A dendrite method for cluster analysis. Communications in Statistics - Theory and Methods vol. 3 1–27 (1974).
7. Dunn†, J. C. Well-Separated Clusters and Optimal Fuzzy Partitions. Journal of Cybernetics vol. 4 95–104 (1974).
8. Pakhira, M. K., Bandyopadhyay, S. & Maulik, U. Validity index for crisp and fuzzy clusters. Pattern Recognition vol. 37 487–501 (2004).
9. Kendall, M. G. A New Measure of Rank Correlation. Biometrika vol. 30 81 (1938).
10. Baker, F. B. & Hubert, L. J. Measuring the Power of Hierarchical Cluster Analysis. Journal of the American Statistical Association vol. 70 31–38 (1975).
11. Hubert, L. J. & Levin, J. R. A general statistical framework for assessing categorical clustering in free recall. Psychological Bulletin vol. 83 1072–1080 (1976).
12. Davies, D. L. & Bouldin, D. W. A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence vol. PAMI-1 224–227 (1979).
13. Halkidi, M., Batistakis, Y. & Vazirgiannis, M. Clustering algorithms and validity measures. Proceedings Thirteenth International Conference on Scientific and Statistical Database Management. SSDBM 2001 doi:10.1109/ssdm.2001.938534.
14. Rohlf, F. J. Methods of Comparing Classifications. Annual Review of Ecology and Systematics vol. 5 101–113 (1974).
15. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics vol. 20 53–65 (1987).
16. Halkidi, M., Batistakis, Y. & Vazirgiannis, M. On Clustering Validation Techniques. J. Intell. Inf. Syst. 17, 107–145 (2001).
17. Song, Y. Class compactness for data clustering. in 2010 IEEE International Conference on Information Reuse & Integration 86–91 (IEEE, 2010)
18. Goel, M. K., Khanna, P. & Kishore, J. Understanding survival analysis: Kaplan-Meier estimate. Int. J. Ayurveda Res. 1, 274–278 (2010).
19. Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P. & Saeys, Y. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics vol. 26 392–398 (2010).
20. Duan, K.-B., Rajapakse, J. C., Wang, H. & Azuaje, F. Multiple SVM-RFE for Gene Selection in Cancer Classification With Expression Data. IEEE Transactions on Nanobioscience vol. 4 228–234 (2005).

21. Granitto, P. M., Furlanello, C., Biasioli, F. & Gasperi, F. Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. Chemometrics Intellig. Lab. Syst. 83, 83–90 (2006).
22. Williams, J. A., Weakley, A., Cook, D. J. & Schmitter-Edgecombe, M. Machine learning techniques for diagnostic differentiation of mild cognitive impairment and dementia. in Workshops at the twenty-seventh AAAI conference on artificial intelligence (2013).
23. Choi, E., Schuetz, A., Stewart, W. F. & Sun, J. Using recurrent neural network models for early detection of heart failure onset. J. Am. Med. Inform. Assoc. 24, 361–370 (2017)
24. Breiman, L. Machine Learning. vol. 45 5–32 (2001).
25. Hosmer, D. W., Jr., Lemeshow, S. & Cook, E. D. Applied Logistic Regression, Second Edition: Book and Solutions Manual Set. (Wiley-Interscience, 2001).
26. Peterson, L. K-nearest neighbor. Scholarpedia J. 4, 1883 (2009).

Chapter 4 - Application of Omada to miRNA profiles

4.1 Background

The expression of genes can reveal a great amount of information about the genetics of a disease. However, as described in Chapter 1 many data types may contribute to a more complete picture of a complex mechanism such as a rare disease. In PH specifically, recent advances have taken advantage of multiple omics to shed light in transcriptional signatures in various cells (Harbaum *et al.*, 2021). In (Sweatt *et al.*, 2019) unsupervised learning was used with proteome data to identify distinct blood cytokine phenotypic patient groups while our transcriptomic work (Kariotis *et al.*, 2021) showed distinct prognosis H/IPAH patient groups. Also, authors of (Rhodes *et al.*, 2017) used metabolomic data to distinguish metabolite profiles associated with survival. The above research highlights the importance of considering multiple omics to discover patient dissimilarities that can lead to risk assessment and potential novel biomarkers associated with clinical outcomes.

In this chapter I describe my unsupervised work on microRNA data included in Manuscript 3 and additional exploratory microRNA clustering analyses. The purpose of this chapter is to assess the utility of clustering on circulating microRNA data in identifying subgroups of patients with associated molecular signatures. A variety of clustering runs were compared and tested based on the output clusters. One of these analyses was used in manuscript 3 to explain heterogeneity within PH. Other runs offered valuable insights in the clustering behaviour of microRNA data and PH classes.

I performed clustering analysis in several subsets of a microRNA dataset described in the following sections. One clustering analysis is part of manuscript 3 and the remaining were exploratory analyses targeting different groups of microRNA activity such as PH categories.

4.2 Contribution

My work in this publication focused on generating microRNA based clusters as part of a methodology to identify molecular signatures associated with different clinical PH classes. I was not the first author of this study and my contribution revolved around creating microRNA patient profiles, so I did not include the full manuscript here, but only the sections I contributed to. In the manuscript sections following the sections/figures not generated by me are in italics and brackets in text.

4.3 Manuscript 3

Introduction

Pulmonary hypertension is a medical condition associated with elevated resting mean pulmonary artery pressure (PAP) and subsequently reduced life expectancy but also a challenging diagnosis with non-exclusive symptoms. PH patients are categorised to one of the five recognised classes, mainly based on their clinical features, which affects treatment decisions that potentially disregard the phenotype complexity of patients. Several approaches, such as circulating biomarkers and NT-proBNP, have recently shown to offer benefits towards informing patient management but were only effective to specific groups. Searching for a different approach, studies have shown miRNAs to be dysregulated in PH contexts (Rhodes *et al.*, 2013; Rothman *et al.*, 2016; Errington *et al.*, 2021) and their potential to be an alternative or additive test towards diagnosis or risk stratification. In this work we consider the diverse cellular origin of miRNAs and hypothesise that circulating miRNAs can be distributed differently across PH types and would inform about molecular PH patient endotypes. Also taking into account the severity spectrum of PH we hypothesised that an miRNA-based unsupervised learning approach would partition patients in groups with similar molecular and clinical phenotypes.

Methods

Study population and clinical data

The study cohort was comprised of 1150 patients with PH and 334 disease controls as summarised in **Table 1**. Patients were recruited from 3 UK national PH referral centres, located at Hammersmith Hospital (Imperial), Royal Hallamshire Hospital (Sheffield) and Royal Papworth Hospital (Cambridge), as summarised in **Supplementary Tables 1** and **2**. All cases were diagnosed between 2008 and 2019 using contemporaneous diagnostic guidelines. All samples were obtained following informed consent to one of three cohorts: the Imperial College Prospective Study of Patients with Pulmonary Vascular Disease cohort (PPVD, UK REC Ref 17/LO/0563), the Sheffield Teaching Hospitals observational study of pulmonary hypertension, cardiovascular and other respiratory diseases (STH-ObS, UK REC Ref 18/YH/0441) or the Royal Papworth cohort (Cambridgeshire East Research Ethics Committee reference 08/H0304/56). All samples were collected as per local standard operating

procedures and stored at -80oC until assayed. All cases/samples were pre-processed into training, interim and validation datasets to balance age, sex, PH classification and recruitment site. The validation samples were analysed separately. [Sample collection *performed by contributed centres*]

Quantification of NT-proBNP and miRNAs

"Total RNA was extracted from 200 µl of serum or plasma using the Maxwell® RSC miRNA Plasma and Serum Kit (Promega, Madison, USA) as per the manufacturer's recommendations with the following modifications: (a) a set of three proprietary spike-in controls (MiRXES, Singapore) was added, representing high, medium, and low levels of RNA, into the lysis buffer C prior to sample RNA isolation. The spike-in controls are 20-nucleotide RNAs with unique sequences (distinct from any of the 2588 annotated mature human miRNAs in miRBase version 21.0, RRID:SCR_003152) and are used to monitor RNA isolation efficiency and normalise for technical variations during RNA isolation; (b) bacteriophage MS2 RNA (Roche, Basel, Switzerland) was added at 0.4ng per sample isolation to improve RNA isolation yield. For biomarker discovery, a highly controlled RT-qPCR workflow was used to quantify the expression of miRNA in each sample. Isolated RNA was reverse transcribed using miRNA-specific reverse transcription (RT) primers according to manufacturer's instructions (ID3EAL Customized Individual miRNA RT Primer, MiRXES) on QuantStudio™ 5 Real-Time PCR System (Applied Biosystems, Foster City, CA, USA). Additional information about the protocol can be found in the Supplementary Methods. NT-proBNP was assayed using coba Elecsys per manufacturer's instructions." [*Performed by miRXES*]

Pre-processing of miRNA expression data

"Data on 326 miRNAs detected in no less than 90% of samples were analyzed further. Missing values were imputed separately in the combined discovery and interim sets, and subsequently the validation set, by replacing missing values with miRNA [mean – 4 standard deviations]. miRNA data were further global normalized. Samples from Cambridge showed higher total miRNA counts than the other centres. To further correct for this batch effect, total miRNA counts were modelled with a LASSO model composed of 11 miRNAs chosen from summed miRNA counts of detected miRNA by increasing lambda arbitrarily to reduce selected miRNAs to a reasonable number with high performance (rho 0.9687). A linear regression using this model was then used to adjust the counts, retaining the mean miRNA levels." [*Written by Niamh Errington and Chris Rhodes*]

Unsupervised Patient Clustering

The clustering methodology required an extra preprocessing step to retain only the relevant information. Additional clinical information was used to filter out samples not in the discovery/interim phase as well as non PH1, PH2 or PH3, generating a finalised dataset of 615 samples and 326 miRNAs. The finalised dataset was used as an input to the heterogeneity clustering methodology described previously (20) with optimization using multiple subsets of the miRNAs, each time increasing the size by 50. Each subset was then used to run multiple spectral clusterings (for k = 2,3,4,5,6) whose stability was measured using a bootstrap approach (package fpc v2.2-3). The mean cluster stability over clusters and ks was calculated for every subset and plotted to discover the size of the most variable miRNA subset that provided the highest stability. The optimal number of patient clusters (range of [2, 10]) was then estimated based on the ensemble voting of internal machine learning indexes and the final clustering run partitioned samples to distinct subgroups based on their miRNA profile. The intra-agreement was calculated using the average of the adjusted Rand index (package fossil v0.3.7) of three pairs of clustering runs per algorithm type (**Supplementary Table 7**).

Supplementary Table 7: *Three examined clustering algorithm types with the parameter pairs used to calculate the intra-agreement of each algorithm type.*

Spectral	**Hierarchical**	**K-means**
rbfdot/polydot	euclidean/manhattan	rbfdot/tanhdot
vanilladot/tanhdot	canberra/minkowski	vanilladot/tanhdot
laplacedot/besseldot	canberra/maximum	rbfdot/vanilladot

Re-classifying patients assigned to PAH, PH-LHD and PH-lung using unsupervised learning from miRNAs

Differentiating patients with PAH, PH-LHD and PH-lung with confidence using routinely collected clinical measurements can be particularly challenging, leading to patients being misclassified. The prospect of shared pathology (particularly the occurrence of pre- and post- capillary PH) might further limit the utility of the current clinical classification for identifying miRNA signatures that represent PH endophenotypes and so limit the full potential of miRNAs to inform on the underlying

molecular drivers. We therefore examined how the distribution of miRNAs, unfettered by the clinical classification, might inform the clinical presentation of patients from a mixed cohort of PAH, PH-LHD and PH-lung (Group 5 PH was excluded because of the small number of patients). The clustering pipeline, utilising the expression of 50 miRNAs, identified 6 stable clusters (Figure 5) that are agnostic to current PH clinical classification (**Supplementary Table 8**). *[Written by Niamh Errington]*

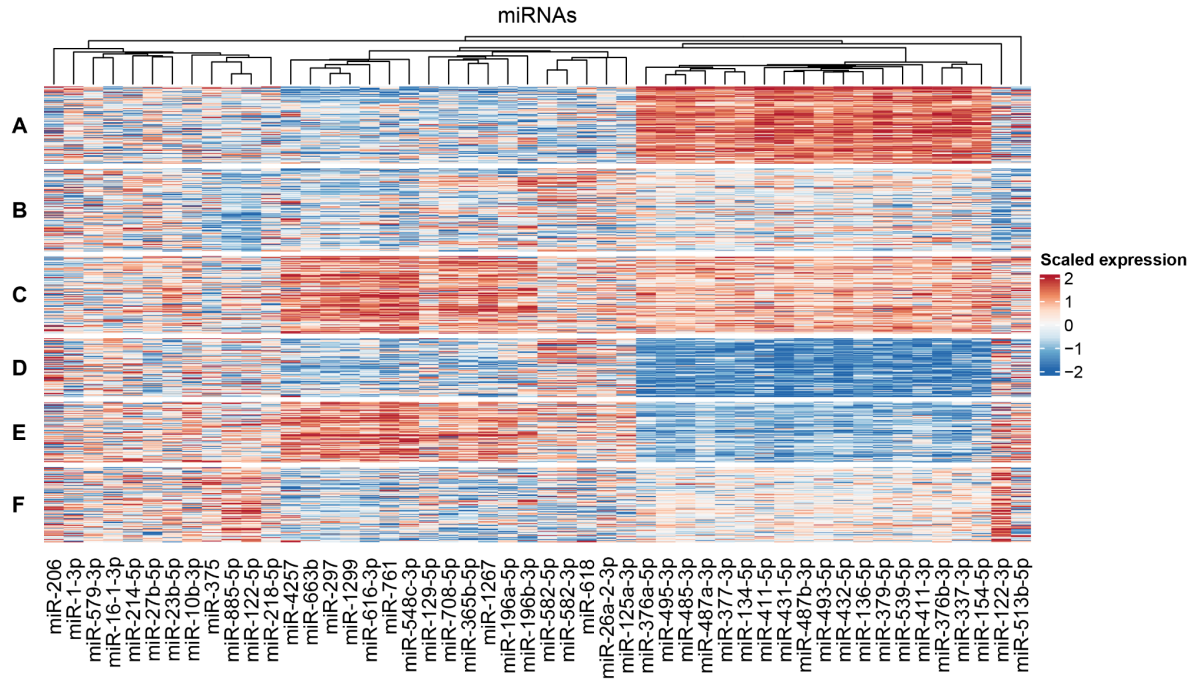


Figure 5: *Unsupervised clustering of patients across PH groups 1, 2 and 3 shows distinct miRNA profiles*

Supplementary Table 8*: Prevalence of PH clinical classifications within clusters*

	Cluster A	Cluster B	Cluster C	Cluster D	Cluster E	Cluster F
PH 1.1	43	49	34	35	34	34
PH 1.2	3	1	4	2	1	1
PH 1.3	1	2	0	1	0	0
PH 1.4	17	24	23	21	27	37
PH 2	36	24	29	12	13	23
PH 3	10	18	20	13	11	11

Results

Six patient clusters discovered by miRNAs have distinct clinical characteristics

Inspection of clinical phenotypes of clusters showed significant differences in clinical phenotypes between the clusters. Interestingly there was no significant difference in the distribution of classification groups (1-3) across the 6 clusters (**Figure 6A**), nor in age, sex, BMI, WHO Functional class or REVEAL risk score. There was, however, a significant difference in 3 or 5-year and all-time survival (**Figure 6B**) with cluster A having the best survival and clusters C and F the worst (Hazard ratios 1, 1.77 and 1.71 respectively). There were also significant differences between the clusters in a number of important clinical variables (**Figure 6C-H**). Most notably cluster A was associated with low mPAP, low PVR but not the lowest NT-proBNP. Patients in clusters C and F were both associated with significantly worse survival than cluster A but patients in cluster F had significantly lower NT-proBNP than cluster C. These characteristics highlight the challenges of using single biomarkers (e.g. NT-proBNP) or haemodynamics (PVR) to attribute risk and define molecular mechanism. *[Written by Niamh Errington]*

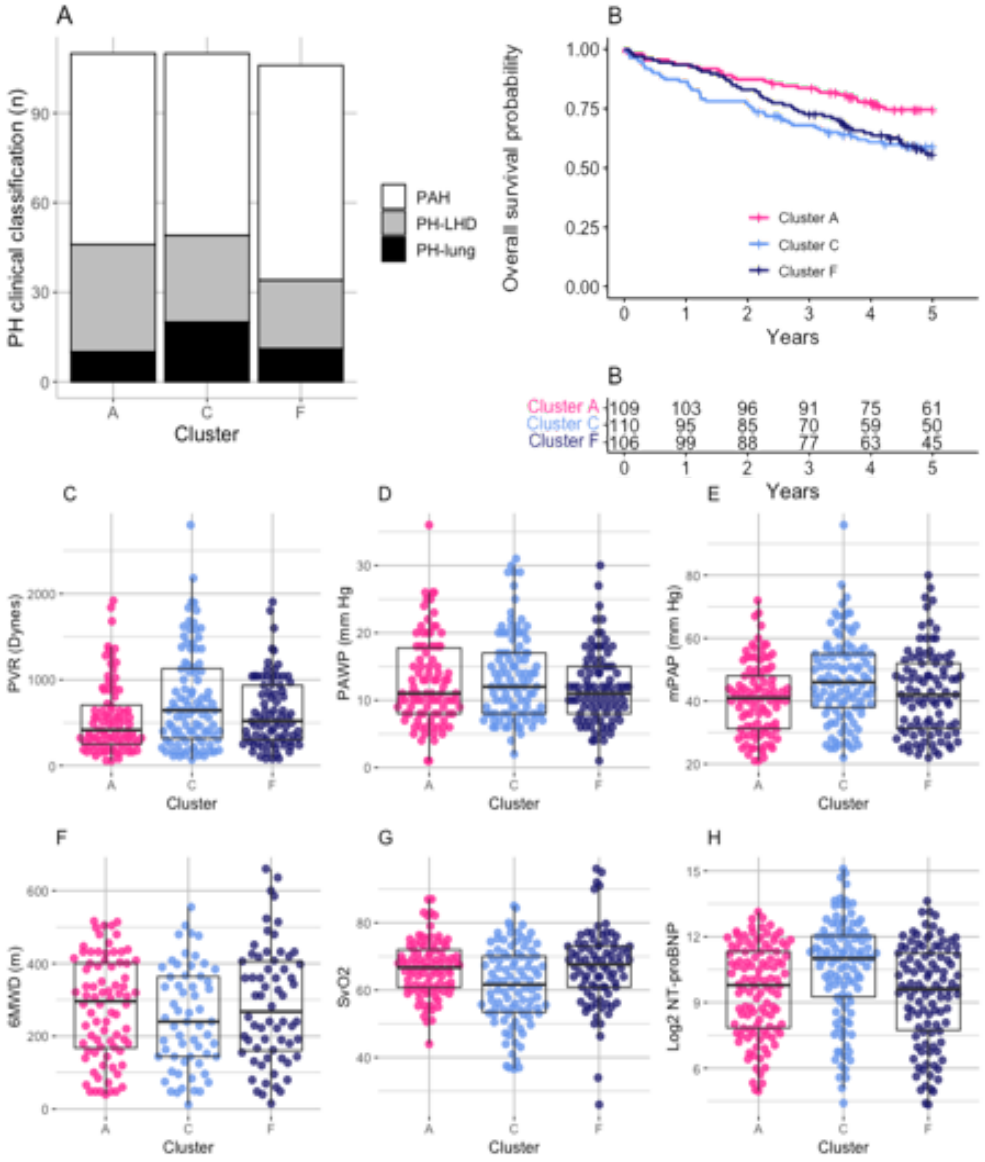


Figure 6: *Clinical Characteristics of the six miRNA clusters. A) Number of patients from each PH Clinical classification group 1-3 within each of the 6 miRNA clusters; B) 5-year survival of patients within each miRNA cluster. Box and Whisker plots show the relative values for C) Serum NT-proBNP levels, D) mean PAP, E) PAWP, F) PVR, G) six-minute walk distance, and H) SvO2 within each miRNA Cluster. * p<0.05, ** p<0.01 on specific comparisons. For H) $ p<0.05 compared to cluster C, and # p<0.05 compared to cluster E following Kruskal-Wallis chi-squared followed by Benjamini-Hochberg post-hoc analysis. [Generated by Niamh Errington]*

4.4 Additional microRNA analysis, signs of heterogeneity in PH

Many PH patients live with undiagnosed or untreated PH since they don't fall under any current clinical PH classification and especially for PH1 where they are treated with vasodilator therapies in absence of PH1 targeted therapies. As an effort to provide relevant insights our analysis of circulating serum miRNAs of over 600 patients of PH1, PH2 and PH3 in manuscript 3 led to 6 distinct clusters whose molecular profiles were extending across clinical classifications i.e. miRNA profiles of PH1 patients resemble those of PH2 or PH3 patients. These clusters were associated with unique miRNA and clinical signatures and demonstrated different survival profiles irrespective of their clinical PH classification. Uniquely enriched pathways were also associated with the clusters along with important PH clinical variables such as PVR, mPAP and NT-proBNP. The above highlighted the potential to identify patients with similar disease drivers that may be benefitted by specific PAH treatments. In light of this information we decided to explore additional subsets of PH patients to assess their utility in identifying insightful clusters. We used a dataset of 1138 patients each belonging to one of the clinical PH categories and a total of 554 miRNAs (generated as described in *Data Introduction* section above). We created subsets of patients to focus on specific PH categories as shown in **Table 1**.

Table 1: *Multiple miRNA analyses with respective PH groups, number of patients as well as the optimal clustering method, number of miRNAs and k. In red are marked the k which received the most votes per analysis.*

Analysis number	IPAH Groups	patients	Clustering method	Optimal miRNAs	K votes
1	PH1	449	spectral	350	**k.2** = 9 k.6 = 5
2	PH1.4	164	spectral	100	**k.2** = 8 k.3 = 2 k.6 = 4
3	PH1, PH2, PH3 (manuscript 3)	615	spectral	50	k.2 = 4 k.4 = 2 k.5 = 4 **k.6** = 5
4	PH4	258	spectral	50	**k.2** = 4 **k.3** = 4 k.4 = 1 **k.5** = 4 k.6 = 1
5	PH4 + PH0_CTED	322	spectral	50	**k.2** = 4 k.3 = 3 k.4 = 3 **k.6** = 4
6	PH1, PH2, PH3, PH4, PH5	900	spectral	50	**k.2** = 8 k.6 = 6
7	All patients	1138	spectral	50	**k.2** = 5 k.3 = 3 k.5 = 4 k.6 = 2

PH1: *Pulmonary Arterial Hypertension*
PH1.4: *systemic diseases such as connective tissue diseases, HIV infection and congenital heart disease*
PH2: *PH due to left heart disease*
PH3: *PH due to lung diseases and/or hypoxia*
PH4: *Chronic thromboembolic pulmonary hypertension (CTEPH)*
PH5: *Pulmonary hypertension with unclear multifactorial mechanisms*

Initially, we investigated PH1 (PAH), the patient class with the most heterogeneity (as shown in (Kariotis *et al.*, 2021). MiRNA analysis 1 focused on the entirety of PH1 and included 449 patients and 554 miRNAs while analysis 2 looked at the more narrow set of patients associated with systemic diseases such as connective tissue diseases, HIV infection and congenital heart disease (PH1.4). Omada's feasibility step showed usable datasets for both analyses 1 and 2 with average stabilities of 0.83 and 0.88, respectively. For all downstream analysis I used a k range of [2, 6]. When tested for the optimal algorithm, spectral clustering achieved a satisfying average of 0.72 (analysis 1) and 0.65 (analysis 2) partition agreement (over all tested k) well above k-means (0.45, 0.46) and hierarchical clustering (0.08, 0.13). Given that

spectral clustering performed best in multiple subsets of the same miRNA dataset (manuscript 3 and analyses 1,2) I expected the remaining analysis to also fit this selection. Looking at the subset of miRNAs that provide stable clusterings, analysis 1 showed that more than half of the most variable miRNAs (350) was the optimal number, an observation that might indicate a weak signal in this specific portion of the data. A weak signal would allow more miRNAs to be included as no miRNAs could provide sufficient value (in our case cluster stability). On the other hand, analysis 2 estimated the top 100 variable miRNAs to provide the most stability, a much smaller number compared to analysis 1 which might stem from the smaller sample size (449 - 164). It is interesting to note that all 100 optimal miRNAs of analysis 2 were also included in the optimal 350 of analysis 1, as seen in **Figure 7**. As a reminder, the most variable miRNAs are recalculated every time across each dataset's samples meaning that different datasets can have very different prioritised miRNAs. Based on this observation and since no miRNAs were unique in analysis 2 we could note that this methodology did not capture any signal related to PH1.4.

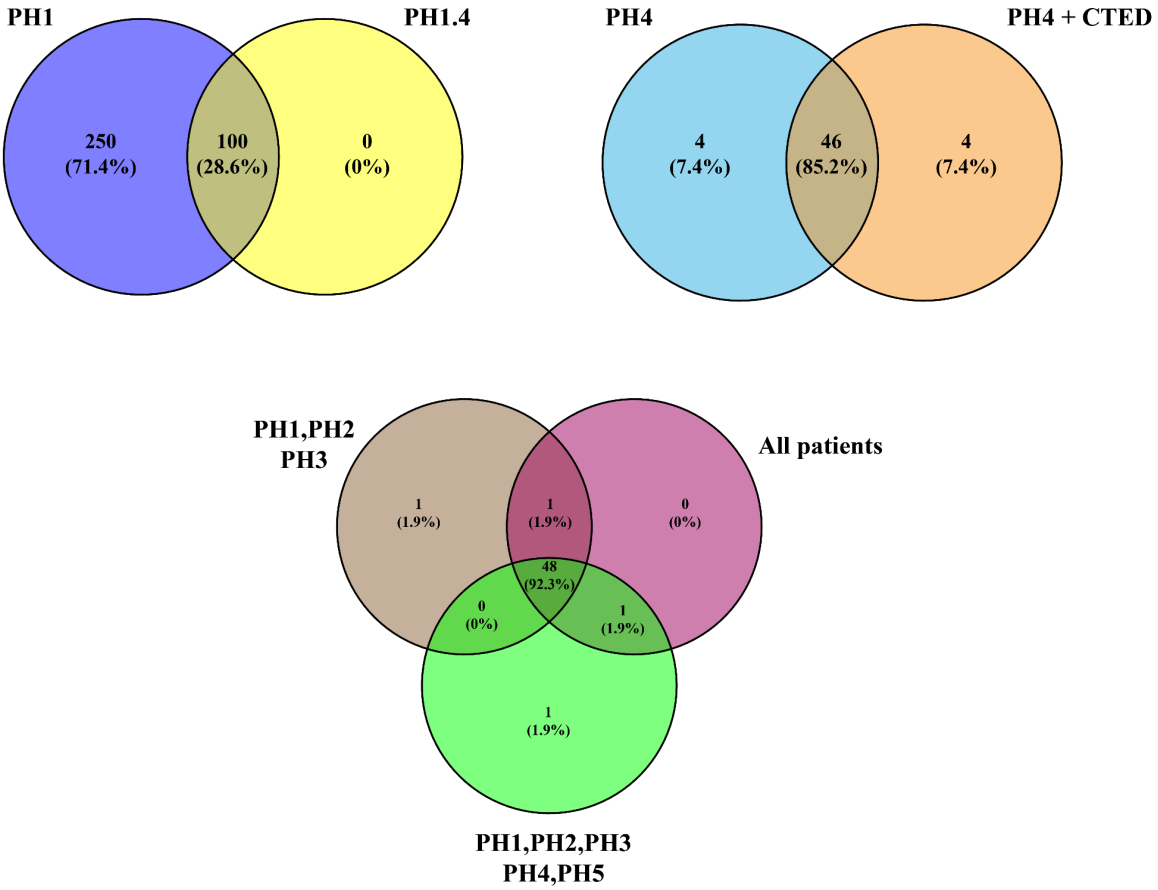


***Figure 7**: Venn diagram of the overlap between the optimal miRNAs estimated for the tested miRNA subset based on their PH categories. In top right we see the overlap of estimated miRNAs when using patients from categories PH1 and PH1.4. Top right venn diagram shows the overlap of estimated miRNAs when using patients from categories PH4 and PH4 + CTED. Bottom venn diagram shows the overlap of miRNAs when comparing the selected miRNAs of (PH1,PH2,PH3) with (PH1,PH2,PH3,PH4,PH5) and all patients used.*

Next, I investigated the number of potential subgroups in the two datasets. Both datasets clearly favoured the existence of two clusters according to our ensemble methodology with over 50% of metrics agreeing (**Table 1**). Then I performed a tsne and PCA analysis to visually inspect whether the two-group partitioning was apparent. In all four plots shown in **Figure 8** two clusters are easily discernible however additional t-sne and PCA analysis showed no groupings according to the official PH1 subcategories (**Figure 9**).

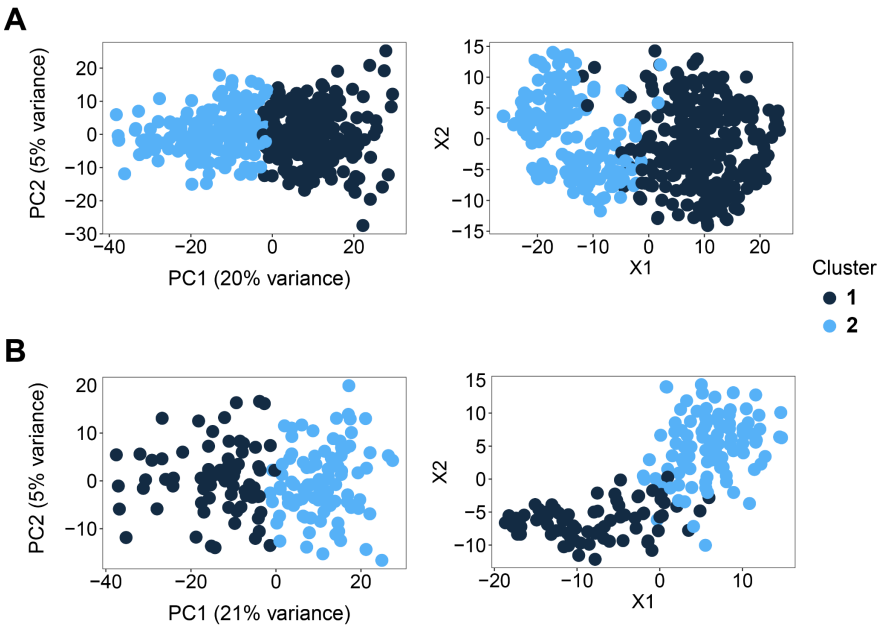


***Figure 8**: PCA (left) and tsne analysis performed on A) analysis 1 and B) analysis 2 showing a distinct two-cluster grouping.*

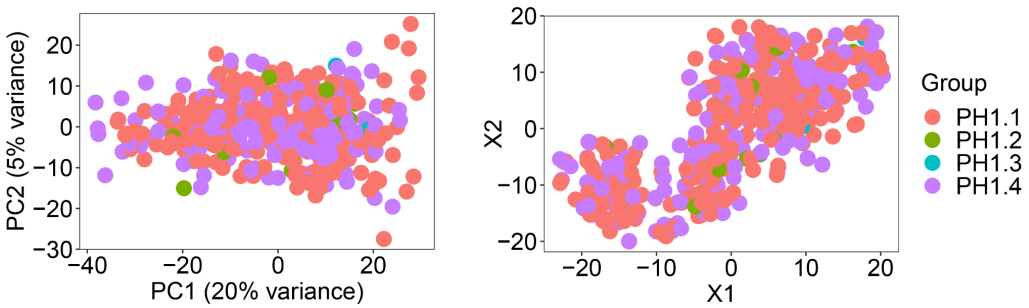


***Figure 9**: PCA (left) and tsne analysis performed on the memberships generated as part of analysis 1. The four colours indicate the PH1 subcategories of patients.*

The latter confirmed that the two clusters were not formed driven by traditional PH diagnosis but rather a different source. To gain some insights on potential discriminating variables I conducted a statistical analysis (Wilcoxon test) on the available numeric clinical variables of the cohort. As demonstrated in **Figures 10**,**11** variables such as BMI and PVR were significantly different between the two clusters (in both analyses 1 and 2) and might offer promising grounds for further analysis.

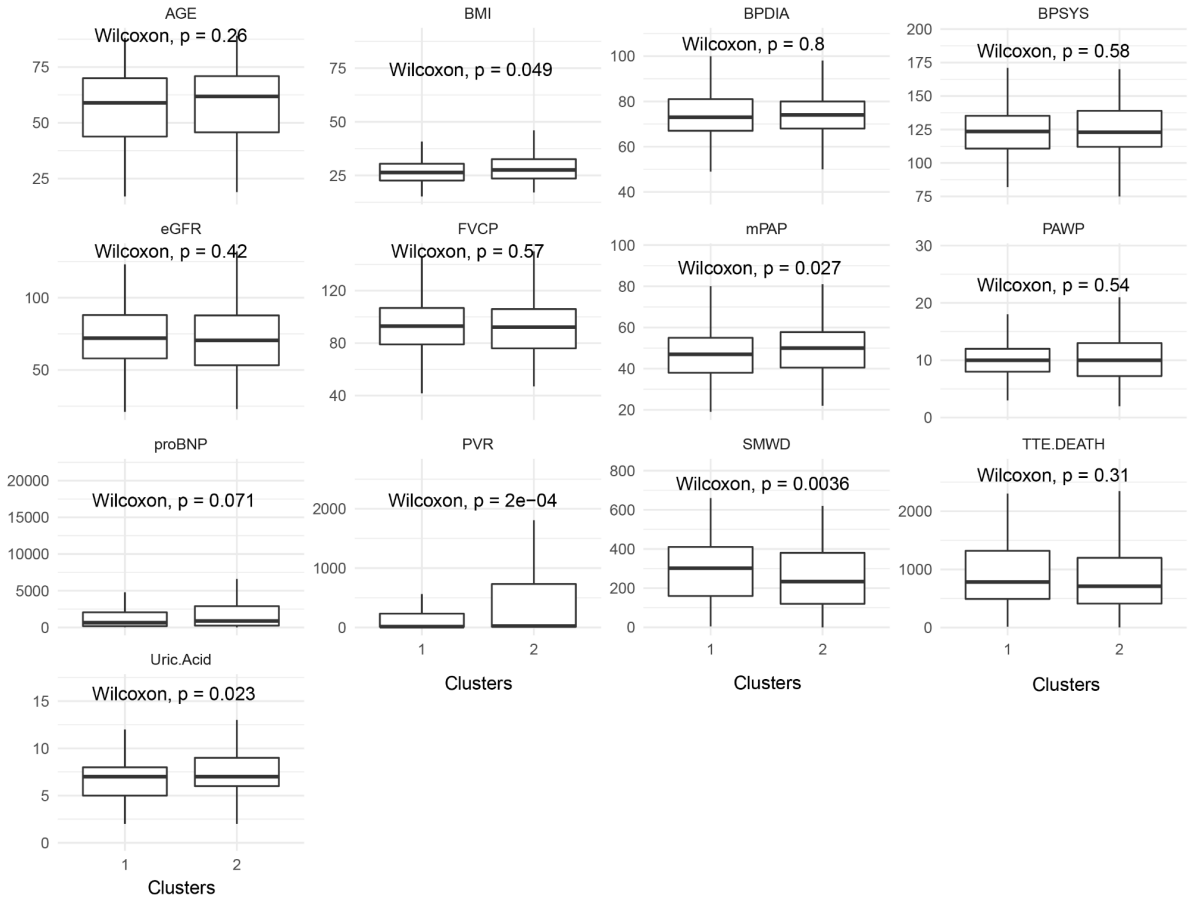


Figure 10: *Boxplots of the clinical variables for clusters 1 and 2 of miRNAs analysis 1.*

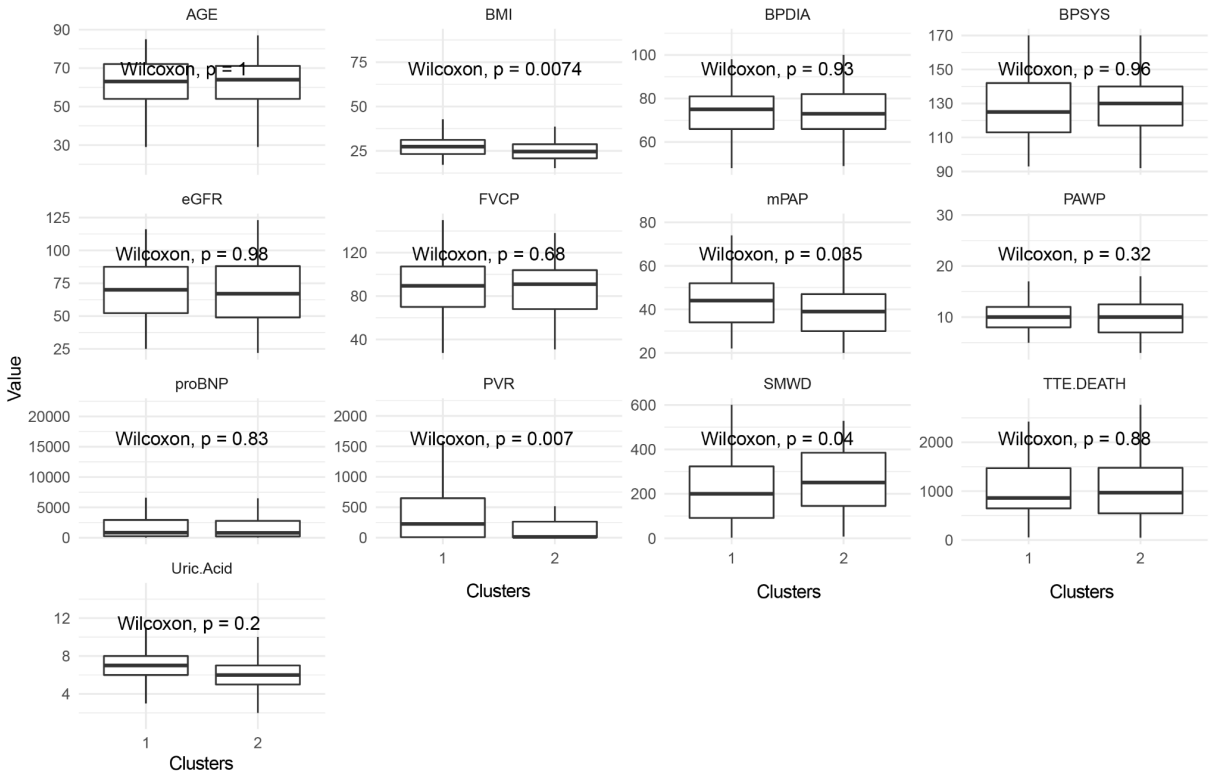


***Figure 11**: Boxplots of the clinical variables for clusters 1 and 2 of miRNAs analysis 2.*

Next I explored PH4 which entails CTEPH patients with chronic thromboembolic pulmonary hypertension. I also included patients with CTED (chronic thromboembolic pulmonary vascular disease to assess whether their miRNA profiles differ as they presumably are variants of the same pathophysiological mechanism(Lang *et al.*, 2021). Analysis 4 focused solely on 258 CTEPH patients and analysis 5 on a total of 322 CTEPH and CTED patients. Both analyses utilised 326 miRNAs as in manuscript 3. The Omada feasibility analysis deemed the datasets of both analyses to have satisfying average stabilities of 0.89 and 0.87, respectively, and spectral clustering was selected as the preferable algorithm. During the miRNA subsets' analysis, Omada estimated 50 miRNAs as the optimal number of features. **Figure 7** shows 85.2% of them were shared between analyses 4 and 5 which makes sense since dataset 4 comprises 80% of dataset 5. Interestingly, there was a spread in selecting the number of clusters in both cases. Analysis 4 (CTEPH) was split between 2,3 and 5 clusters each getting 26% of the ensemble votes (**Table 1**). To investigate how these clusters and their profiles look, I generated t-SNE, PCA plots as well as the miRNA heatmap which shows the expression profiles per k (**Figure 12**). For every k, according to the heatmaps there seem to be small sets of miRNAs that differentiate between the clusters (usually places on the left side of the heatmap due to the internal column/miRNA clustering) agreeing with the previous step where 50 miRNAs showed to provide the most stable results. PCA and t-SNE also seem to support some cluster distinction for each k but with low variance explained (16% on the first

component). Further clinical statistics were generated for those clusters by Niamh Errington showing some clinical variable differences such as PVR, survival and NTproBNP mostly for the two-cluster run. However, a cox hazard ratio model showed that these differences were associated with patient age and sex rather than cluster membership (global p-value of 3.596e-07).

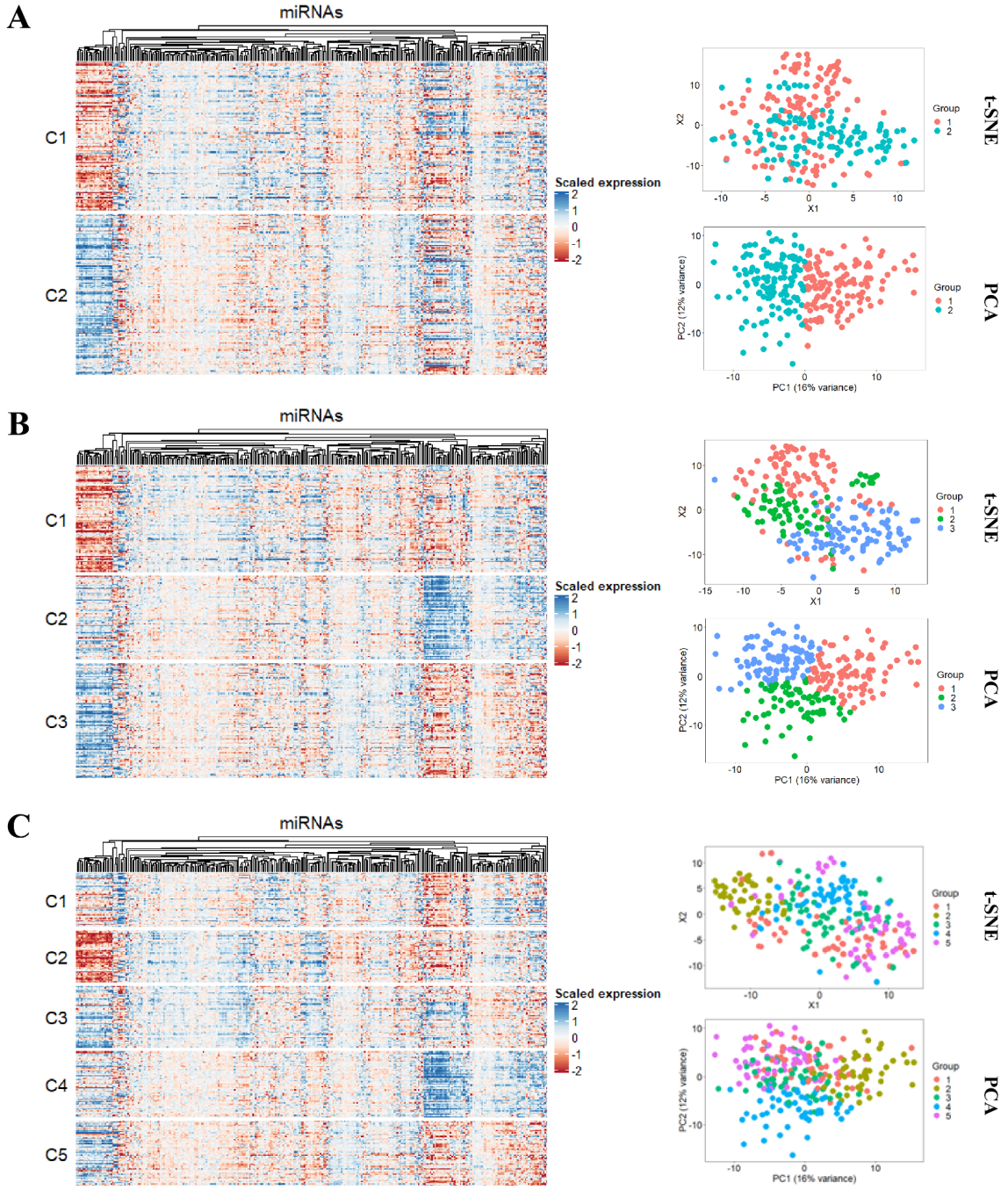


***Figure 12**: miRNA profiles, t-SNE and PCA analysis for the three possible k clusters for miRNA analysis 4 (CTEPH patients). A) refers to two clusters B) refers to three clusters and C) refers to five clusters*

Analysis 5 showed a slightly different picture since the ensemble votes were split between several k values (**Table 1**). t-SNE and PCA analyses showed some distinction between the clusters for the most voted ks of 2 and 6. Interestingly, when considering six clusters different sets of miRNAs seemed to differentiate specific clusters. In **Figure 13B**, clusters C2 and C3 seem to have large expression differences for a small number of miRNAs while clusters C5 and C6 differ in a different set of miRNAs. Statistical analysis (Fisher's Exact and Chi-squared tests) [*performed by Niamh Errington*] showed that when we consider six clusters there are some significant differences between pairs of clusters in various clinical variables, however we didn't note a difference in survival that wasn't attributed to age.

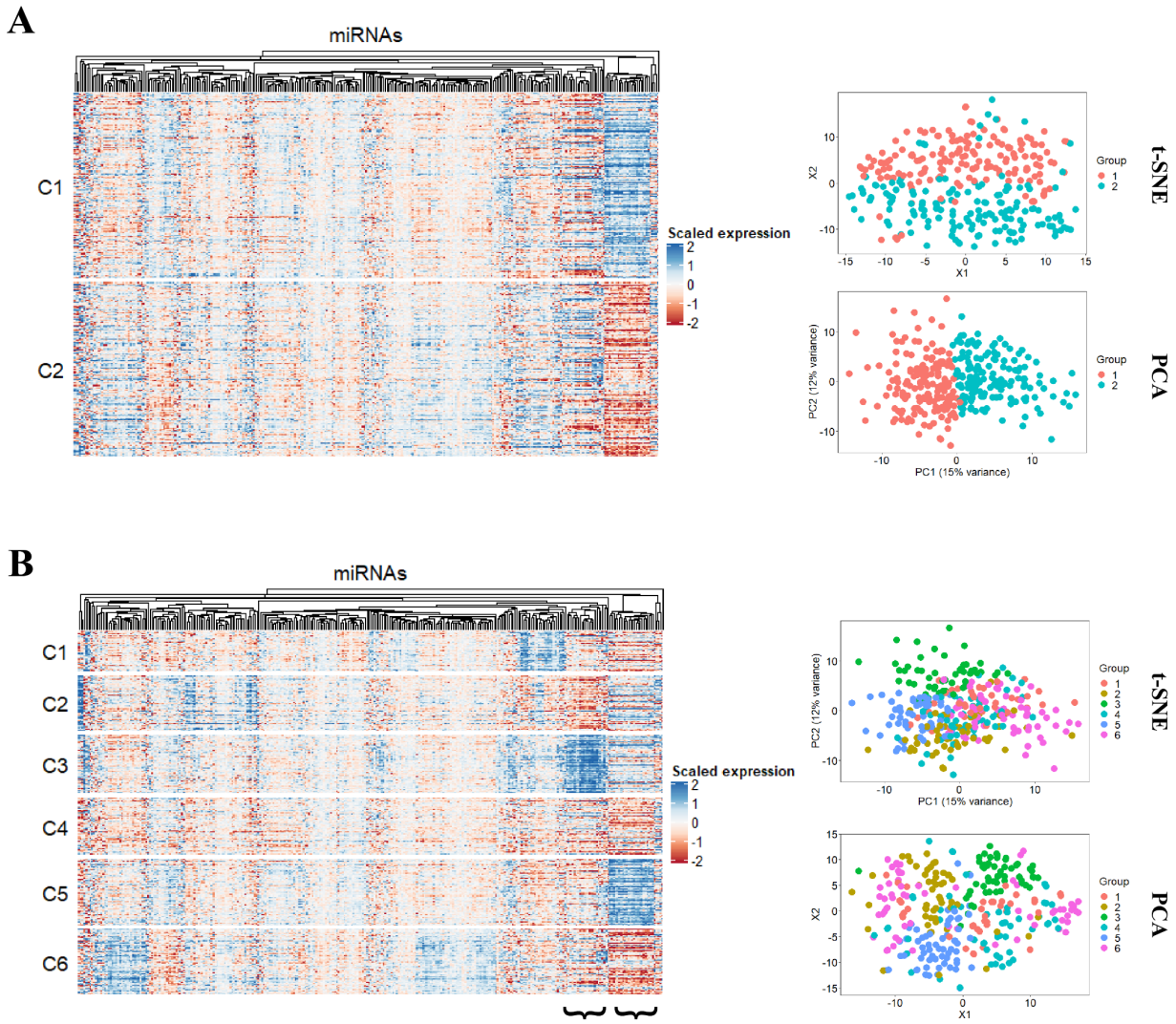


***Figure 13**: miRNA profiles, t-SNE and PCA analysis for the two estimated k for miRNA analysis 5 (CTEPH+CTED patients). A) refers to two clusters and B) refers to six clusters.*

Finally, to investigate purely PH patients, analysis 6 includes only the five PH groups. Analysis 7 looked at the full cohort which includes the five PH classifications as well as CTED(81) and patients without PH(254). The goal of these two analyses was to explore if miRNA-based clustering was able to discriminate between all or some of the PH classes and the patients without PH. As expected, both datasets were producing more stable clusters when spectral clustering was applied and they were both good candidates for Omada analysis with average stabilities of 0.91 and 0.89. As in previous analyses 50 miRNAs promoted stable clusterings with average scores of 0.83 and 0.86. **Figure 7** shows a selected miRNA overlap of 92.3% for analyses 3(manuscript 3), 6 and 7 possibly indicating PH1, PH2 and PH3 to have the strongest signatures as discovered in manuscript 3.

When estimating the number of clusters (k), analysis 6 is split between two and six possible clusters with similar vote percentages. However, looking at their respective miRNA profiles and t-SNE/PCA analyses for six clusters (**Figure 14B**) there seem to be no correspondence between the PH classes and the formed clusters despite the visible cluster separation.

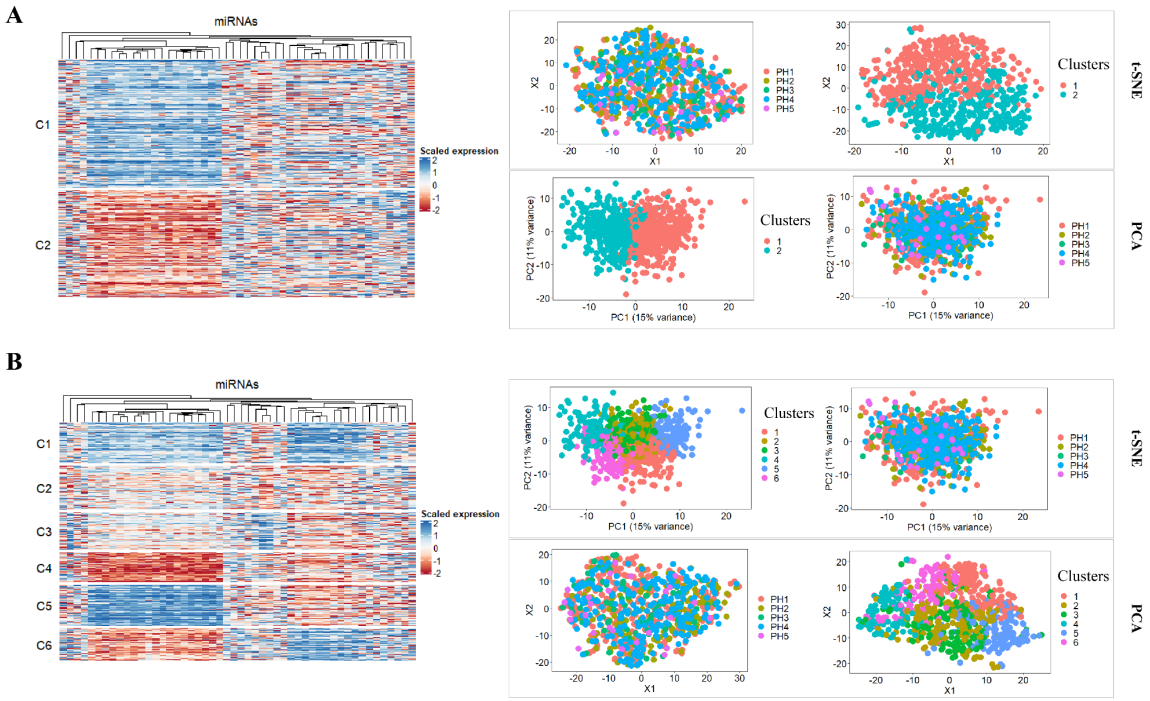


***Figure 14**: miRNA profiles, t-SNE and PCA analysis for the two estimated k for miRNA analysis 6 (all PH patients). A) refers to two clusters and B) refers to six clusters.*

Additionally, none of the PH classes concentrates in any one of the clusters as evident in **Table 2**. To get an idea of why the clusters differentiated, post-hoc tests were performed by Niamh Errington based on the cluster memberships. A number of interesting clinical variables, such as NTproBNP, mPAP and PVR, were found to be

significantly different between pairs of six clusters, fully described in **Table 3**, indicating possible differences in the mechanisms driving these groups of patients. Survival did not show any significant deviations based on cluster membership when age was also considered. Despite its inability to match patient clusters with PH classes this analysis showed the potential of this miRNA-based patient partitioning to capture phenotypically distinct groups stressing the need for further analysis and validation of the groups.

***Table 2**: Six cluster memberships for analysis 6 containing all five PH categories along with PH1 subcategories*

	A	B	C	D	E	F	Sum
PH1.1	29	54	29	34	45	39	230
PH1.2	4	1	1	2	3	1	12
PH1.3	0	2	0	1	1	0	4
PH1.4	23	24	32	23	18	29	149
PH2	26	27	23	11	35	15	137
PH3	19	21	8	13	9	13	83
PH4	55	43	52	30	50	28	258
PH5	6	4	3	7	2	5	27
Sum	162	176	148	121	163	130	**900**

***Table 3**: Clinical variables statistically tested with their adjusted p-values and the specific post-hoc tests deemed significant*

Clinical variable	Adjusted p-value	Significant post-hoc tests
NTproBNP	0	A/B, A/C, A/D, A/E, B/F, C/F, D/F
Treatment	0	A & D
sPAP	0.0012	A/E, B/E, B/F, C/F, D/E, D/F, E/F
6MWD	0.0014	A/D, A/F, B/D, B/F, C/F, D/F, E/F
mPAP	0.0014	A/E, B/F, C/F, E/F
Site	0.0025	A,B,C,D,F
Creatinine	0.0025	A/B, A/C, A/D, A/E, D/F
SvO2	0.0026	A/C, A/D, A/E, B/F, C/F, D/F, E/F
Uric Acid	0.0074	A/B, A/D, A/E
REVEAL risk group	0.0080	-
eGFR	0.0080	A/D, D/F
mRAP	0.0093	A/C, A/D, A/E
dPAP	0.0094	A/C, A/E, C/F, E/F
PVR	0.0198	E/F

Adding non-PH and CTED patients for analysis 7 the selection of k showed a different picture by spreading the votes. Non PH and CTED patients were spread across clusters with no significant enrichment anywhere (**Table 4**).

***Table 4**: Five cluster memberships for analysis 7 containing all five PH categories along with PH1 subcategories, Non-PH and CTED patients*

	A	B	C	D	E	Sum
CTED	8	16	26	14	6	70
No PH	36	54	34	21	35	180
PH1.1	66	52	29	44	38	229
PH1.2	3	4	0	2	4	13
PH1.3	0	2	0	0	2	4
PH1.4	25	25	27	34	38	149
PH2	13	42	25	33	19	132
PH3	19	13	14	19	15	80
PH4	28	58	67	57	45	255
PH5	4	4	2	7	9	26
Sum	202	270	224	231	211	**1,138**

4.5 Discussion

Gene expression studies have shown the presence of heterogeneity within PH, such as in (Kariotis *et al.*, 2021) where we showed its biological manifestation specifically on IPAH patients through exploratory machine learning. Looking from a different perspective, our learning models on circulating miRNA of manuscript 3 revealed the strength of a miRNA signature by matching or surpassing the performance of established clinical biomarkers in discriminating PH patients and/or controls. Ouf belief that patients characterised by the same signature might share pathologies was strengthened by miRNA clusters with variable prognosis and associations with distinct molecular pathways which according to literature are implicated in PAH subgroups. That was in agreement with (Yao *et al.*, 2021) where molecular signatures derived from lung tissue showed potential towards understanding PAH mechanisms. The additional miRNA analysis investigated subsets of the miRNA cohort and identified potential miRNAs that can play a role in differentiating categories which do not necessarily coincide with PH classifications. As demonstrated in (Sessa and Hata, 2013), dysregulation of miRNAs can influence the function of tissues with pathological conditions and in PH specifically where it can affect metabolic reprogramming, and enhanced proliferative capacity. Our different additional analyses showed results ranging from the inability to detect any differentiating signal

to promising phenotypic differences based on clinical variables underlying the complexity that characterises such work. However, as we showed, clustering can provide a way to explore patient heterogeneity purely based on miRNA signal and independently of any clinical information and pre-existing classifications. However, since clustering is an exploratory tool, further analysis is always needed to validate the generated clusters and associate them with clinical characteristics and outcomes such as survival.

Chapter 5 - Application of Omada to metabolite profiles

5.1 Introduction

In this chapter I worked on metabolomic data which were available for a large subset of the PAH patients presented in Chapter 3 and 4. Metabolomic data have recently been associated with PH in various ways. Metabolites showed dysregulated metabolism pathways on PH patients compared to their healthy counterparts (Chen *et al.*, 2020). Moreover, specific metabolite profiles could distinguish CTEPH patients from controls and disease comparators in (Swietlik *et al.*, 2021) showing the implication of metabolites in PH. To explore their potential, I applied Omada to a dataset of 1072 samples which overlapped with those of the dataset used for the miRNA clustering analysis in the previous section. The current dataset includes 1522 metabolites drawn from blood serum as a readout of the whole body metabolism state. This dataset was generated using liquid chromatography mass spectrometry (LC-MS) as described in Chapter 1 section *Metabolomic data*.

5.2 Methods

The current analysis aimed to partition the dataset into patient subgroups showing distinct metabolic profiles, therefore multiple combinations of PH categories were used as input. Metabolomic data collection and processing is described in the methods section of (Rhodes *et al.*, 2017). Metabolomics analysis 1 was performed for 571 patients in groups PH1, PH2 and PH3 aiming to show whether metabolites can discriminate between patients of the three categories. Metabolomics analysis 2 included 246 patients from PH4 and was performed to assess whether there is heterogeneity within CTEPH that is detectable by metabolism profile differences. Lastly, all 1072 patients were considered in metabolomics analysis 3 to assess if metabolite differences are detectable between all PH classes in this dataset. For all three clustering analyses we tested a cluster range *k* of [2, 6] for selecting the clustering method, optimal metabolite subset (with the additional parameter of comparisons set to 3, please refer to manuscript 1) and determining the number of clusters, *k*. Finally, clusters were generated for all $k \in [2, 6]$ but we calculated statistics only for the top 3 voted *k [statistics were generated by Niamh Errington]*. For the calculations, each parameter was assessed for normality, followed by the appropriate statistical test to look for associations to cluster membership. In cases of unequal variance, the data were log transformed. When the initial p-value was significant, post hoc tests were applied between groups.

5.3 Results

5.3.1 No distinct metabolism subgroups found

Table 5 details the PH categories participating in each clustering analysis along with the main Omada results for each tool. I went through with the analysis as all three datasets showed clustering potential based on the Omada feasibility analysis (using the default 3 classes) with average stability scores of 0.89, 0.88 and 0.85, respectively.

Table 5: *The three metabolite clustering analyses with the PH groups and samples sizes considered. The clustering method, optimal number of metabolites and the votes for each number of clusters (k) estimated from the Omada application are also noted.*

Analysis number	PH Groups	Number of patients	Clustering method selected	Optimal number of metabolites	k votes
1	PH1, PH2, PH3	571	spectral	50	**k.2** = 12 k.3 = 1 k.6 = 2
2	PH4	246	spectral	50	**k.2** = 12 k.5 = 1 k.6 = 2
3	All groups	1072	spectral	500	**k.2** = 13 k.3 = 1 k.6 = 1

PH1: *Pulmonary Arterial Hypertension*
PH2: *PH due to left heart disease*
PH3: *PH due to lung diseases and/or hypoxia*
PH4: *Chronic thromboembolic pulmonary hypertension (CTEPH)*
PH5: *Pulmonary hypertension with unclear multifactorial mechanisms*

The second step also showed promising signs that spectral clustering was the appropriate algorithm for the dataset as it was selected in all three cases with average (over all tested *k*) partition agreement of 0.68, 0.65 and 0.74, respectively (**Figure 15**).

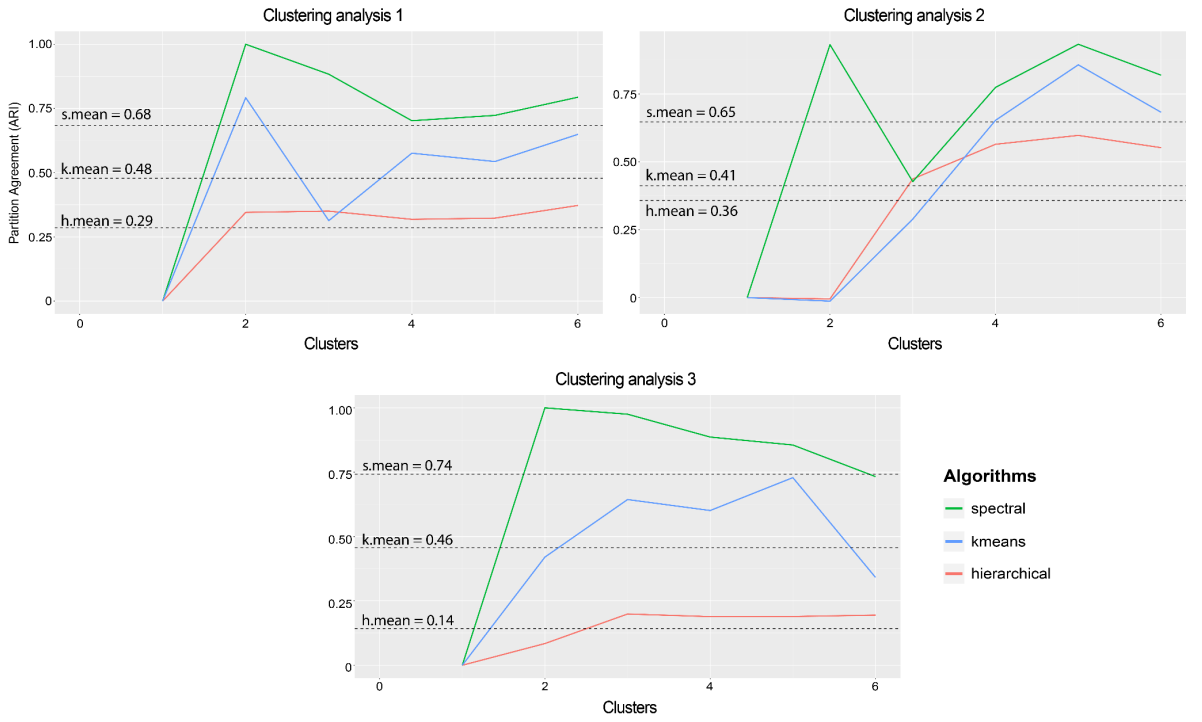


***Figure 15**: Average partition agreements for the three clustering analyses and the three algorithms tested. In all cases spectral clustering surpassed the Adjusted Rand Index threshold of 60% while the remaining algorithms showed considerably less robustness.*

Analysis 1 (PH1, PH2, PH3) and analysis 2 (PH4) showed a relatively smooth stability for different sets and selected the same number of metabolites (50) to provide the most stable clusters as shown in **Table 5** and **Figure 16**. 25 metabolites overlapped (50%) as the most variable in the two analyses. When considering analysis 3, where all patients were included, the unsupervised model estimated a much larger set of metabolites (500) and a very sharp drop when more were considered.

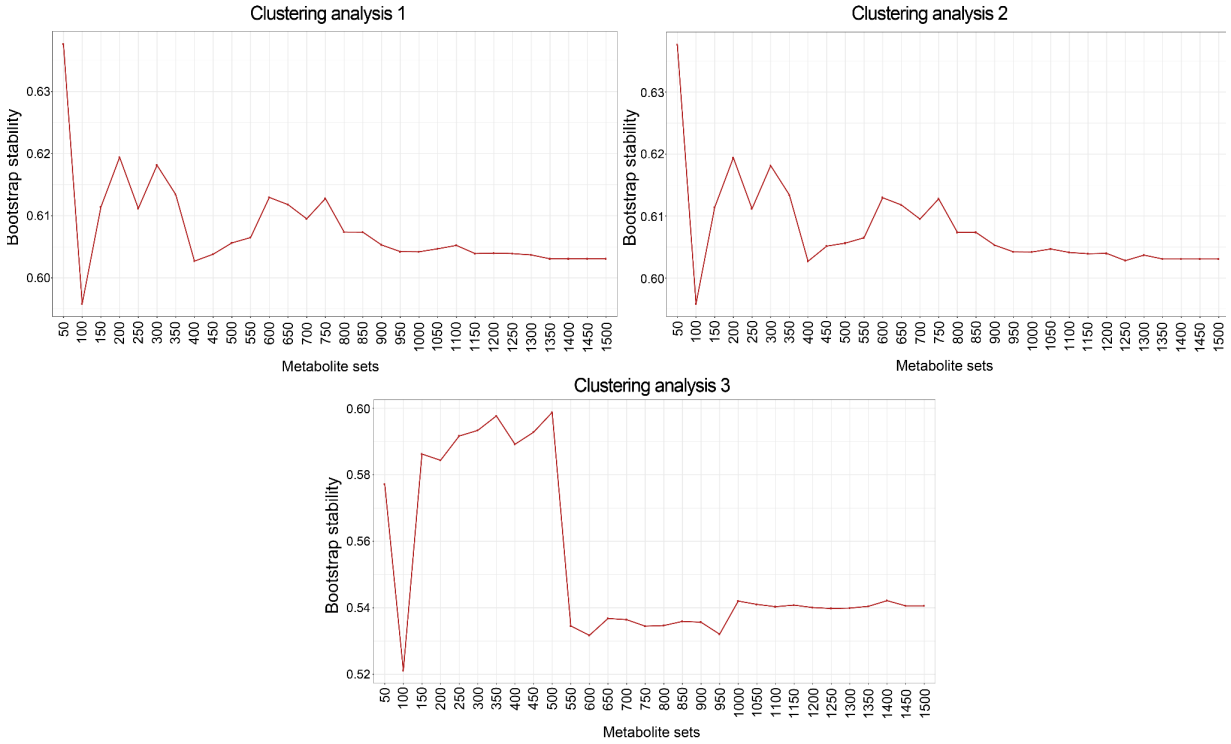


***Figure 16**: Average bootstrap stabilities for the three clustering analyses including each subset of the most variable (across patients) metabolites. Analyses 1 and 2 showed almost identical stabilites both selecting 50 metabolites (only 50% of which were overlapping). Analysis 3 showed a much different picture selecting 500 metabolites and showing a very sharp drop when additional metabolites were added.*

When estimating the number of clusters in each dataset all analyses overwhelmingly indicated the existence of two subgroups (**Table 5**) even when considering the diverse nature of the 15 internal indexes. That would indicate a very confident estimation on *k*, however since two is the smallest possible number of clusters it is very important to consider the possibility of no present clusters. First, I checked the cluster sizes for every case (**Table 6**). The large differences in sample sizes, especially in the 2-cluster cases, indicated that the algorithm was not picking up a strong signal between real subgroups and might only be forming 2 clusters due to minimal differences (otherwise the clusters would have more similar sizes). Further analysis, such as tSNE plots, show no discernible groupings and no overlap with any PH classes therefore implying the potential lack of structure in the data.

***Table 6**: The first three estimates for the number of clusters for the three analyses*

Analysis 1 (PH1, PH2, PH3)						
k	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	**Cluster 6**
2	567	5	-	-	-	-
3	29	54	489	-	-	-
6	42	305	31	48	107	39
Analysis 2 (PH4)						
k	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	**Cluster 6**
2	247	2	-	-	-	-
5	27	6	31	22	163	-
6	158	31	27	22	9	2
Analysis 3 (PH1, PH2, PH3, PH4, PH5)						
k	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	**Cluster 6**
2	3	1069	-	-	-	-
3	3	1059	10	-	-	-
6	26	394	501	1118	317	320

5.4 Discussion

It is shown by current research that metabolite changes occur under PH conditions and those changes have been utilised to identify PH categories from healthy individuals (Rhodes *et al.*, 2017; Chen *et al.*, 2020). However, no current studies present models able to discriminate between PH classes. We explored this potential from a machine learning (stability) perspective and the metabolite datasets we used showed promise in being used as inputs for clustering based on dimensionality and specific algorithm robustness. Despite our model's inability to form distinct metabolism clusters, the 25 metabolites that overlapped in variability in analyses 1 and 2 might be an interesting target for further research. Since the two analyses did not have overlapping patients the fact that 25 metabolites were selected in both means that their selection was purely based on their variable expression levels (in these two PH contexts) instead of being dependent on specific patients. Consequently, this set of metabolites might potentially play a role in general PH metabolism since the two analyses collectively include four PH categories (1: PH1, PH2, PH3 and 2: PH4). When considering all patients, the unsupervised model

estimated a much larger set of metabolites that provided stable clusters (~60%). Although this observation only serves as an indication with further analysis needed, it might imply sets of metabolites that differentiate PH subgroups in this dataset, not necessarily following the established PH classes, as in (Carlsen *et al.*, 2022) where authors demonstrated different metabolic phenotypes independent of PH diagnosis. All three analyses/datasets pointed towards the existence of two clusters underlying however a huge cluster size discrepancy. Our tSNE results for the various analyses showed no visible groupings of patients and thus no metabolism driven clusters, an idea also backed by the lack of significant differences from the statistical analysis of clinical variables associated with the clusters. Several factors might influence the ability of this metabolite dataset to find clinically relevant clusters. From a machine learning perspective, in some of these analyses the partitioning process generated clusters of very different sizes drastically reducing the statistical power of any association with clinical data. Moreover, blood serum metabolites might not present the ideal source of metabolic difference discovery as literature has shown that metabolic abnormalities in PH show variations by organ or cell type (Xu, Janocha and Erzurum, 2021). Finally, additional data might be needed to interpret or help create any metabolic clusters, such as pathway related data which have shown to differentiate within PH (Xu *et al.*, 2004; Xu and Erzurum, 2011) since the current dataset was not able to reveal any systematic structure to be captured by unsupervised learning.

Chapter 6 - Longitudinal exploration of activity during COVID

6.1 Introduction

In recent years, SARS-CoV-2 infections have caused a diverse set of symptoms varying in duration and severity (Sudre *et al.*, 2021). This large degree of heterogeneity in symptoms and the long-term outcome of infected patients with milder/asymptomatic COVID-19 that were not hospitalised remains unknown. The potential of cardiovascular activity measurements has been studied on hospitalised patients, such as heart rate variability (HRV) in (Mol *et al.*, 2021), however mild and asymptomatic cases have not been explored as of yet. Moreover, the use of wearable devices for relevant data capture during the pandemic has increased considerably (Hijazi *et al.*, 2021). The work presented in this section, which is part of manuscript 6, builds on the above information and attempts to answer the following questions:

- *Can activity-based longitudinal clustering uncover distinct COVID-19 patient groups?*
- *Can physical activity COVID19 longitudinal clusters be associated with symptom trajectories?*

During the COVID-19 pandemic frontline healthcare workers (HCW) were at a high risk of infection resulting in several local and national studies to monitor infection and symptoms. Several studies from UK NHS trusts have shown SARS-CoV-2 seroprevalence rates of 20-50% in frontline HCW after the first pandemic wave, at a time when the estimated seropositivity in the general population was only 6% (Eyre *et al.*, 2020; Houlihan *et al.*, 2020). A longitudinal study of the long-term symptoms of COVID in 38 HCW indicated that 55% indicated at least one continuous symptom, with fatigue being the most abundant symptom (57%) six months after COVID-19 diagnosis. Within the population there is a large degree of heterogeneity in COVID both in terms of severity and duration of symptoms (Sudre *et al.*, 2021), and the long-term outcome of those infected with COVID-19 who had milder or asymptomatic COVID-19 and therefore not-hospitalised remains unknown. [*Written by co-authors Varsha Gupta and Elham Alhathli*]

Previous studies have highlighted the potential for heart rate variability (HRV) to predict survival in hospitalised patients with COVID-19 (Mol *et al.*, 2021) and there

has been heightened interest in the use of wearable devices and smartphones to capture data during the pandemic (Hijazi *et al.*, 2021) but little is known about the utility of HRV, or other measurements of cardiovascular/physical activity in relation to mild or asymptomatic people infected with SARS-CoV-2. [*Written by co-authors Varsha Gupta and Elham Alhathli*]

We present data with >1 year follow up from a cohort of HCWs enrolled into a clinician-led COVID-19 symptom monitoring study who also possessed compatible iOS devices and enrolled into the MyHeart Counts iOS App study. Using unsupervised learning methods for longitudinal data we identified two trajectory patterns of COVID-19 symptoms, and two trajectory patterns representing different levels of physical activity. We observe associations of long and short COVID-19 symptom patterns with some of the physical activity features. This study highlights the importance of further monitoring of COVID symptoms and physical activity levels through wearables to understand the long-term impact of COVID-19 as well as other forms of cardiovascular and respiratory diseases. [*Written by co-authors Varsha Gupta and Elham Alhathli*]

Focusing on a 121 frontline health care worker cohort, the first part of this study generates long and short patient classes of symptom trajectories over different time-points. Additionally, activity-based patient clusters were created by longitudinal unsupervised learning over a maximum of 596 time points. Long COVID trajectories were associated with a higher number of symptoms at baseline and symptoms showing longer duration evidence. We then build association models between the classes and clusters to investigate their relationship and associate them with physical activity, health and demography data, reinfections and biological markers.

6.2 Methods

Participant recruitment

Participants were recruited into the Sheffield Teaching Hospitals NHS Foundation Trust Observational study of pulmonary hypertension, cardiovascular and respiratory diseases study (STH-ObS, 18/YH/0441). All participants provided written informed consent. Eligible participants were adults aged 18 years or older, currently working as health-care workers, including allied support and laboratory staff, and possessing an iPhone 6 or later were offered an Apple Watch Series 4. Procedures were done in compliance with the principles of the Declaration of Helsinki (2008) and the International Conference on Harmonisation Good Clinical Practice guidelines.

Of the 204 participants recruited to the study between July 2020 and July 2021 138 participants had either been PCR or were sero-positive for COVID-19 infection at the time of consent. The remaining 65 Participants were negative for COVID-19 and 1 participant did not have information about PCR or sero-positive test. Participant demographics and data on COVID-19 testing, symptoms, vaccination status were recorded during clinician led clinics. 140 participants owned compatible iOS devices and were recruited into the MyHeart Counts Study. [*Written by co-author Allan Lawrie*]

MyHeart Counts

MyHeart Counts is an iOS smartphone app that collects information about an individuals cardiovascular health, wellbeing, diet, smoking history and can perform a six-minute walking test (McConnell *et al.*, 2017; Hershman *et al.*, 2019). In addition to these self-reported questionnaires, the app can pull heart rate data from compatible wearables through Apple HealthKit and has been used as a platform for a randomised control interventional study of physical activity (Shcherbina *et al.*, 2019). During the early stage of the COVID-19 pandemic MyHeart Counts was updated (Version 2.3.0) to include self-reported COVID-19 symptoms and testing. Each participant was provided with a pseudonymous identifier to link data obtained from the App to their clinical data and participants were asked to complete questionnaires on their cardiovascular health status and history, physical activity levels, diet, sleep, wellbeing and risk perception and fortnightly updates on COVID-19 testing and symptoms. [*Written by co-authors Allan Lawrie and Emmanuel Jammeh*]

Clustering of physical activity trajectories

Longitudinal activity data of 73 patients were imported to R (version 4.2.0) and filtered based on the availability of relevant dates of interest. For each patient the start of the time series was derived using the following priorities. The date of onset (first symptoms) was prioritised followed by the date of a potential positive PCR test, followed by a potential positive serology test and lastly the date of the first physician visit. Based on the availability of the above, trajectories of 31-34 patients were created depending on the specific activity and varied in number of timepoints ranging between 42 and 596. The trajectories were utilised by a longitudinal clustering approach (KmlShape R package) to generate activity profiles for each variable. To reduce the complexity and computational requirements of longitudinal clustering as required by the algorithm, the trajectories were first minimally merged based on the Fretchet distance (Montero and Vilar, 2015). The distance between each pair of trajectories was calculated and used to generate a smaller number of senator trajectories (30-33) which represent the actual trajectories but reduce the number of calculations needed for the following clustering. In turn, the senator trajectories were

reduced to 100 time points for the purposes of calculating shape similarities. Multiple clusterings were carried out with selected k - number of clusters - ranging between two and five creating activity profiles for every type of activity registered. Due to sample sizes, the two-cluster run was retained for each activity and the differences between the mean representative trajectories of each cluster were calculated by averaging the trajectories of their members, smoothed by local polynomial regression fitting (loess in R Stats package).

The continuous distance between each trajectory and the representative trajectory was calculated using the Fretchet distance (TSdist R package) which allows the computation of distances even with unequal number of times points. The representative trajectory was calculated as the mean of the trajectories of the generated clusters based on their difference on each variable value. Each individual trajectory was signed as positive or negative depending on its position relative to the representative trajectory. Distances closer to zero represent patients closer to the representative trajectory demonstrating less clear distinction between the clusters while higher absolute values represent samples further away from the representative trajectory and as an extension from the opposite cluster.

6.3 Results

Physical activity trajectories during and after COVID-19 symptoms

Two clusters of participants were formed for each physical activity with cluster sizes and the statistics for the time-points used (overall average 487 time-points) presented in **Supplementary Table 1**. Depending on the activity measure, 31-34 patients were split between two clusters with the low activity clusters being on average 3 times larger in each case. 'Walking distance generated the most similar sized clusters ($ratio_{low\text{-}high}$ = 1.38) while the number of flights climbed showed a $ratio_{low\text{-}high}$ of 7. Walking heart-rate average, flights climbed and step count considered the most timepoints on average across patients (496.5, 493.5 and 493.5 respectively) with basal energy burned using the least (472). When checked for activity levels the pairs of clusters showed significant differences for every activity (p-values <= 2.858e-07) except the average Walking Heart Rate (p-value = 0.3593). All p-values, calculated by two-sided t.tests, are presented in **Supplementary Table 2**. The mean representative trajectories were plotted in **Figure 1** showing a distinction of high and low mean activity of the trajectory clusters accompanied by the individual time-points of the patient trajectories they represent. More prominently, patients in the high activity cluster (green colour in **Figure 1**) showed double walking/running distance (10,955m) as well as energy burned (949,650) compared to the low cluster members (4964.5m and 538,190, respectively with p-values of 2.2e-16). The highest difference

was noted at the number of flights climbed where high activity patients climbed 243 more flights, especially during the first 30 days according to **Figure 1**. Heart rate was as an average number almost identical between the two clusters (around 1.4), however high activity patients showed lower heart rate early on until the point of 160 days and higher afterwards. Heart rate variability (SDNN) differentiated considerably between the two clusters at around 200 days while energy burned demonstrated a peak at 250 days after the date of onset/PCR/serology/first visit.

Supplementary Table 1*: Timepoint statistics and details for the clusters generated for each activity variable*

	Patients	Cluster 1 (low)	Cluster 2 (high)	Min	Mean	Median	Max
heartRate	33	25	8	56.0	414.4	487.0	593.0
basalEnergyBurned	34	25	9	75.0	406.3	472.0	593.0
stepCount	32	21	11	67.0	421.3	493.5	593.0
heartRateVariabilitySDNN	32	24	8	87.0	429.2	489.0	593.0
flightsClimbed	32	28	4	88.0	434.5	496.5	596.0
distanceWalkingRunning	31	18	13	42.0	411.1	487.0	593.0
walkingHeartRateAverage	32	19	13	86.0	428.2	493.5	593.0
energyBurned	34	27	7	62.0	426.6	481.0	593.0

***Supplementary Table 2**: The clusters formed by longitudinal clustering for each of the 8 activity measures along with their sizes and the significance of the difference between their means*

Activity variables	**Cluster 1(low)**		**Cluster 2 (high)**		**p-value ($mean_{c1-c2}$)**
	n	**Mean**	**n**	**Mean**	
heart Rate	25	1.4 (0.8-2.7)	8	1.42 (0.8-2.7)	2.858e-07
basal Energy Burned	25	1,716,256.0 (14,896.0-32,318,588.0)	9	2,251,475.0 (14,979.0-63,452,416.0)	2.2e-16
step Count	21	6763.2 (4.0-64,102.0)	11	11,753.8 (2.0-218,935.0)	2.2e-16
heart Rate Variability SDNN	24	35.9 (7.8-164.0)	8	60.6 (14.1-212.7)	2.2e-16
flights Climbed	28	61.1 (1.0-489.0)	4	304.2 (2.0-1,062.0)	2.2e-16
distance Walking Running	18	4964.5 (1.3-43,860.9)	13	10,955.7 (1.51-481,780.4)	2.2e-16
walking Heart Rate Average	19	1.71 (0.8-2.9)	13	1.7 (0.9-3.0)	0.3593
energy Burned	27	538,190.8 (21.0-5,267,592.0)	7	949,650.0 (37.0-25,279,771.0)	2.2e-16

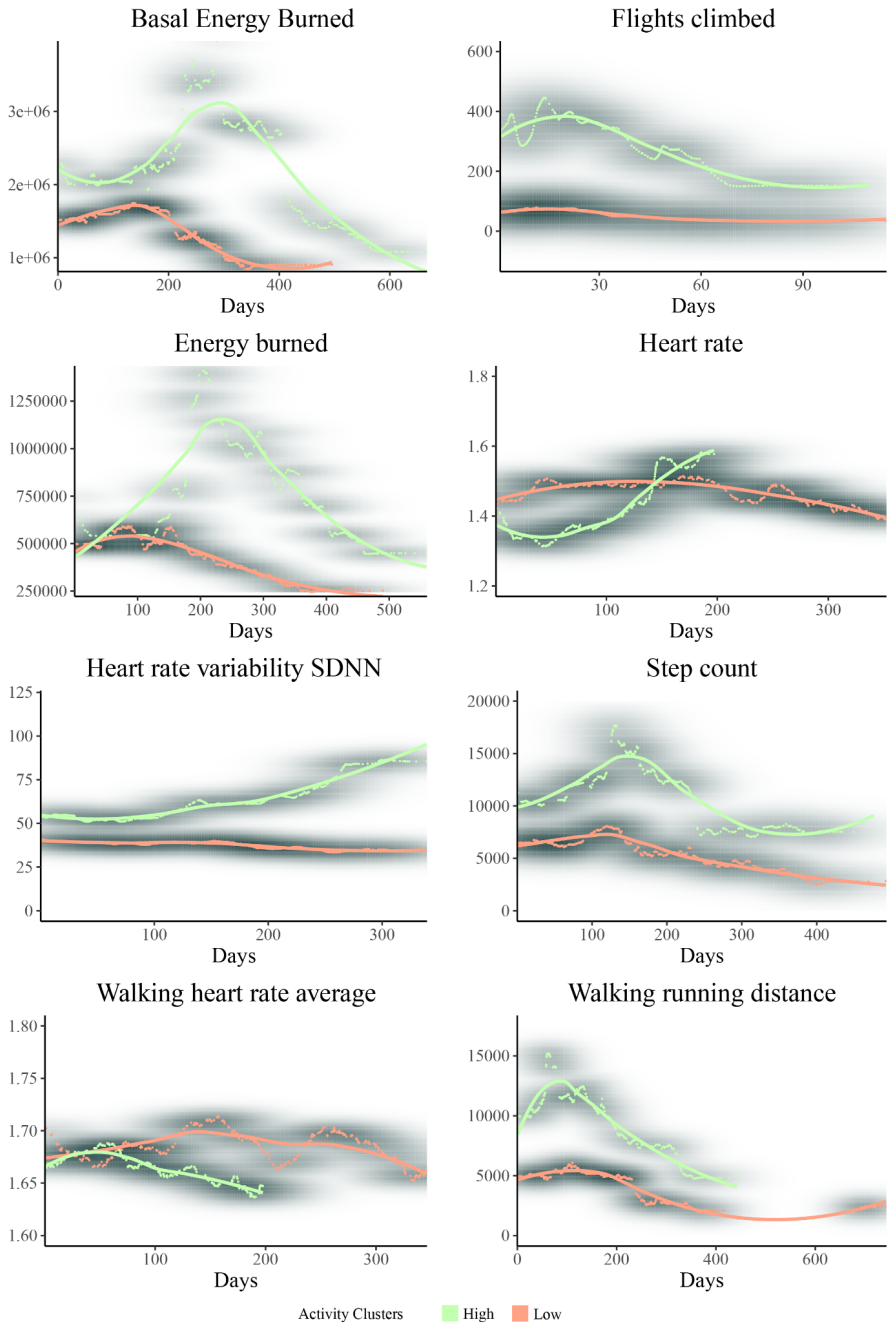


***Figure 1**: The smoothed mean representative trajectories of each generated cluster for the physical activity measures. The individual time points of the cluster members are visible around the curves along with a density estimation to help visualise their presence.*

Association of symptoms and physical activity trajectories

There were only 21 participants with both COVID symptoms and physical activity longitudinal patterns (**Figure 2**). Amongst these 21 participants, 6 participants also had re-infection. The association of high/low activity vs short/long covid using chi-square test (**Table 1**) revealed significant association of short covid with low activity cluster of ‘flight climbed’ and ‘distance walking running’ ($p<0.05$). [*Paragraph written by co-authors Elham Alhathli and Niamh Errington*]

There is association between symptoms trajectories and some baseline daily activity data, that started from the first date of symptoms onset and was averaged at different time points of 3 days, 1 week, 2 weeks, 1 month and 3 months. The activity variable distance walking, running, flight climbed (starting from 1 week average), and steps count (starting from 1 month average) have significant associations with any symptoms of long/short covid ($p <0.05$) . The difference in means between short/long covid are represented in **Figure 3** and in all the three activity variables (distance walking running, flight climbed and steps count) long covid has lower mean activity level compared to short covid (**Figure 3**). [*Paragraph written by co-authors Elham Alhathli and Niamh Errington*]

We studied the distribution of biological markers in long and short trajectories corresponding to 21 patients (9 short, 12 long trajectories) who also had reported physical activity. Although, there is no statistically significant difference in biological markers between long and short trajectories, there is elevation of ORF8 with both long covid and high activity trajectories. [*Written by co-authors Elham Alhathli and Niamh Errington*]

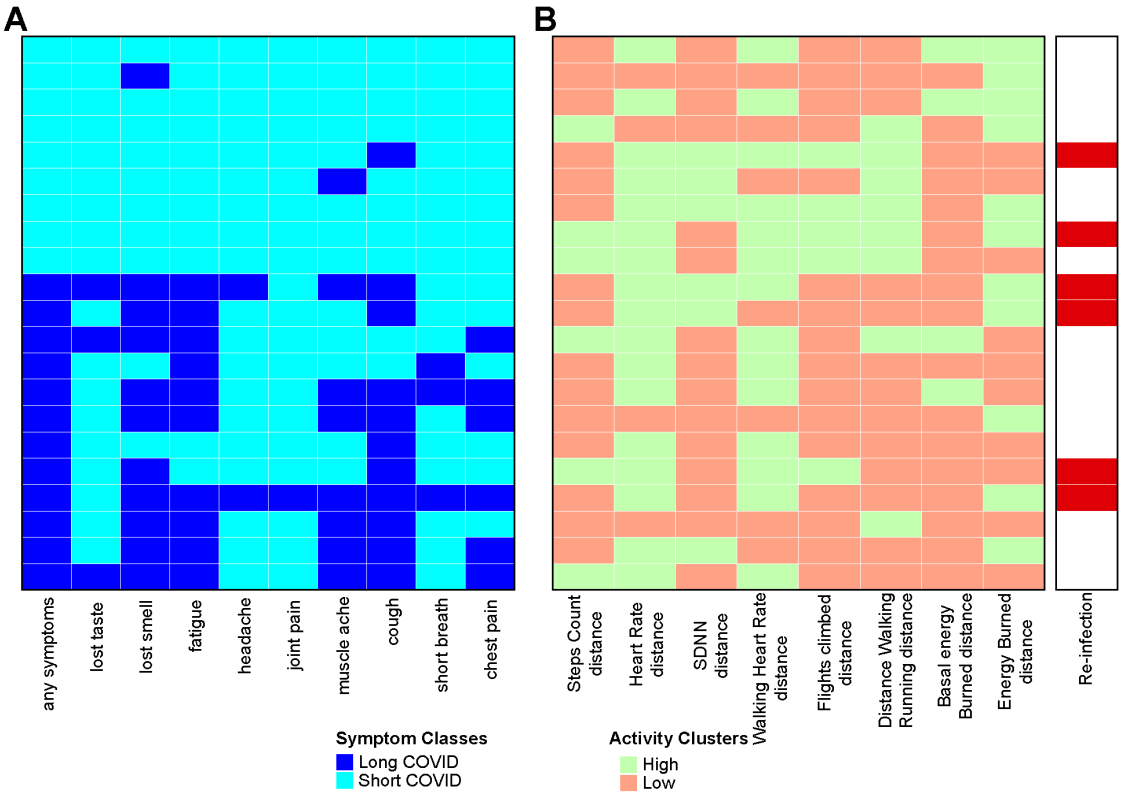


***Figure 2**: Heatmaps of unsupervised classes of 21 patients that have complete symptom and activity data. A) This heatmap shows the binary symptom classes with cyan denoting patients classified in short covid profiles and dark blue denoting patients with a long covid symptom profile. B) This heatmap shows the Frechet distance (log2 scaled) from the centre point between the two cluster centres(low and high activity). Green cells represent members of the high activity cluster while red cells represent patients of the low activity cluster. Numbers closer to zero represent patients closer to the centre point demonstrating less clear distinction between the clusters. Numbers with higher absolute values represent patients further away from the centre point and as an extension from the opposite cluster. On the right side with dark red are denoted the patients that were re-infected*

***Table 1**: Association between long and short covid trajectories vs high and low activity cluster in in any symptoms using chi-sq test*

Low v High Activity Cluster	Long vs Short COVID Symptoms	
	chi-squared	P-value
Steps Count	5.36e-32	1
Heart Rate	0.1	0.75
Heart Rate Variability	0.17	0.68
Walking Heart Rate	0	1
Flights Climbed	3.7	0.05
Distance Walking Running	5.5	0.02
Basal Energy Burned	0.1	0.75
Energy Burned	1.28	0.26

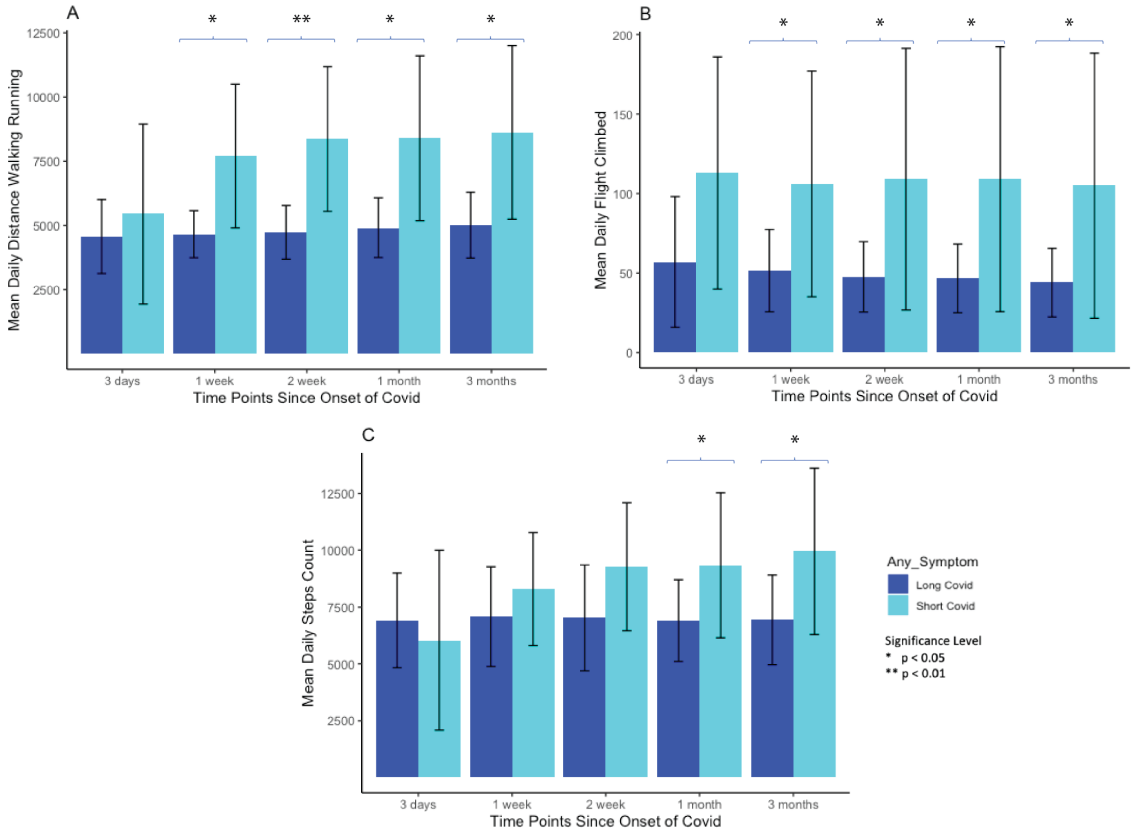


***Figure 3**: Boxplots to show distribution of biological markers in long and short COVID patients. [Figure generated by co-authors]*

6.4 Discussion

Recent COVID literature has shown that infections can cause short and long lasting symptoms and that activity measurements can be used to partition patients. Specifically, wearable device readouts currently generate physical activity measurements of adequate quality and depth to be used in healthcare research such as in (O'Regan *et al.*, 2021) where cluster analysis showed distinct patient subgroups associated with morbidity and healthcare utilisation. Looking from a longitudinal perspective and using activity data, three phenotypes (clusters) of different physical activity patterns and quality of life have been previously identified (Carvalho da Silva *et al.*, 2022). In this work, we used activity data from wearable devices of 121 healthcare workers along with registered symptoms to generate symptom trajectories (short and long covid) and longitudinal physical activity trajectories/clusters (high and low activity). We observed some associations between short COVID and high activity in the form of higher levels of variables like walking, flights climbed and step count. The clustering approach we used for activity trajectories provided a flexible method that allowed patient trajectories with different amounts of timepoints to be included so that we avoid the shrinkage of our dataset. However, having a small number of patients (~32) reduced the capabilities of our longitudinal model to more accurately partition patients and generate representative trajectories per cluster. Moreover, an even smaller set of patients overlapped with the symptom dataset (21) further reducing the statistical power of our associations. The activity gaps observed due to work limitations relative to the wearable devices (healthcare worker restrictions of wearing devices) also created a mismatch of registered data at different timepoints resulting in the disqualification of clustering algorithms that require input time series vectors of identical length.

Chapter 7 - Final discussion and future directions

This work aimed to explore the potential of unsupervised learning used along with various omics data to uncover patient subgroups present due to disease heterogeneity. The main focus was to build a toolkit for non-machine-learning experts that largely automates unsupervised analysis of RNA-sequencing data and provides valuable insights. This will directly enable researchers to overcome the challenge of using complex machine learning algorithms without extensive training to investigate heterogeneity through molecular big data. To validate its utility, I applied the Omada toolkit to a real heterogeneous dataset and discovered biologically and clinically distinct patient subgroups through additional analysis. To assess the expansion capabilities and robustness of Omada (and as an extension the potential of clustering analysis) I also tested on more omics data types and across a variety of clinical classes. Furthermore, I experimented with clustering trajectories of times series activity data to stratify a COVID cohort longitudinally.

Omada was designed to make the most impactful decisions required during any clustering analysis, overcoming the underpowered approaches of using default clustering algorithms and parameters. Such decisions include selecting the most robust algorithm for a dataset, which set of genes offers the highest cluster stability as well as determining how many clusters we should be looking for in our data. Utilising multiple widely used and established algorithms, metrics and measures of clustering quality this toolkit aims to especially enable non-experts to work through a complete clustering analysis justifying each decision along the way allowing more time to be invested in interpreting insights. Notably, these tools were created with core machine learning values in mind so they can be used, extended and maintained effectively. Firstly, every result, intermediate or final, can be justified by established machine learning theory and formulas and in turn provide an explanation for each decision (i.e. explainability). Moreover, the tools allow for testing hypotheses, such as the existence of patient subgroups in a dataset, and results can be directly interpreted (i.e interpretability) towards valuable insights (e.g. lower algorithm agreement and stability scores indicate the lack of multiple clusters/heterogeneity). As far as the results are concerned, all outputs are provided in an easy to find, collect and understand manner and no decisions are hidden making them findable and accessible. This pipeline also has an interoperable character as each step feeds the next one and the final step provides data that can be used in further standard analyses. Additionally, each result can be re-used by other methods due to their standard format.

All the above were studied in the context of gene expression and were shown to output invaluable insights concerning disease heterogeneity. However, additional useful knowledge was extracted from this exploratory work. Omada's application on simulated data showed when a dataset consists of a single class we can observe

significantly lower scores when we are selecting the number of clusters. This observation can serve as a hint towards the limited utility of the input dataset. Furthermore, when tested on a dataset without visible (or expected) differences among participants (i.e. all healthy maternal RNA, GUSTO) our tools can provide indications towards the lack of subgroups (consistently low stabilities) and the presence of potential biases which becomes more transparent when there is no underlying signal. Omada's tools attempt to directly assist with several clustering analysis challenges such as reducing dimensions to improve specificity and computational efficiency. Crucially, utilising and comparing multiple approaches (clustering algorithms, cluster quality metrics etc) also helps reduce the complexity of making informed and dependable analysis decisions.

Pulmonary Hypertension (PH) and its understudied structure is the focus of my work's heterogeneity exploration and a complex disease with several unique pathologies characterised by increased pressure within pulmonary circulation. Its development has been associated with numerous genes (such as BMPR2), biochemical and molecular pathways (TGF-β) establishing a genetic link early on. In (Kariotis *et al.*, 2021) we described our methodology to discover H/IPAH patient clinically-agnostic subgroups based on gene expression (RNA-seq) and presented our results that associated this heterogeneity with prognosis levels, immune gene profiles and relevant variants. In (Kariotis *et al.*, 2021) we validated the utility of Omada by partitioning H/IPAH patients in three distinct subgroups independently of any PH clinical classification aiming for novel insights. Using additional clinical data we demonstrated different prognosis (survival), expression and immune profiles across the subgroups by revealing the predictive power of immunoglobulin genes, NOG and ALAS2 emphasising their potential to determine patient outcome and potentially be future drug targets or patient treatment response identifiers. In general, this work highlights the importance and utility of RNA-sequencing data in reflecting certain aspects of disease mechanisms which can then be perceived and examined as systems of reduced complexity. Such systems can then be investigated and provide more direct information on disease related hypotheses.

To showcase that multiple omics can be used by unsupervised learning to show different aspects of a disease I also used microRNAs, non-protein coding RNA sequences that play a very important role in gene regulation by RNA silencing or post-transcriptional regulation within the nucleus. A large number of microRNAs have been measured in blood and their role has been assessed in various contexts (O'Brien *et al.*, 2018), including as promising biomarkers in cancer diagnostics (Mishra, 2014). Especially in PAH, (Zhou, Chen and Raj, 2015) has reviewed the role of microRNAs in various cell types and their future contribution to treatment strategies. PH has a specific clinical classification (that includes PAH, as described in Chapter 1, Pulmonary Arterial Hypertension) based on which patients are diagnosed and subsequently treated. However, this classification is based on clinical observations which might limit the identification of microRNA signatures that can potentially drive

existing patient heterogeneity. In this work I focused on exploring this heterogeneity on a dataset of 1138 samples and 554 microRNAs in line with the current need to stratify patients towards targeted medicine. The main body of this section concerned a subset of patients (615) and microRNAs (326) and challenged their categorisation in three PH categories PH1(Pulmonary Arterial Hypertension), PH2 (PH due to left heart disease), PH3(PH due to lung diseases and/or hypoxia). Our Omada analysis identified six clusters based on 50 microRNAs with distinct expression profiles and clinical characteristics such as phenotypes with low survival/mPAP/PVR but higher than expected NT-proBNP. The aforementioned fact coupled with the observation that no PH classes were fitting in any generated clusters highlighted the existence of currently undetected phenotypic patient differences that might provide integral disease knowledge. Of interest was the observation that these groups of patients were characterised by combinations of clinical variables rather than a single measurement (e.g. PVR) showing that single biomarkers might possess insufficient identifying power. Next I chose to widen the scope of research questions and utilised multiple combinations of patients belonging to PH classes. I discovered that the various microRNA datasets were feasible inputs for Omada (clustering) and spectral clustering seemed to be the most robust algorithm to use. I generally detected a weak signal when examining single PH classes (PH1) and subclasses (PH1.1, systemic diseases such as connective tissue diseases, HIV infection and congenital heart disease). However, some significant clinical differences in the generated clusters were detected in more than one of these microRNA analyses and might provide an interesting prospect for further research between cluster phenotypes. Analysis of CTEPH and CTED was unable to indicate distinct phenotypes based on microRNAs as patient age and sex drove the partitioning, a phenomenon unsurprisingly (risk factors in PAH (Schachna *et al.*, 2003)) also observed during RNA-seq clustering analysis. Lastly, when the whole cohort was examined 50 microRNAs generated the most stable clusters and they completely overlapped with the microRNAs found in manuscript 3 enhancing the confidence in the finding of manuscript 3. Interestingly, the associations of manuscript 3 (PH1, PH2, PH3) could not be replicated in our full cohort analysis (PH1, PH2, PH3, PH4, PH5) potentially showing a stronger signal in the first three PH classes which weakens when additional PH4, PH5 patients are included. All the above phenotypic differences across patient subgroups constitute a first step towards building representative molecular subtypes which in turn can be the basis of new diagnostic tests.

Following the latest research where metabolites are being implicated with dysregulated metabolism in PH and certain profiles distinguishing CTEPH patients, I investigated a metabolite PAH cohort overlapping with the aforementioned microRNA dataset. I focused on the potential heterogeneity in clinical PH classes and patients with CTEPH. The datasets were stable enough and spectral clustering was sufficiently robust but in all cases the minimum *k* (=2) was selected with extremely differently sized clusters. If we also consider the unclear t-SNE results when looking for PH classes, the two clusters are most probably a selection of exclusion and not

based on a real metabolite signal. However, during the analysis 25 metabolites were independently selected when exploring different PH classes showing potential to influence PH functions and providing a hint for future research.

Studies have shown that clinical measurements and COVID-19 can be used in predictive models as in (Mol *et al.*, 2021) where heart rate variability was used to predict survival of hospitalised patients. Wearable devices were also reviewed in terms of physical activity changes pre and post COVID-19 and showed high percentages of reduction (Panicker and Chandrasekaran, 2022). Following such leads, our longitudinal COVID-19 unsupervised methods identified associations between symptoms and short/long COVID-19 groups with time series clustering generating independent high and low activity groups. Although the activity groups showed weak association with short/long COVID groups they still provide an indication of low activity participants more likely to have long COVID-19. Moreover, we showed increased presence of the ORF8 protein in both high activity and long COVID-19 groups, agreeing with recent literature (Flower *et al.*, 2021) about the contribution of ORF8 to COVID-19 pathogenesis. This longitudinal work showed the utility of a different clustering approach in grouping patient trajectories and therefore the overall potential of unsupervised models to provide insights in various health contexts.

For users with new datasets, we recommend pre, during and post application tips for using Omada or similar approaches. Although preprocessing steps are not part of this toolkit we have found that, as a clustering pipeline, potential input data need to be processed to increase the chances of identifying strong relations between data points of the same cluster. Proper scaling and normalisation, such as the hyperbolic arcsine used in manuscript 1 will allow algorithms to access each gene equally and reduce the impact of outliers and data noise present in every real life medical dataset. Also, missing values in gene expression should either be imputed or the genes excluded depending on the percentage of missingness. Post processing can also be applied after the generation of the first clustering results (memberships) when the user can identify obvious genes driving clusters, such as sex exclusive genes (nearly) perfectly separating patients into male and female clusters. The feasibility preparation step Omada offers can and must be used only as an indication of dataset's fitness for clustering analysis. Since this step assesses a dataset based on its dimensions it serves as a checkpoint towards the main clustering analysis and not as a reassurance that a real signal exists in the input data. After acquiring a dataset fit for clustering analysis, selecting the appropriate algorithm through Omada will yield the most stable and robust methodology across multiple cluster requirements and parameters ensured by the randomised component of the function. Our testing showed spectral clustering to work very well with RNA-sequencing data due to its ability to globally consider a large dataset, in contrast to hierarchical clustering which includes some arbitrary localised decisions. Regarding the question of which genes to include in an analysis, Omada's feature selection step is expected

to filter down a gene set considerably (from tens of thousands to maximum hundreds). The relevant scores can be a strong indication of the gene set that can drive the differences between patient groups and the stability trajectory should be examined carefully to decide the most inclusive set of genes that might be of interest. Following that assessment, the number of clusters can be estimated through Omada's ensemble step. It is advised not to automatically adopt the highest estimate but rather observe the top two or three most voted options. In case these estimates are close in number (i.e. 3 and 4 clusters) the highest one should be selected so as not to lose potentially valuable information. Lastly, as part of a meta-analysis, Omada's Lasso coefficients can serve as a hint towards the association (and its directionality) between specific genes and patient groups since they result from the optimised clustering model generated by the above steps. It is advised to complement these associations with further analysis of biological nature.

7.1 Limitations

Clustering is a valuable tool to uncover relationships between samples and form groups with unique characteristics irrespectively of external information, but since it is not based on truth labels its accuracy can be lesser than supervised models. Also, an integral and challenging part of using unsupervised learning in scientific research is the interpretation of partitioning results as we did in (Kariotis *et al.*, 2021) (manuscript 2) through a large number of analyses. In the same study we used whole blood as a surrogate which might have hampered our findings' power, avoiding however invasive sample collection. The totality of this work was hindered up to a point by the amount of data wrangling required to run any part of the analysis. Data generated in bulk cannot be used without, sometimes rigorous, preparation. In our case, although we undertook several preprocessing steps, in both RNA-sequencing and microRNA datasets, a number of samples were excluded due to low quality or uncertain clinical information effectively reducing the statistical power of our models or analyses. Although this did not harm our presented results it might have hidden weaker associations that we would otherwise have picked up. Additionally, although we used various excellent data this research could delve deeper if the overlap of patients across our available datasets (RNA-seq, microRNA, metabolomic) was higher in order to combine the different dimensions of patient groups and look for signs of more complex disease mechanisms possibly forming networks. Lastly, the inherent differences between RNA datasets did not allow for the creation of an additional Omada modality that would systematically tackle input data preprocessing. However, indications of well tested and used preprocessing techniques and suggestions are presented throughout this work to assist users with their dataset.

7.2 Future directions

Unsupervised learning is gaining traction with the increased availability of data, the constant conception of new advanced techniques and the inability of people to label generated data. This creates an interesting future opportunity to increase the scope of this work by adding newly developed clustering methods as well as more data types. Due to the character of clustering, where no solution can be validated without a doubt, examining and comparing more approaches can add an extra degree of result confidence or even provide a way to tackle an exploratory problem where other angles have failed. Furthermore, various hypotheses can be tested in the clustering agnostic way when new data types (omics and non-omics) are utilised showing new perspectives to look at solutions. In principle, most current clustering approaches require normally distributed continuous numeric inputs (but definitely not exclusively) which is also the format modern omics measurements are generated as potentially allowing for transferable use of the same algorithms. This work looked at several omic types individually to provide insights but to take it a step further, omics integration may offer a more robust and complete picture of cell mechanisms and therefore contain a stronger biological signal, with our RNA-seq, microRNA and metabolite datasets serving as a potential example of integration. However, omics (and non-omics) integration can prove to be a very challenging piece of work (López de Maturana *et al.*, 2019) requiring intricate data structures, such as tensors, and specialised algorithms (Garali *et al.*, 2018). Longitudinal analysis could also be an additional tool to explore datasets (Domanskyi, Piermarocchi and Mias, 2020) where the interest lies in the development of a disease rather than snapshots of expression. The second part of an exploratory analysis aims to take advantage of sample partitions to extract useful insights in conjunction with clinical characteristics. Based on that, another exciting future focus of this work/toolkit can be additional meta-analysis, such as survival, pathway, clinical associations and expression-clinical correlation networks. With adequate data, agnostically generated memberships can be connected to patient clinical features and implicate pathways while combining expression patterns with the enrichment of clinical variables. The above would provide a wider picture of the disease in discussion and more opportunities for further targeted research.

In summary, this work is a comprehensive exploration of the application of clustering on gene expression (and other omics) data in a disease context. It attempts to test hypotheses related to the need of patient stratification due to the numerous sources of disease heterogeneity. A large number of decisions and machine learning model tuning was implemented while handling omics data revealing (and dealing with) a lot of the difficulties researchers face when unsupervised learning is applied. As a main contribution, this work resulted in a toolkit tailored to assist researchers in their clustering analysis and more importantly provide the initial insights on which further research can be based on to illuminate underlying drivers of patient phenotypic

groups. The potential future extension of this toolkit, as clustering algorithms advance, can serve as the first step for disease studies that target personalised medicine.

Bibliography

Aalen, O.O. *et al.* (2015) 'Understanding variation in disease risk: the elusive concept of frailty', *International journal of epidemiology*, 44(4), pp. 1408–1421.

Adams, G. (2020) 'A beginner's guide to RT-PCR, qPCR and RT-qPCR', *The biochemist*, 42(3), pp. 48–53.

Ameur, A., Kloosterman, W.P. and Hestand, M.S. (2019) 'Single-Molecule Sequencing: Towards Clinical Applications', *Trends in biotechnology*, 37(1), pp. 72–85.

Atkin, A.J. *et al.* (2017) 'Harmonising data on the correlates of physical activity and sedentary behaviour in young people: Methods and lessons learnt from the international Children's Accelerometry database (ICAD)', *International Journal of Behavioral Nutrition and Physical Activity*. Available at: https://doi.org/10.1186/s12966-017-0631-7.

Balan, T.A. and Putter, H. (2020) 'A tutorial on frailty models', *Statistical methods in medical research*, 29(11), pp. 3424–3454.

Bayat, A. (2002) 'Science, medicine, and the future: Bioinformatics', *BMJ* , 324(7344), pp. 1018–1022.

Bazan, I.S. and Fares, W.H. (2015) 'Pulmonary hypertension: diagnostic and therapeutic challenges', *Therapeutics and clinical risk management*, 11, pp. 1221–1233.

Berrig, C., Andreasen, V. and Frost Nielsen, B. (2022) 'Heterogeneity in testing for infectious diseases', *Royal Society open science*, 9(5), p. 220129.

Bisserier, M. *et al.* (2020) 'Targeting epigenetic mechanisms as an emerging therapeutic strategy in pulmonary hypertension disease', *Vascular biology (Bristol, England)*, 2(1), pp. R17–R34.

Blankley, S. *et al.* (2014) 'The application of transcriptional blood signatures to enhance our understanding of the host response to infection: the example of tuberculosis', *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 369(1645), p. 20130427.

Burgel, P.-R., Paillasseur, J.-L. and Roche, N. (2014) 'Identification of clinical phenotypes using cluster analyses in COPD patients with multiple comorbidities', *BioMed research international*, 2014, p. 420134.

Burrell, R.A. *et al.* (2013) 'The causes and consequences of genetic heterogeneity in cancer evolution', *Nature*, 501(7467), pp. 338–345.

Cannell, I.G., Kong, Y.W. and Bushell, M. (2008) 'How do microRNAs regulate gene expression?', *Biochemical Society transactions*, 36(Pt 6), pp. 1224–1231.

Carè, A. *et al.* (2007) 'MicroRNA-133 controls cardiac hypertrophy', *Nature medicine*, 13(5), pp. 613–618.

Carlsen, J. *et al.* (2022) 'An explorative metabolomic analysis of the endothelium in pulmonary hypertension', *Scientific reports*, 12(1), p. 13284.

Carvalho da Silva, M.M. *et al.* (2022) 'Health-Related Quality of Life and Daily Physical Activity Level in Patients with COPD- a Cluster Analysis', *COPD*, 19(1), pp. 309–314.

Chen, C. *et al.* (2020) 'Metabolomics reveals metabolite changes of patients with pulmonary arterial hypertension in China', *Journal of cellular and molecular medicine*, 24(4), pp. 2484–2496.

Chen, X. *et al.* (2008) 'Integration of external signaling pathways with the core transcriptional network in embryonic stem cells', *Cell*, 133(6), pp. 1106–1117.

Chen, Y.-Y. and Assefa, Y. (2021) 'The heterogeneity of the COVID-19 pandemic and national responses: an explanatory mixed-methods study', *BMC public health*, 21(1), p. 835.

Choi, E. *et al.* (2017) 'Using recurrent neural network models for early detection of heart failure onset', *Journal of the American Medical Informatics Association: JAMIA*, 24(2), pp. 361–370.

Cho, J.H. and Feldman, M. (2015) 'Heterogeneity of autoimmune diseases: pathophysiologic insights from genetics and implications for new therapies', *Nature medicine*, 21(7), pp. 730–738.

Chung, R.-H. and Kang, C.-Y. (2019) 'A multi-omics data simulator for complex disease studies and its application to evaluate multi-omics data analysis methods for disease classification', *GigaScience*, 8(5). Available at: https://doi.org/10.1093/gigascience/giz045.

Cimmino, A. *et al.* (2005) '*miR-15* and *miR-16* induce apoptosis by targeting BCL2', *Proceedings of the National Academy of Sciences*, pp. 13944–13949. Available at: https://doi.org/10.1073/pnas.0506654102.

Clarke, R. *et al.* (2008) 'The properties of high-dimensional data spaces: implications for exploring gene and protein expression data', *Nature reviews. Cancer*, 8(1), pp. 37–49.

Clish, C.B. (2015) 'Metabolomics: an emerging but powerful tool for precision medicine', *Cold Spring Harbor molecular case studies*, 1(1), p. a000588.

Coene, K.L.M. *et al.* (2018) 'Next-generation metabolic screening: targeted and untargeted metabolomics for the diagnosis of inborn errors of metabolism in individual patients', *Journal of inherited metabolic disease*, 41(3), pp. 337–353.

Corchete, L.A. *et al.* (2020) 'Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis', *Scientific reports*, 10(1), p. 19737.

Costa, V. *et al.* (2013) 'RNA-Seq and human complex diseases: recent accomplishments and future perspectives', *European journal of human genetics: EJHG*, 21(2), pp. 134–142.

Das, J. *et al.* (2014) 'Clustering-based recommender system using principles of voting theory', in *2014 International Conference on Contemporary Computing and Informatics (IC3I)*, pp. 230–235.

Davidson, S.B., Overton, C. and Buneman, P. (1995) 'Challenges in integrating biological

data sources', *Journal of computational biology: a journal of computational molecular cell biology*, 2(4), pp. 557–572.

Dawes, T.J.W. *et al.* (2017) 'Machine Learning of Three-dimensional Right Ventricular Motion Enables Outcome Prediction in Pulmonary Hypertension: A Cardiac MR Imaging Study', *Radiology*, 283(2), pp. 381–390.

Devi, S.G., Gayathri Devi, S. and Sabrigiriraj, M. (2018) 'Feature Selection, Online Feature Selection Techniques for Big Data Classification: - A Review', *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)* [Preprint]. Available at: https://doi.org/10.1109/icctct.2018.8550928.

van Dijk, E.L. *et al.* (2018) 'The Third Revolution in Sequencing Technology', *Trends in genetics: TIG*, 34(9), pp. 666–681.

Domanskyi, S., Piermarocchi, C. and Mias, G.I. (2020) 'PyIOmica: longitudinal omics analysis and trend identification', *Bioinformatics* , 36(7), pp. 2306–2307.

Drury, R.E., O'Connor, D. and Pollard, A.J. (2017) 'The Clinical Application of MicroRNAs in Infectious Disease', *Frontiers in immunology*, 8, p. 1182.

Ehret, G. (2018) *Complex cardiovascular diseases: the genetics of arterial hypertension*. Oxford University Press.

Elinoff, J.M. *et al.* (2018) 'Challenges in Pulmonary Hypertension: Controversies in Treating the Tip of the Iceberg. A Joint National Institutes of Health Clinical Center and Pulmonary Hypertension Association Symposium Report', *American journal of respiratory and critical care medicine*, 198(2), pp. 166–174.

Ellsworth, R.E. *et al.* (2017) 'Molecular heterogeneity in breast cancer: State of the science and implications for patient care', *Seminars in cell & developmental biology*, 64, pp. 65–72.

Errington, N. *et al.* (2021) 'A diagnostic miRNA signature for pulmonary arterial hypertension using a consensus machine learning approach', *EBioMedicine*, 69, p. 103444.

Esteva, A. *et al.* (2017) 'Dermatologist-level classification of skin cancer with deep neural networks', *Nature*, 542(7639), pp. 115–118.

Eyre, D.W. *et al.* (2020) 'Differential occupational risks to healthcare workers from SARS-CoV-2 observed during a prospective observational study', *eLife*, 9. Available at: https://doi.org/10.7554/eLife.60675.

Feachem, R.G.A. *et al.* (2010) 'Shrinking the malaria map: progress and prospects', *The Lancet*, 376(9752), pp. 1566–1578.

Feng, F.Y. *et al.* (2021) 'Association of Molecular Subtypes With Differential Outcome to Apalutamide Treatment in Nonmetastatic Castration-Resistant Prostate Cancer', *JAMA oncology*, 7(7), pp. 1005–1014.

Fiers, M.W.E.J. *et al.* (2018) 'Mapping gene regulatory networks from single-cell omics data', *Briefings in Functional Genomics*, pp. 246–254. Available at: https://doi.org/10.1093/bfgp/elx046.

Firth, A.L., Mandel, J. and Yuan, J.X.-J. (2010) 'Idiopathic pulmonary arterial hypertension', *Disease models & mechanisms*, 3(5-6), pp. 268–273.

Fisher, R., Pusztai, L. and Swanton, C. (2013) 'Cancer heterogeneity: implications for targeted therapeutics', *British journal of cancer*, 108(3), pp. 479–485.

Flower, T.G. *et al.* (2021) 'Structure of SARS-CoV-2 ORF8, a rapidly evolving immune evasion protein', *Proceedings of the National Academy of Sciences of the United States of America*, 118(2). Available at: https://doi.org/10.1073/pnas.2021785118.

Fradkov, A.L. (2020) 'Early History of Machine Learning', *IFAC-PapersOnLine*, 53(2), pp. 1385–1390.

Freedson, P. *et al.* (2012) 'Assessment of physical activity using wearable monitors: recommendations for monitor calibration and use in the field', *Medicine and science in sports and exercise*, 44(1 Suppl 1), pp. S1–4.

Füzéki, E., Engeroff, T. and Banzer, W. (2017) 'Health Benefits of Light-Intensity Physical Activity: A Systematic Review of Accelerometer Data of the National Health and Nutrition Examination Survey (NHANES)', *Sports medicine* , 47(9), pp. 1769–1793.

Garali, I. *et al.* (2018) 'A strategy for multimodal data integration: application to biomarkers identification in spinocerebellar ataxia', *Briefings in bioinformatics*, 19(6), pp. 1356–1369.

García-Guerrero, V.M. and Beltrán-Sánchez, H. (2021) 'Heterogeneity in Excess Mortality and Its Impact on Loss of Life Expectancy due to COVID-19: Evidence from Mexico', *Canadian studies in population*, 48(2-3), pp. 165–200.

Genomic profiling for prostate and bladder cancers (2019) *Decipher*. Decipher Urologic Cancers, subsidiary of Veracyte Inc. Available at: https://decipherbio.com/ (Accessed: 2 August 2022).

Gerdes, M.J. *et al.* (2014) 'Emerging understanding of multiscale tumor heterogeneity', *Frontiers in oncology*, 4, p. 366.

Gieger, C. *et al.* (2008) 'Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum', *PLoS genetics*, 4(11), p. e1000282.

Gligorijević, V. and Pržulj, N. (2015) 'Methods for biological data integration: perspectives and challenges', *Journal of the Royal Society, Interface / the Royal Society*, 12(112). Available at: https://doi.org/10.1098/rsif.2015.0571.

Gräf, S. *et al.* (2018) 'Identification of rare sequence variation underlying heritable pulmonary arterial hypertension', *Nature communications*, 9(1), p. 1416.

Gundem, G. *et al.* (2015) 'The evolutionary history of lethal metastatic prostate cancer', *Nature*, 520(7547), pp. 353–357.

Gurovich, Y. *et al.* (2019) 'Identifying facial phenotypes of genetic disorders using deep learning', *Nature medicine*, 25(1), pp. 60–64.

Haffner, M.C. *et al.* (2013) 'Tracking the clonal origin of lethal prostate cancer', *The Journal of clinical investigation*, 123(11), pp. 4918–4922.

Haffner, M.C. *et al.* (2021) 'Genomic and phenotypic heterogeneity in prostate cancer', *Nature reviews. Urology*, 18(2), pp. 79–92.

Hamid, A.A. *et al.* (2021) 'Transcriptional profiling of primary prostate tumor in metastatic hormone-sensitive prostate cancer and association with clinical outcomes: correlative analysis of the E3805 CHAARTED trial', *Annals of oncology: official journal of the European Society for Medical Oncology / ESMO*, 32(9), pp. 1157–1166.

Harbaum, L. *et al.* (2021) 'The application of "omics" to pulmonary arterial hypertension', *British journal of pharmacology*, 178(1), pp. 108–120.

Hawley, J.A. *et al.* (2014) 'Integrative biology of exercise', *Cell*, 159(4), pp. 738–749.

Hennig, C. *et al.* (2015) *Handbook of Cluster Analysis*. CRC Press.

Hernandez-Gonzalez, I. *et al.* (2020) 'Clinical heterogeneity of Pulmonary Arterial Hypertension associated with variants in TBX4', *PloS one*, 15(4), p. e0232216.

Hershman, S.G. *et al.* (2019) 'Physical activity, sleep and cardiovascular health data for 50,000 individuals from the MyHeart Counts Study', *Scientific data*, 6(1), p. 24.

Hijazi, H. *et al.* (2021) 'Wearable Devices, Smartphones, and Interpretable Artificial Intelligence in Combating COVID-19', *Sensors* , 21(24). Available at: https://doi.org/10.3390/s21248424.

Hong, M.K.H. *et al.* (2015) 'Tracking the origins and drivers of subclonal metastatic expansion in prostate cancer', *Nature communications*, 6, p. 6605.

Houlihan, C.F. *et al.* (2020) 'Pandemic peak SARS-CoV-2 infection and seroconversion rates in London frontline health-care workers', *The Lancet*, 396(10246), pp. e6–e7.

Howe, K. *et al.* (2013) 'The zebrafish reference genome sequence and its relationship to the human genome', *Nature*, 496(7446), pp. 498–503.

Hurvitz, N. *et al.* (2021) 'Establishing a second-generation artificial intelligence-based system for improving diagnosis, treatment, and monitoring of patients with rare diseases', *European journal of human genetics: EJHG*, 29(10), pp. 1485–1490.

InterAct Consortium *et al.* (2012) 'Validity of a short questionnaire to assess physical activity in 10 European countries', *European journal of epidemiology*, 27(1), pp. 15–25.

Jaksik, R. *et al.* (2015) 'Microarray experiments and factors which affect their reliability', *Biology direct*, 10, p. 46.

Jamal-Hanjani, M. *et al.* (2015) 'Translational implications of tumor heterogeneity', *Clinical cancer research: an official journal of the American Association for Cancer Research*, 21(6), pp. 1258–1266.

Jansen, M.P.H.M. *et al.* (2005) 'Molecular classification of tamoxifen-resistant breast carcinomas by gene expression profiling', *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*, 23(4), pp. 732–740.

Jia, H. *et al.* (2014) 'The latest research progress on spectral clustering', *Neural computing*

& applications, 24(7), pp. 1477–1486.

Jiang, L. *et al.* (2016) 'GiniClust: detecting rare cell types from single-cell gene expression data with Gini index', *Genome biology*, 17(1), p. 144.

Kallenberg, M. *et al.* (2016) 'Unsupervised Deep Learning Applied to Breast Density Segmentation and Mammographic Risk Scoring', *IEEE transactions on medical imaging*, 35(5), pp. 1322–1331.

Kamoun, A. *et al.* (2020) 'A Consensus Molecular Classification of Muscle-invasive Bladder Cancer', *European urology*, 77(4), pp. 420–433.

Kariotis, S. *et al.* (2021) 'Biological heterogeneity in idiopathic pulmonary arterial hypertension identified through unsupervised transcriptomic profiling of whole blood', *Nature communications*, 12(1), p. 7104.

Keenan, B.T. *et al.* (2018) 'Recognizable clinical subtypes of obstructive sleep apnea across international sleep centers: a cluster analysis', *Sleep*, 41(3). Available at: https://doi.org/10.1093/sleep/zsx214.

Kheradpisheh, S.R., Ganjtabesh, M. and Masquelier, T. (2016) 'Bio-inspired unsupervised learning of visual features leads to robust invariant object recognition', *Neurocomputing*, 205, pp. 382–392.

Kieler, H. *et al.* (2012) 'Selective serotonin reuptake inhibitors during pregnancy and risk of persistent pulmonary hypertension in the newborn: population based cohort study from the five Nordic countries', *BMJ* , 344, p. d8012.

Koirala, B. *et al.* (2021) 'Heterogeneity of Cardiovascular Disease Risk Factors Among Asian Immigrants: Insights From the 2010 to 2018 National Health Interview Survey', *Journal of the American Heart Association*, 10(13), p. e020408.

Kozomara, A. and Griffiths-Jones, S. (2014) 'miRBase: annotating high confidence microRNAs using deep sequencing data', *Nucleic acids research*, 42(Database issue), pp. D68–73.

Kreso, A. and Dick, J.E. (2014) 'Evolution of the cancer stem cell model', *Cell stem cell*, 14(3), pp. 275–291.

Kukurba, K.R. and Montgomery, S.B. (2015) 'RNA Sequencing and Analysis', *Cold Spring Harbor protocols*, 2015(11), pp. 951–969.

Lamichhane, S. *et al.* (2018) 'Chapter Fourteen - An Overview of Metabolomics Data Analysis: Current Tools and Future Perspectives', in J. Jaumot, C. Bedia, and R. Tauler (eds) *Comprehensive Analytical Chemistry*. Elsevier, pp. 387–413.

Lamot, L. *et al.* (2015) '[MICROARRAY AND GENE EXPRESSION ANALYSIS]', *Lijecnicki vjesnik*, 137(5-6), pp. 188–195.

Lane, K.B. *et al.* (2000) 'Heterozygous germline mutations in BMPR2, encoding a TGF-β receptor, cause familial primary pulmonary hypertension', *Nature Genetics*, pp. 81–84. Available at: https://doi.org/10.1038/79226.

Lang, I.M. *et al.* (2021) 'Chronic Thromboembolic Disease and Chronic Thromboembolic Pulmonary Hypertension', *Clinics in chest medicine*, 42(1), pp. 81–90.

Lan, N.S.H. *et al.* (2018) 'Pulmonary Arterial Hypertension: Pathophysiology and Treatment', *Diseases (Basel, Switzerland)*, 6(2). Available at: https://doi.org/10.3390/diseases6020038.

Laurinavicius, A. *et al.* (2021) 'Machine-Learning–Based Evaluation of Intratumoral Heterogeneity and Tumor-Stroma Interface for Clinical Guidance', *The American journal of pathology*, 191(10), pp. 1724–1731.

Lee, D., Park, Y. and Kim, S. (2021) 'Towards multi-omics characterization of tumor heterogeneity: a comprehensive review of statistical and machine learning approaches', *Briefings in bioinformatics*, 22(3). Available at: https://doi.org/10.1093/bib/bbaa188.

Lee, Y. *et al.* (2018) 'Gene-gene interaction analysis for quantitative trait using cluster-based multifactor dimensionality reduction method', *International journal of data mining and bioinformatics*, 20(1), pp. 1–11.

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

Li, Y. *et al.* (2017) 'Systematic review regulatory principles of non-coding RNAs in cardiovascular diseases', *Briefings in bioinformatics*, 20(1), pp. 66–76.

Li, Y. *et al.* (2018) 'Mobile Phone Clustering From Speech Recordings Using Deep Representation and Spectral Clustering', *IEEE Transactions on Information Forensics and Security*, 13(4), pp. 965–977.

Li, Y. and Kowdley, K.V. (2012) 'MicroRNAs in common human diseases', *Genomics, proteomics & bioinformatics*, 10(5), pp. 246–253.

Loi, S. *et al.* (2008) 'Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen', *BMC genomics*, 9, p. 239.

López de Maturana, E. *et al.* (2019) 'Challenges in the Integration of Omics and Non-Omics Data', *Genes*, 10(3). Available at: https://doi.org/10.3390/genes10030238.

Lu, C., Meyers, B.C. and Green, P.J. (2007) 'Construction of small RNA cDNA libraries for deep sequencing', *Methods* , 43(2), pp. 110–117.

Manchia, M. *et al.* (2013) 'The impact of phenotypic and genetic heterogeneity on results of genome wide association studies of complex diseases', *PloS one* [Preprint]. Available at: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0076295.

Mao, B. *et al.* (2019) 'Metabolic profiling reveals the heterogeneity of vascular endothelial function phenotypes in individuals at extreme cardiovascular risk', *RSC advances*, 9(52), pp. 30033–30044.

Martin, J.-C. *et al.* (2015) 'Can we trust untargeted metabolomics? Results of the metabo-ring initiative, a large-scale, multi-instrument inter-laboratory study', *Metabolomics:*

Official journal of the Metabolomic Society, 11(4), pp. 807–821.

Marx, V. (2013) 'Biology: The big challenges of big data', *Nature*, 498(7453), pp. 255–260.

McConnell, M.V. *et al.* (2017) 'Feasibility of obtaining measures of lifestyle from a smartphone app: the MyHeart Counts Cardiovascular Health Study', *JAMA cardiology*, 2(1), pp. 67–76.

McGill, H.C., Jr, McMahan, C.A. and Gidding, S.S. (2008) 'Preventing heart disease in the 21st century: implications of the Pathobiological Determinants of Atherosclerosis in Youth (PDAY) study', *Circulation*, 117(9), pp. 1216–1227.

Michor, F. and Polyak, K. (2010) 'The origins and implications of intratumor heterogeneity', *Cancer prevention research* , pp. 1361–1364.

Mishra, P.J. (2014) 'MicroRNAs as promising biomarkers in cancer diagnostics', *Biomarker research*, 2, p. 19.

Mok, A. *et al.* (2019) 'Physical activity trajectories and mortality: population based cohort study', *BMJ* , 365, p. l2323.

Mol, M.B.A. *et al.* (2021) 'Heart-rate-variability (HRV), predicts outcomes in COVID-19', *PloS one*, 16(10), p. e0258841.

Montani, D. *et al.* (2010) 'Long-term response to calcium-channel blockers in non-idiopathic pulmonary arterial hypertension', *European heart journal*, 31(15), pp. 1898–1907.

Montani, D. and Simonneau, G. (2012) 'Updated Clinical Classification of Pulmonary Hypertension', *Pulmonary Vascular Disorders*, pp. 1–13. Available at: https://doi.org/10.1159/000334959.

Montero, P. and Vilar, J.A. (2015) 'TSclust: An R Package for Time Series Clustering', *Journal of statistical software*, 62, pp. 1–43.

Morrell, N.W. *et al.* (2019) 'Genetics and genomics of pulmonary arterial hypertension', *The European respiratory journal: official journal of the European Society for Clinical Respiratory Physiology*, 53(1). Available at: https://doi.org/10.1183/13993003.01899-2018.

Moura *et al.* (2016) 'A predictive model for prognosis in motor neuron disease', *Journal of neurological disorders* [Preprint]. Available at: https://www.researchgate.net/profile/Mirian-Moura/publication/312036236_A_Predictive_Model_for_Prognosis_in_Motor_Neuron_Disease/links/5876556408ae6eb871cf60c2/A-Predictive-Model-for-Prognosis-in-Motor-Neuron-Disease.pdf.

Musunuru, K. and Kathiresan, S. (2019) 'Genetics of Common, Complex Coronary Artery Disease', *Cell*, 177(1), pp. 132–145.

Neufer, P.D. *et al.* (2015) 'Understanding the Cellular and Molecular Mechanisms of Physical Activity-Induced Health Benefits', *Cell metabolism*, 22(1), pp. 4–11.

Nikolaidis, A. *et al.* (2022) 'Heterogeneity in COVID-19 pandemic-induced lifestyle stressors predicts future mental health in adults and children in the US and UK', *Journal of*

psychiatric research, 147, pp. 291–300.

O'Brien, J. *et al.* (2018) 'Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation', *Frontiers in endocrinology*, 9, p. 402.

O'Donnell, M.J. *et al.* (2016) 'Global and regional effects of potentially modifiable risk factors associated with acute stroke in 32 countries (INTERSTROKE): a case-control study', *The Lancet*, 388(10046), pp. 761–775.

O'Regan, A. *et al.* (2021) 'A cluster analysis of device-measured physical activity behaviours and the association with chronic conditions, multi-morbidity and healthcare utilisation in adults aged 45 years and older', *Preventive Medicine Reports*, 24, p. 101641.

Panicker, R.M. and Chandrasekaran, B. (2022) '"Wearables on vogue": a scoping review on wearables on physical activity and sedentary behavior during COVID-19 pandemic', *Sport sciences for health*, pp. 1–17.

Papageorgiou, N. (2016) *Cardiovascular Diseases: Genetic Susceptibility, Environmental Factors and their Interaction*. Academic Press.

Park, S. and Zhao, H. (2018) 'Spectral clustering based on learning similarity matrix', *Bioinformatics* , 34(12), pp. 2069–2076.

Peng, J. *et al.* (2021) 'Machine Learning Techniques for Personalised Medicine Approaches in Immune-Mediated Chronic Inflammatory Diseases: Applications and Challenges', *Frontiers in pharmacology*, 12, p. 720694.

Pezoulas, V.C. *et al.* (2021) 'Machine Learning Approaches on High Throughput NGS Data to Unveil Mechanisms of Function in Biology and Disease', *Cancer genomics & proteomics*, 18(5), pp. 605–626.

Pinto, M.F. *et al.* (2020) 'Prediction of disease progression and outcomes in multiple sclerosis with machine learning', *Scientific reports*, 10(1), p. 21038.

Pita-Juárez, Y. *et al.* (2018) 'The Pathway Coexpression Network: Revealing pathway relationships', *PLoS computational biology*, 14(3), p. e1006042.

Planet, E. *et al.* (2012) 'htSeqTools: high-throughput sequencing quality control, processing and visualization in R', *Bioinformatics* , 28(4), pp. 589–590.

Potla, P., Ali, S.A. and Kapoor, M. (2021) 'A bioinformatics approach to microRNA-sequencing analysis', *Osteoarthritis and Cartilage Open*, 3(1), p. 100131.

Prasad, A. *et al.* (2021) 'Next Generation Sequencing', in V. Singh and A. Kumar (eds) *Advances in Bioinformatics*. Singapore: Springer Singapore, pp. 277–302.

Qiang-long, Z. *et al.* (2014) 'High-throughput Sequencing Technology and Its Application', *The journal of Northeast Agricultural University*, 21(3), pp. 84–96.

Quackenbush, J. (2007) 'Extracting biology from high-dimensional biological data', *The Journal of experimental biology*, 210(Pt 9), pp. 1507–1517.

Qureshi, A. *et al.* (2014) 'VIRmiRNA: a comprehensive resource for experimentally validated

viral miRNAs and their targets', *Database: the journal of biological databases and curation*, 2014. Available at: https://doi.org/10.1093/database/bau103.

Reis-Filho, J.S. *et al.* (2010) 'Molecular profiling: moving away from tumor philately', *Science translational medicine*, 2(47), p. 47ps43.

Režen, T. *et al.* (2022) 'Integration of omics data to generate and analyse COVID-19 specific genome-scale metabolic models', *Computers in biology and medicine*, 145, p. 105428.

Rhodes, C.J. *et al.* (2013) 'Reduced microRNA-150 is associated with poor survival in pulmonary arterial hypertension', *American journal of respiratory and critical care medicine*, 187(3), pp. 294–302.

Rhodes, C.J. *et al.* (2017) 'Plasma Metabolomics Implicates Modified Transfer RNAs and Altered Bioenergetics in the Outcomes of Pulmonary Arterial Hypertension', *Circulation*, 135(5), pp. 460–475.

Rich, J.N. (2016) 'Cancer stem cells: understanding tumor hierarchy and heterogeneity', *Medicine*, 95(1 Suppl 1), pp. S2–S7.

Rivera-Andrade, A. and Luna, M.A. (2014) 'Trends and heterogeneity of cardiovascular disease and risk factors across Latin American and Caribbean countries', *Progress in cardiovascular diseases*, 57(3), pp. 276–285.

Robertson, A.G. *et al.* (2017) 'Comprehensive Molecular Characterization of Muscle-Invasive Bladder Cancer', *Cell*, 171(3), pp. 540–556.e25.

Rodebaugh, T.L. *et al.* (2021) 'Acute Symptoms of Mild to Moderate COVID-19 Are Highly Heterogeneous Across Individuals and Over Time', *Open forum infectious diseases*, 8(3), p. ofab090.

Roessner, U. and Beckles, D.M. (2009) 'Metabolite Measurements', *Plant Metabolic Networks*, pp. 39–69. Available at: https://doi.org/10.1007/978-0-387-78745-9_3.

Rokach, L. and Maimon, O. (2005) 'Clustering Methods', in O. Maimon and L. Rokach (eds) *Data Mining and Knowledge Discovery Handbook*. Boston, MA: Springer US, pp. 321–352.

Rothman, A.M.K. *et al.* (2016) 'MicroRNA-140-5p and SMURF1 regulate pulmonary arterial hypertension', *The Journal of clinical investigation*, 126(7), pp. 2495–2508.

Saeed, M.M., Al Aghbari, Z. and Alsharidah, M. (2020) 'Big data clustering techniques based on Spark: a literature review', *PeerJ. Computer science*, 6, p. e321.

Santos-Ferreira, C.A. *et al.* (2020) 'Micro-RNA Analysis in Pulmonary Arterial Hypertension: Current Knowledge and Challenges', *JACC. Basic to translational science*, 5(11), pp. 1149–1162.

Sarwar, A. and Agu, E. (2021) 'Passive COVID-19 Assessment using Machine Learning on Physiological and Activity Data from Low End Wearables', in *2021 IEEE International Conference on Digital Health (ICDH)*, pp. 80–90.

Saxena, A. *et al.* (2017) 'A review of clustering techniques and developments', *Neurocomputing*, 267, pp. 664–681.

Schachna, L. *et al.* (2003) 'Age and risk of pulmonary arterial hypertension in scleroderma', *Chest*, 124(6), pp. 2098–2104.

Schaefer, J. *et al.* (2020) 'The use of machine learning in rare diseases: a scoping review', *Orphanet journal of rare diseases*, 15(1), p. 145.

Schatz, M.C. (2015) 'Biological data sciences in genome research', *Genome research*, 25(10), pp. 1417–1422.

Schork, N.J. (2019) 'Artificial Intelligence and Personalized Medicine', *Cancer treatment and research*, 178, pp. 265–283.

Scossa, F. *et al.* (2018) 'The Integration of Metabolomics and Next-Generation Sequencing Data to Elucidate the Pathways of Natural Product Metabolism in Medicinal Plants', *Planta medica*, 84(12-13), pp. 855–873.

Sessa, R. and Hata, A. (2013) 'Role of microRNAs in lung development and pulmonary diseases', *Pulmonary circulation*, 3(2), pp. 315–328.

Shahapure, K.R. and Nicholas, C. (2020) 'Cluster Quality Analysis Using Silhouette Score', in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 747–748.

Sharma, D.K. *et al.* (2021) '7 - Early detection and diagnosis using deep learning', in V.E. Balas, B.K. Mishra, and R. Kumar (eds) *Handbook of Deep Learning in Biomedical Engineering*. Academic Press, pp. 191–217.

Shashi, V. *et al.* (2014) 'The utility of the traditional medical genetics diagnostic evaluation in the context of next-generation sequencing for undiagnosed genetic disorders', *Genetics in medicine: official journal of the American College of Medical Genetics*, 16(2), pp. 176–182.

Shcherbina, A. *et al.* (2019) 'The effect of digital physical activity interventions on daily step count: a randomised controlled crossover substudy of the MyHeart Counts Cardiovascular Health Study', *The Lancet Digital Health*, 1(7), pp. e344–e352.

Simonetto, C. *et al.* (2022) 'Heterogeneity in coronary heart disease risk', *Scientific reports*, 12(1), p. 10131.

Simonneau, G. *et al.* (2019) 'Haemodynamic definitions and updated clinical classification of pulmonary hypertension', *European Respiratory Journal*, p. 1801913. Available at: https://doi.org/10.1183/13993003.01913-2018.

Skovbjerg, Honoré and Mechlenburg (2022) 'Monitoring Physical Behavior in Rehabilitation Using a Machine Learning–Based Algorithm for Thigh-Mounted Accelerometers: Development and Validation …', *JMIR Bioinformatics* [Preprint]. Available at: https://bioinform.jmir.org/2022/1/e38512.

Statello, L. *et al.* (2021) 'Gene regulation by long non-coding RNAs and its biological functions', *Nature reviews. Molecular cell biology*, 22(2), pp. 96–118.

Steinegger, M., Mirdita, M. and Söding, J. (2019) 'Protein-level assembly increases protein

sequence recovery from metagenomic samples manyfold', *Nature methods*, 16(7), pp. 603–606.

Sudre, C.H. *et al.* (2021) 'Attributes and predictors of long COVID', *Nature medicine*, 27(4), pp. 626–631.

Sumner, L.W. *et al.* (2007) 'Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI)', *Metabolomics: Official journal of the Metabolomic Society*, 3(3), pp. 211–221.

Sweatt, A.J. *et al.* (2019) 'Discovery of Distinct Immune Phenotypes Using Machine Learning in Pulmonary Arterial Hypertension', *Circulation research*, 124(6), pp. 904–919.

Swietlik, E.M. *et al.* (2021) 'Plasma metabolomics exhibit response to therapy in chronic thromboembolic pulmonary hypertension', *The European respiratory journal: official journal of the European Society for Clinical Respiratory Physiology*, 57(4). Available at: https://doi.org/10.1183/13993003.03201-2020.

Thenappan, T. *et al.* (2018) 'Pulmonary arterial hypertension: pathogenesis and clinical management', *BMJ* , 360, p. j5492.

Thrash, A., Arick, M., 2nd and Peterson, D.G. (2018) 'Quack: A quality assurance tool for high throughput sequence data', *Analytical biochemistry*, 548, pp. 38–43.

Trauer, Dodd and Gomes (no date) 'The importance of heterogeneity to the epidemiology of tuberculosis', *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America* [Preprint]. Available at: https://academic.oup.com/cid/article-abstract/69/1/159/5154892.

Turashvili, G. and Brogi, E. (2017) 'Tumor Heterogeneity in Breast Cancer', *Frontiers of medicine*, 4, p. 227.

Udrescu, L. *et al.* (2016) 'Clustering drug-drug interaction networks with energy model layouts: community analysis and drug repurposing', *Scientific reports*, 6, p. 32745.

Vachiéry, J.-L. and Gaine, S. (2012) 'Challenges in the diagnosis and treatment of pulmonary arterial hypertension', *European respiratory review: an official journal of the European Respiratory Society*, 21(126), pp. 313–320.

Vallée, A. (2022) 'Heterogeneity of the COVID-19 Pandemic in the United States of America: A Geo-Epidemiological Perspective', *Frontiers in public health*, 10, p. 818989.

VanderWaal, K.L. and Ezenwa, V.O. (2016) 'Heterogeneity in pathogen transmission: mechanisms and methodology', *Functional ecology* [Preprint]. Available at: https://agris.fao.org/agris-search/search.do?recordID=US201700208398.

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

Westeneng, H.-J. *et al.* (2018) 'Prognosis for patients with amyotrophic lateral sclerosis: development and validation of a personalised prediction model', *Lancet neurology*, 17(5), pp. 423–433.

Woolhouse, M.E. *et al.* (1997) 'Heterogeneities in the transmission of infectious agents: implications for the design of control programs', *Proceedings of the National Academy of Sciences of the United States of America*, 94(1), pp. 338–342.

Wu, Y., Duan, H. and Du, S. (2015) 'Multiple fuzzy c-means clustering algorithm in medical diagnosis', *Technology and health care: official journal of the European Society for Engineering and Medicine*, 23 Suppl 2, pp. S519–27.

Xiao, C. and Rajewsky, K. (2009) 'MicroRNA control in the immune system: basic principles', *Cell*, 136(1), pp. 26–36.

Xiao, G., Xie, L. and Lian, G. (2018) 'A1511 Identification of CERB as a master transcription factor in monocrotaline–induced pulmonary arterial hypertension', *Journal of hypertension*, 36, p. e15.

Xu, W. *et al.* (2004) 'Increased arginase II and decreased NO synthesis in endothelial cells of patients with pulmonary arterial hypertension', *FASEB journal: official publication of the Federation of American Societies for Experimental Biology*, 18(14), pp. 1746–1748.

Xu, W. and Erzurum, S.C. (2011) 'Endothelial cell energy metabolism, proliferation, and apoptosis in pulmonary hypertension', *Comprehensive Physiology*, 1(1), pp. 357–372.

Xu, W., Janocha, A.J. and Erzurum, S.C. (2021) 'Metabolism in Pulmonary Hypertension', *Annual review of physiology*, 83, pp. 551–576.

Yao, X. *et al.* (2021) 'Molecular Characterization and Elucidation of Pathways to Identify Novel Therapeutic Targets in Pulmonary Arterial Hypertension', *Frontiers in physiology*, 0. Available at: https://doi.org/10.3389/fphys.2021.694702.

Yin, X. *et al.* (2020) 'A Comparative Evaluation of Tools to Predict Metabolite Profiles From Microbiome Sequencing Data', *Frontiers in microbiology*, 11, p. 595910.

Zanwar, S. and Kumar, S. (2021) 'Disease heterogeneity, prognostication and the role of targeted therapy in multiple myeloma', *Leukemia & lymphoma*, 62(13), pp. 3087–3097.

Zapf, M.P. *et al.* (2018) 'Data-Driven, 3-D Classification of Person-Object Relationships and Semantic Context Clustering for Robotics and AI Applications', in *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 180–187.

Zheng, W., Chung, L.M. and Zhao, H. (2011) 'Bias detection and correction in RNA-Sequencing data', *BMC bioinformatics*, 12, p. 290.

Zhou, G., Chen, T. and Raj, J.U. (2015) 'MicroRNAs in pulmonary arterial hypertension', *American journal of respiratory cell and molecular biology*, 52(2), pp. 139–151.

Zhou, Q. and Wang, J. (2016) 'SparkSCAN: A Structure Similarity Clustering Algorithm on Spark', in *Big Data Technology and Applications*. Springer Singapore, pp. 163–177.

Zito Marino, F. *et al.* (2019) 'Molecular heterogeneity in lung cancer: from mechanisms of origin to clinical implications', *International journal of medical sciences*, 16(7), pp. 981–989.

List of publications

Chapter 2

Manuscript 1

Kariotis, S. et al. ***Omada: Robust clustering of transcriptomes through multiple testing*** (https://doi.org/10.1101/2022.12.19.519427)

Repositories:
Bioconductor: https://bioconductor.org/packages/release/bioc/html/omada.html
Github: https://code.bioconductor.org/browse/omada/RELEASE_3_16/

Chapter 3

Manuscript 2

Kariotis, S., Jammeh, E., Swietlik, E.M. et al. ***Biological heterogeneity in idiopathic pulmonary arterial hypertension identified through unsupervised transcriptomic profiling of whole blood***. Nat Commun 12, 7104 (2021). https://doi.org/10.1038/s41467-021-27326-0

Conference abstract 1

Pablo Otero-Núñez, Christopher Rhodes, John Wharton, Emilia Swietlik, **Sokratis Kariotis**, Lars Harbaum, Mark Dunning, Jason Elinoff, Niamh Errington, Roger Thomson, James Iremonger, Gerry Coghlan, Paul Corris, Luke Howard, David Kiely, Colin Church, Joanna Pepke-Zaba, Mark Toshner, Stephen Wort, Ankit Desai, Marc Humbert, William Nichols, Laura Southgate, David-Alexandre Trégouët, Richard Trembath, Inga Prokopenko, Stefan Gräf, Nicholas Morrell, Dennis Wang, Allan Lawrie, Martin Wilkins. ***Multi-omic profiling in pulmonary arterial hypertension*** European Respiratory Journal Sep 2020, 56 (suppl 64) 4458; DOI: 10.1183/13993003.congress-2020.4458 (Presented at ERS International Congress 2020, session “Respiratory viruses in the “pre COVID-19” era”)

Conference abstract 2
Sokratis Kariotis, Emmanuel Jammeh, Allan Lawrie, Dennis Wang**. *Best practices for unsupervised machine learning to stratify patients from high-dimensional molecular profiles*.** (Poster presented at PVRI 2022, 15th Annual World Congress on PVD in Athens, Greece)

Conference abstract 3
Sokratis Kariotis, Niamh Errington, Emmanuel Jammeh, Dennis Wang, Allan Lawrie. ***Time series clustering on RNA-seq data to identify patients progression between iPAH groups*.** (Poster presented at ATS 2022, The American Thoracic Society's International Conference, San Francisco, USA)

Conference abstract 4
Sokratis Kariotis, Allan Lawrie, Dennis Wang**. *Machine learning of whole-blood gene expression to uncover IPAH heterogeneity.*** (Presented at Symposium: Machine Learning for Genomics in Precision Medicine, Lancaster, UK)

Chapter 4

Manuscript 3

Niamh Errington, **Sokratis Kariotis**, Christopher J Rhodes, Emmanuel Jammeh, Yiu-Lian Fong, Zhou Lihan, Cheng He, Timothy Jatkoe, Tatiana Vener, John Wharton, A A Roger Thompson, Robin Condliffe, David G Kiely, Luke Howard, Eileen Harder, Aaron Waxman, Mark Toshner, Dennis Wang*, Allan Lawrie* Martin R Wilkins*, ***Diagnostic miRNA signatures for treatable forms of pulmonary hypertension highlight challenges with clinical classification*** (under submission)

Manuscript 4

Rhodes CJ, Otero-Núñez P, Wharton J, Swietlik EM, **Kariotis S**, Harbaum L, Dunning MJ, Elinoff JM, Errington N, Thompson AAR, Iremonger J, Coghlan JG, Corris PA, Howard LS, Kiely DG, Church C, Pepke-Zaba J, Toshner M, Wort SJ, Desai AA, Humbert M, Nichols WC, Southgate L, Trégouët DA, Trembath RC, Prokopenko I, Gräf S, Morrell NW, Wang D, Lawrie A, Wilkins MR. ***Whole-Blood RNA Profiles Associated with Pulmonary Arterial Hypertension and Clinical Outcome.*** Am J Respir Crit Care Med. 2020 Aug 15;202(4):586-594. doi: 10.1164/rccm.202003-0510OC. PMID: 32352834; PMCID: PMC7427383.

Manuscript 5

Niamh Errington, James Iremonger, Josephine A. Pickworth, **Sokratis Kariotis**, Christopher J. Rhodes, Alexander MK Rothman, Robin Condliffe,Charles A. Elliot, David G. Kiely, Luke S. Howard, John Wharton, A. A. Roger Thompson, Nicholas W Morrell, Martin R. Wilkins, Dennis Wang, Allan Lawrie. ***A diagnostic miRNA signature for pulmonary arterial hypertension using a consensus machine learning approach***. EBioMedicine 69, 103444 (2021).

Chapter 6

Manuscript 6

Varsha Gupta, **Sokratis Kariotis**, Mohammed Rajab, Niamh Errington, Elham M Alhathli, Emmanuel Jammeh, Martin Brook, Naomi Meardon, Paul Collini, Joby Cole, James Wild, Steven Hershman, Roger Thompson, Thushan de Silva, Euan Ashley, Dennis Wang, Allan Lawrie, ***Unsupervised machine learning identifies and associates trajectory patterns of COVID-19 symptoms and physical activity measured via a smart watch*** (under submission)

List of datasets

Single-class simulated dataset

A single-class dataset of 100 samples and 100 genes drawn from a single distribution, generated by Omada's function *feasibilityAnalysis*. Utilised in **Chapter 2**.

Multi-class simulated dataset

A multi-class dataset of 359 samples and 300 genes based on the contents and dimensions of the RNA-seq dataset (see below). Composed of five groups of samples drawn from five different distributions representing the five classes and generated by Omada's function *feasibilityAnalysisDataBased*. Utilised in **Chapter 2**.

Multi-tissue Pan-cancer RNA-seq dataset

An RNAseq expression dataset of 2244 samples and 253 genes representing three types of cancers: breast (n=1084), lung (n=566) and colon/rectal (n=594). The mRNA expression was in the form of z-scores relative to normal samples. Accessible through cBioPortal from the TCGA PanCancer Atlas (details in Chapter 2). Utilised in **Chapter 2**.

Whole blood RNA-seq IPAH/HPAH dataset

An IPAH/HPAH RNAseq dataset drawn from whole blood of 359 patients/samples and 25,955 genes concerning idiopathic and heritable pulmonary arterial hypertension cases. The transcriptomic data can be found in the European Genome-phenome Archive database. The raw sequencing data were processed into the final TPM(transcript per million) format (details in Chapter 3). This dataset's samples partially overlap with the blood serum/plasma microRNA and blood serum metabolite datasets. Utilised in **Chapter 2** and **Chapter 3**.

Whole blood RNA-seq (GUSTO)dataset

An RNA dataset from the whole blood of 238 mothers/samples during midgestation representing 24,070 genes. Read counts were extracted from GEO with further preprocessing (details in Chapter 2). Utilised in **Chapter 2**.

Blood serum/plasma microRNA dataset

A dataset of 1138 patients each belonging to one of the clinical Pulmonary Hypertension classes and a total of 554 microRNAs from blood serum or plasma (details in Chapter 4). This dataset's samples partially overlap with the whole blood RNA-seq IPAH/HPAH and blood serum metabolite datasets. Utilised in **Chapter 4**.

Blood serum metabolite dataset

A metabolite dataset of 1072 samples and 1522 metabolites drawn from blood serum generated using liquid chromatography mass spectrometry (LC-MS). It contains five classes of Pulmonary Hypertension (PH) patients (details in Chapter 5). This dataset's samples partially overlap with the whole blood RNA-seq IPAH/HPAH and blood serum/plasma microRNA datasets. Utilised in **Chapter 5**.

COVID-19 activity dataset

A longitudinal dataset of 34 patients and 8 activity measures with a maximum of 596 timepoints. The patients all contracted COVID-19 and their physical activity measures were captured by smart watches. Utilised in **Chapter 6**.