A Systems Medicine Approach for Heart Valve Diseases

Addressing the Proteomic Landscape and Differential Expression Software

Dissertation

submitted to the
Digital Engineering Faculty of the University of Potsdam
for the attainment of the degree
Doktor der Naturwissenschaften (Dr.rer.nat.)
in the scientific discipline Digital Health – Personalized Medicine.

by: Dipl.-Ing. Sara Milena Kraus

Published online on the
Publication Server of the University of Potsdam:
https://doi.org/10.25932/publishup-52226
https://nbn-resolving.org/urn:nbn:de:kobv:517-opus4-522266

Summary

In Systems Medicine, in addition to high-throughput molecular data (*omics), the wealth of clinical characterization plays a major role in the overall understanding of a disease. Unique problems and challenges arise from the heterogeneity of data and require new solutions to software and analysis methods. The SMART and EurValve studies establish a Systems Medicine approach to valvular heart disease – the primary cause of subsequent heart failure.

With the aim to ascertain a holistic understanding, different *omics as well as the clinical picture of patients with aortic stenosis (AS) and mitral regurgitation (MR) are collected. Our task within the SMART consortium was to develop an IT platform for Systems Medicine as a basis for data storage, processing, and analysis as a prerequisite for collaborative research. Based on this platform, this thesis deals on the one hand with the transfer of the used Systems Biology methods to their use in the Systems Medicine context and on the other hand with the clinical and biomolecular differences of the two heart valve diseases. To advance differential expression/abundance (DE/DA) analysis software for use in Systems Medicine, we state 21 general software requirements and features of automated DE/DA software, including a novel concept for the simple formulation of experimental designs that can represent complex hypotheses, such as comparison of multiple experimental groups, and demonstrate our handling of the wealth of clinical data in two research applications - *DEAME* and *Eatomics*. In user interviews, we show that novice users are empowered to formulate and test their multiple DE hypotheses based on clinical phenotype. Furthermore, we describe insights into users' general impression and expectation of the software's performance and show their intention to continue using the software for their work in the future. Both research applications cover most of the features of existing tools or even extend them, especially with respect to complex experimental designs. *Eatomics* is freely available to the research community as a user-friendly R Shiny application.

Eatomics continued to help drive the collaborative analysis and interpretation of the proteomic profile of 75 human left myocardial tissue samples from the SMART and EurValve studies. Here, we investigate molecular changes within the two most common types of valvular heart disease: aortic valve stenosis (AS) and mitral valve regurgitation (MR). Through DE/DA analyses, we explore shared and disease-specific protein alterations, particularly signatures that could only be found in the sex-stratified analysis. In addition, we relate changes in the myocardial proteome to parameters from clinical imaging. We find comparable cardiac hypertrophy but differences in ventricular size, the extent of fibrosis, and cardiac function. We find that AS and MR show many shared remodeling effects, the most prominent of which is an increase in the extracellular matrix and a decrease in metabolism. Both effects are stronger in AS. In muscle and cytoskeletal adaptations, we see a greater increase in mechanotransduction in AS and an increase in cortical cytoskeleton in MR. The decrease in proteostasis proteins is mainly attributable to the signature of female patients with AS. We also find relevant therapeutic targets.

In addition to the new findings, our work confirms several concepts from animal and heart failure studies by providing the largest collection of human tissue from *in vivo* collected biopsies to date. Our dataset contributing a resource for isoform-specific protein expression in two of the most common valvular heart diseases. Apart from the general proteomic landscape, we demonstrate the added value of the dataset by showing proteomic and transcriptomic evidence for increased expression of the SARS-CoV-2- receptor at pressure load but not at volume load in the left ventricle and also provide the basis of a newly developed metabolic model of the heart.

Zusammenfassung

In der Systemmedizin spielt zusätzlich zu den molekularen Hochdurchsatzdaten (*omics) die Fülle an klinischer Charakterisierung eine große Rolle im Gesamtverständnis einer Krankheit. Hieraus ergeben sich Probleme und Herausforderungen unter anderem in Bezug auf Softwarelösungen und Analysemethoden. Die SMART- und EurValve-Studien etablieren einen systemmedizinischen Ansatz für Herzklappenerkrankungen – die Hauptursache für eine spätere Herzinsuffizienz.

Mit dem Ziel ein ganzheitliches Verständnis zu etablieren, werden verschiedene *omics sowie das klinische Bild von Patienten mit Aortenstenosen (AS) und Mitralklappeninsuffizienz (MR) erhoben. Unsere Aufgabe innerhalb des SMART Konsortiums bestand in der Entwicklung einer IT-Plattform für Systemmedizin als Grundlage für die Speicherung, Verarbeitung und Analyse von Daten als Voraussetzung für gemeinsame Forschung. Ausgehend von dieser Plattform beschäftigt sich diese Arbeit einerseits mit dem Transfer der genutzten systembiologischen Methoden hin zu einer Nutzung im systemmedizinischen Kontext und andererseits mit den klinischen und biomolekularen Unterschieden der beiden Herzklappenerkrankungen. Um die Analysesoftware für differenzielle Expression/Abundanz, eine häufig genutzte Methode der System Biologie, für die Nutzung in der Systemmedizin voranzutreiben, erarbeiten wir 21 allgemeine Softwareanforderungen und Funktionen einer automatisierten DE/DA Software. Darunter ist ein neuartiges Konzept für die einfache Formulierung experimenteller Designs, die auch komplexe Hypothesen wie den Vergleich mehrerer experimenteller Gruppen abbilden können und demonstrieren unseren Umgang mit der Fülle klinischer Daten in zwei Forschungsanwendungen – *DEAME* und *Eatomics*. In Nutzertests zeigen wir, dass Nutzer befähigt werden, ihre vielfältigen Hypothesen zur differenziellen Expression basierend auf dem klinischen Phänotyp zu formulieren und zu testen, auch ohne einen dedizierten Hintergrund in Bioinformatik. Darüber hinaus beschreiben wir Einblicke in den allgemeinen Eindruck der Nutzer, ihrer Erwartung an die Leistung der Software und zeigen ihre Absicht, die Software auch in der Zukunft für ihre Arbeit zu nutzen. Beide Forschungsanwendungen decken die meisten Funktionen bestehender Tools ab oder erweitern sie sogar, insbesondere im Hinblick auf komplexe experimentelle Designs. Eatomics steht der Forschungsgemeinschaft als benutzerfreundliche R Shiny-Anwendung frei zur Verfügung.

Eatomics hat weiterhin dazu beigetragen, die gemeinsame Analyse und Interpretation des Proteomprofils von 75 menschlichen linken Myokardgewebeproben aus den SMART- und EurValve-Studien voran zu treiben. Hier untersuchen wir die molekularen Veränderungen innerhalb der beiden häufigsten Arten von Herzklappenerkrankungen: AS und MR. Durch DE/DA Analysen erarbeiten wir gemeinsame und krankheitsspezifische Proteinveränderungen, insbesondere Signaturen, die nur in einer geschlechtsstratifizierten Analyse gefunden werden konnten. Darüber hinaus beziehen wir Veränderungen des Myokardproteoms auf Parameter aus der klinischen Bildgebung. Wir finden eine vergleichbare kardiale Hypertrophie, aber Unterschiede in der Ventrikelgröße, dem Ausmaß der Fibrose und der kardialen Funktion. Wir stellen fest, dass AS und MR viele gemeinsame Remodelling-Effekte zeigen, von denen die wichtigsten die Zunahme der extrazellulären Matrix und eine Abnahme des Metabolismus sind. Beide Effekte sind bei AS stärker. Zusätzlich zeigt sich eine größere Variabilität zwischen den einzelnen Patienten mit AS. Bei Muskel- und Zytoskelettanpassungen sehen wir einen stärkeren Anstieg der Mechanotransduktion bei AS und einen Anstieg des kortikalen Zytoskeletts bei MR. Die Abnahme von Proteinen der Proteostase ist vor allem der Signatur von weiblichen Patienten mit AS zuzuschreiben. Außerdem finden wir therapierelevante Proteinveränderungen.

Zusätzlich zu den neuen Erkenntnissen bestätigt unsere Arbeit mehrere Konzepte aus Tierstudien und Studien zu Herzversagen durch die bislang größte Kollektion von humanem Gewebe aus in vivo Biopsien. Mit unserem Datensatz stellen wir eine Ressource für die isoformspezifische Proteinexpression bei zwei der häufigsten Herzklappenerkrankungen zur Verfügung. Abgesehen von der allgemeinen Proteomlandschaft zeigen wir den Mehrwert des Datensatzes, indem wir proteomische und transkriptomische Beweise für eine erhöhte Expression des SARS-CoV-2- Rezeptors bei Drucklast, jedoch nicht bei Volumenlast im linken Ventrikel aufzeigen und außerdem die Grundlage eines neu entwickelten metabolischen Modells des Herzens liefern.

Acknowledgements

First and foremost, I want to thank Prof. Hasso Plattner and Prof. Erwin Boettinger to create the opportunity to work and do research in the incredibly fruitful environment of HPI and DHC – especially Prof. Boettinger, who sharpened the focus of my thesis and provided valuable feedback. I want to thank Dr. Matthieu-P. Schapranow for putting faith in my capabilities coming from a different profession, for guiding me throughout the thesis and for fostering my independence.

Collaborators I very much appreciated to work with Sarah Nordmeyer, who became an idol in her work ethic, in being assertive without being intrusive, in giving constructive feedback and in just being a fun person to talk to. Prof. Titus Kuehne, Prof. Vera Regitz-Zagrosek, Prof. Philipp Mertins, Dr. Matthias Ziehm and Dr. Marieluise Kirchner were always available for fruitful discussions on our joint research, provided guidance and let me evolve into a confident, valued member of the team.

Frollueages – I am so thankful to be a part of such a great team of both **fr**iends and c**olleagues** (Harry Freitas da Cruz et al. 2019). Especially Cindy Perscheid, who was the first to guide me around and with whom I shared so many blockades, both work-associated and private. Who I could always approach, who had an active part in the shape of this thesis and my personal growth. I want to thank Tamara Slosarek for her great effort as student in my courses and her Master's thesis, for her work as a HiWi, for her perfectionism in reviewing my text artefacts and in welcoming me in the new neighborhood. I want to thank Claudia Schurmann for showing up at DHC just about right to guide me through final analysis efforts, for spending hours of one-on-one sessions in critically, but formost constructively reviewing my work, paper drafts and helping me with the publishing process.

Special appreciation is given to Susanne, Ariane, Harry, JP, Ralf and Lin for reading parts of the thesis and providing valued feedback and helpful suggestions and Anja for her input on user testing sessions. All current and former frolleagues from the EPIC and Personalized Medicine chair for our constructive working atmosphere, fun lunch breaks, great retreats, private and business trips, evening events, celebrations and Zoom meetings throughout the pandemic.

Family & Friends Throughout the years, the Victoria Lacrosse club was a place not only to stay active and achieve athletic success, but to feel true team spirit and company, to find long-lasting friendships, especially Marleen and Jan, and having a lot of fun at celebrations and tournament trips. My former biotech fellow students, Marie, Pina, Renée, Sophie, Janine and Jessi have provided me with solid grounds for the past 10 years and will last for the years to come. Last, but not least, I want to thank my family, Ingrid, Marie, Alex, Paula, Benni, Nils, Leonie, Luca and Dirk, for their support, cheers and their

provision of family retreat. Finally, I want to thank Malou for her refreshing views, ideas and her sunny temper and Christian for his endless patience and resilience to frustration outbursts, for staying with me throughout the rocky journey. Both of them for being my closest friends, lockdown companions, and for our calming, loving family hours.

Contents

List of Figures vii

List of Tables x

1 Introduction 1

1.1 Research Questions and Objectives 6

1.2 Contributions 8

1.3 Thesis Outline 10

2 Background and Preliminaries 12

2.1 Data and queries in Systems Medicine settings 12

2.1.1 Data in Systems Medicine 12

2.1.2 Differential expression/abundance and enrichment analysis 15

2.2 SMART - A Systems Medicine consortium 21

2.2.1 Challenges consortium work 23

2.2.2 Implementation of the SMART IT platform 23

2.2.3 Summary on SMART IT platform 28

2.3 The human heart and heart valve disease 28

2.3.1 Cellular and molecular setup of myocardial tissue 29

2.3.2 Left ventricular heart valve disease 31

3 DE/DA Analysis Software in Systems Medicine 34

3.1 Motivation 34

3.2 Related Work 35

3.2.1 DE tools for transcriptomic HTQ data 35

3.2.2 DA tools for proteomic HTQ data 36

3.3 Methods 37

3.3.1 Requirements engineering and feature definition 37

3.3.2 Personas 38

3.3.3 User studies 39

3.4 Requirements and Features 41

3.4.1 Requirement space: Pre-processing 44

3.4.2 Requirement space: DE/DA in Systems Medicine 45

3.4.3 Requirement space: Annotation and interpretation 48

3.5 Implementation 51

3.5.1 DEAME 51

3.5.2 Eatomics 56

3.6 Evaluation 65

3.6.1 Comparison of functionality 65

3.6.2 Examples of utility 72

3.6.3 User studies 74

3.7 Discussion 77

4 Molecular and clinical characterization of AS and MR 83

4.1 Motivation 83

4.2 Related Work 85

4.3 Methods and Study Setup 86

4.3.1 Biopsy acquisition and distribution 86

4.3.2 Assessment of the clinicome 88

4.3.3 Transcriptome analysis 89

4.3.4 Proteome analysis 91

4.3.5 Visualization 95

4.4 Results 96

4.4.1 Clinicome - Study cohort 96

4.4.2 Overview on the myocardial transcriptome 97

4.4.3 Overview on the myocardial proteome 98

4.4.4 Disease- and sex-specific effects in AS and MR 101

4.5 Discussion 121

5 Conclusion and Outlook 128

Author Contributions and Credits 132

List of Acronyms 134

Bibliography 137

A Supporting information on DE/DA analysis software 152

A.1 Clinically motivated questions towards assessed molecular data (German) . 152

A.2 Model of the preprocessing requirements space 153

A.3 Eatomics: Installation and dependencies 153

A.4 Examples of input files needed for Eatomics. 154

A.5 User Interview Testing Artefacts 155

B Quality control reports and and additional analyses 168

C Exploratory comparison of clustering strategies on a Systems Medicine data set 177

D Data tables 184

List of Figures

1.1 Publications in Systems Medicine. 1

1.2 Setting in Systems Biology and Systems Medicine. 2

1.3 Effects of hemodynamic load in heart valve disease. 5

1.4 Summary of research artefacts. 9

2.1 Process steps of differential expression analysis. 16

2.2 Visualization of a GLM of a simple control vs. treated experiment. 19

2.3 Process of data assessment, analysis, and communication within SMART. . 22

2.4 Software architecture of the SMART platform. 24

2.5 Entity relationship diagram depicting SMART data representation. 25

2.6 RNA sequencing pre-processing pipeline. 26

2.7 User interface components of the SMART platform. 28

2.8 Schematic drawing of the human heart. 29

2.9 Details of cardiac tissue composition. 30

2.10 Normal and disturbed hemodynamic flow in the left human heart. 31

2.11 Effects of mechanical stress caused by heart valve diseases. 32

3.1 Differential expression/abundance in Systems Medicine feature space. . . . 46

3.2 Schematic representation of model parameter selection and modification. . . 49

3.3 Schematic representation of design formula creation 50

3.4 DEAME's software system architecture. 52

3.5 Implemented RNAseq pre-processing pipeline. 53

3.6 Overview of DEAME's UI. 54

3.7 Screenshot of DEAME's experimental design panel. 55

3.8 General setup of Eatomics' user interface (UI). 58

3.9 Load and Prepare tab panel overview. 60

3.10 Process and message flow in Eatomics' first tab panel. 61

3.11 Differential Abundance tab panel. 62

3.12 Results from User Interviews. 75

4.1 Intersection of SMART generated data artefacts. 84

4.2 Study design to derive a proteomic landscape. 87

4.3 Illustration of result subsets for effects in differential abundance results. . . 94

4.4 Cardiac transcriptome summary. 99

4.5 Cardiac proteome coverage and distribution. 100

4.6 Quantitative analyses of disease- and sex-specific differences in protein abundance. 102

4.7 Quantitative analyses of sex-specific differences in protein abundance. . . . 103
4.8 Disease and sex-specific differences in abundance of proteins related to extracellular matrix composition. 104
4.9 Fibrous tissue content and ECM protein intensity. 106
4.10 Overview on metabolic protein expression and GO enrichment results. . . . 107
4.11 Mitochondrial protein abundance. 108
4.12 Summary of metabolic processes in AS and MR. 109
4.13 Comparison of ejection fraction in AS and MR 110
4.14 Overview on proteostasis protein expression. 111
4.15 Combined results of GO term enrichment analysis for proteostasis. 112
4.16 Sex-stratified analysis of ribosomal subunits and translation IF in AS. . . . 113
4.17 Sex-stratified comparison of myocardial mass and pressure gradient in AS. . 114
4.18 Functional and organellar assignments of preoteins in the muscle and cytoskeleton category. 115
4.19 Protein abundance in calcium handling proteins. 116
4.20 Specific increase in desmosomal proteins. 117
4.21 AS-specific changes in in the cytoskeleton and sarcomere. 117
4.22 Ratio of myosin heavy chains in AS, MR and CON. 118
4.23 Cortical cytoskeleton in MR. 119
4.24 ACE2 abundance in heart valve disease. 120

A.1 SCRM diagram of the preprocessing pipeline requirements space. 153
A.2 Examples of input files needed for Eatomics. 154

B.1 Overview on clinical parameters from magnetic resonance imaging stratified to conditions. 168
B.2 Overview on clinical parameters from magnetic resonance imaging stratified to sex in AS and MR. 169
B.3 Age in conditions stratified to sex. 170
B.4 ummary reports created by MultiQC for all 33 paired end RNA sequencing FASTQ files after trimming. 171
B.5 Examination of proteins belonging to GO terms which were assigned to the other terms category. 172
B.6 Fold changes of proteasomal subunits in condition vs. control. 173
B.7 LFQ intensity as measured in the reference sample. 174
B.8 Down-sampled analyses to exclude bias of having more samples in AS. . . . 175
B.9 Summary of sample exclusion because of duplicate measurements and strong blood contamination. 176

C.1 Experimental setup for multi-omics experiments. 178
C.2 SNF multi-omics heatmap. 181

C.3 iCB multi-omics heatmap. 182

List of Tables

3.1 DE/DA application features and requirements. 42
3.1 DE/DA application features and requirements. 43
3.2 Parameter type specification. 47
3.3 Differences of DEAME and Eatomics 51
3.4 Comparison of RNAseq analysis applications. 67
3.4 Comparison of RNAseq analysis applications. 68
3.4 Comparison of RNAseq analysis applications. 69
3.5 Comparison of MaxQuant output analysis applications. 71
3.6 Comparison of the design feature details. 72

4.1 Overview on the definition of result subsets from the comparison of conditions. 93
4.2 Clinical cohort description. 96

C.1 Quantitative metrics for multi-omics experiments. 180

D.1 Assignments of Gene Ontology (GO) terms to five categories. 184

1 Introduction

Systems Medicine's main objective is to integrate all relevant biological and medical data levels to broaden our understanding of pathophysiological mechanisms, prognosis, diagnosis, and treatment of diseases [1]. The branch of research is similar to Systems Biology and both are fueled by the emergence of high-throughput data acquisition techniques. Because of the similarities in objectives, another common short definition is “Systems Medicine is the implementation of Systems Biology approaches into medical research.” [2]. Systems Biology gained momentum approximately 10 years before Systems Medicine did, as can be inferred from the number of publications on the topics as indexed in PubMed (Figure 1.1). It is assumed that the advancements and solutions for data acquisition and computational capabilities from Systems Biology can be transferred, adapted, or extended to the challenges unique to medical research.

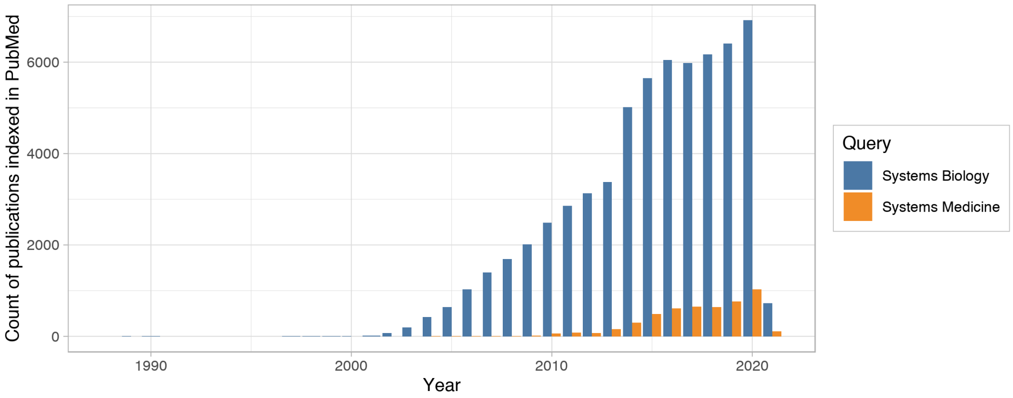


Figure 1.1: Count of publications indexed in PubMed from 1988 to today. Queries to PubMed used to retrieve result counts are "systems medicine"[All Fields] and for reference the related term "systems biology"[All Fields] on Jan 27, 2021 at `https://pubmed.ncbi.nlm.nih.gov/`.

To do so, we need to understand the difference and similarities in both research settings, as depicted in Figure 1.2. In biomolecular research, a molecular biologist sets up *in vitro* experiments, which serve as the basis for high-throughput measurements. For example, cell or bacterial cultures are treated or serve as control while keeping external conditions identical (control vs. treated experiment – two-group comparison). As such, the perturbation to the system under investigation is well-controllable. The metadata to describe the experiment is limited and structured. The effort is theoretically only limited by human and financial resources and an experiment may be repeated several times to assess sufficient evidence. In contrast, in medical research, the clinical scientist plays an additional and crucial role and the investigation shifts focus towards human tissue or fluids. The specimens are taken after undergoing various *in vivo* perturbations, e.g., high blood pressure. As a result, molecular profiles are less controllable. The influences are captured in the clinical information and examination results, which by nature are diverse in the assessed parameters and in the data type they are stored in, i.e., the digital version. In any case, there are categorical and numerical parameters that can be grouped into subcategories, or the data consists of fully unstructured clinical notes. The heterogeneity poses challenges on data storage, handling, access and also needs to be considered in analyses. Although the term is not common in the community, we refer to the clinical data as the "clinicome". The setup in medical research is limited to the accessibility of human specimen.

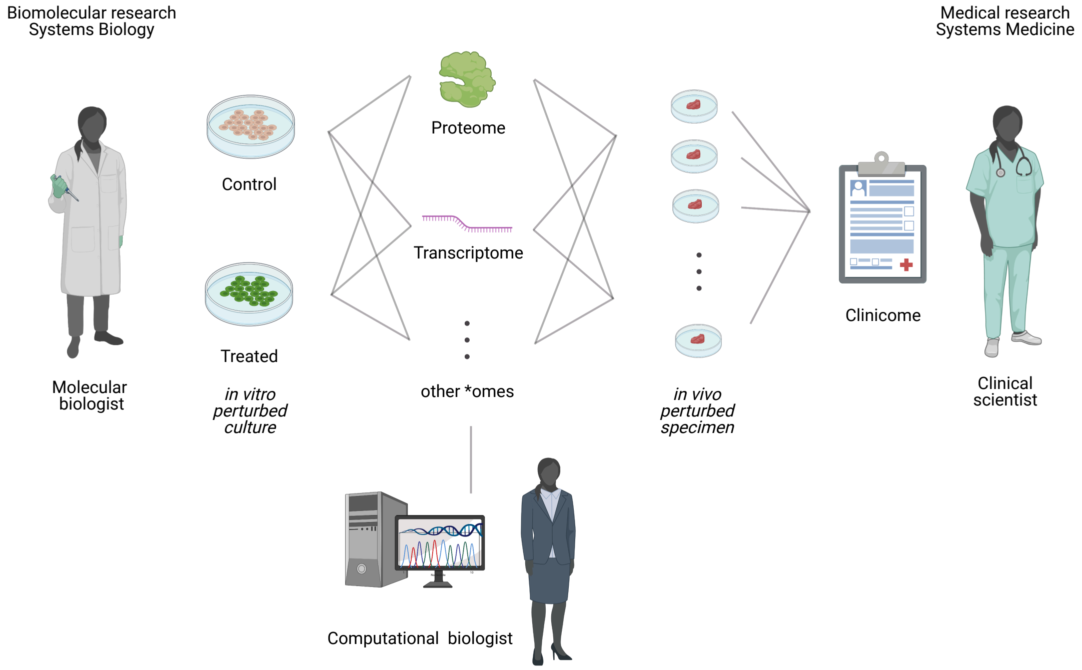


Figure 1.2: Overview in the setting in Systems Biology and Systems Medicine. Both research areas are fueled by high-throughput data (*omes). In biomolecular research, the data is derived from *in vitro* perturbed cultured cells or bacteria and is performed by the molecular biologist in a controlled environment. A dedicated computational biologist or easy-to-use software solutions are needed for data analysis. In medical research, the focus shifts towards human specimens, perturbed by the human system's various influences before biopsy extraction. Influences are captured in the clinical examination data (= clinicome) assessed by the clinical scientist.

In both settings, high-throughput data is assessed, which primarily refers to the molecular *omes (greek for totality), i.e., data artefacts that describe the totality of all molecular features by reporting their sequence/chemical setup or quantity for one or more biologic samples. Examples are the exome or the whole genome, which describe the qualitative variation in a specific subject's genetic code, or the transcriptome, i.e., the quantitative profile of gene transcripts that define the cells' major functions. Both are valuable molecular information resources that next-generation sequencing (NGS) techniques have enabled.

Similarly, the proteome, i.e., the abundance of proteins in a specific cell or tissue, is becoming increasingly available for broad assessment in modern mass spectrometry setups [3]. For the remainder of the thesis, we refer to the quantitative measurements as high-throughput quantification (HTQ) data, i.e., data that describe the quantity for every feature of a molecular level for one or more biologic samples.

Through increasing availability and decreasing costs, HTQ has become a cornerstone in modern life science research groups [4]. However, the large amounts of molecular data require computational power and expertise in analysis. For example, raw high-throughput measurements result in several Gigabytes of data per sample and undergo intensive pre-processing, e.g., analysis of spectra from mass spectrometry experiments [5] for proteomics or mapping of RNA sequencing reads in transcriptomics [6]. Interpretation can be performed within one level of molecular data, e.g., in differential expression analysis, or by spanning multiple omes, e.g., in network analysis or multi-omics subtype detection. Furthermore, external knowledge bases can be exploited to support the analysis process or help in interpretation, e.g., through gene set/pathway enrichment analysis. To fully exploit the wealth of molecular data, a lab relies on a dedicated computational biologist or analysis tools easy enough to be operable by the molecular biologist [4].

A common approach to derive insights from HTQ data in Systems Biology, and thus a starting point for method transfer is differential expression (DE) analysis, also called differential abundance (DA) analysis in the proteomics field. Inherent to DE/DA analysis are at least two groups of cell- or tissue samples assumed to show differences in gene/protein quantity. In the classical sense, the differences result from a single controlled *in vitro* perturbation of one group of samples and thus the system under investigation. The scientific question can be stated as how to find genes/proteins that are differentially expressed/abundant between the groups under investigation. Initially, large differences in gene expression measured with microarrays were detected visually. However, when the technology became more mature and scaled out from less than 50 quantified genes to many thousands, accurate statistical methods to define difference needed to be developed [7] and expanded to the various singularities of HTQ data. Many state-of-the-art solutions adopted generalized linear models (GLM), which provide great flexibility in modeling the measurement data's characteristics [8, 9]. Additionally, a large collection of work has been invested in optimizing data transformation, variance-mean dependency correction, information sharing across genes, shrinkage, and many more aspects. In biomolecular research, GLMs are used to model the two-group comparison but are well suited to model more complex experimental setups.

In medical research, the clinicome describes the causes, influences, and consequences of the disease. As a primary stakeholder and expert on the clinical data, the clinical scientist has the deepest knowledge and understanding of patient and disease. From the clinical scientist's perspective, there are many ideas and hypotheses on how parameters from the clinical representation may influence the affected tissue's molecular profile. As a result, we need to consider the clinicome for DE/DA model setup. Many of the clinicians' questions can be rephrased as "What are the differences in group A vs. group B?" or "Are there any relationships between protein expression and a continuous clinical parameter?". Theoretically, it is possible to answer most questions using DE/DA analysis. However, it usually requires a computational biologist's work to translate the hypothesis into a valid design formula, to set up and execute the calculations, and to refine the results into a condensed overview for the clinical scientist. In contrast to the two-group comparison in the Systems Biology setup, in Systems Medicine, the proper definition of the design formula is aggravated by the necessity to include and thus model all reasonable confounding clinical parameters. As a result, the DE/DA analysis models may likely become very complex, including the influence of interest and several confounding effects or interactions.

While the aim of method transfer usually refers to statistical methods and machine learning [10, 11], it is also true for the development of computational platforms and appropriate medical information technology [12]. Systems Medicine depends on powerful information technology (IT) platforms that can integrate, process, and analyze the multiple, heterogeneous biomedical datasets and simultaneously support the workflow of research consortia and the needs of interdisciplinary teams [13].

In Europe, the Systems Medicine research area profited from CASyM (Coordinating Systems Medicine across Europe) [14] and EASyM (European Association of Systems Medicine) [15]. . The two initiatives developed a roadmap to pave the way to distribute more than 24 million euros to promote Systems Medicine approaches and demonstrate their utility. Fueled by these efforts and other international initiatives [16], newly founded research institutes [17], dedicated conferences [18, 19], a journal [20], and training courses [21, 22] emerged soon after.

Recently, successful examples of Systems approaches in cardiovascular conditions are emerging [23]. For example, Schlotter et al. (2018) developed a spatio-temporal molecular atlas of the human aortic valve. The integration of post-operative molecular imaging and pathology with proteomics, transcriptomics, and network analysis reveals disease networks driving calcific aortic valve disease. As there is no treatment available, the authors argue that the similarity to inflammatory diseases may be exploited to search for a potential treatment [24].

Due to the complexity of heart failure (HF) and cardiovascular diseases in general with numerous risk factors, e.g., genetic predisposition or physical inactivity, the disease promises a strong susceptibility to a patient-specific, Systems Medicine treatment approach [25–27].

The Systems Medicine Approach for Heart Failure (SMART) study is a demonstrator within the German e:Med initiative to explore and analyze the complex regulatory network that triggers the course and onset of HF. Similarly, the EurValve study aimed to model multiple clinical and molecular data modalities to find the best time point for surgical intervention in heart valve diseases [28]. The common ground for both SMART and EurValve is their focus on heart valve diseases as the leading cause of HF. The left ventricle's proper function relies on mitral and aortic valves to close and open appropriately. Valve stenosis denotes the valve's improper opening to ensure blood flow, while valve regurgitation denotes improper closing and leakage of the valve. The disturbance of hemodynamic flow in aortic valve stenosis (AS) and mitral valve regurgitation (MR) causes chronic cardiac pressure or volume overload, which triggers distinctive forms of cardiac remodeling. One very prominent adaptation mechanism is left ventricular hypertrophy (Figure 1.3), i.e., an increase in myocardial mass, typically concentric in pressure and eccentric in volume overload [29, 30]. In aging populations, the incidence is increasing drastically and is becoming a serious health burden. AS and MR are the most frequent types of valve diseases and have reached an incidence of more than 12% of the population older than 65 years for AS and 9% for MR [31, 32]. In an adapted compensated state, patients can remain asymptomatic for years; however, once there is a transition into HF and patients become symptomatic, the prognosis is poor in both patient groups if they remain untreated [33]. Furthermore, significant sex differences have been reported regarding left ventricular hypertrophy, heart failure progression, and valvular heart disease in general [34, 35]. Most knowledge about cardiac adaptation mechanisms in valve disease is currently available at the organ scale where clinical parameters like ventricular function or myocardial mass and fibrosis can be investigated with non-invasive imaging methods [36]. Much less is known about human cardiac adaptation mechanisms at the cellular or protein expression level.

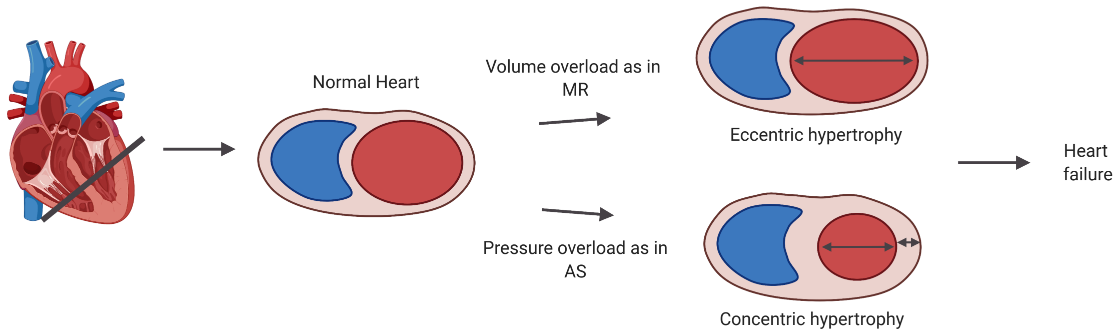


Figure 1.3: The disturbance of hemodynamic flow in AS and MR causes chronic cardiac pressure or volume overload, which triggers cardiac hypertrophy, i.e., an increase in myocardial mass, especially in the left ventricle (red). In pressure overload, concentric remodeling is characterized by a thickening of the left ventricle wall and a decrease in inner diameter. In eccentric remodeling, as in volume overload, the left ventricle's inner diameter widens. MR - mitral valve regurgitation, AS - aortic valve stenosis.

SMART and EurValve represent common characteristics and challenges in (cardiovascular) Systems Medicine research as depicted by Gietzelt et al. (2016) [10] and summarized by Kramer et al. (2018) [26]:

1. Both studies provide valuable human specimens from the left ventricle (LV). Therefore, they are eligible to support the translation of findings from animal models to humans.

2. In order to characterize heart valve diseases, both studies exploit the new possibility of feasibly measuring multiple *omes in the medical research setting. Molecular

levels include the genome derived by whole genome sequencing, the transcriptome derived by RNA sequencing (RNAseq), the proteome derived by shot-gun, label-free mass spectrometry measurements, and the clinicome in an observational fashion. In total, 124 individuals are characterized in at least one *ome (Figure 1.4). Of these, 17 individuals constitute a healthy control cohort; 60 subjects were diagnosed with aortic stenosis and 47 with mitral valve regurgitation.

3. Collaborators and stakeholders in a Systems Medicine consortium form an interdisciplinary team of researchers involving clinicians and clinical scientists, mathematical modelers, computational and molecular biologists as well as software engineers. The diversity in expertise shall provide a comprehensive picture to gather all necessary information for clinical decision making, similar to a tumor board. Such a setup requires strategic efforts and a common ground of basic understanding of each other's discipline and formation of shared research hypotheses.

4. Both studies are dependent on new, powerful computational platforms for data storage, data handling, flow of information, and methodologies for statistical analyses and integration of heterogeneous data sets.

Both studies are summarized in Figure 1.4. Throughout this thesis, they serve as representative examples of Systems Medicine approaches.

In summary, the wealth of data in Systems Medicine is large and is characterized by omics and clinical data. SMART and EurValve are Systems Medicine approaches committed to exploring heart valve diseases. The transfer of methods and approaches from Systems Biology is a dedicated aim of Systems Medicine and needs careful consideration of the unique setup in medical research. Because of the field's recency, hurdles in pursuing a Systems Medicine approach and thereby deriving meaningful biomedical results are high. Additionally, a detailed picture of the molecular setup in healthy human hearts and the difference in cardiac tissue of hearts under pressure and volume load would help translate findings from animal models and broaden our general understanding and establish relationships between the clinical phenotype and the molecular setup. Therefore, our central aim is to **progress towards Systems Medicine by utilizing and advancing Systems Biology approaches in general and on the specific use case of heart valve diseases**. As such, the thesis is divided into a first part that considers the transfer of Systems Biology software to enable Systems Medicine approaches and a second part that disseminates characteristics in heart valve diseases across the molecular and clinical phenotype.

1.1 Research Questions and Objectives

In this thesis, we address two distinct research questions, one relating to transferring Systems Biology methods to be used in Systems Medicine from a software development perspective and the other addressing particular biomedical insights in human heart valve disease.

The first question addresses the need for DE/DA analysis software adaptations towards their use in Systems Medicine settings. The clinical scientist usually has a broad knowledge of the clinical phenotype and expresses many questions and hypotheses on molecular processes. The computational biologist knows how to perform the analysis and answer the questions, but not the resources to work on all of them. In Systems Biology, the lack of a computational biologist in life science research groups has been addressed by a plethora of tools or platforms that cover many steps of raw data pre-processing and calculation [4], e.g., for DE [37–42], DA [43–47] or even differential methylation [48] calculation. However, they are tailored to the origin of DE/DA analysis in Systems Biology, which assumes minimal experimental metadata to take into account. As a result, solutions that run all computation steps in one run are appropriate and there is no dedicated requirement to handle the extensive clinical/phenotypic data as common in medical research. Furthermore, no approach has ever been subject to (published) user testing. Usage of current solutions promotes redundant processing of data and/or restriction of clinical hypotheses towards few simple designs, thus not leveraging the full potential of both the molecular quantification data and the clinical phenotype. Additionally, a software solution developed without user testing may in fact not be valuable to the actual user after all. Therefore, the first research question is as follows:

Research Question 1: *How can automated DE/DA calculation software be adapted to the wealth of data and the exploratory nature of data analysis in an observational Systems Medicine setting?*

The first part of this thesis aims to determine and implement features and requirements of a research application, which

(i) is dedicated to handling the clinicome,

(ii) qualifies novice user to interactively define and configure complex hypotheses to be tested on HTQ data, and which

(iii) automates processing steps while adhering to the scientific standards of best practice procedures to receive publication-ready results on DE/DA results.

We use a hybrid approach of Design Thinking, scientific software engineering, and literature research to define features and requirements of DE/DA software for Systems Medicine. We implement prototypes of such a software and utilize user testing to evaluate the specific Systems Medicine adaptations concerning the user's perception and intention to use.

In the second part of the thesis, we address the biomedical insights with regard to our Systems Medicine data set from the SMART and EurValve study:

Cardiac hypertrophy caused by pressure overload (as in AS) or volume overload (as in MR) result in distinct forms of cardiac remodeling. Because of the difficulty to obtain human biopsy samples and the only recently emerging large-scale proteomics measurements, the changes in protein abundance in the myocardium of AS and MR could not be measured in well-powered studies yet [49]. Furthermore, sex-differences play a role in cardiac disease [34, 35] but are seldom considered in molecular studies. The majority of knowledge is gathered from clinical data assessed with non-invasive methods [36]. Attempts at elucidating the molecular setup in heart valve diseases and the impact of the differing hemodynamic loads are either based on animal models [50], are very small with regard to sample size and balance of sexes [49, 51], or cover only a small subset of proteins expressed in the myocardium [49, 50]. A larger body of extensive human evidence extends our general knowledge base on the topic and eventually helps to develop new targeted therapy approaches and avoid interventions that are not efficient in a specific condition or sex.

Research Question 2: *How does the myocardial proteome in heart valve diseases differ from the normal state and between conditions and how do these changes relate to known clinical characteristics? What are sex-specific changes?*

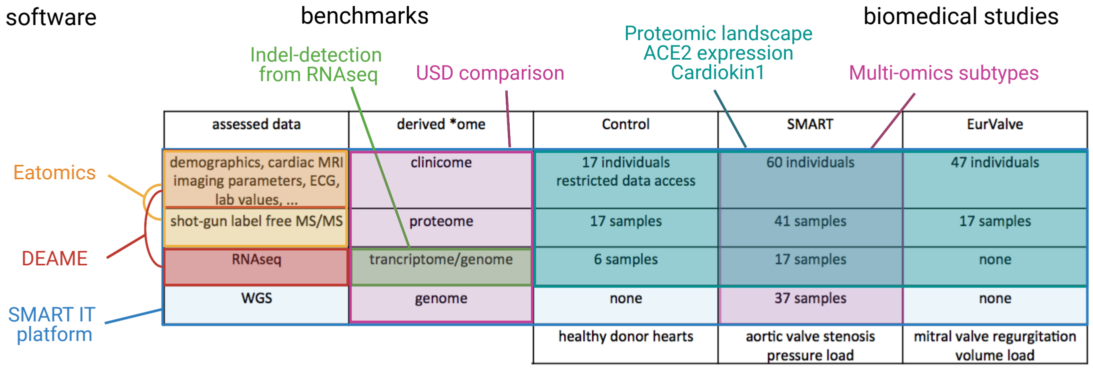
The second part of the thesis aims to obtain deeper insight into condition- and sex-specific differences in human heart valve disease and relate the extensive proteomic data to clinical parameters in a well-powered study of human tissue. We gather clinical parameters as well as deep proteomic measurements from collected tissue samples from the SMART and EurValve context. We apply differential abundance analysis, gene set enrichment, and a sophisticated approach of combining the results from different comparisons to derive shared and condition- and sex-specific proteome alterations and their relationship to clinical parameters.

1.2 Contributions

The contributions result from collaborative efforts and address parts of the research questions with regard to software, benchmarks, and biomedical findings. All relevant research artefacts are summarized in Figure 1.4 and shares in authorship and work effort are addressed in a dedicated section at the end of the thesis.

The SMART and EurValve projects serve as representative examples. Two benchmarks, one on indel-detection from RNAseq data and one on unsupervised subgroup detection (USD) methods, are not the main subjects considered in this thesis but can be acknowledged in Slosarek et al. (2018) [52] and Appendix C. Although the multi-omics subtypes in AS found through USD algorithms are not robust, the results influenced analysis directions and follow-up studies.

As a prerequisite, and as such described in subsection 2.2.1, we define and implement requirements of an IT platform for Systems Medicine tailored to the specific use case of



assessed data	derived *ome	Control	SMART	EurValve
demographics, cardiac MRI imaging parameters, ECG, lab values, ...	clinicome	17 individuals restricted data access	60 individuals	47 individuals
shot-gun label free MS/MS	proteome	17 samples	41 samples	17 samples
RNAseq	trancriptome/genome	6 samples	17 samples	none
WGS	genome	none	37 samples	none
		healthy donor hearts	aortic valve stenosis pressure load	mitral valve regurgitation volume load

Figure 1.4: Overview on SMART and EurValve data as representatives for a Systems Medicine setting, derived research artefacts and how they relate to different data sources. Contributions are grouped to belong to software, benchmarks or being biomedical studies. MS/MS - tandem mass spectrometry, USD -unsupervised subgroup detection, RNAseq - RNA sequencing, WGS - whole genome sequencing.

the SMART and EurValve projects. The platform serves as a central hub for data entry, upload, and storage of Systems Medicine data. It provides infrastructure and tools for processing NGS raw data and serves as a starting point for embedding the DE/DA research applications. Principles of the platform are described in:

> Milena Kraus and Matthieu-P Schapranow. An in-memory database platform for systems medicine. In *Proceedings of the 9th Int'l Conf. on Bioinformatics and Computational Biology*. ISCA, 2017. [53]

Our research and analysis applications are published in the following articles:

> Milena Kraus et al. DEAME – Differential Expression Analysis Made Easy. In *44th Int'l Conf. on Very Large Data Bases, Workshop on Heterogeneous Data Management, Polystores, and Analytics for Healthcare*, pages 162-174, Springer, 2018 [54]

> Milena Kraus, Mariet Mathew Stephen and Matthieu-P Schapranow. Eatomics: Shiny exploration of quantitative proteomics data. *Journal of Proteome Research*, 20(1):1070-1078, 2020 [55]

In the thesis, we provide details on the definition and implementation of general requirements for DE/DA software. For use in the Systems Medicine context, we develop the novel flexible experimental design concept and demonstrate our handling of the wealth of clinicome data within our research applications. We show that novice users, like clinical scientists, are qualified to explore and test their manifold hypotheses on changes in the transcriptome based on the samples' clinical phenotype and gather unique insights into user impressions, performance expectancy, and intentions to use. Our research applications DEAME and Eatomics cover or outnumber the majority of functionalities of existing tools. Eatomics is freely available to the research community as an easy-to-use R Shiny application.

Insights contributing to the biomedical knowledge base are described in (first authors with equal contribution are marked with an asterisk):

> Johannes Stegbauer*, Milena Kraus*, Sarah Nordmeyer* et al. Proteomic analysis reveals upregulation of ACE2, the putative SARS-CoV-2 receptor in pressure-but not volume-overloaded human hearts. *Hypertension*, 76(6):e41–e43, 2020 [56]
>
> Sarah Nordmeyer*, Milena Kraus*, Matthias Ziehm*, Marieluise Kirchner* et al. Myocardial proteome profiling reveals disease- and sex-specific alterations in patients with aortic valve stenosis and mitral valve regurgitation. *Circulation* (submitted)
>
> Nikolaus Berndt et al. Cardiokin1: Computational assessment of myocardial metabolic capability in healthy controls and patients with valve diseases. *Circulation* (accepted for publication)

In the SMART and EurValve projects, we assemble, process, and curate data from a large cohort of living human patients and healthy control donor hearts. Our curated data set provides quantification for more than 3500 cardiac proteins and more than 80 isoforms, accompanied by 120 clinical parameters for, in total, 75 human subjects. For a subset of 21 subjects, there is additional data on RNA expression available. The resource is used to sketch the landscape of disease- and sex-specific alterations in the proteome, transcriptome, and clinical phenotype in AS and MR. We describe alterations in the proteomic profile regarding proteins involved in extracellular matrix composition, metabolism, cytoskeleton, and cell adhesion shared in AS and MR. In general, we show that changes are more pronounced in AS. We present evidence from human tissue of living individuals for many observations previously only described in animal models or on the transcriptomic level. Additionally, we make novel observations in the proteostasis machinery being drastically reduced in female AS patients and relate the changes to less myocardial mass in females. Apart from the general landscape, we show the added value of the data set in providing evidence of increased levels of the putative SARS-CoV-2 virus receptor (ACE2) in pressure, but not volume loaded myocardial tissue in the proteome and transcriptome. Additionally, the data enabled the development of a novel cardiac metabolism model (Cardiokin1).

1.3 Thesis Outline

The remainder of the thesis is structured as follows: In chapter 2, we provide the necessary background information and definitions on Systems Medicine in general as well as common data sources and queries. Furthermore, the SMART project, which provides the context of all implementation efforts, is depicted. A short introduction to the biomedical foundations of the human heart and heart valve diseases is also part of the chapter.

Chapter 3 provides more detail on the motivation of differential expression software in Systems Medicine and why related approaches do not suffice the current needs. We explain our rationale behind the flexible experimental design feature in general and how we implemented two instances of the feature for DE and DA analysis. We evaluate our work within a functional comparison to other tools and user testing sessions. We discuss how our prototypes advance DE/DA analysis towards Systems Medicine by qualifying clinical and life scientists to perform flexible hypothesis testing using the rich clinical phenotype.

In chapter 4, we elaborate more on factors that have hindered research of human cardiac proteome alterations in human heart valve disease. We then share our approach and experimental setup to measure and analyze the clinical, transcriptomic, and proteomic data. We report disease- and sex-specific changes in protein abundance and clinical imaging parameters and discuss how these results confirm findings from animal models or open up new perspectives of human heart valve disease.

Finally, we summarize all findings in chapter 5, answer our two research questions, and present ideas for future directions in Systems Medicine.

2 Background and Preliminaries

In this chapter, we provide the necessary background information and definitions on Systems Medicine in general as well as common data sources and queries. Furthermore, the SMART project, which provides the context of all implementation efforts, is depicted. A short introduction to the biomedical foundations of the human heart and heart valve diseases is also part of the chapter.

2.1 Data and queries in Systems Medicine settings

To understand the full spectrum of a Systems Medicine approach, this chapter introduces background information on biomedical high-throughput measurements and systems biology methods, with their potential for reuse in Systems Medicine as well as the representation of Systems Medicine efforts in current IT infrastructures based on the example of the SMART consortium.

2.1.1 Data in Systems Medicine

Data in Systems Medicine consortia are mostly molecular data, stemming from high-throughput omics technologies in combination with a thorough clinical phenotype and description [11]. This introduction will therefore first establish an overview on the different molecular and clinical data modalities and their biomedical meaning. After that, every modality is described in terms of how it is commonly assessed or measured and preprocessed. Finally, this section helps to understand the digital representation of Systems Medicine data and therefore input and output characteristics of data used within this thesis. However, the reader should be very aware of the vast amount of different laboratory instrumentation and bioinformatics tools available for the creation of data. The exact inputs and outputs may differ across pipelines and this summary is limited to a very broad, simplified version of the processes.

Biological meaning of the molecular data

The genome comprises all genetic material of the in our context human cell, i.e., deoxyribonucleic acid (DNA). Chemically, the DNA is a large molecule comprised of two strands assembled into a double helix and consisting of a sugar-phosphate backbone and four nucleobases: adenine (A), thymine (T), guanine (G), and cytosine (C). The sequence of nucleobases encodes instructions for cell functions, development, growth, and reproduction. The general flow of information is drafted by the general dogma of molecular biology, i.e., how the instructions are realized to compose a living being. The sequence of the DNA can be copied to DNA (DNA replication) or be transcribed to (messenger) ribonucleic acid (RNA), which in turn is a template to synthesize proteins (translation). Therefore, mRNA abundance can be seen as a proxy for protein abundance, although correlation is less pronounced as expected due to massive post-transcriptional modification and regulation of mRNA, partially through small- or long non-coding RNAs. Proteins consist of amino acids (AA) and perform all enzymatic processes and functions in cells, tissues, and organs. While the DNA is constant across cells of an organism, mRNA, and protein abundance are tissue or cell type specific. An alteration of sequence in the DNA may result in a change of sequence of mRNA and thus in the alteration of the protein sequence, structure, and function. In another mechanism, a genomic variant can lead to higher or lower mRNA abundance (expression quantitative trait loci, eQTL) and/or altered protein abundance (protein quantitative trait loci, pQTL). Other popular molecular levels that are not further concerned in this work are the epigenome, i.e., chemical modifications on the DNA molecule itself or on proteins engaged in DNA structure, such that mRNA transcription is enabled or hindered. The metabolome describes the totality and/or abundance of low-molecular chemical components serving in metabolic processes. The molecular components interact heavily with each other in a non-random fashion forming large networks of processes, called pathways, needed for proper functions of the cells, tissues, organs, and ultimately the whole organism.

Genomic variation

Genomic variation describes any difference of deoxyribonucleic acid (DNA) sequence when compared to a reference sequence. The genome of a human is approximately 3.2 Giga bases long. While genotyping arrays are another way of retrieving genomic variation, here we focus on describing genome sequencing.

Biologic input: To assess genomic variation, a blood or saliva sample is sufficient to extract DNA from the cells. The DNA is sheared into smaller fragments, multiplied and chemically modified for subsequent analysis in a sequencer. The sequencer is capable of detecting the four different bases A, C, T, and G in all multiplied fragments in parallel. The detected base is written to a file, supplemented with a quality score, i.e. describing the certainty of the measurement.

Bioinformatics processing: Depending on the sequencing technology the smaller fragments are of very different length but usually need assembly in order to represent the original genome. The assembly of fragments, from now on called reads, is guided by the reference genome, which describes the most common sequence representation of a large population sample, e.g., the human genome project. The comparison of any given read to the reference sequence will lead to predominantly perfect matches in sequence. However, in a process called variant calling, the mismatches are analyzed and classified as being the result of measurement errors or a true difference in the sample.

Common output: Regardless of the origin of variation, a common digital representation of genomic variation is the variant call format (VCF). Essentially, the format contains a header of metadata describing the fields of the body. Every row of the body represents a variant, described by 8 mandatory fields of the vcf: CHROM and POS describing the genomic location, ID containing one or more identifiers, REF and ALT being the reference and alternative allele sequence, QUAL describing the quality of the call and FILTER harbouring a flag for the variant to have passed specified filter criteria. The INFO field holds key value pairs of information on the variant, whereas the FORMAT, the ninth filed, contains key value pairs describing the information given in the samples' columns. Further information can be found at `https://github.com/samtools/hts-specs/blob/master/VCFv4.3.pdf`. The size of the VCF depends on multiple factors, e.g., if multiple samples are listed, which results in more columns, how much additional information on the variants and samples is given, the handling of large blocks representing the reference sequence (gVCF) and the origin of variation.

Origin of variation: Germ line variation results from a comparison of the measured sequence against the reference genome. The variant must have been present in the germ line of an individual and thus was propagated in all cells of the organism. Approximately 3-5 million germ line variants are found in a whole genome sequencing analysis. In contrast, somatic variants are the result of a sequence alteration within the genome of a single cell. In the course of cell division, the alteration is passed to all next generations of that cell. Multiple alteration events (mutations) can lead to the development of a tumor. In many tumors, DNA repair mechanisms are corrupted leading to an accumulation of somatic variants in the tumor tissue. The number of somatic variants differs across cancer types, but with a range between five to a few thousand when compared to normal tissue of the same individual, it is still a couple of orders of magnitude lower than germ line variants.

Quantitative transcriptomics

The transcriptome describes the set of all RNA molecules in a cell at a specific point in time. This includes sequence information, i.e., which parts of the DNA are transcribed, and quantitative information, i.e., the abundance of RNA molecules. Measurement of the transcriptome can be realized via microarrays or RNA sequencing. For the remainder of this thesis we focus primarily on the latter as it is becoming more popular due to higher accuracy and dropping prices.

Biologic input: RNA is extracted from cells or tissue of interest and fragmented into smaller pieces. Reverse transcription into more stable cDNA and further chemical modification is needed to prepare the sample for sequencing, which follows the same process as described for DNA sequencing.

Bioinformatics processing: The sequence reads need assembly to represent the transcribed regions of the genome (transcripts) guided by a reference genome or transcriptome sequence. All reads are summarized into a quantity across a defined genomic region, e.g., a gene, which results in one abundance vector per sample.

Common output: In RNA sequencing, it is most favourable to analyze all samples of a comparison that originate from one experimental run with the same protocol. Therefore, the abundance vectors can be summarized in a matrix of counts of reads detected per genomic feature and sample.

Quantitative proteomics

As per definition, the proteome is the entirety of all proteins expressed in a cell, tissue or organism at a specific time and under specific conditions. Depending on the detection method, also peptide sequence and isoform states may be inferred. Label-free shotgun proteomics is a variant of quantitative proteomics, i.e., measurement of the abundance of thousands of proteins from a peptide mixture, replacing the classical method of a Western Blot by being a lot more sensitive and offering a large coverage of proteins.

Biologic input: In a first step, proteins are extracted from cells or a tissue sample of interest. A cleaving enzyme digests the proteins into smaller fragments at specific sites. The peptide mixture is then fractionated based on the peptides chemical properties to yield an eluted mixture in liquid chromatography. Within the mass spectrometry (MS) peptides are ionized and undergo further fragmentation in tandem MS. The second MS detects mass spectra of the peptides. A reference sample containing all potential peptides of the samples under investigation can be used to guide subsequent detection and increase coverage.

Bioinformatics process: Peptide spectra function as a fingerprint, which in turn can be assigned to one or more specific peptides based on a database search. Peptides are mapped to amino acid sequences of proteins or more specifically assigned to protein groups sharing sequences.

Common output: The quantitative analysis of proteomics resutls in an intensity matrix of protein groups vs. samples, which is calculated from spectra per sample as a measure of peptide/protein abundance.

Clinicome

The clinical phenotype in this thesis serves as the broad term of all patient characteristics that are assessed at the German Heart Center Berlin and are part of the patient's health record at the clinic as necessary for the SMART and EurValve study. However, the clinical phenotype itself is very diverse in the sense of parameters assessed as well as in the type of data it is stored in, i.e., its digital version. In any case, the clinicome consists of categorical and numerical parameters that can be grouped into subcategories or in many cases of fully unstructured clinical notes.

2.1.2 Differential expression/abundance and enrichment analysis

Gietzelt et al. (2016) state that there are no methods specific for Systems Medicine as the new field draws ideas from Systems Biology and machine learning [10]. Relating to this aspect, a typical Systems Biology analysis of transcriptomic and proteomic data includes DE or DA analysis as well as pathway analysis (PA). Therefore, we introduce the general process of DE analysis. The mathematical foundation inherent to many state-of-the-art

DE calculation tools is laid out in greater detail in this section, as it is crucial for the realization of explorative hypothesis testing. The coarse example may easily be replaced by a representation of quantitative proteomics data and thus the process is applicable to DA as well. Pathway analysis and/or gene set enrichment is part of DE/DA result interpretation and is introduced shortly as well.

DE/DA Analysis Principles

The development of NGS techniques have enabled the usage of microarrays and RNAseq data as a source for DE analysis. Equivalently, the advances in high-throughput label-free MS shot gun proteomics are a source of large-scale quantification of proteins. Analysis of DE is the process of identifying genes that have an altered level of expression in a group of samples, which is statistically significant when compared to another. Congruently, DA identifies differences in protein abundance. The differences in expression or abundance levels may be the result of a disease or other perturbations on the examined cells or tissues. Therefore, the identification of the differences can lead to biomarkers of a disease [57] or a transcriptomic or proteomic profile that may be reversed through a new or existing treatment. For this thesis, we summarize proteomic and transcriptomic quantification data as HTQ data.

General process of DE analysis

In the following, we briefly describe the process of DE based on an RNAseq experiment. The high-level process is modelled in Figure 2.1 in Business Process Modeling Notation (BPMN) and is based on [6]. We need to consider five process steps, namely the wet lab experiment design and execution, bioinformatics processing of raw data, DE calculation, and visualization, annotation, and interpretation steps.

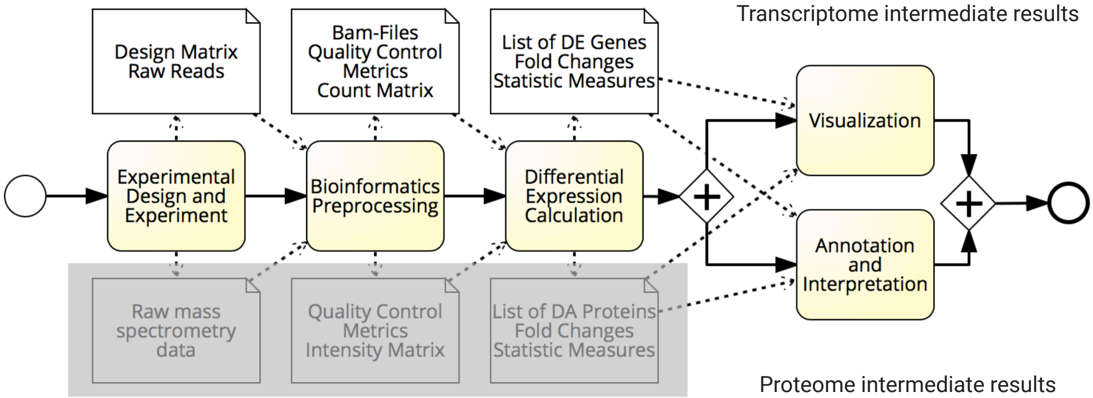


Figure 2.1: Generic differential expression process steps, i.e., activities (rounded boxes) and their resulting data artefacts modelled in BPMN 2.0. Intermediate results of equivalent processing steps for proteome analysis are also displayed in a greyed out fashion for more clarity.

Experimental design and experiment

Inherent to DE analysis are at least two groups of samples that are assumed to show differences in gene expression. These groups need to be specified before *in vitro* testing in order to plan and design the wet lab process, such as treatment with a specific chemical or drug. As a result, the researcher needs to define a design formula, which resembles the research hypothesis and is the basis of any DE experiment.

Bioinformatics pre-processing

The sequencing process results in raw reads, i.e., short fragments of the actual genomic sequence section 2.1.1. Raw reads go through quality control and in some cases need to be trimmed from adapter sequences prior to alignment. All reads are aligned to a reference genome or transcriptome. In the best case, all genomic ranges, such as a gene, an exon or a coding region, are covered by multiple reads after the alignment step. Counting tools calculate the exact quantity of reads per given genomic range.

DE calculation

DE calculation is the statistical process of finding significant expression differences of two or more groups as defined in the experimental design. In short, all counts of a genomic range in one group are compared to the counts of the same range in another group of samples. The calculation provides information about the fold change, i.e., how much more counts where found in one group when compared to the other. Additionally, p-values are given, which are adjusted for multiple testing, as many data sets comprise 10-20 k genomic regions to compare. A more detailed description of the underlying mathematical foundations are given in section 2.1.2.

Visualization

Visualization of results is a critical part in DE analysis, as raw and transformed data as well as DE results are usually high in dimension and therefore need to be displayed in a comprehensive format. Frequently used techniques are principal component analysis (PCA) and clustering of data. Both give an impression of similarity between the analyzed samples. For example, plotting samples on their corresponding first and second principal component (dimension of largest variation) should result in scatters of samples grouped according to the experimental design formula. Accordingly, clustering algorithms should be able to find clusters and a dendrogram resembling the desired study groups. Clustered heatmaps are specifically popular as they can display sample-to-sample as well as gene-to-gene relationships and the corresponding normalized and log transformed count values in a single diagram. Volcano plots depict the p-value versus expression fold change between two conditions. DE genes are usually highlighted and therefore the plot gives a good overview of all results. Many more diagnostic plots are used, e.g., as depicted in a bioconductor workflow [58].

Annotation and Interpretation

Annotation and interpretation of results is a critical and complex part of the analysis. Typically, more than 100 genes are found differentially expressed between patient groups. Regarding the most relevant expression changes, a manual search for function and involved pathways is performed. GO term annotation and gene set enrichment analysis (GSEA) help to find perturbed anatomical structures, biochemical processes or pathways in an automated manner.

Generalizability

Although the process and diagram are tailored for RNAseq data, the general concept can be transferred to many other quantitative omics measurements. For example, in a label-free shot-gun MS proteomics setting, the laboratory procedures, measurement, and also bioinformatics processing differs remarkably; however, a count matrix or a matrix of intensities can be considered equivalent and can, with the exception of changes in specific default parameters, be analyzed using the same DE calculation tools, e.g., Limma [8].

Mathematical foundation for DE calculation

For the remainder of the thesis, we define the output of a HTQ experiment, e.g., RNAseq or label-free MS, as a matrix Y of i rows representing genomic ranges, e.g., genes or proteins, and n columns of samples. The matrix entry $Y_{i,n}$ indicates the quantification of the genomic range i for sample n.

Furthermore, we define a matrix Xf as being meta information accompanying the quantification experiment. The matrix entry $Xf_{n,j}$ denotes the value or level of an assessed parameter j (columns) for sample n (rows). Selecting parameters and samples of interest from Xf derives the model matrix X. In the simplest case, e.g., control vs. treated, X can be reduced to being a vector of zeroes and ones denoting the assignment of samples to the two groups. Similarly, the simplest case for matrix Y would be a vector of expression values for one gene for all or a subset of samples. A t-test to assess the truth of the null hypothesis (H_0), i.e., there is no difference in expression between groups of samples, would suffice in the case that the assumption of normal distribution and equal variances is true. However, to gain more flexibility with regard to these assumptions and towards including reasonable amounts of metadata for more accurate statistics, many current implementations of DE software use GLM in combination with a moderated t-test.

Generalized linear models

The advantage of using GLM is being able to include more sources of variation into the analysis, as is often useful in the analysis of expression data and to let the distribution of expression values differ from a normal distribution. The basic linear model in DE analysis is displayed in Equation 2.1. It consists of a response variable $Y_{1,1}$ in $\log_2$ scale, e.g., the abundance of a gene product or protein, dependent on an explanatory variable X, e.g., belonging to the control or treatment group, β_0 being the constant intercept or baseline expression in case of X being zero and β_1 the coefficient describing the slope by which Y in- or decreases dependent on X. ϵ is an individual error term denoting the variation of the measured data point Y from the group mean.

$$\log_2(Y_{1,1}) = \beta_0 + \beta_1 X + \epsilon \tag{2.1}$$

In the simple case of a control vs. treated experiment, the two β coefficients correspond directly to the group means and the difference of the two specifies the $\log_2$ of the fold change (FC) between them (Figure 2.2).

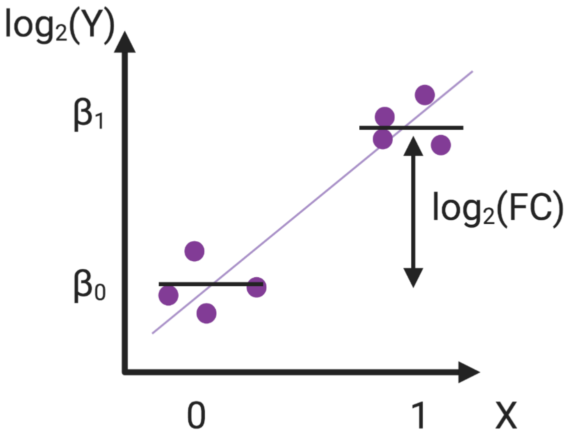


Figure 2.2: Graphical representation of a generalized linear model in the simple setup of a control (=0) vs. treated (=1) experiment. The y-axis denotes the expression strengths in log2 scale, whereas the two β coefficients correspond directly to the group means and the difference of the two specifies the $\log_2$ of the FC between them.

Further explanatory variables, i.e., parameters from the meta information Xf, and coefficients can be added to describe more complex relationships. They can either accompany the model as covariates that are of no particular interest to the researcher or they function as additional variables of interest. Every observation of a sample Y gives rise to one more linear equation, every additional explanatory variable will add another X to the equation system (Equation 2.2).

$$log_2(\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}) = \begin{bmatrix} 1 & X_{1,1} & \cdots & X_{1,j} \\ 1 & X_{2,1} & \cdots & X_{2,j} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n,1} & \cdots & X_{n,j} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_j \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \tag{2.2}$$

The equation may also be written as a shorter formula to be used as input in linear regression algorithms.

$$Y \sim X_1 + X_2 + \cdots + X_j \tag{2.3}$$

The formula gives rise to the actual model matrix, which includes the split of categorical variables with more than two levels ($m > 2$) into m dummy variables when β_0 is set to zero.

As Y and X are known, β coefficients need to be calculated to obtain an optimal fit of a line through the given data points. Methods for the calculation of the coefficients are manifold and matured over time in the field of DE analysis to adapt to the very specific characteristics of HTQ data sets. Relevant examples are the use of maximum likelihood estimation (MLE) in DESeq2 and weighted or generalized least squares methods in Limma. Additionally, both methods utilize Empirical Bayes methods to adjust for high variance in the quantification values Y across genomic ranges and utilize sophisticated assumptions on the distribution of Y.

Test for statistical significance and multiple testing correction

Similar to the sophisticated methods on how to treat Y for best results, DE methods are optimized to use mean-variance correction or Empirical Bayes methods to correct fold changes and enable a well-informed statistical test, e.g., moderated t-test or Wald test. Furthermore, as the number of genomic ranges in many experiments exceeds the thousands or ten thousands, an effective correction of p-values for testing multiple hypotheses is employed commonly by using the Benjamini Hochberg (BH) procedure [59].

Pathway analysis and gene set enrichment.

PA is used to map genes found in high throughput molecular experiments to meaningful categories, i.e., pathways, that facilitate interpretation of the oftentimes overwhelming results from DE analysis. Input to PA are a knowledge bases, i.e., a collection of genes biologically linked to each other (gene sets or pathways), output is a list of gene sets relevant to the condition under study. Further input are gene lists from the experiment, preprocessed in dependence on the chosen PA method. PA methods are loosely categorized into over-representation analysis (ORA), functional class scoring (FCS), and pathway topology methods [60]. ORA and FCS methods are used in different parts of this thesis and therefore require explanation:

In FCS methods, the full list of genes together with a gene-level statistic is used as input. In the original sense of GSEA [61] the statistic is based on, e.g., the t-statistic from DE analysis, which is used to rank the genes in the list. The ranked list is then used to generate a pathway-level statistic, e.g., an enrichment score (ES). Based on a specific gene set, the ES is increased relative to the rank, e.g., higher increase when further away from the median, and decreased if the gene is not contained in the given gene set. As a result, if genes of a set are spread randomly across the ranked list, the ES will be close to zero, whereas an accumulation either on top or bottom of the list will result in ES being closer to 1. p-values and, in the case of multiple tests, false discovery rate (FDR) is calculated via a random permutation of the ranked list.

In ORA methods, the gene list is trimmed to only contain genes meeting an arbitrary cutoff, e.g., being above an absolute fold change threshold or being below a p-value threshold as derived from DE analysis. The overlap of the trimmed list and a given gene set is then compared to the overlap of a background list of genes. Fisher's exact test is used to calculate if a gene set is over-represented in the trimmed gene set and FDR can be used to control for multiple testing. No method has been proven to be superior in all aspects and thus it is up to the user's preference and use case [62]. GSEA, a very popular FCS method, was recently extended to being used in a single sample mode (single sample gene set enrichment analysis (ssGSEA)) [63]. In ssGSEA the ranked list is based on the actual (relative) expression value of a gene while the output list of significantly enriched terms is calculated by comparing the mean single sample enrichment score (ssES) of one group of samples against another. This approach confers the advantage of postponing the actual comparison of groups to being after pathway analysis and allows for more flexibility. Furthermore, the ssES can potentially be used as a means of normalization between omics levels. Please note that there is some confusion in wording and names of PA methods. Throughout the thesis we use definitions as introduced in Kathri et al. (2012) [60].

2.2 SMART - A Systems Medicine consortium

In the SMART consortium, interdisciplinary experts establish methods for interrelating parameters and modelling of HF to improve patient care. In a joint effort, the transcriptome, proteome, cell function, regulating hormones, tissue composition, hemodynamics, and whole organ function, up to a whole body description of patients suffering from AS are derived by the consortium members [64]. The process of data assessment and utilization is modelled in Figure 2.3.

Data is acquired in the course of a dedicated observational study of patients undergoing a replacement of the left ventricle valve at the heart center. Clinical scientists assess a large body of parameters from general patient assessment, imaging, hemodynamic evaluation, and surgery. Biopsy samples, taken at time of surgery, are sent to molecular biology laboratories to generate data on the transcriptome and proteome. Cell-, whole-organ-, and multi-scale modellers, located at individual institutes, depend on the parameters assessed by other consortium members.

As such, Systems Medicine in general is dependent on powerful platforms that can integrate, process, and analyze the multiple, heterogeneous biomedical datasets and at the same time support the workflow of research consortia [13]. At the time of project start in 2015 IT platforms aiming at supporting Systems Medicine need were sparse: The Georgetown Database of Cancer (G-DOC) was one of the few research platforms already available, but could not suffice requirements of the SMART project, mainly because design decisions in G-DOC were tailored for the specific use case of cancer [65]. In contrast, the San Raffaele Systems Medicine Platform for Non-Communicable Diseases (SR-NCD) was launched in 2013 with the aim to cover many pre-defined non-communicable diseases for Systems Medicine [66]. However, at the time of start in SMART the SR-NCD was still in an early planning stage and thus no usable prototype was available. Similar in name, the transSMART project is the most promising platform in its aim to enable collaboration

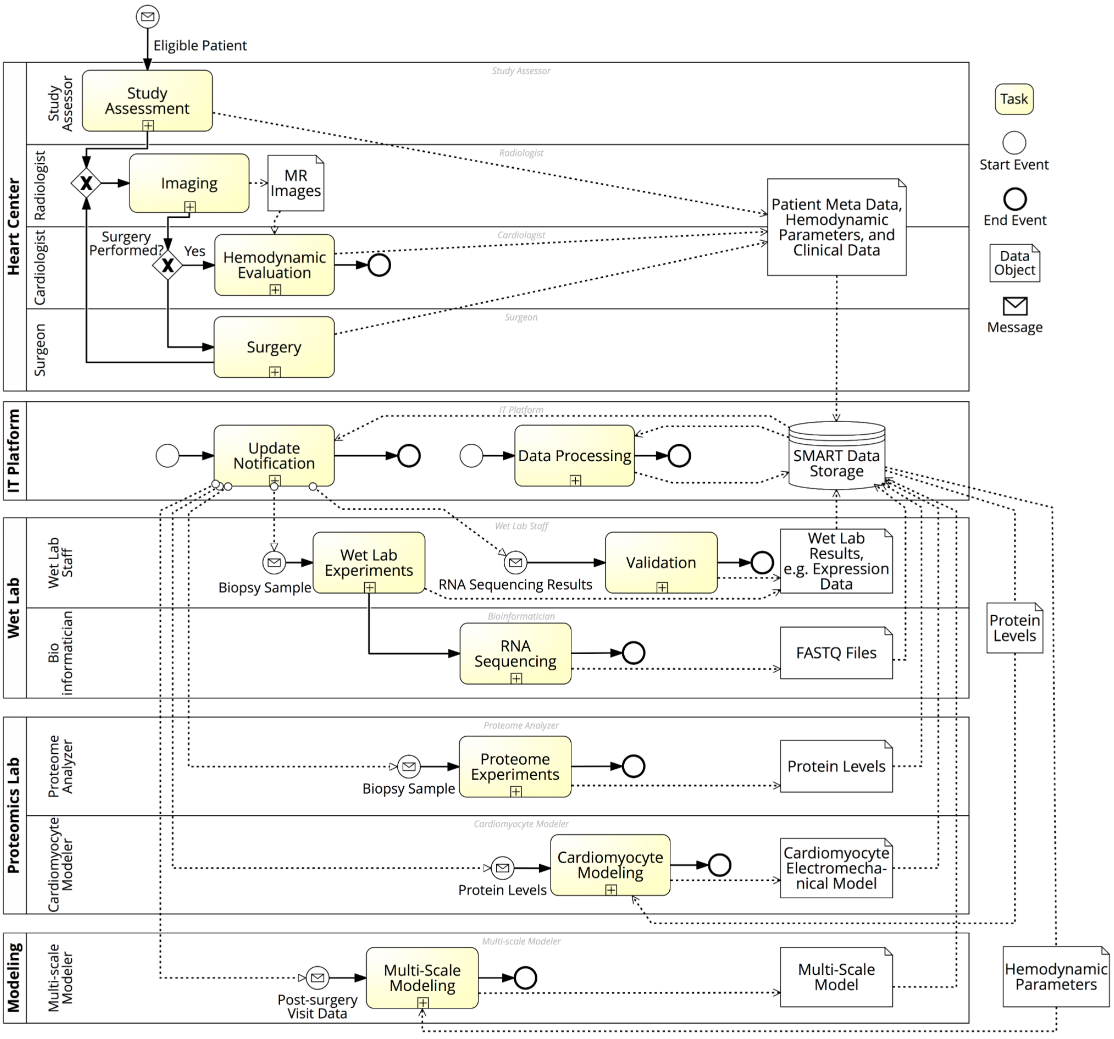


Figure 2.3: Process of data assessment, analysis, and communication within the SMART consortium. Key stakeholders and roles are shown as swim lanes and rounded boxes show process steps, which may be nested. Data is acquired in the course of a dedicated observational study of patients undergoing a replacement of the left ventricle valve at the heart center. Clinical scientists assess a large body of parameters, such as hemodynamic parameters. Biopsy samples, taken at time of surgery, are sent to molecular biology laboratories to generate data on the transcriptome and proteome. Cell-, whole-organ, and multi-scale modellers depend on the parameters assessed by other consortium members. An IT platform serves as a central hub for data storage and processing.

for precision medicine, through sharing, integration, standardization, and analysis of heterogeneous data from healthcare and research [67]. Unfortunately, the foundation only gained momentum after the merge of the i2b2 and the transSMART projects. With the release of version 17.1, which was developed by a professional IT company, the software took a major step towards being a reasonable choice for a Systems Medicine platform. As a result, we needed to provide our own solution within the SMART context. The SMART IT platform provides the basis for the research applications as described in chapter 3.

2.2.1 Challenges consortium work

Within the SMART consortium we have identified the following challenges in collaborative work, which relate well to challenges found by Kramer et al. (2018) [26]:

1. Sharing of Data: Individual consortium members require input data from others, process it, create new output data, and provide it to other researchers. For example, the heart center collects clinical parameters during a visit of a patient in spread sheets and sends them to other consortium members, e.g. the proteomics lab, via email. Changes in these spread sheets are not under version control.

2. Communication: Interdependencies between consortium members require immediate communication of new data. In the SMART consortium, an electrochemical model of a cardiomyocyte will be calculated based on proteome and clinical data. Communication between at least three members of the consortium, e.g., the heart center, the wet lab, and the proteomics lab, must be established to enable the cardiomyocyte modeller to work. Any delays in communication decelerate the generation of modelling results.

3. Reproducible Data Processing: Data processing is mostly performed at local lab sites, which acquire data from samples and process them directly. As a result, the reproducibility of results is shielded from other consortium members. For example, differentially expressed genes are computed from raw RNA sequencing data at the bioinformatics facility. The processing requires domain knowledge, tools, and computational power. Thus, the processing can not be reproduced by the other consortium members.

These challenges guided overall platform development.

2.2.2 Implementation of the SMART IT platform

An IT infrastructure facilitates collaborative work within the consortium in various aspects, such as sharing of data, communication between partners and reproducible processing of data. In the following we provide details on the data, platform and application layer and their interplay as shown in the software architecture diagram in Figure 2.4.

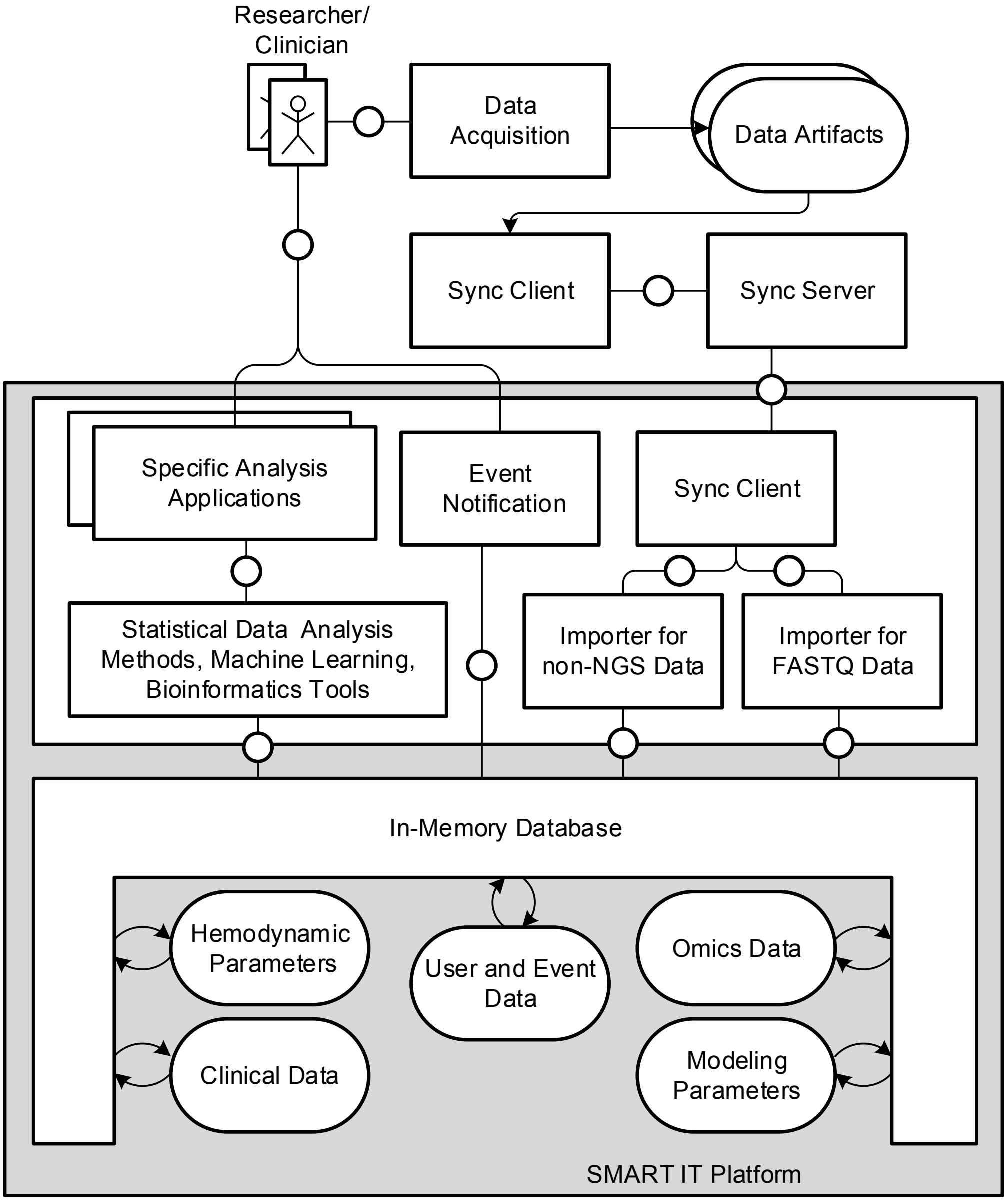


Figure 2.4: Software architecture of the SMART IT platform. An In-Memory Database holds hemodynamic parameters and clinical data in general. Additionally, it stores omics data, modelling parameters, and user and event data. The application layer holds an event notification routine, a sync client, and importers for different data modalities. Specific research applications establish data pre-processing pipelines incorporating various bioinformatics tools and analysis procedures. Researchers/clinicians can access the applications and trigger data integration.

Data Layer

The purpose of our SMART IT platform is the integration of selected medical and molecular data sources. In-Memory Databases (IMDBs) have been shown to be suitable to integrate various biomedical data sources in a single system, which provides features for the real-time analysis of data mainly by holding all data in the main memory [68, 69]. An IMDB serves as the central unit to store data and to enable flexible real-time data analysis.

Amongst others, we combine patient metadata, clinical parameters, and different kinds of *omics data provided by the consortium partners. We utilized a star schema to model individual entities within the database and allowed their combination as depicted in the Entity Relationship Diagram in Figure 2.5.

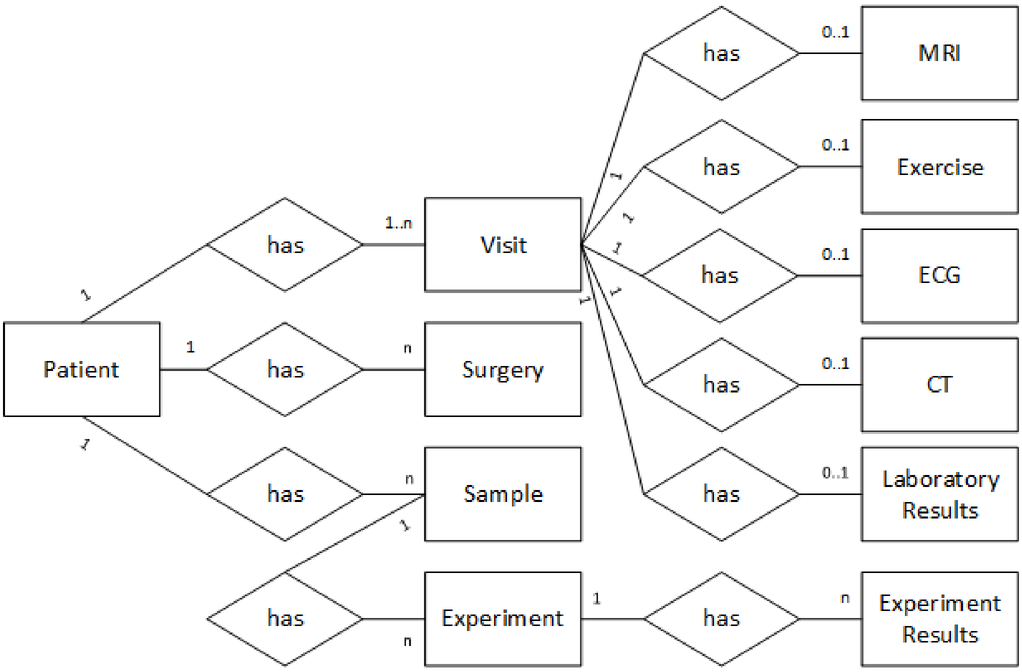


Figure 2.5: Entity relationship diagram depicting SMART data representation. The central entity of our model is the `Patient` that performs at least two `Visits` in the course of the treatment. At every visit, multiple examinations, and diagnostic procedures, such as magnetic resonance imaging (MRI) or electro cardiogram (ECG) are performed. At time of `Surgery` a `Sample` is taken, which is used for multiple `Experiments` by the laboratory team. `Experiments` represent any molecular biology lab or modelling procedure, e.g., Western Blots, RNA sequencing experiments or multi-scale models.

The central entity of our model is the `Patient` that performs at least two `Visits` in the course of the treatment. At every visit, multiple examinations, and diagnostic procedures, such as MRI or ECG, are performed. At time of `Surgery` a `Sample` is taken, which is used for multiple `Experiments` by the laboratory team. `Experiments` represent any molecular biology lab or modelling procedure, e.g., Western Blots, RNA sequencing experiments or multi-scale models, where the corresponding procedure is defined as an attribute. In our data model, only common attributes, such as an experiment ID, a time point, and a link to the source file, are directly appended to the experiment. All additional attributes, such as specific experiment conditions and results, are stored as attribute value pairs to allow flexible extension of our system for further use cases. The majority of data is stored in columnar format to be able to fully exploit the advantages of using an IMDB. Additionally, all tables are history tables, which can be used to reproduce any previous state of the database. History tables are especially useful in an event-driven notification system.

Platform Layer

The data processing logic was transferred from specific tools operated by individual researchers to the platform layer of our SMART IT platform. Event notifications and real-time data integration as well as specific processing logic for the SMART data are the foundation of our platform layer.

The bioinformatics research community has developed a set of software tools for processing RNAseq data, which need to be executed in a pipeline. For demonstration purpose, we selected the Tuxedo protocol, which was extended with widely used quality control and trimming tools [70]. To execute the pipeline automatically by the AnalyzeGenomes worker framework, we modelled it using the Business Process Modelling Notation (BPMN) as depicted in Figure 2.6. The notation determines configurable parameters and the order in which the individual steps are executed. The boxes represent the incorporated bioinformatics tools (activities), e.g., TopHat and Trimmomatic. For traditional research, each of these steps would be executed separately, i.e., each tool loads input data from the hard disk into memory and writes the results to an output file, which in turn again is stored on the hard drive. In our approach, input data, intermediate results, and final results are stored in the database. The user is requested to select patients that will be analyzed through the user interface. As all the other parameters are pre-configured once by an expert, e.g., reference genome, number of threads, and the succession of steps is determined by a pipeline model, the data processing is executed automatically.

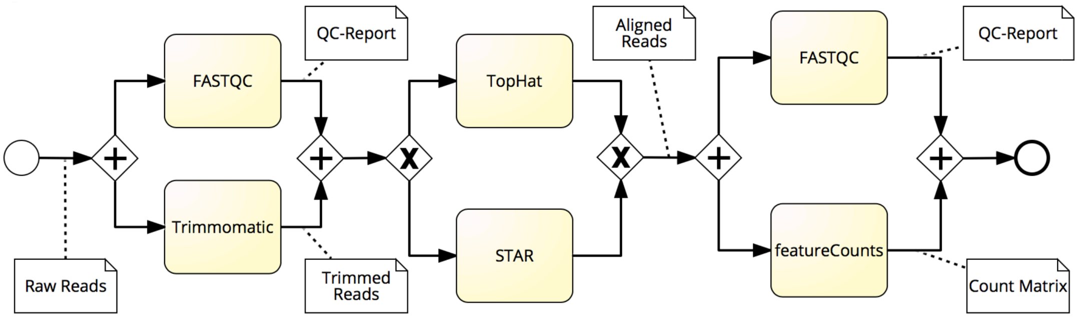


Figure 2.6: A typical bioinformatics pre-processing pipeline for RNAseq raw reads modelled in BPMN to be executed by the AG worker framework. The notation determines the execution order of jobs (rounded boxes) and resulting data artefacts.

Data, which was acquired by individual consortium partners at local sites is synchronized automatically to the SMART IT platform using a private ownCloud server instance, which requires consortium partners to simply store it in a pre-defined folder. As a result, data integration does not require personal intervention after configuring it once. A cron job regularly checks for updates in the folder and calls corresponding import jobs to load data into the database. Clinical parameters are generated and entered throughout the time of study. Therefore, either an initial bulk load option can be used via the `sync client` or changes and additional parameters can be entered via the StudyPI App. Several stakeholders within the consortium are dependent on timely data provision by other partners. In our system, an automated notification informs the user in case relevant data has been added, while limiting the number of separate notifications to a minimum. Depending on the given roles and permissions, a user can choose to register for notifications for specific events, e.g., new visit data added, as depicted in Figure 2.7A.

To identify relevant changes per user, the last login time of the user and the time of the last event occurrence are checked by means of history tables. It is assumed that the user has acknowledged all relevant data changes after having read the email or having logged in to the application.

Application Layer

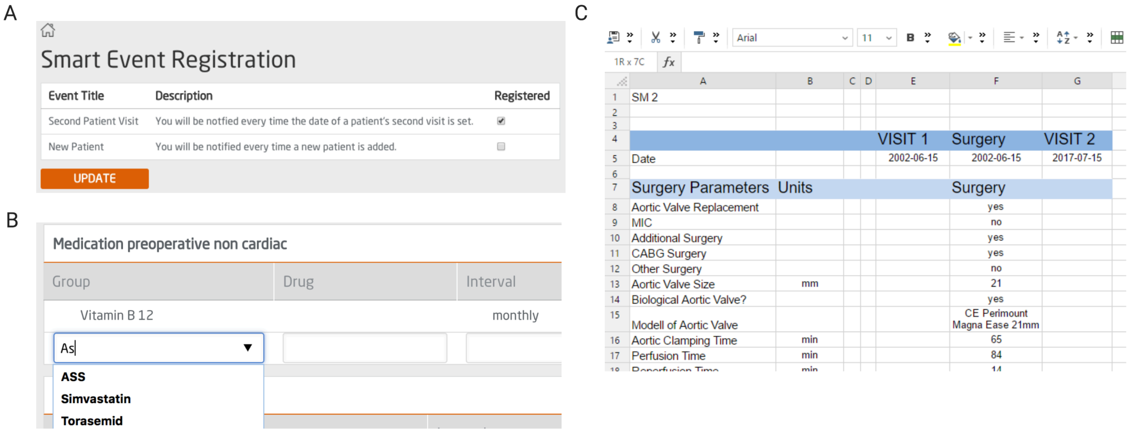
The application layer was designed together with subject matter experts to provide an easy-to-use, intuitive UI for researchers and clinicians. We followed the guidelines for a responsive UI to make the platform accessible on desktop and mobile devices equally.

All data artefacts are automatically integrated into the SMART IT platform through a `sync client`, which periodically checks for new files and uploads changes to keep all data up to date. Raw RNA sequencing reads are stored in the FASTQ file format [71] and are loaded into the database at time of processing.

We created specific application views per user role that have been designed to resemble the distinct research processes of the consortium member. Additionally, there is a management view, which presents different editing options depending on the rights within a role. Every partner signs up to the platform and has a personal user profile. The UI enables several specific features, such as triggering RNA processing tasks. The entities `Patients`, `Samples`, and `Experiments` as well as their attributes can be added, edited, and deleted through our UI in a structured format. Graphical data exploration of phenotypic data is leveraged through SAP Lumira in an additional app.

The StudyPI app is one of multiple apps that will be available in our platforms. It resembles the workflow and tasks of the the study's principal investigator. The principal investigator assumes primarily management tasks, such as adding and editing new patients and the corresponding metadata. Furthermore, it includes constant study monitoring, e.g., deriving enrolled patients in the respective study groups, and evaluating heart specifics, e.g., development of myocardial mass after surgery.

The formerly used spreadsheet representation of data was integrated as an online form, which can, for means of backward compatibility, be downloaded again as a spreadsheet



Smart Event Registration

Event Title	Description	Registered
Second Patient Visit	You will be notfied every time the date of a patient's second visit is set.	☑
New Patient	You will be notified every time a new patient is added.	□

Medication preoperative non cardiac

Group	Drug	Interval
Vitamin B 12		monthly
As		

ASS
Simvastatin
Torasemid

	A	B	C	D	E	F	G
1	SM 2						
2							
3							
4					VISIT 1	Surgery	VISIT 2
5	Date				2002-06-15	2002-06-15	2017-07-15
6							
7	Surgery Parameters	Units				Surgery	
8	Aortic Valve Replacement					yes	
9	MIC					no	
10	Additional Surgery					yes	
11	CABG Surgery					yes	
12	Other Surgery					no	
13	Aortic Valve Size	mm				21	
14	Biological Aortic Valve?					yes	
15	Modell of Aortic Valve					CE Perimount Magna Ease 21mm	
16	Aortic Clamping Time	min				65	
17	Perfusion Time	min				84	
18	Reperfusion Time	min				14	

Figure 2.7: Selected elements to illustrate the UI and specific functions from the SMART IT platform. A) Event registration via frontend. B) Auto-completion of medication options. C) Exported patient data as spreadsheet.

(Figure 2.7C). The principal investigators view on the data is not restricted. A bulk upload through the synchronization client reduces the need for manual entry of data. Additional parameters may be added throughout the course of the study. Manual entry is facilitated through range and type checking of parameters and auto-completion through the UI (Figure 2.7B).

2.2.3 Summary on SMART IT platform

Our proposed architecture contributes by reducing media breaks, using a central computing infrastructure, streamlining research communication by instant notifications, and guaranteeing reproducible research by performing data processing within the central computing infrastructure. We developed the platform on the basis of a deep understanding of the Systems Medicine process in the SMART project. We addressed challenges in basic consortium work on the setting of the SMART project and deliver a profound basis to implement a viable computing infrastructure. The platform serves as the basis to develop and evaluate further research applications as needed in Systems Medicine.

2.3 The human heart and heart valve disease

The human heart provides the whole organism with blood and with it oxygen and nutrients. Blood circulation is ensured through a complex anatomy consisting of two atria and two ventricles separated by the septum and connected to the rest of the body by vessels (Figure 2.8). Blood from the venous system of the body, i.e., low in oxygen, enters the right atrium via the superior vena cava, flows through the tricuspid valve into the right ventricle. At contraction the blood is ejected through the pulmonary valve and pulmonary arteries to the lungs, where it is loaded with oxygen. Coming from the lungs, the blood flows through the pulmonary veins into the left atrium and through the mitral valve into the left ventricle. Again at contraction, the blood is thrown out into the aorta, passing the aortic valve and thus entering a new round through the body. One heart cycle consists of a diastole and systole. During diastole, the heart muscle relaxes and widens to allow enough blood volume to enter the ventricles while pulmonary and aortic valve are closed. Contraction of the heart muscle leads to a reduction of inner ventricle volume and forces blood through the pulmonary and aortic valves, while tricuspid and mitral valves are closed to ensure blood does not flow back towards the lung or venous system.

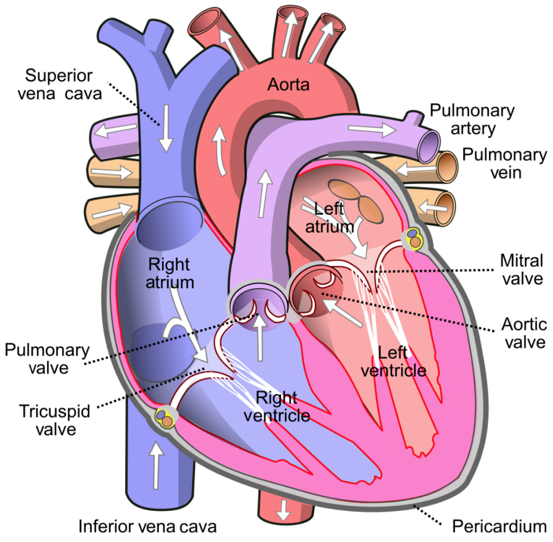


Figure 2.8: Schematic drawing of the human heart with a description of anatomical structures and arrows indicating blood flow through the heart. Created by Wapcaplet in Sodipodi and licensed by GNU Free Documentation License, Version 1.2. `https://en.m.wikipedia.org/wiki/File:Diagram_of_the_human_heart_(cropped).svg` accessed on 10.02.2020

2.3.1 Cellular and molecular setup of myocardial tissue

Myocardial tissue, as in the two ventricles and septum mainly consists of cardiomyocytes, cardiac fibroblasts, vessels and capillaries, and extracellular matrix (Figure 2.9). Cardiomyocytes are responsible for muscle contraction, vessels and capillaries provide the tissue with blood, oxygen, and nutrients, while fibroblasts produce and maintain the extracellular matrix (ECM), which provides structure to the tissue.

Cardiomyocytes and muscle contraction

The outer cell membrane of cardiomyocytes is called sarcolemma and encloses multiple strands of myofibrils interspersed with other cell organelles. Cardiomyocytes are connected via intercalated discs, where adherence junctions and desmosomes are located. Together they form the area composita, which is responsible for force transmission along the

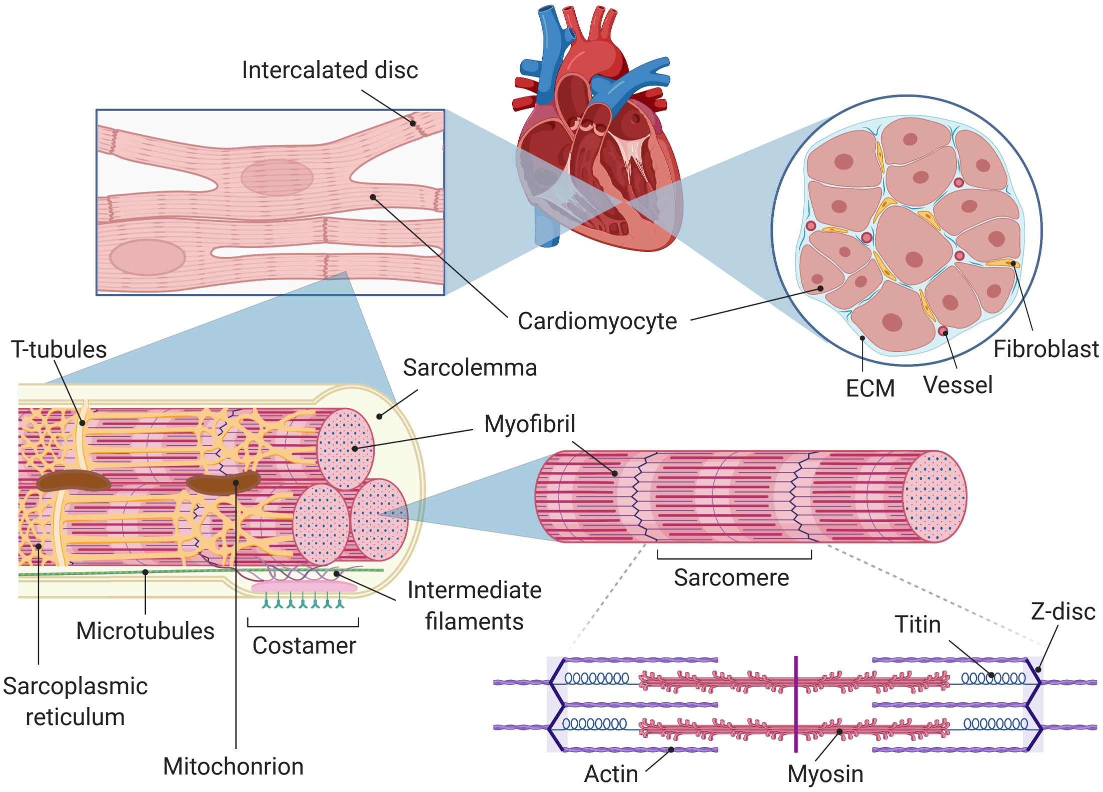


Figure 2.9: Zoom in on the different components of cardiac tissue from the septum. Cellular components are shown in a cross-section. Connected cardiomyocytes are shown in an overview as a starting point to zoom into details on cellular organizationion and myofibril as well as the sarcomere molecular structure.

cardiomyocytes. Costameres represent the point of attachment of cardiomyocytes towards the ECM. Within the cell membrane, adhesion points are linked to the cytoskeleton, which consist of cytoplasmic actin, intermediate filaments, and microtubules. The cytoskeleton plays a crucial role in maintaining cellular stability and reacting to external perturbations through signal transmission and subsequent remodelling. T-tubules are the main route for incoming excitation and are in connection with the sarcoplasmic reticulum, which is responsible for calcium release toward the myofibrils to trigger contraction. Myofibrils consist of multiple sarcomeres aligned in sequence. As such, sarcomeres represent the smallest contractile unit of a myofibril. Sarcomeres are flanked by Z-discs, which serve as anchor points for thick and thin filaments. Thin filaments mainly consist of actin, tropomyosin, troponin, and nebulin. Thick filaments consist of myosin, which is bundled through C-proteins and is anchored to the Z-line via the elastic protein titin. Briefly, muscle contraction is realized through tropomyosin, which is wound along the myosin binding sites on actin covering them up. Troponin acts on increased calcium levels in releasing tropomyosin from the binding sites. Myosin to actin bridges can then be formed, which undergo a conformational change while consuming ATP. Actin and myosin filaments slide together in parallel resulting in a shortening of the sarcomere. Mitochondria are providing the energy necessary for cell metabolism and muscle contraction. The metabolism of normal cardiac tissue is reliant on fatty acid utilization, to a much lesser extent on glucose.

Fibroblasts and ECM

Within cardiac tissue, fibroblasts are the major producers and maintainers of ECM. ECM serves as a scaffold to provide structure to a tissue and provides a reservoir of signalling proteins, which may be released through ECM degradation. Cardiac ECM mainly consist of collagen type I and III, which in addition to conferring tensile strength and elasticity, contribute to the transmission of contractile forces. Furthermore, the ECM contains elastic fibers, fibronectin, proteoglycans, and glycosaminoglycans. Fibroblast transdifferentiation into secretory and contractile cells, termed myofibroblasts, leads to fibrotic remodelling of cardiac tissue. Fibrosis is characterized by an increase of collagen I and III.

ECM, but also sarcomeres are subject to constant turnover, i.e., a fragile homeostasis of degradation and synthesis. Many cardiac pathologies are the result of disturbances in this balance.

2.3.2 Left ventricular heart valve disease

The left ventricle's proper function relies on mitral and aortic valves to close and open appropriately. A valve stenosis denotes improper opening of the valve to ensure blood flow, while valve regurgitation denotes improper closing and thus a leakage of the valve (Figure 2.10).

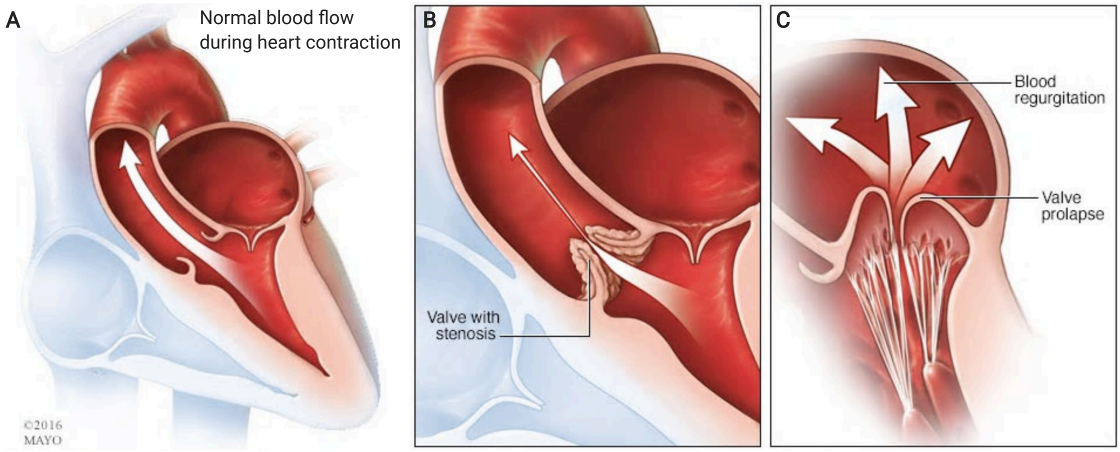


Figure 2.10: Blood flow in the left heart during contraction under normal conditions (A) and in aortic stenosis (B) and mitral regurgitation in (C). Used with permission of Mayo Foundation for Medical Education and Research, all rights reserved.

Regardless of which left ventricle valve is affected, stenosis and regurgitation will result in hemodynamic overload and mechanical stress, which in turn will result in cardiac hypertrophy, i.e. an increase in left ventricular myocardial mass. In the case of mitral regurgitation, hemodynamic stress manifests in volume overload leading to an increase of left ventricle inner diameter and an addition of sarcomere units increasing mainly the length of the cardiomyocyte, i.e., eccentric hypertrophy. In contrast, aortic stenosis leads to concentric hypertrophy, as the pressure overload leads to addition of sarcomeres in length and width, an increase in wall thickness, and a smaller left ventricle inner diameter. Eventually, both conditions may develop towards irreversible heart failure. On the molecular basis, mechanisms that drive pathological hypertrophy of the heart are manifold and are distinct of those driving physiological hypertrophy as response to, e.g., physical exercise [72].

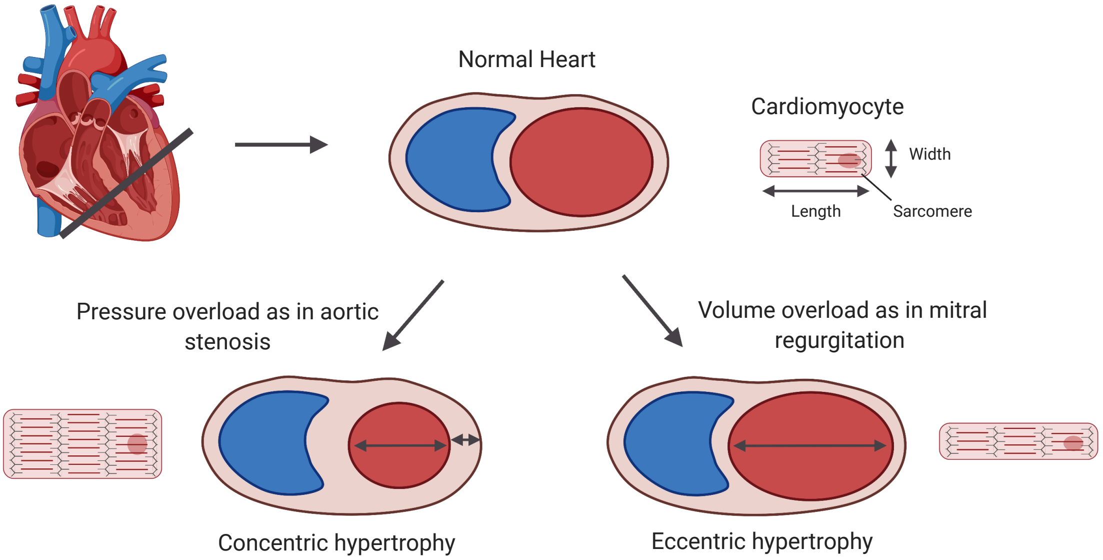


Figure 2.11: Overview on the hypertrophic effects of mechanical stress caused by heart valve diseases. A cross section of the normal heart shows schematic proportions of the left (red) and right ventricle (blue) and a cardiomyocyte with normal length and width in relation to sarcomeres. Aortic stenosis leads to pressure overload and mainly concentric hypertrophy, i.e., a larger width of the ventricle wall and a smaller inner diameter of the left ventricle, through addition of sarcomeres in length and width. In contrast, mitral regurgitation leads to volume overload and thus eccentric hypertrophy, characterized by addition of sarcomeres in length resulting in a larger inner diameter of the left ventricle. Figure inspired by [72]

Aortic valve stenosis

Aortic valve stenosis is the most common valvular heart disease and describes a narrowing of the exit of the left ventricle, such that blood flow is abnormal [30]. AS results in the need of higher pressure generated by the left ventricle during the ejection phase to ensure blood flow through the aorta into the body (pressure overload). The pressure gradient across the aortic valve can increase from few mmHG to more than 100. Therefore, the muscular walls of the left ventricle need to thicken to be able to generate more force, a process called myocardial hypertrophy. The process includes an increase of cardiomyocyte size and protein content, whereas the cells do not necessarily proliferate [29]. In AS, the walls of the ventricle thicken approximately equally, which is known as concentric hypertrophy [30]. At this stage, the mechanical stress induces fibroblast-mediated production of collagen and anti-proteases to avoid ECM degradation. ECM deposition results in stiffening of the ventricle and diastolic dysfunction (or heart failure with preserved ejection fraction HFpEF). The molecular and cellular changes in left ventricular hypertrophy may eventually lead to another remodelling process through ECM degradation including LV dilation (widening) and impaired function which increases the risk for congestive heart failure [73, 74].

Mitral valve regurgitation

In mitral valve regurgitation the mitral valve, i.e., the inlet from the left atrium to the left ventricle, is impaired. In the most common cause of mitral valve insufficiency the valve leaflets do not close properly during the blood ejection phase of the heart. Remodelling or change in dimension of the left ventricle are reasons for the valves to not close properly. As a result, blood will flow back into the left atrium during the ejection phase causing elevated preload. Due to the increased volume, the left ventricle will stretch up to a point at which cross bridges between myosin and actin filaments needed for muscle contraction cannot form properly. The effect is also called volume overload in which contractile efficiency is impaired especially during diastole (diastolic dysfunction).

As we have now established background knowledge on Systems Medicine/Biology, the human heart, and the IT platform that forms the preliminaries to our project work, we may now continue with the description of DE/DA analysis software in Systems Medicine.

3 DE/DA Analysis Software in Systems Medicine

In this chapter, we explain how we addressed the research questions regarding DE/DA analysis software in a Systems Medicine context. In this notion, we elaborate in more detail on the motivation that led us to explore the topic in greater depth. We then introduce existing software for automated DE and DA analysis, which are mainly developed from a Systems Biology perspective. To progress towards Systems Medicine, we identify general software requirements, user groups and personas. The key requirement of flexible experimental design is explained conceptually in greater detail. The concept is implemented in two research applications - DEAME and Eatomics - one tailored to transcriptomic, one to proteomic data, respectively. The concept is evaluated in user interviews and technology acceptance studies. We compare how the two applications differ from and advance over related tools in terms of their functionality. Results are discussed with regard to our stated research questions, their generalizability, and limitations.

3.1 Motivation

The initial motivation of a DE/DA software was ignited by questions and queries stated by clinical scientists of the SMART consortium, as for example the following:

> Are there differences [in gene expression/protein abundance] in [cardiac tissue of] patients with and without cardiac hypertrophy? In relation to indexed myocardial mass in g/BSA? Are the effects different in sex? – *a clinical scientist from the SMART consortium*

This question and the remaining questions in section A.1 neatly exemplify the definition of Systems Medicine. In recent years, many studies, e.g., in the context of Systems Medicine, included a detailed clinical examination of patients, supported by a molecular characterization via omics technologies [11]. Oftentimes these studies have an observational, i.e., a non-interventional character and do not include the effect of an active perturbation, e.g., testing a new drug or therapy in a controlled environment. Thus, effects on the molecular level, e.g., in gene expression, are the result of many *in vivo* factors. These factors may be of interest or can be regarded as confounding factors, such as batch effects or other patient specific clinical parameters, which need to be taken into account when analyzing DE/DA results.

Clinical scientists, i.e., physicians that work in part as a physician but also conduct research on specific patients, observe these *in vivo* factors, such as sex or previous diagnoses, but only have a limited understanding and capability to conduct DE/DA analysis. Contrary, computational biologists have little insights into clinical practice and thus, their research hypotheses are mainly motivated by literature. In order to find and validate a joint research hypothesis the clinical scientist and the computational biologist must interact and communicate efficiently. While the computational biologist has little insights with respect to the patients studied and the resulting hypotheses, the clinical scientist cannot perform the needed computational processes steps on their own.

As such, the wealth of data as assessed in Systems Medicine settings poses new challenges to the analysis. Within medical research, it is of interest how the clinical phenotype relates to the molecular setup of the diseased tissue. To answer the questions, it is required to use the Systems Biology approach – DE/DA analysis. The generic steps needed for DE/DA analysis as detailed in section 2.1.2 are well suited for automation. As a result, the lack of a computational biologist in life science research groups has been addressed by a plethora of tools or platforms that cover many steps of raw data pre-processing and calculation as described in the following.

3.2 Related Work

Related approaches are separated into those specialized on processing transcriptomic and proteomic data. For transcriptomic data, we considered all tools that offer a pre-processing pipeline, quality control, DE calculation and offer a graphical user interface. We considered all analysis applications relying on the popular MaxQuant algorithm output and performing quality control, DA, and enrichment analysis and offer a graphical user interface.

3.2.1 DE tools for transcriptomic HTQ data

Gaur et al. (2017) provide an overview about automated RNAseq analysis platforms and a short description of their utility [75].
The main aim of RAP [37] is to provide an RNAseq tool that does not need to be installed on the client side. The web interface provides possibility for data submission and a browsing facility for results exploration. While the overall appearance seems more user friendly than command line tools, the platform is suited for users with bioinformatics knowledge that are able to configure pipelines and interpret results. Furthermore, RAP offers a great variety of possibilities for analyzing RNAseq data, but does not focus on DE analysis. Especially visualizations and plots are not available so far.

RNAminer [38] provides three different fully parameterized pipelines that work simultaneously while results are consolidated among the pipeline. However, the resulting DE genes are given as text files and any new hypothesis requires an upload of files and a manual specification of two groups of samples offering no flexibility in experimental design.
QuickNGS [39] provides many options to analyze a variety of NGS data. As a tool with no focus on specific use cases, it lacks visualizations and functions that are specific for RNAseq analysis. Plots are limited to a static clustered heatmap and a PCA plot. Additionally, experimental design is static and as described within the publication only possible for two groups (sample and control) plus batch effects. NGS-trex is available through a web

interface. The tool allows the user to pre-process raw data and to calculate DE results [40]. From what can be discerned from the documentation, there are no visualizations available and no information on possible experimental design configuration could be found. TRAP is a web service tailored to analyze RNAseq data from time series experiments [41]. Although TRAP covers all analysis steps, there is no graphical visualization of results. Wolfien et al. (2016) implemented TRAPLINE for automated analysis of RNAseq data, evaluation, and annotation within the Galaxy framework [42, 76]. The TRAPLINE workflow is built to enable experimentalists to analyze data without requiring programming skills [42]. In addition to pre-processing and DE calculation, it provides several lists of results and help or links for visualizing data. Additionally, links to annotation and interpretation tools are given.

3.2.2 DA tools for proteomic HTQ data

Tools for the user-friendly analysis of MS-based proteomics shotgun measurements have been emerging quickly in recent years as the data has become available to a wider research community.

Perseus is one of the first and surely one of the hallmarks of proteomics data analysis platforms, and covers a broad variety of pre-processing and analysis features [43]. Perseus handles sample annotations in a flexible manner, as many annotation types are supported and differential testing can be based on these. While model setup for DA analysis allows multi-group and also a continuous setup, Perseus is written as a stand-alone desktop application in the C# programming language and is limited to the Windows operating system.

In order to be platform independent, similar analysis workflows to perform MS proteomics data analysis are written in R statistical language. Differential Enrichment analysis of Proteomics data (DEP) is an R package that provides an integrated workflow analysis of raw MS data as generated by quantitative analysis such as MaxQuant or IsobarQuant [44]. DEP is tailored to suffice a bioinformaticians analysis workflow and addresses the growing need for user interaction by wrapping the analysis into an R Shiny application.

LFQ Analyst is the most recent addition to Shiny-based applications and wraps many DEP functions into an automated and interactive workflow. The authors show how the use of Limma advances over statistics as used in Perseus [45].

iMetaShiny evolved from the iMetaLab project focusing on metaproteomics analysis, but covers all crucial analysis steps in dashboard configuration instead of a complete workflow [47].

All four R-based applications rely on strictly predefined meta information inputs, which is sufficient mainly for *in vitro* scenarios.

While these approaches are more or less mature, they are designed in the scope of Systems Biology. Thus, current analysis applications offer default solutions for simple research hypotheses, e.g., case vs. control, as they are very common in *in vitro* perturbation studies. More complex designs, as they arise in Systems Medicine, are either not covered in the standard implementation of a tool or require cumbersome reconfiguration and redundant calculations.

As a result, we set out to create a research application that (i) enables the novice user to interactively define and configure complex hypotheses to be tested on HTQ data, and which (ii) automates processing steps while adhering to the scientific standards of best practice procedures to receive publication ready results on DE/DA results.

3.3 Methods

We use a hybrid approach of Design Thinking, scientific software engineering, and literature research to define the DE/DA software for Systems Medicine. We utilize user testing to evaluate specific features of the software with regard to the user's perception and intention to use.

3.3.1 Requirements engineering and feature definition

Many software development methodologies imply that a user already has extensive domain knowledge to guide the requirements engineering process. In our case, we aim at making a scientific method available to users that are new to the field. Therefore, we make use of established methods like Design Thinking and a method specifically tailored to scientific requirements engineering: Li et al. (2011) developed Domain specific ReqUirements Modeling for Scientists (DRUMS) model, which introduces the scientific knowledge into the requirements specification and defines `features` to describe desirable properties that are end-user visible and represent an abstract view of the expected solution. As such, a `feature` serves as an additional preliminary acquisition layer for requirements engineering. A feature can be further refined and be detailed by the realizing requirements, which together form the requirement space [77].

As such, we followed the DRUMS model to outline DE/DA analysis features, the corresponding features spaces and the comprised requirements. The requirement spaces are modelled through the Scientific Computing Requirements Model (SCRM) specification.

When mapped to the DRUMS model, the scientific problem and numerical solution of DE and DA overlap heavily as described in section 2.1.2. As a result, we are able to define many joint features. However, specific requirements tailored to the underlying data sources are needed as well. Requirement subclasses, e.g., processes and data flow, are shaped by best practices and workflow examples reviewed by the scientific community. Conesa et al. (2016) provide an exhaustive overview of processing steps in RNAseq analysis [6]. Similarly, Poplawski et al. (2015) summarize two well-established workflows from Anders et al. (2013) [78] and Trapnell et al. (2012) [70] and additionally define evaluation criteria for a systematic evaluation of user interfaces for RNAseq analysis from a life scientist perspective [4]. The evaluation criteria serve as a starting point for the definition of features, i.e., desirable properties that are end-user visible and represent an abstract view of the expected solution. In the case of proteome measurements, we derived functional requirements on processing pipelines from two publications by Tyanova et al. (2016) [5, 43] and informed by two further publications [79, 80]. Based on our literature review, we compiled a first list of features and requirements.

In order to validate our list of functional requirements, we conducted informal phone interviews with experts that focus on the analysis of RNAseq data and DE analysis from different research institutes. We discussed all steps of the technical pipeline to determine the acceptance of tools within the user community and also assess subjective advantages or shortcomings of selected programs. This step is crucial to not miss important developments in the scientific community. The expert interviews guided the selection of tools in cases where multiple tools suffice the objective quality metric as described in literature.

Furthermore, the new setting of Systems Medicine and many complex hypotheses necessitated the addition of further features. Throughout ideation and development of the application we discussed and evaluated several UI prototypes containing the features within the SMART consortium. The consortium consists, among others, of clinical scientists and computational biologists. The prototypes are based on RNAseq and clinical data as raised within the SMART observational study on heart failure patients. As such, we adapted requirements in an iterative fashion based on user feedback. In this process, we utilized the ideas of Design Thinking, which provide a process framework asking for constant communication between developing team, stakeholders, and targeted end users throughout the software development process [81]; thereby, the user's perspective essentially shapes the system to be of actual value.

3.3.2 Personas

We identified and characterized two stakeholders and thus main users of the application: The **clinical scientist** who is interested in (i) testing own hypotheses based on daily observations and assessed clinical parameters and (ii) interpretation of DE/DA results in the clinical context, e.g., if results point to a disease, a potential treatment or interesting research directions. All of that should not require any programming skills. Furthermore, although the clinical scientist usually has many complex hypotheses towards the data, statistical training to define proper designs is lacking frequently.

The **computational Biologist** is primarily interested in an accurate and fast preprocessing pipeline and calculation of DE/DA results. The execution of the pipeline should require minimum input, configuration and manual tasks. It should allow ad-hoc exploration and analysis of DE/DA experiment results. Furthermore, the computational biologist would like to get publication-ready result reports. While the computational biologist has little insights with respect to the patients studied and the resulting hypotheses, the clinician cannot properly analyze the data alone. Frequently, the clinician has no experience with *omics data and therefore does not know what information can be obtained from it.

3.3.3 User studies

User interviews aim at assessing data to answer research question 2. Testing it with clinical scientists and computational biologists in an observational Systems Medicine setting assessed the utility of the working prototype for exploratory DE analysis. For our analysis, we focus on verifying the achievements of stated software requirements with a focus on the DE experimental design feature. Namely, R11 – Rapid Experimental Design Creation, R13 – Interactive Visualization of Results, and general usability are the primary requirements of interest in our user interviews. The requirements R11 and R13 are covered via the correct completion of given tasks. The actual empowerment of the user, i.e., the ability of conducting an experiment and interpreting the results is concluded from the completion of the given tasks. The users' acceptance of the Differential Expression Analysis Made Easy (DEAME) prototype, i.e., the intention to use, is assessed through The Unified Theory of Acceptance and Use of Technology (UTAUT) questions [82]. We adopted the UTAUT model due to the unique research environment, in which there is less social pressure on the use of a tool and additionally the facilitating conditions are part of the SMART IT infrastructure and thus a new condition in itself. As a result, we limited the model to test the performance expectancy, effort expectancy, and the intention to use DEAME. Assessment of all items is achieved through a mixture of quantitative and qualitative questions which are acquired via a user questionnaire, testing notes filled in by the interviewers during the interviews and a screencast of the application, mouse movements, and voice of the testers following the advice given in Anderson et al. (2010) [83]. Informed consent is obtained from all participants. The testing materials and the interview procedure are pre-tested for functionality, comprehensibility, and time required for response.

Interview procedure

The interviews are estimated to last at maximum one hour. A one-page study description (section A.5) and the consent form (section A.5) are sent out to all testers prior to the interview via email to enable them to read them thoroughly. Both documents are also provided at the interview site to collect signed consent in person. Interviews are conducted in calm rooms at the working sites of testers or at Hasso Plattner Institute (HPI) with stable WiFi connection. Next to the tester, two interviewers are attending the interviews, one moderator, and one assistant. After a short welcome and introduction of the interviewers, the testers are asked to select the language of the interview they would feel most comfortable with to express thoughts and comments while testing the application. Choices are English or German. All other research artefacts are written and filled in in English. As a first action, the testers are given room to ask questions on the consent form and after clarification asked to sign the form and to fill in the first part of the questionnaire. Thereafter, the testers watch a short introductory video[1] explaining the differential expression analysis and are again given the opportunity to ask questions. The interviewers put a special emphasis

[1] The video is available for download at: `https://www.dropbox.com/s/ltaylxl4skez7ep/TutorialVideo_Final.mov?dl=0`.

on questions during testing: Any uncertainties regarding the understanding of the tasks and questions are explicitly asked to be addressed to the interviewers directly, whereas uncertainties of how to use the application should be solved through the usage of the help pages within the application. The testers then read the first task given in the questionnaire and start exploring the application to be able to complete the tasks and answer questions. After the completion of all tasks, there is room for further questions and remarks

Questionnaire

The questionnaire (see appendix section A.5) comprises three parts. In **part I** a general assessment of demographics and background data to characterize the tester population and their fit into user groups is performed. Testers are selected to fit into the definition of the two user groups *clinical scientist* and *computational biologist*. However, the actual criteria for inclusion are broader in the sense that anyone with a university or college course level on gene expression is eligible to be included. Furthermore, the user specifies his/her profession being a clinician/medical expert, a computational biologist, or other, which had to be specified.

Part II: The second part of the questionnaire leads the user to the first task to complete within the application. The task is given as verbalized instructions manner to mimic the actual hypothesis instead of giving technical instructions. Questions II.1-5 ask for analysis results, which can only be answered correctly if specific functions within the application are found, executed, and interpreted correctly. Reasons for incorrect answers are noted by the interviewers or identified retrospectively in the screen casts (details in section 3.3.3). Question II.6 tests the understanding of the design matrix in a backward manner, i.e. if the users can translate the design matrix into a correctly verbalized hypothesis.

Part III tests the translation of a user specified hypothesis. The user is asked to write down their hypothesis and is then asked to explore the setup of the design matrix and explore the results. The interviewers note further details regarding the exploration down. Within Part III, the user also answers the UTAUT questions with a focus on the variables performance expectancy and effort expectancy using six questions for each item. Furthermore the intention to use is assessed in four items. More specifically, tick mark questions in part III.1-14 resemble a Likert scale (-2 = strongly disagree, -1 = disagree, 0 = neutral, 1 = agree, 2 = strongly agree). Questions 1-4 and 9-11 are formulated positively while 5-8 and 12-14 are formulated negatively. For example, “The app is easy to use.” versus "It was difficult to use the app.“. The intention to use is assessed using two more qualitative questions, giving room for explanation of reasons to use the application or not.

Interview notes

The interview notes refer to a guided sheet, in which the interviewers assessed further information to characterize the testing (section A.5). Artefacts from the testing notes include the time needed to complete the tasks, if the tester completed subtasks on their own or if they needed assistance, the design the tester created to compare to the design the tester expressed to want to test and further notes on questions and complications. Furthermore, we captured the actual computation time to rule out network inconsistencies.

Screencasts and audio recordings

Screencasts and audio are recorded using QuickTime Player (Version 10.4 (855)) and are only referred to in cases where the interviewers are not able to directly write down all aspects of the interview, e.g., when the testers are very active or needed help during app exploration.

3.4 Requirements and Features

Based on our literature and user research, we extracted and summarized features and requirements of a DE/DA application as listed in Table 3.1. In the following section we elaborate more on all items in the table and provide greater details for our approach to the DE/DA design feature.

Table 3.1: Overview on features, requirements and their detailed specification for a DE/DA application aimed at enabling analysis in a Systems Medicine setting grouped into categories for better comprehension.

	ID	Item	Specification
System	R1	Straight-forward installation	Time required <15 min and very simple or unnecessary (e.g., all-in-one installer package, feasible just by clicking)
	R2	Platform independence	The software should be available to the major operating systems (Windows, MacOS and Linux)
	R3	Data security	Clinical and expression data, either in a protected remote environment or the user's local environment
Pre-processing	R4	Default Configurations	Very simple/no configuration (e.g. just file path has to be set by clicking), default configurations available
	R5	Full-fledged pre-processing	Major part of pre-processing should be covered
	R6	Acknowledged tools	Tools should meet scientific state-of-the-art
	R7	Independent tools	Bioinformatics tools within the processing pipeline need to be independent from the experimental design
	R8	Automated execution	Pre-configured pipelines should run automatically
DE/DA design	R9	Handling of meta data	The system needs to accept and process an arbitrary amount of meta data
	R10	Intuitive formulation of design	The translation of the clinician's hypothesis into an experimental design matrix needs to be easy
	R11	Rapid design creation	The clinical scientist should be enabled to create a design fast
	R12	Complex designs	Complex designs including continuous variables, stratification and covariate inclusion need to be possible in addition to the simple case vs. control design

Table 3.1: Overview on features, requirements and their detailed specification for a DE/DA application aimed at enabling analysis in a Systems Medicine setting grouped into categories for better comprehension.

	ID	Item	Specification
Visualization, Annotation & Interpretation	R13	Interactive visualization of results	Results of DE calculation are of high dimension and need proper and interactive visualization.
	R14	Enrichment analysis	Enrichment analysis should be used as a key strategy to interpret results
	R15	Actionable information on results	Additional information on DE results need to be provided within the application context, i.e., publications on regulated genes
Wrap-up	R16	Report generation	Detailed report with many intermediate results and graphics
	R17	Data download	Result and intermediate data needs to be accessible in a structured format
	R18	Reproducibility	Processing and calculation should be reproducible in a fully automated/ scriptable way
Overall	R19	Documentation	Documentation needs to be comprehensive, focused and clear, e.g., help pages, tutorials and introductory videos.
	R20	Example and test data	A detailed example and testing data for comprehension needs to be provided
	R21	No IT skills needed	No or very simple, well-documented command line execution

Several non-functional requirements, i.e., criteria that shape and constrain the operation of a system, and therefore the specification of our DE/DA features and functional requirements. A common non-functional requirement is Usability. It is important to note, that "Usability" in the context of software development is mainly defined towards the user UI design [84] and in many cases assumes that the software is already set up and functional, but also that the potential user is familiar with the software's purpose and the usual steps needed to accomplish a specific goal. As such, the UI design is the main contributor to a users success and thus satisfaction. In scientific software this might not be the case, especially when a new user group is supposed to be enabled to perform a task. This need is well reflected in the evaluation criteria defined by Poplawski et al. (2016) [4]. As such, they highlight the need for straightforward installation procedure and represent the Learnability of the scientific problem, it's mathematical model, and the numerical method (not the UI) in multiple items, such as help pages, details on when to use which algorithm and an exhaustive example of usage based on demo data. Additionally, "Usability" is mainly defined as the software being usable, without dedicated IT skills, such as command line execution. In fact, a proper, intuitive UI should expand beyond simply avoiding command line interfaces. Furthermore, scientific standards need to be met in terms of proper documentation of methods, reporting, and reproducibility. The balancing act between a visually appealing representation of data and scientifically correct content needs to be performed. The application's features need to solve the demands as stated by the clinical scientists, yet need to fulfill scientific standard with regard to processing requirements. Starting at step two of the generic process as described in section 2.1.2, necessary features can be mapped to concern bioinformatics pre-processing, DE analysis, visualization and annotation, and interpretation. However, the context of Systems Medicine has been included neither in the development of related platforms, nor in the reviews on current pipelines and user interfaces. The Systems Medicine context defines the new scientific problem, which constraints the definition of the pre-processing pipeline and requires a new concept for the experimental design feature. With regard to annotation and interpretation, there is a remarkable amount of ready-to-use application programming interfaces (APIs) to choose from, which provide results that can directly be visualized by the application in the Systems Medicine context.

3.4.1 Requirement space: Pre-processing

The joint pre-processing requirement space for transcriptomic as well as proteomic raw data is visualized in Figure A.1. The feature's focus lies in the definition of the data handling methods, which differ tremendously between the two data sources and are thus detailed in the implementation parts of the two applications (see section 3.5.1). However, both data flows handle raw measurement values, i.e., FASTQ files or spectra files (.RAW, .wiff, .baf, .dat), a reference file in FASTA format and configuration parameters and are restricted to tools that are independent from the experimental design and can be executed in an automated fashion, e.g., by a job execution/scheduling framework. The UI needs to provide an interface to allow data upload, configure pipeline parameters and to view quality control plots.

Hardware and performance definitions are hard to specify – Poplawski et al. (2016) [4] define, e.g., a very good run time for their test data set to be below 48 hours. Furthermore, they reduce hardware claims to main memory and disc space. A "good" solution would be remote data handling and analysis and thus requiring no local hardware. However, the external handling counteracts the definition for data safety, which is best ensured in the local environment. Both assumptions are not necessarily true. We thus define our requirements in this regard with a setup on either a safe remote environment, e.g., at institutional servers with secure data transfer, to allow good performance of heavy computations as they are common in raw data pre-processing or a local desktop installation, which most probably would not accomplish computations in the given time frame, but is under full control of the user.

3.4.2 Requirement space: DE/DA in Systems Medicine

The SCRM model to define the DE/DA in Systems Medicine requirement space shown in Figure 3.1 contextualizes requirements listed in Table 3.1 with the corresponding feature and scientific knowledge space. Although the knowledge space already provides numerical methods for the solution of the DE/DA in Systems Medicine problem, a proper configuration of these methods is neglected and reduced to the most basic solution in most published applications. We therefore specify a new data process space including requirements with regard to data flow and data definition (lower right hand side in Figure 3.1). Here, the data flow is expanded to accept the full matrix Xf in addition to the HTQ measurement matrix Y. The data process space is explained in greater detail in the next paragraph.

All metadata needs to be accessible to the user for cohort definition and design setup from the user interface. The translation of the users hypothesis into an experimental design matrix needs to be easy and fast. As many established solutions rely on no ad-hoc or only static visualization of results, we required our solution to visualization of intermediate steps for quality control and for result inspection to be manifold and interactive.

The Systems Medicine DE/DA feature requires a software interface to the precedent bioinformatics pre-processing as it is reliant on the calculated HTQ measurements. Furthermore, the feature provides the results to be used for annotation and interpretation. Hardware is not necessarily a crucial factor to consider in this feature, as the computational load is rather limited. However, computational power, and parallelization is needed for the pre-processing pipeline and in the case of many client requests to the application. The most popular numerical solutions to DE/DA calculation are implemented in R statistical language, restricting the feature language to be R or to be able to execute R code.

Concept of flexible experimental design matrix creation

The definitions and concepts of GLM and how to calculate DE or DA are laid out in detail in section 2.1.2. They provide the basis for the flexible design matrix creation feature.

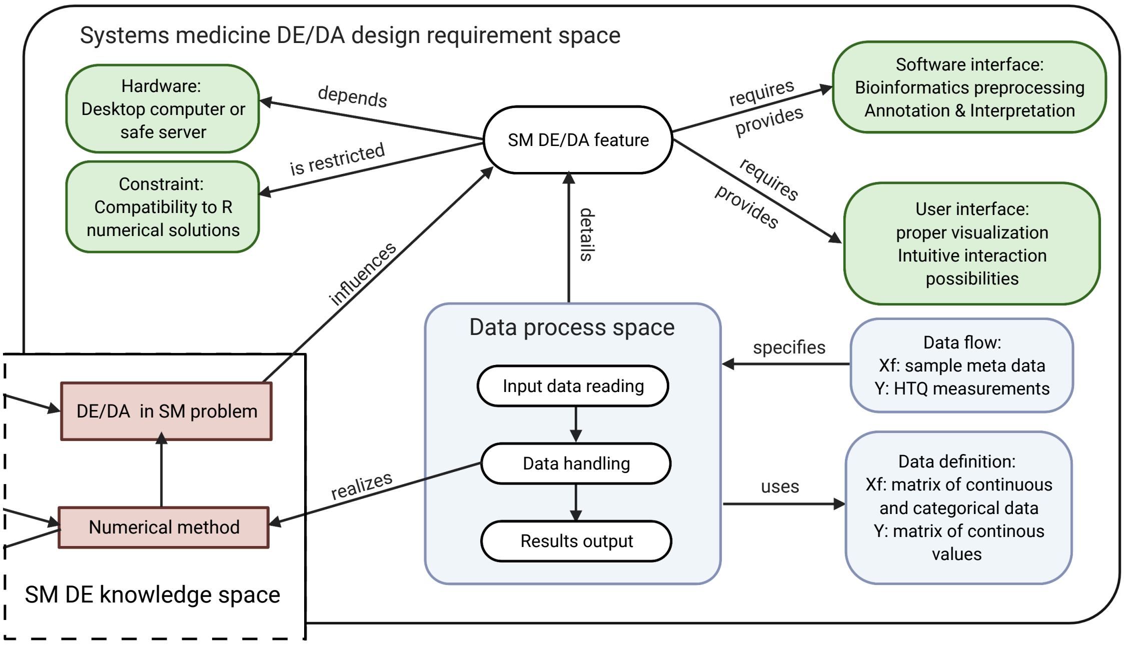

Systems medicine DE/DA design requirement space
Hardware: Desktop computer or safe server
Constraint: Compatibility to R numerical solutions
depends
is restricted
SM DE/DA feature
requires
provides
Software interface: Bioinformatics preprocessing Annotation & Interpretation
requires
provides
User interface: proper visualization Intuitive interaction possibilities
details
influences
Data process space
Input data reading
Data handling
Results output
specifies
Data flow: Xf: sample meta data Y: HTQ measurements
uses
Data definition: Xf: matrix of continuous and categorical data Y: matrix of continous values
realizes
DE/DA in SM problem
Numerical method
SM DE knowledge space

Figure 3.1: SCRM diagram of the new requirement space of the DE/DA in Systems Medicine (SM) analysis feature. The model contextualizes requirements listed in Table 3.1 with the corresponding feature and scientific knowledge space.

We defined matrix Y as being the result of an HTQ experiment and Xf being the meta information accompanying the experiment. In terms of the DRUMS specification, Xf and the design matrix represent the new data flow. Parameters from Xf define the sample cohort and are used to create the experimental design matrix X that corresponds to the hypothesis of interest. The parameters from Xf may be of numeric or categorical nature and may exceed the parameters needed for a specific design. The flexibility of the theoretical possibility to include all parameters given in the meta information Xf into the design matrix X is encompassed by some disadvantages:

1. Many experimental setups may not need a complex design – simple relationships are to be tested,
2. A complex design complicates interpretation of results – sometimes dramatically – and thus would not be suitable for our user group and
3. A complex design with many stratification options tends to produce many small-sample-number groups, which do not have the statistical power to detect significant differences.

Therefore, we introduced the constraint of allowing **a maximum of two parameters of Xf needing to define the cohort and contrast of interest**. We define the two parameters as Xf_1 and Xf_2. The constraint enables us to reduce the possibilities of design formula setup dramatically while avoiding biologically less relevant or unreasonable configurations.

The design formula is generated by (i) selecting and modifying Xf_1 and Xf_2 to build the `mainParameter` and the `filterParameter` as illustrated in Figure 3.2 and, if needed, (ii) filtering of samples and adding a reasonable number of covariates, if needed (see Figure 3.3). Possible initial data types for Xf_1 and Xf_2 and examples of their modification to are shown in Table 3.2. Additionally, all categorical parameters and thus also dichotomized representations of numeric parameters are represented in two or `levels`. For example, a `level` of the parameter Sex is male. Modification in this context means either dichotomization of the parameter or selection levels of interest. The design matrix X is derived from the formula and the specification of contrast. We provide a detailed description of possible ways through the algorithm shown in Figure 3.2 and Figure 3.3, i.e., the process requirement, in the following paragraphs.

Table 3.2: Description of possible initial data types for Xf_1 and Xf_2, a full range example of a parameter, and a corresponding binary representation.

Data type	**Parameter** **Full range example**	**Binary representation/`level`** **Example**

Categorical binary	Sex
	Male/Female
		Male = all male patients
		Female = all female patients

Numerical	Age
	0-90 years
		Below_x = [0 - x)
		AboveAnd_x = [x - 90]

Categorical exclusive	Blood group
	A, B, AB, 0
		Blood_1 = A, B, AB
		Blood_2 = 0

Case 1: Numeric $\mathbf{Xf_1}$

Xf_1 being and staying numeric throughout the modification process towards `mainParameter` means that a linear relationship between the expression or abundance value and the numeric parameter, e.g., the age, is to be tested. At this point `mainParameter` may be fully defined as equivalent to Xf_1 and `filterParameter` as being 0 (follow the left-most path in Figure 3.2). Further, specification of the cohort can be accomplished through specification of Xf_2, which represents a filter. Thus Xf_2 must be categorical, however a numeric Xf_2 may be dichotomized (= transformed to a binary representation) by selecting a cut-off value to divide the cohort into two parts. `filterParameter` is then equivalent to Xf_2.

Case 2: Categorical $\mathbf{Xf_1}$

In the case of Xf_1 being categorical before or after modification (= `mainParameter`) a linear relationship between two `levels` of the `mainParameter` is assumed. While `levels` are natural in the case of binary data, categorical parameters with more than two `levels` require either a merge of several `levels` to form a new `level`. If no further stratification of the cohort is needed, `mainParameter` and `filterParameter` are fully defined. Choosing to stratify, enables the selection of Xf_2, which again needs to be dichotomized based on a cut-off value in order to then be united with Xf_1 to form a `mainParameter` comprising all possible combinations of `levels` from Xf_1 and Xf_2. In this case `filterParameter` is always defined as being zero. The difference between the merge loop in the beginning and the final unite operation is the variable it relates to. The merge loop relates to a combination of different levels of Xf_1 into one new level, the unite function combines levels from Xf_1 and Xf_2.

Completing the formula

In Figure 3.3 we show the completion of the formula after the previous definition of the `mainParameter` and the `filterParameter`. In the numeric case, the cohort needs reduction to those samples within the selected filter groups. In addition to `mainParameter`, p covariates may be selected from Xf to be added to the formula. Samples with missing data for `mainParameter` or within the covariates are removed from the cohort at this point and the formulas are assembled to $\sim 1 + mainParameter + \sum_{l=0}^{p} C$ in the case of `mainParameter` being numeric and $\sim 0 + mainParameter + \sum_{l=0}^{p} C$ in the case of `mainParameter` being categorical. The latter is mainly necessary for easy sorting and definition of contrast.

Contrast specification and design matrix translation

The formula is translated into a design matrix, which in the case of a categorical `mainParameter` will result in the creation of a dummy variable for every `level` of `mainParameter`. For covariates the amount of dummy variables can be reduced to one less than the amount of `levels` to achieve a full rank design matrix. The last step is the definition of contrasts, i.e., which groups to calculate the difference of *beta* coefficients on. In the case of a numerical `mainParameter`, no dedicated contrast needs to be specified as the second coefficient $beta_1$ of the linear model represents the change of the abundance in relation to one unit of Xf_1.

3.4.3 Requirement space: Annotation and interpretation

Enrichment analysis is one of the most common and established methods to summarize results into comprehensible biological terms. The analysis is based on lists of genes, sometimes accompanied by a numerical value indicating a ranking as resulting from previous analysis steps and lists of genes as provided by annotation databases. Actionable information could be achieved through an interface to a document retrieval system including natural language processing. The data flow would then consist of key words, e.g., the gene name of interest, or of full queries representing the context of analysis as given by the analysis results and clinical meta information. Additionally, interviewed clinical scientists asked specifically for actionable, additional information on results to be provided within the application context, i.e., publications on regulated genes. Both requests do not needs adaption to the Systems Medicine context. Therefore, either implementation of current libraries or the usage of an application programming interface would be reasonable requirements. In any case, visualization of results and intuitive interaction possibilities need to be realized through the user interface.

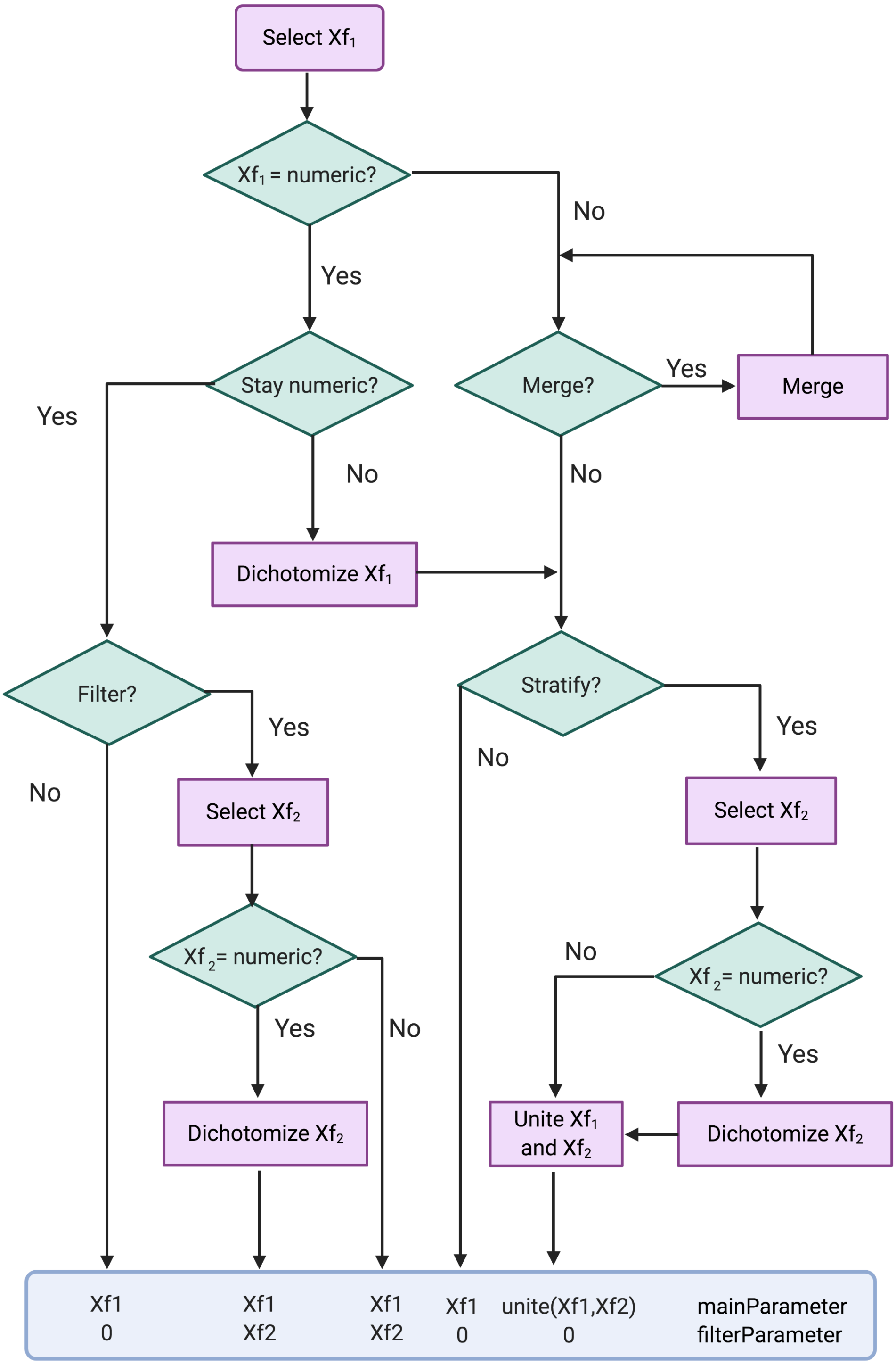


Figure 3.2: Schematic representation of model parameter selection and modification. The selection of Xf_1 and Xf_2 in dependency to their data type being and staying numeric or not in combination with other modification option like merging and dichotomizing leads to a fully defined `mainParameter` and `filterParameter`.

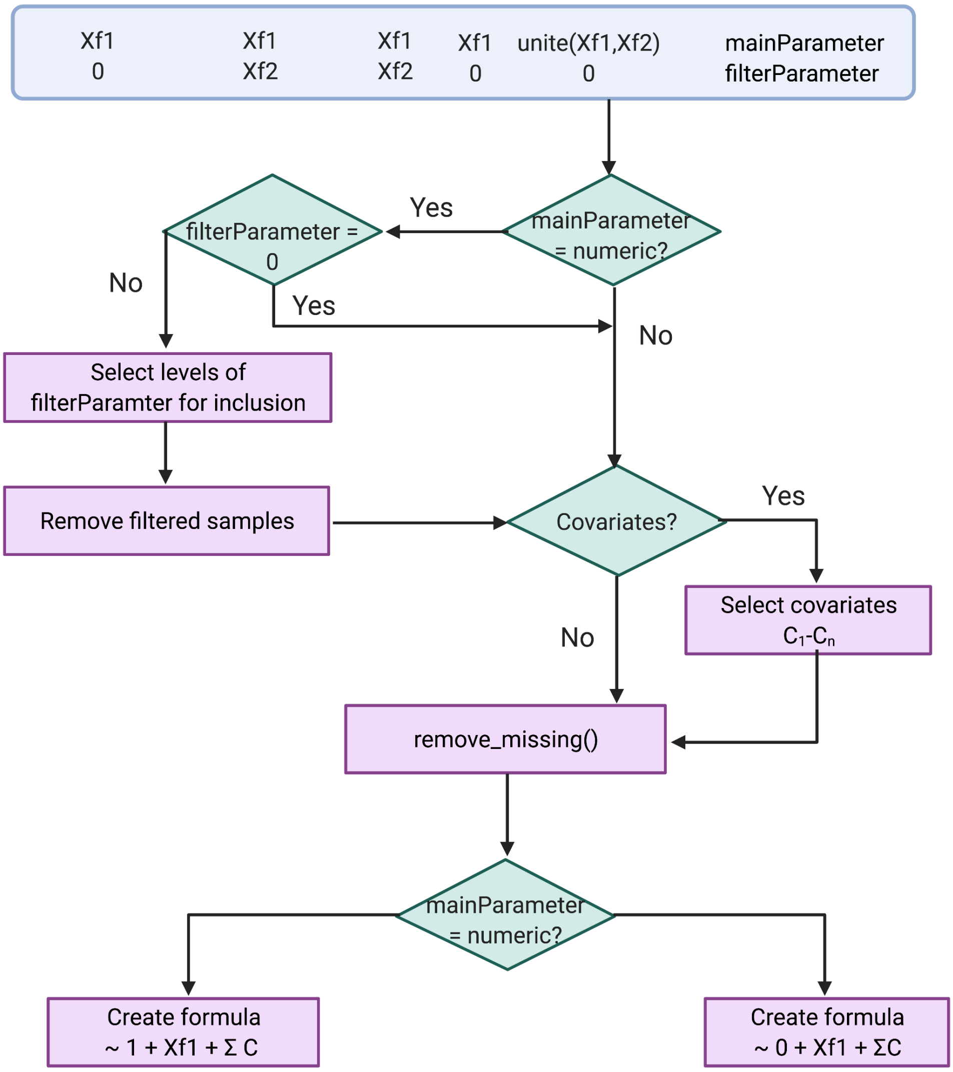


Figure 3.3: Flowchart showing how `mainParameter`, `filterParameter` and further specifications are used to specify, reduce, and complete the definition of the design formula.

3.5 Implementation

Throughout the iterative development process, we constructed several mockups and prototypes to reach two viable research applications: DEAME for DE and Eatomics for DA analysis. While both applications conceptually are implementations of the same software requirements and especially the concept of rapid and flexible experimental design setup, they differ in a few regards.Table 3.3 summarizes the main differences in application usage and implementation details.

The primary HTQ data source analyzed in DEAME is RNAseq data. As DEAME is tailored to be used within the SMART consortium work, active development and code availability are restricted to the SMART project definition. Implementation decisions are considering the framework of the SMART IT platform [53] and in a broader sense also utilized existing software artefacts from the AnalyzeGenomes platform [69].

As the concept of rapid and flexible experimental design creation proved to be well accepted, we decided to transfer our insights to the implementation of Eatomics, and thus to make code available publicly and usable for all researchers.

We switch the primary source of HTQ data to come from label-free MS-based shotgun proteomics, because other parties are working on very mature RNAseq platforms with large man power (e.g., Chipster). Moreover, proteome data analysis applications are sparse and as the data source becomes more available the need is growing. Furthermore, despite the end of the SMART project, the proteomic data assessed until then did not undergo proper analysis yet and thus Eatomics would benefit the project retrospectively. Details on how we implement the two applications are given in the following sections.

Table 3.3: Differences and capabilities of the two research applications DEAME and Eatomics.

	DEAME	Eatomics
Tool availability	SMART/EurValve project members with data access	free
Code availability	private	public
Access via	SMART IT platform, AnalyzeGenomes	R studio, Shiny
Data storage	in-memory database	file-based
Backend	R serve, python, ruby	R Shiny modules, R helper functions
Frontend	React	R Shiny
Primary HTQ data source	RNAseq	shot-gun, label-free proteomics

3.5.1 DEAME

The DEAME application is part of the SMART IT platform described in subsection 2.2.1 and in Kraus et al. (2017) [53] and uses resources, such as the worker framework, provided by the AnalyzeGenomes platform [69]. In Figure 3.4, the overall software architecture of the DEAME application and relevant parts of the SMART platform are modelled using Fundamental Modeling Concepts (FMC) notation. Our React front-end communicates with the Python back-end via a RESTful API implemented with Flask. The back-end has a connection to the In-Memory Database that contains data and can execute R scripts in form of stored procedures. The pre-processing pipeline is executed through the AnalyzeGenomes worker framework and stores results in the In-Memory Database. A thorough explanation of all components, i.e., the data, platform, and application layers, is given in the following sections.

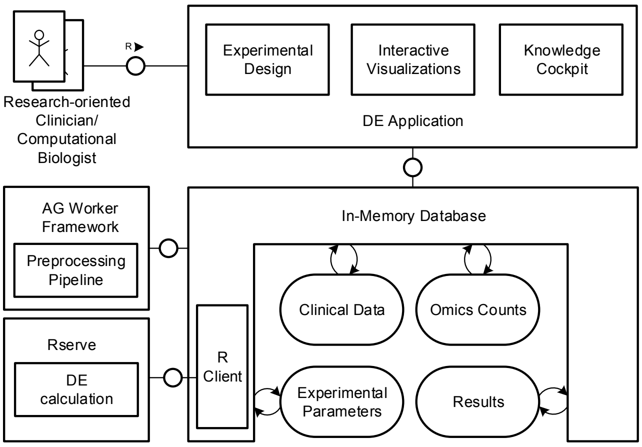


Figure 3.4: Software system architecture of the DEAME application including parts provided by the SMART and AnalyzeGenomes IT infrastructures [53, 69].

Data Layer

An In-Memory Database contains all frequently accessed data: The patient-centric star schema of the SMART platform is expanded with a section to accommodate the experimental data (please refer to [53] for further details on the clinical data). Tables for counts, as they are produced within the pre-processing, are added as well as tables for experimental parameters, and results of DE calculation. Furthermore, an R client is established to perform DE calculation within an Rserve instance.

Platform Layer

The platform layer contains the pre-processing pipeline, experimental setup information, and DE calculation functionality. The split into pre-processing and experimental design plus DE calculation is a design decision that limited the selection of tools to be used within the pipeline when compared to the traditional setup as outlined in section 2.1.2. The split resembles the need given within a clinical setting, where many hypotheses may be tested and thus, the experimental design for DE calculation is not known before pre-processing of raw data. As a result, pre-processing and DE calculation are independent from each other.

Pre-processing pipeline In our architecture the pre-processing is embedded within the worker framework of AnalyzeGenomes. In Figure 3.5 we describe the pipeline, input, and output of the individual steps and the order in which they are executed. The boxes represent applications, i.e., python wrappers around the incorporated bioinformatics tools, such as TopHat. These programs could be extended and interchanged when new tools need to be introduced. In our literature review and after interviewing experts in the field, we identified the following tools to be suitable for our first prototype: FastQC [85] for quality control before and after trimming of reads with trimmomatic [86], Tophat [87] or STAR [88] for alignment of reads to the reference genome, and featureCounts [89] for creating count tables from alignment files. In this setup, we avoid redundant pre-processing, as it is implemented in some of the related tools.

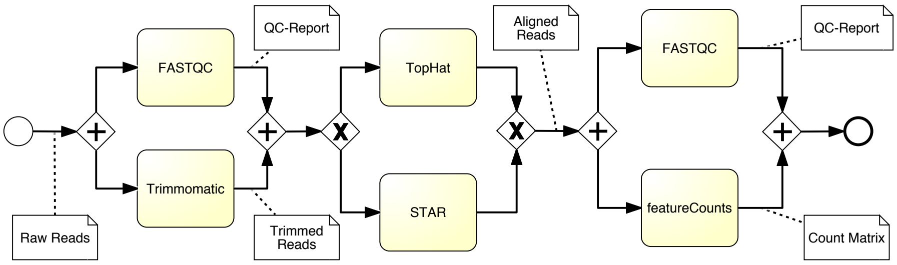


Figure 3.5: Specific implementation of the automatic RNAseq pre-processing pipeline to yield count matrices.

Interactive Visualization and Annotation Many results and intermediate results are of interest for both the clinician and computational biologist. Quality control as done by FastQC produces an HTML-file for every sample, which is stored and accessed for display within the application. Additionally, results from DESeq2, i.e., the list of DE genes, their p-values, and the complete normalized and transformed count matrix, are visualized. Interactive heatmaps are implemented via the Clustergrammer software and its biology-specific extensions to show gene/protein names, cluster statistics and GSEA [90]. Further plots are implemented using the D3 JavaScript library.

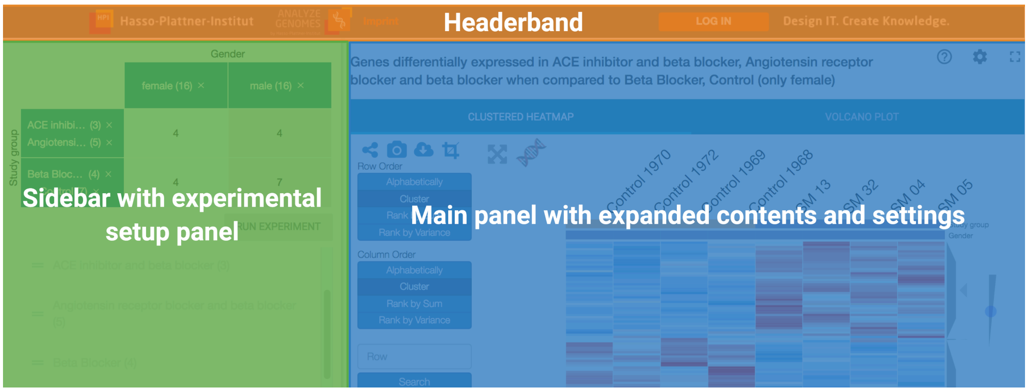
Differential expression calculation Differential expression calculation follows the general pattern as described in section 3.4.2. However, there are specific adaptions: For DEAME we did not implement the continuous case for Xf_1 and Xf_2. As a result, continuous parameters always require dichotomization. Furthermore, we did not implement the option to add covariates to the design formula in the DEAME prototype. Clinical parameters corresponding to Xf and the expression count matrix (Y) are stored within the database for reproducibility.

DE calculation is done via DESeq2 [9] within our Rserve instance. DESeq2 is called from a stored procedure within our in-memory database and requires the raw count table as generated by our pre-processing pipeline. Furthermore, the stored procedure also receives Xf_1 and Xf_2 for filtered samples and unites them to represent the `mainParameter`. The `mainParameter` is translated into the design formula. All reasonable contrasts, i.e., $level_2 - level_1$ in the binary case and $level_2 - level_1$, $level_3 - level_1$, $level_2 - level_4$, $level_3 - level_4$ in the categorical case are calculated and send back to the database.

Application Layer

Users can access the DEAME front-end using a web browser. To built the web application we used React, which is a JavaScript library that controls the mounting and rendering of components. Additionally, we use Redux to manage the application state. The styling is mostly defined by React Material-UI and custom styles using the styled components module. React is a JavaScript library that controls the mounting and rendering of components. Each component implements different life cycle methods that can manipulate its state. Information between components is passed as `props`, whereby one-way data binding is enforced. Redux introduces a global state that acts as a single source of truth, to which every component can connect. A component can dispatch an action that possibly makes a back-end request and propagates the desired change to reducers. Reducers are pure functions that return the new state or state slice, which is the part of the state they are responsible for. Components are notified about changes that are accessible in their props.

DEAME's UI consists of three parts: the experimental design panel is located in the sidebar, a visualization panel, and a knowledge panel are situated in the main panel.

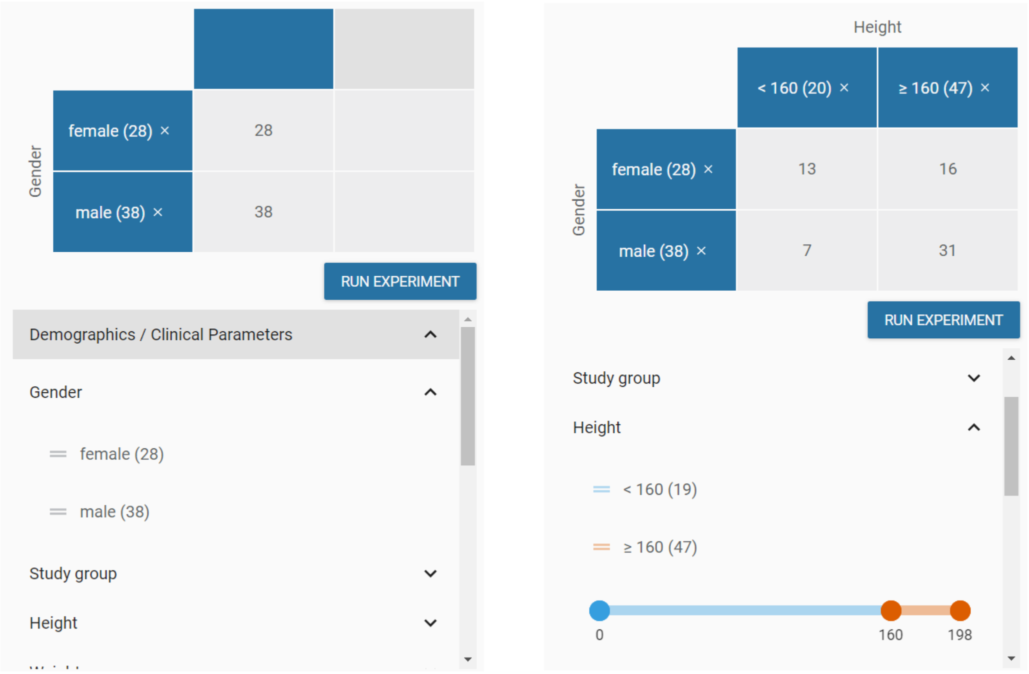


Genes differentially expressed in ACE inhibitor and beta blocker, Angiotensin receptor blocker and beta blocker when compared to Beta Blocker, Control (only female)

Figure 3.6: Overview of DEAME's UI consisting of a sidebar harbouring the experimental design setup (left) and a main panel showing the visualization panel (right).

Experimental Design Panel The experimental design panel is the main part of the application as it enables to dynamically choose interesting clinical patient data categories to be studied in DE analysis (Figure 3.7).

The overall goal is to split the patient population into at least two subgroups based on the patients' characteristics. For demonstration purposes, we use data from the SMART study. Subjects are characterized by a plethora of clinical variables (e.g., sex, height, blood pressure) that are grouped in categories (e.g., demographics or MRI derived measurements). Binary and categorical variables can be dragged into the design matrix directly. Categorical



Gender		
female (28) ×	28	
male (38) ×	38	

Gender	Height: < 160 (20) ×	Height: ≥ 160 (47) ×
female (28) ×	13	16
male (38) ×	7	31

Figure 3.7: User interface of the experimental design panel in two configured versions. Left: The parameter `Gender` is expanded to show the available `levels`. The `levels` are dragged from the list of available parameters into the upper matrix and as such resemble a proper design. The numbers in brackets denote the count of subjects in belonging to the respective level. Right: An additional numeric parameter `Height` is expanded, a threshold for dichotomization is set to 160 and is used to split the previously two groups of female/male into four groups based on the subjects gender and height.

variables may be combined within one column of the design matrix. Continuous variables are split by the user via a slider over the full range of possible values. The design matrix displays the parameters and `levels` and calculates group sizes similar to a contingency table. After the creation of a valid design, i.e., at least three samples in every group, DE calculation is triggered via the `Run Experiment` button. Technically, the sidebar is an own React component that includes the matrix and categories components. Categories are a list of category components. A category is in turn a list of parameter components, while a parameter either is a `CategoricParameter` or a `NumericParameter` that contains a `NumberSlider`. A parameter contains a list of multiple `level` components. A `level` can be dragged and dropped into the `HeaderCells` of the matrix, implemented using React DND.

Visualization Panel The main content is a collection of `Components`, mainly the `Header`, `Diagrams`, `Settings`, and the `Footer`. The `Header` provides options, such as showing settings and switching to full-screen mode and back. The `Diagrams` component receives the experiment and quality control data and renders different diagrams that are selectable from tabs, currently either `ClusteredHeatmap` using Clustergrammer or `VolcanoPlot` using canvasJS, which shows to be performant with many data points as given in DE analysis. A hover over gene names displays a short description. The title that describes which contrast is currently shown is added above the diagrams. The `Settings` component that can be opened from the `Header` enables the user to adapt different significance thresholds, which is then reflected in the diagrams. Currently, download of results is restricted to the capabilities of the Clustergrammer library.

Knowledge Panel Especially the clinical scientists need additional external information on analysis results. While the Clustergrammer library already supports the retrieval of gene product definitions, a more complex knowledge panel is envisioned. Instead of querying for mere gene names, further relationships, e.g., effects of up-regulation or the selected disease context, should be included in the query to find actionable insights. Examples for external resources that can be leveraged are search engines such as Olelo [91] for intelligent PubMed [92] queries or DisGeNet [93] for gene-disease associations.

3.5.2 Eatomics

Eatomics is designed to enable users from a wide range of backgrounds to comfortably perform (i) quality control of MaxQuant-generated proteomics data and (ii) differential abundance analysis comparing the difference between any clinical observation of choice. In this section we provide an overview on how persistent and session data is handled, how user interface objects and the server function interact according to Shiny's reactive programming model and the appearance of Eatomic's UI.

Data Layer

The application is primarily developed as a standalone R Shiny application that can be launched from RStudio locally or a Shiny Server. Data is loaded from text files at run time and may persist in the form of downloadable data tables and the report. To make session data available between different application components, we made heavy use of Shiny reactive values objects that essentially are single variables or lists of objects calculated, refreshed, and stored throughout the analysis; for example, the PCA plot is build on the first tab panel, then stored as a reactive value and can then be accessed from the download handler and the report generation function. At the end of a session all temporary data is deleted.

Application logic

R Shiny applications consist of a UI object and a server function. Both are handed over to the `shinyApp()` function which creates the application out of the given UI/server pair. Reactive sources, such as user input, can be used within reactive conductors, i.e., functions to calculate results based on the input, and will then inform output objects to be rendered to the user. In addition to the main application components, there are several modules and helper functions. Modules are application components, which are reused within the application several times. They follow the structure of UI/server function pairs and consist of reactive elements. For example, the user may want to change text elements, such as title, subtitle, and caption of a plot before download. The module collects all text inputs and returns them for application to the plot element. The module is reused for every diagram. Helper functions are called from within the main application's server function and do not require own UI components. Examples are the assembly of plots or data transformation tasks.

Visualization For the majority of plots, we exploited the great flexibility of the ggplot2 and adjunct libraries. Manipulation on plots and diagrams, e.g., a change in title or subtitle, only requires a new definition of the text layer instead of recalculating the whole plot. For the volcano plot we also made use of the plotly library, which wraps accessory data for further interaction in plots in the UI. For example, as the gene name can be retrieved via hovering over the data point, there is no need to put labels and thus a lot of noise into the plot.

Model setup DA analysis is the key component in Eatomics and requires the translation of a given research hypothesis into a model that well describes the data and a difference of interest, i.e., the contrast. For Eatomics, we re-iterated on and consequently re-implemented the model setup as described in section 3.4.2 and applied the logic to be used twice: The experimental setup module is used to define DA and also to find differences in gene set and pathway enrichment on the basis of single sample (ss) enrichment scores. ssGSEA is an extension of conventional GSEA [61, 94]. Each ssES represents the degree to which the genes, i.e., in our context the respective gene product, in a particular gene set are coordinately up- or down-regulated within a single sample. As such, ssESs represent a transformation for a given protein abundance data set. Therefore, ssES are equivalent to an HTQ experiment and may serve as input Y.

In Eatomics, the flow of model setup deviates from the one shown in Figure 3.2 in that we have not implemented a dedicated merge function. However, it is possible to merge groups by exploiting the unite function, i.e., select $Xf_2 = Xf_1$, and then select the groups of interest from the resulting combinations of groups. As such, there is only one contrast defined per model fit. By omitting the merge step and introducing only one contrast per calculation run, we do not have to filter samples from the model matrix, which results in the inclusion of all measured samples into the design even if they are not in the specified contrast. As a result, variance trend correction can be used for a finer estimation of variance of expression across all samples and the calculation of differences of multi-group comparisons. However, this is not true for continuous explanatory variables. Samples missing meta information of the selected variable are excluded. With regard to the results, contrasts are generated automatically, based on the user-defined groups of interest. In practice, the first selected group resembles the reference `level`, while the second group will be subtracted to calculate fold changes.

For the core of differential abundance calculation, we decided to use the R package

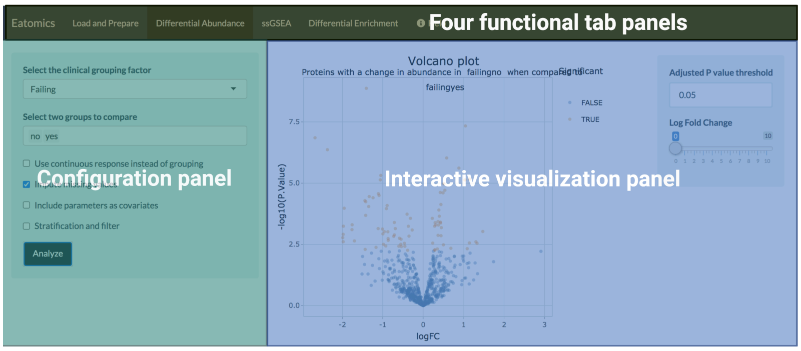
Limma [8], which is widely used in the differential expression analysis of microarray and RNAseq data. As stated before, Limma uses generalized linear models to calculate the relationship of expression values and explanatory variables. As such, it comes with a dedicated capability of modelling continuous parameters and thus it is perfectly suited for a wide range of explanatory variables. Although Limma is developed to perform well on gene expression data, it has been shown to be superior over traditional approaches on quantitative proteomics data as well [95, 96].

User interface

The Eatomics user interface is structured according to the four functional units:

1. **Load and prepare** sample metadata and MaxQuant output, as well as perform quality control.
2. Conduct **differential abundance** analysis.
3. Calculate enrichment scores per sample using **ssGSEA**.
4. Conduct **differential enrichment** analysis.

These functional units are represented in the four tab panels In general, all tab panels enable to set overall analysis parameters in a left hand sidebar panel, while further plotting parameters can be set and interactive visualization can be conducted in the main panel (Figure 3.8). A fifth tab panel holds help pages and a detailed step-by-step tutorial using a demo data set.



Select the clinical grouping factor: Failing

Select two groups to compare: no yes

☐ Use continuous response instead of grouping

☑ Impute missing values

☐ Include parameters as covariates

☐ Stratification and filter

Adjusted P value threshold: 0.05

Log Fold Change: 0 – 10

Figure 3.8: General setup of Eatomics' UI. Four tab panels establish the core functionalities for the application. Every tab panel is structured into a left hand side, which contains configuration modules and a main panel for interactive visualization of results. The user can navigate between functional tab panels through the header band.

Eatomics process flow and user interaction

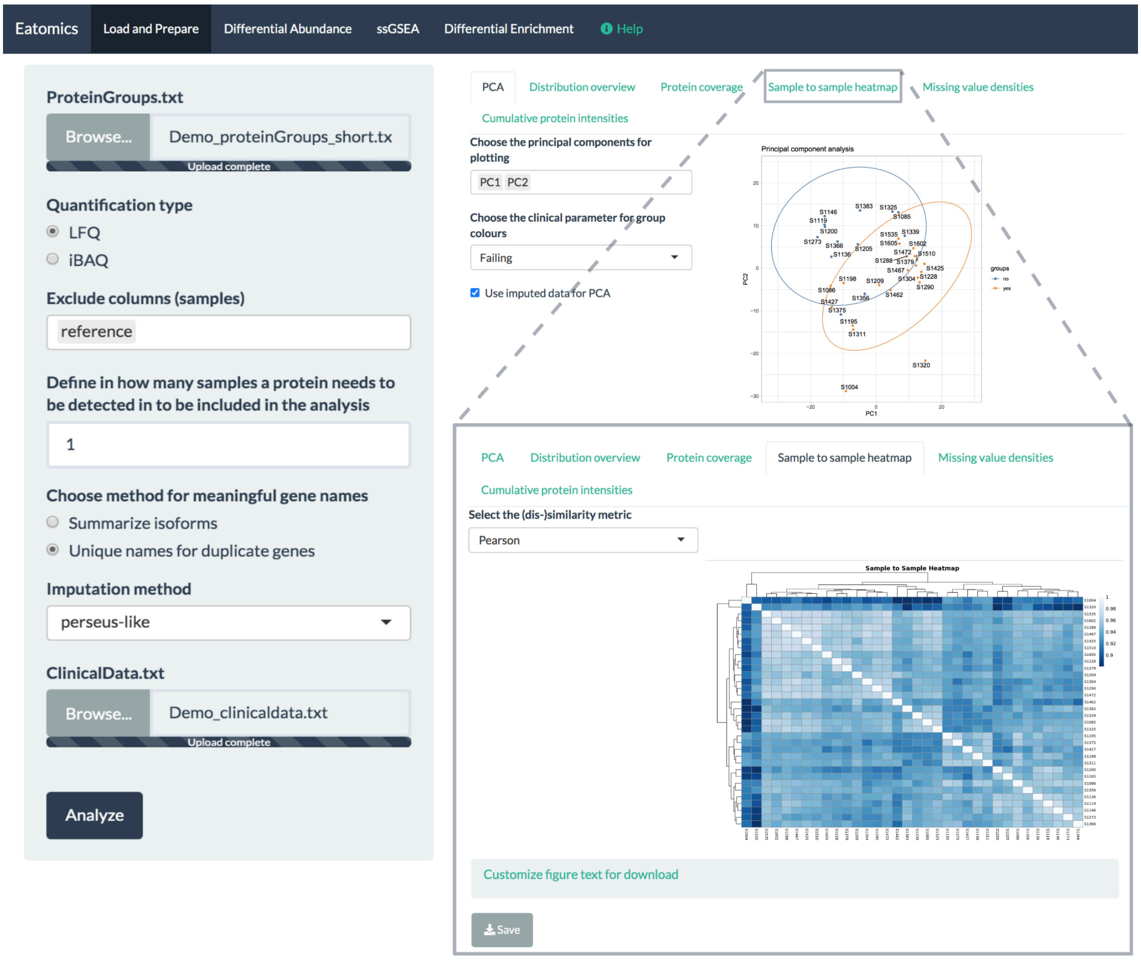
In this paragraph we want to give a detailed impression on the processes in and user interaction with Eatomics' first two tab panels. We only briefly describe the other two, as the third on ssGSEA calculation is mainly an interface to configure the ssGSEA algorithm embedded in our application and the fourth is an instance of the second tab panel with only minor changes related to the new input data.

In Figure 3.10 we modelled the process and message flow on data loading, pre-processing, and performing quality control in BPMN 2.0. A screen shot of the UI showing the first panel is displayed in Figure 3.9 to accompany the process flow. The first part of the process relates to the load and configure side bar panel of the application, while the expanded sub-processes depict view, manipulation, and download of plots and diagrams. The user first selects the protein evidence (`proteinGroups`) file, which corresponds to the standard output of the MaxQuant algorithm. It contains at least label-free quantification (LFQ) intensities or intensity Based Absolute Quantification (iBAQ) values and other information related to the MS measurements and the algorithms annotations. The file selection triggers the load function and the provided data is scanned for the "Reverse", "Potential contaminant", and the "Only identified by site" column per default. Rows containing a "+" in these columns are removed as they should not be considered in the analysis. Further filtering requires user interaction: LFQ or iBAQ columns together with the columns `ProteinIDs`, `Majority ProteinIDs`, and `Gene names` are extracted. Samples can be excluded by name, or a threshold for a minimum of valid values measured per protein can be set to exclude proteinGroups not meeting a certain coverage. In the case of multiple `proteinGroup` entries per gene name, the user can choose to create unique names for duplicate gene names by extending them with `.x` and `x` being the count of the duplicate or by summing up intensities of multiple entries. Missing gene names are replaced with the respective majority protein ID.

Zeroes in the protein evidence are set to `NA`, i.e., being missing. The result is stored in the reactive values list as `proteinAbundance - original`. Missing values can be imputed using four selected methods: `perseus-like` resembling Perseus' `ReplaceMissingFromGaussian` [43] function re-implemented in R, `knn` for k-nearest-neighbour imputation, or `MinDet`, and `QRILC` from the imputeLCMD package [97]. The imputed data set is also stored for further analysis.

In a next step, the user selects the sample or patient metadata file which is a tab-separated $m \times n$ matrix containing additional information on samples and experimental design. The metadata file needs to contain a `PatientID` column, which matches the sample IDs from the `proteinGroups` header and one or more named columns with parameters, i.e. textual/factual/logical or continuous/integer values.

The `proteinAbundance` and the `clinData` object are now available throughout the application for further analysis. They are used to create quality control (QC) plots: a PCA plot to visualize the main sources of variation, a bar plot for protein coverage across samples, a box plot for intensity distribution per sample, a sample-to-sample similarity or correlation heatmap, a missing value density plot, and a cumulative intensities plot highlighting highly abundant proteins. For plot configuration and update we implemented several modules as explained in section 3.5.2. For sample-wise plots, the user can select a parameter to define colors of samples from the `clinData` object. Additionally, custom titles, subtitles, and captions may be entered and thus added to the plots. Every plot can be downloaded as single portable document file (PDF). In addition to the manipulation of aesthetics and text of plots, the user can also manipulate plot data, e.g., by selecting specific principal components for display or configuring the distance measure for the heatmap (functions are not shown in Figure 3.10).



ProteinGroups.txt

Browse... Demo_proteinGroups_short.tx

Upload complete

Quantification type

● LFQ

○ iBAQ

Exclude columns (samples)

reference

Define in how many samples a protein needs to be detected in to be included in the analysis

1

Choose method for meaningful gene names

○ Summarize isoforms

● Unique names for duplicate genes

Imputation method

perseus-like

ClinicalData.txt

Browse... Demo_clinicaldata.txt

Upload complete

Analyze

Choose the principal components for plotting

PC1 PC2

Choose the clinical parameter for group colours

Failing

☑ Use imputed data for PCA

Select the (dis-)similarity metric

Pearson

Figure 3.9: Load and Prepare tab panel overview. Within the configuration module on the left two input files can be uploaded separately. General configurations and adaptions of the protein abundance data can be set in advance. Examples are the exclusion of outliers based on coverage or PCA analysis or the selection of an imputation strategy for missing values. On the right, interactive diagrams, showing for example a PCA or a sample correlation heatmap (zoom in), can be displayed and manipulated.

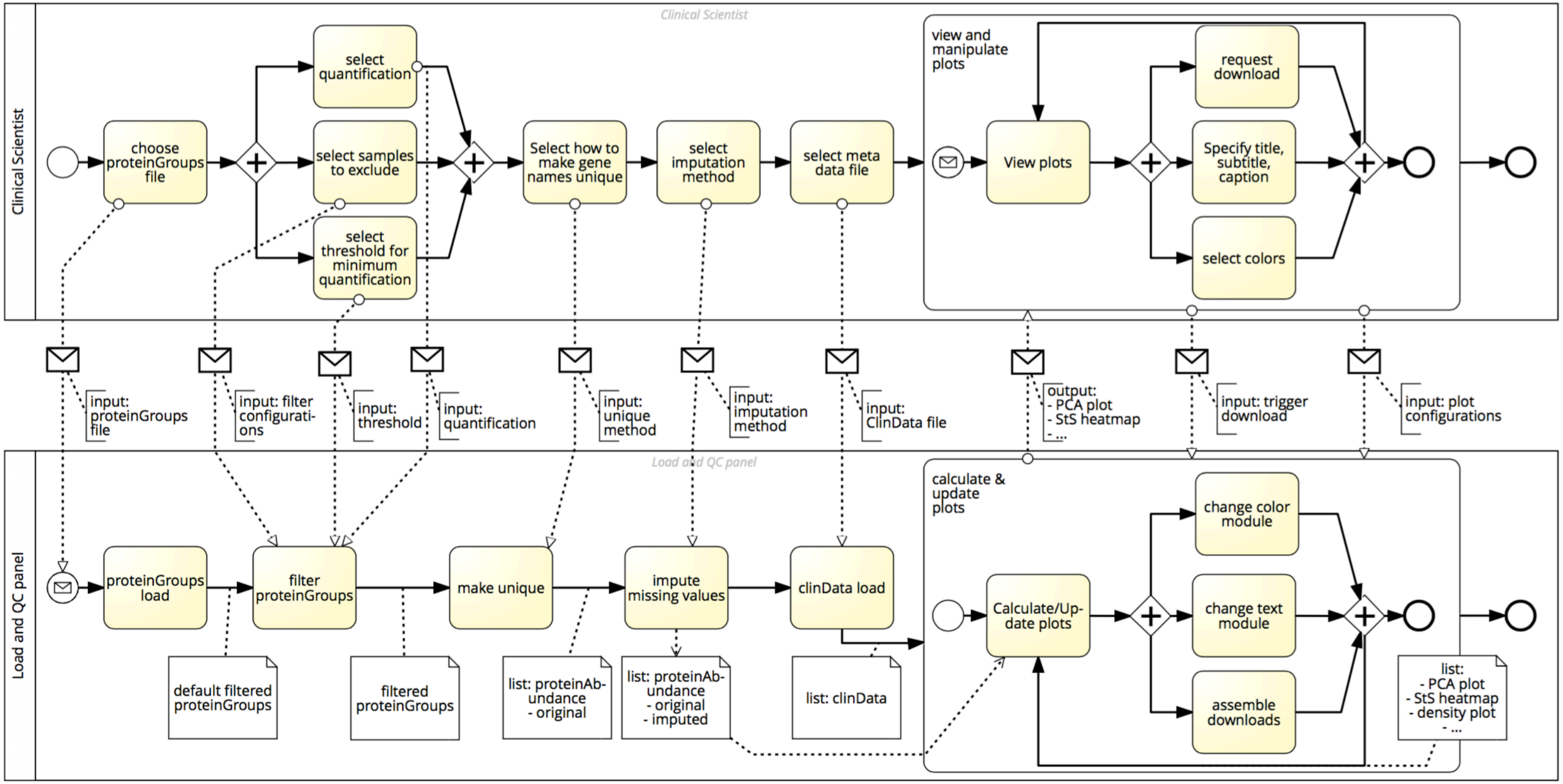
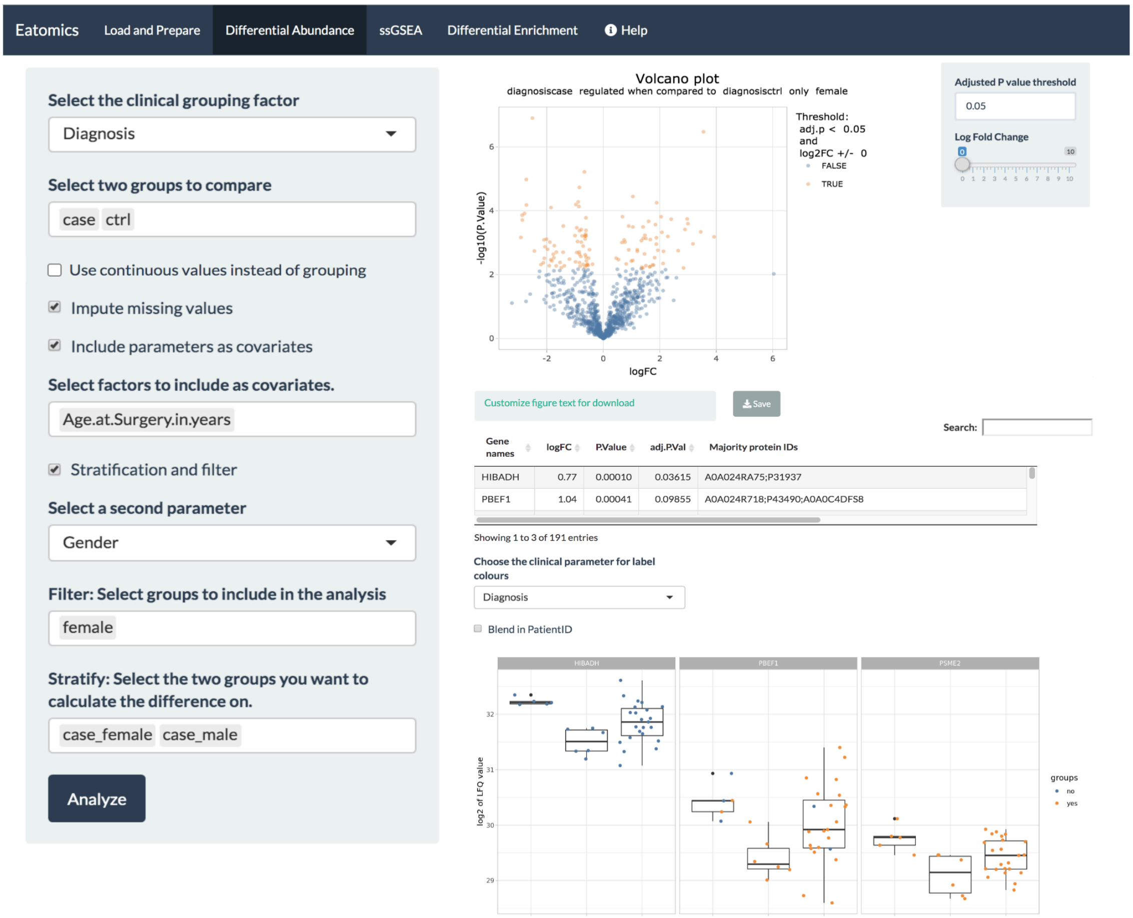


Figure 3.10: Process and message flow between the clinical scientist and Eatomics' first panel on data load and pre-processing main process (left) and quality control and visualization (expanded sub-processes on the right). Messages contain input to the calculation as selected by the user and output objects to be displayed in the UI. Tasks represented as yellow boxes resemble functions implemented in Eatomics.

When satisfied with data quality and sample selection the user may proceed to the differential abundance tab panel. The panel enables the user to translate a given hypothesis on the data into an experimental design and to test the hypothesis.

Our example hypothesis is illustrated in Figure 3.11. We want to investigate the difference in protein abundance in female case and male case subjects, while an effect of age may be possible but is not under investigation. The user selects the first grouping parameter (Xf_1) in this example the the case/control assignment and the sex as second variable (Xf_2). Both are categorical parameters and subject to the schema as shown in section 3.4.2 and are thus used to define the `mainParameter`. The respective combined groups (`case_female` and `case_male`) are then selected as contrast specification from a drop down list. Eatomics automatically relevels the `mainParameter` and removes samples that lack information from the `clinData` object. Together with the covariates, in our example the age at surgery in years, Eatomics creates the formula and consequently the model matrix as input for the DA calculation procedure (refer to section 3.4.2 for further details). In addition, the user



Select the clinical grouping factor
Diagnosis

Select two groups to compare
case ctrl

☐ Use continuous values instead of grouping

☑ Impute missing values

☑ Include parameters as covariates

Select factors to include as covariates.
Age.at.Surgery.in.years

☑ Stratification and filter

Select a second parameter
Gender

Filter: Select groups to include in the analysis
female

Stratify: Select the two groups you want to calculate the difference on.
case_female case_male

Adjusted P value threshold
0.05

Log Fold Change
0 – 10

Search:

Gene names	logFC	P.Value	adj.P.Val	Majority protein IDs
HIBADH	0.77	0.00010	0.03615	A0A024RA75;P31937
PBEF1	1.04	0.00041	0.09855	A0A024R718;P43490;A0A0C4DFS8

Showing 1 to 3 of 191 entries

Choose the clinical parameter for label colours
Diagnosis

☐ Blend in PatientID

Figure 3.11: Differential Abundance tab panel. On the left hand side the clinical parameters and levels of interest can be selected from drop down menus for main effect, as a filter or for stratification. The main panel shows an interactive visualization of a volcano plot, lists of significant proteins as specified by user-defined thresholds. An in-depth view of protein abundance can be generated as a box plot.

specifies if imputed or original data should be used in DA calculation. Please note that the same design would have resulted from selecting the sex as first variable and the group assignment as second parameter. The actual comparison is defined by the selection of groups, i.e., the contrast, in the last drop down menu. In the case of continuous values for Xf_1 or Xf_2 the UI would display a slider to specify a numeric cut-off value.

A volcano plot displays log_2 fold changes of proteins versus the respective adjusted p-values of the defined group comparison with colors used for the distinction between significant and non-significant results. The significance thresholds can be adjusted directly. For the visualization of abundance of specific proteins in relation to the sample description, either box or scatter plots are displayed. For these, we can re-utilize coloring modules as described for the previous tab panel. In addition to the download of single plots, the user can now also decide to render a full report containing all experiment details and plots as well as data tables.

The third tab panel offers an interface to configure ssGSEA algorithm and calculate ssESs. As mentioned in section 3.5.2 the result of the algorithm represents a transformation of the input `proteinAbundance` data towards the selected gene set or pathway. As such, the calculation is mainly a prerequisite to perform differential enrichment in step 4. The user selects a gene set file from the list of MsigDB Hallmark, Gene Ontology (GO), all terms or subsets of GO molecular function, GO biological process, or GO cellular compartment, Reactome, Biocarta or Kegg to calculate the enrichment score. Alternatively, the user may provide a custom gene set file in `.gmt` format by pasting it into the `Data/GeneSetDBs` folder of the application. As in the original code, an expert user may set many parameters. However, for a quick setup the default options from the original publication are implemented. The resulting ssGSEA scores are stored as files in the user's local app directory and are available for calculating differential enrichments.

Using the ssGSEA procedure to calculate single sample enrichment scores enables us to re-use the DA logic from the second tab panel for differential enrichments. As such, tab panel four allows the user to apply the research hypothesis directly to the enrichment scores and find gene sets and pathways that are significantly enriched. In practice, the UI and process flow is almost identical to the second tab panel. As a difference, the user has to choose the enrichment score file from those prepared on the ssGSEA tab panel that then serve as input instead of the `poteinAbundance` data.

Installation and technology dependencies

We summarize the most important packages Eatomics depends on in section A.3. A full list can be found in the repository at `https://github.com/Millchmaedchen/Eatomics/archive/master.zip`.

The installation procedure for the local instance requires R and we would recommend also Rstudio, which are available for a multitude of different operating systems. A single command is needed to start Eatomics from within R studio. The application opens in the users standard browser similar to a web application. For institutional use, Shiny applications are well suited for server installations and scale to serve more users.

3.6 Evaluation

In this section, we elaborate on the features implemented in DEAME and Eatomics with regard to the assessed software requirements and in comparison to related approaches. Furthermore, we evaluate the new concept to introduce the common practice of DE calculation into the Systems Medicine context in user interviews tailored to test if clinical scientists and computational biologists are enabled to perform DE analysis.

3.6.1 Comparison of functionality

In this section, we provide an evaluation of DEAME's and Eatomics' functionality, i.e., which requirements are met so far by both applications and how their compare to other available tools as introduced in section 3.2.

DEAME

We considered all tools that offer a pre-processing pipeline, DE calculation, and a graphical user interface as described in section 3.2. RAP, RNAMiner, and NGS-trex are closed source web applications, while TRAP, QuickNGS, TRAPLINE (as a Galaxy workflow), and DEAME offer the possibility to setup a private instance. However, to do so the user needs a considerable amount of programming skills. Thus, **R21** only refers to the IT skills need to perform analysis, not for setup of the application. Due to the large amount of data and the need for databases to store and efficiently access the data prohibits straightforward installation processes (**R1**). As a result, web services are the main option for inexperienced users. For the user, these web services are platform independent, as they can be accessed via their browser. Remarkably, the vast majority of the web services do not provide any information on how data is stored and security is assured. DEAME is populated with data from the SMART IT platform (subsection 2.2.1), which is secured behind institutional firewalls. Only members of the consortium were granted access after identity check (**R3**).

For DEAME, bioinformatics processing of raw RNA reads is completed automatically from within the SMART IT platform to yield count matrices (**R4-R5, R8**). The user is therefore in control to configure, however can decide to use default values if he/she lacks the domain knowledge. All other tools accomplish the necessary pre-processing in a similar fashion. However, only the count-based approach, as implemented by RAP, QuickNGS, NGS-trex, Chipster, and DEAME, renders the tools used in the pipeline to be independent from the experimental design. This choice of count based pipeline tools does not necessarily reduce the time to test a single hypothesis, but it avoids redundant pre-processing and thus eliminates computational overhead as soon as multiple hypotheses are tested.

However, all platforms use tools to meet scientific acknowledged within the research community (**R6**).

With regard to DE design DEAME and Chipster are the only tools offering an interface to configure complex designs. In Chipster, metadata has to be entered by hand and transformed to numbers to avoid confusion in the statistical evaluation rendering **R9** more cumbersome and error-prone for the clinical scientist. However, simple and complex designs (**R12**) including multi-group comparisons, continuous explanatory variables can be configured in DEAME and Chipster likewise. Chipster even extends towards the inclusion of covariates and interaction terms. Visualization options for both tools are comparably well established, although there is no direct interaction in Chipster as plots are statically rendered and send to be displayed in the front end (**R13**). Although stated in the manual, we could not reproduce the enrichment analysis function in Chipster. RNAMiner, QuickNGS, TRAP, and DEAME offer a direct option for enrichment analysis (**R14**). Report generation, data download, and reproducibility are best supported in QuickNGS and Chipster as they provide a fine granular solution to download data, results, and images as well as the used workflows and code chunks at any step of the analysis (**R16-R18**). Similarly, **R19-R21** are well supported in these tools.

Please note, that Chipster's and QuickNGS's capabilities were extended in parallel to the development of DEAME. Both platforms were well funded and equipped with a whole team of developers. Initially, Chipster was developed to analyze only microarray data it was equipped with a large range of tools after the first publication in 2011 [98].

The capabilities of these very mature tools however were not suitable for our application within the SMART consortium at time of active project work.

As a result, DEAME was equipped with similar functionality, covering an interactive volcano plot and heatmap (**R13**), which also accomplishes enrichment analysis (**R14**). Furthermore, Olelo may be used to retrieve actionable information via literature search (**R15**). Report generation (**R16**) is not possible in DEAME, however result data tables and images of visualizations can be downloaded (**R17**). Reproducibility is mainly covered in the pre-processing pipeline as workflows and configurations are saved in workflow diagrams. DEAME documentation may be found in a popup window within the application and concerns usage of the application, rather than an explanation of the statistical methods used (**R19**). Example and test data is provided and not IT skills are needed to use the application.

Table 3.4: Comparison of applications for RNAseq pre-processing, quality control, differential expression and enrichment analysis, ● fully supported ◐ partially supported ○ not supported.

	ID	Item	RAP [37]	RNA-miner [38]	QuickNGS [39]	NGS-trex [40]	TRAP [41]	TRAPLINE/ Galaxy [42]	Chipster [98]	DEAME [54]
System	R1	Straight-forward installation	web service	web service	◐	web service	◐	web service	web service	web service
	R2	Platform independence	●	●	only linux	●	●	●	●	●
	R3	Data security	remote	remote	●	remote	local	remote	remote	remote
Pre-processing	R4	Default configurations	●	●	●	●	●	●	●	●
	R5	Full-fledged pre-processing	●	●	●	●	●	●	●	●
	R6	Acknowledged tools	●	●	●	●	●	●	●	●
	R7	Independent tools	●	○	●	●	○	○	●	●
	R8	Automated execution	●	●	●	●	●	●	●	●
	R9	Handling of meta data	○	○	○	○	○	○	◐	●

Table 3.4: Comparison of applications for RNAseq pre-processing, quality control, differential expression and enrichment analysis, ● fully supported ◐ partially supported ○ not supported.

	ID	Item	RAP [37]	RNA-miner [38]	QuickNGS [39]	NGS-trex [40]	TRAP [41]	TRAPLINE/ Galaxy [42]	Chipster [98]	DEAME [54]
DE design	R10	In-app formulation of design	◐	○	○	○	○	○	●	●
	R11	Rapid design creation	○	○	○	○	○	○	◐	●
	R12	Complex designs	◐	○	○	○	◐	○	●	◐
Visualization, Annotation & Interpretation	R13	Interactive visualization	○	○	○	○	○	○	◐	◐
	R14	Enrichment analysis	○	●	●	○	●	○	◐	●
	R15	Actionable information	○	◐	○	○	◐	◐	●	◐
Wrap-up	R16	Report generation	◐	○	●	○	○	○	●	○
	R17	Data download	●	●	●	◐	●	●	●	◐
	R18	Reproducibility	◐	○	●	◐	?	◐	●	◐
Overall	R19	Documentation	●	◐	●	◐	◐	◐	●	◐
	R20	Example and test data	●	●	●	◐	◐	◐	●	●

Table 3.4: Comparison of applications for RNAseq pre-processing, quality control, differential expression and enrichment analysis, ● fully supported ◐ partially supported ○ not supported.

	ID	Item	RAP [37]	RNA-miner [38]	QuickNGS [39]	NGS-trex [40]	TRAP [41]	TRAPLINE/ Galaxy [42]	Chipster [98]	DEAME [54]
Overall	R21	No IT skills needed	●	●	●	●	●	●	●	●
		source code availability	○	○	●	○	●	●	●	○

Eatomics

In this section, we compare functions of Eatomics as they relate to our software requirements and to other available tools as described in section 3.2. Perseus is by far the most mature related analysis platform, however is only available for the Windows operating system and thus is not platform independent. In contrast, the Shiny framework aims at providing easy-to-use web applications and making it inherently easy to deploy the applications at institutional servers. The concept solves the problem of local and remote instances with regard to data security. The user may choose to either use the remote server, thus giving data away to a potentially not trustworthy party, or to stay secure with a local instance (**R3**). In addition, an institutional installation may provide secure access to multiple users. Currently only the source code of SAM and Eatomics are openly available and thus supporting the straightforward local and platform independent installation (**R1, R2**).

The pre-processing feature could not be implemented in Eatomics and all other tools, as they all depend on the popular MaxQuant algorithm, for which the license restricts incorporation of the algorithm into third party software. Some functions of pre-processing, e.g., logarithmic transformations and missing value imputation are covered by all tools (**R4-R8**).

Visualization, annotation, and interpretation features (**R13**, **R14**) are fully supported by all tools with the only exception of **R15**, which is actionable information on results. As for DEAME, Olelo would be a good candidate to retrieve documents, however an interface was not established yet. Results and intermediate data may be downloaded at least partially in all applications (**R17**), while a full report is only available in LFQ Analyst and Eatomics (**R16**). Example and test data is available in for all tools (**R20**), while documentation varies from being one paragraph in iMetaShiny to detailed manuals and workflows for the other tools (**R19**).

Table 3.5: Comparison of applications for quality control, differential abundance and enrichment analysis of MaxQuant proteomics output data ● fully supported ◐ partially supported ○ not supported.

	ID	Item	LFQ Analyst	SAM Shiny	Perseus	iMetaShiny	Eatomics
System	R1	Straight-forward installation	web service	●	●	web service	●
	R2	Platform independence	●	●	○	●	●
	R3	Data security	remote	local/remote	local	remote	local/remote
Pre-processing	R4-R8		MaxQuant dependency causes license restrictions				
DE/DA design	R9	Handling of meta data	◐	◐	◐	◐	●
	R10	In-app formulation of design	◐	○	◐	◐	●
	R11	Rapid design creation	◐	○	◐	◐	●
	R12	Complex designs	◐	◐	◐	○	●
Visualization, Annotation & Interpretation	R13	Interactive visualization	●	●	●	●	●
	R14	Enrichment analysis	●	●	●	●	●
	R15	Actionable information	○	○	○	○	○
Wrap-up	R16	Report generation	●	○	○	○	●
	R17	Data download	●	●	◐	◐	●
	R18	Reproducibility	◐	◐	◐	◐	◐
Overall	R19	Documentation	●	●	◐	◐	●
	R20	Example and test data	●	●	●	●	●
	R21	No IT skills needed	●	●	●	●	●
		source code available	◐	●	◐	○	●

Metadata handling and design creation feature (**R9-R12**) are not or only partially supported in most other tools and resemble the heart of the application. Therefore, we prepared another table to provide further functional details in Table 3.6.

Table 3.6: Comparison R9-R12 of the DE/DA design feature details of applications for quality control, differential abundance and enrichment analysis of MaxQuant proteomics output data, ● fully supported ◐ partially supported ○ not supported.

		LFQ Analyst [45]	SAM Shiny [46]	Perseus [43]	iMetaShiny [47]	Eatomics [55]
Complex design	Two-group	●	●	●	●	●
	Multi-group	◐	◐	●	◐	●
	Continuous	○	●	◐	○	●
	Time series	○	◐	○	○	○
	Covariates	○	○	○	○	●
	Filter and stratification	○	○	○	○	●
	Interactions	○	○	○	○	○
Meta data handling	Filter on samples	○	○	●	●	●
	Filter on rows	○	○	●	●	●
	Accepts more data than needed for design	○	○	●	○	●
	Supports major meta data types	○	○	●	○	●

3.6.2 Examples of utility

In the following, we show and discuss how simple, but also more complex designs can be configured and executed in Eatomics. We use the demo protein evidence and metadata set derived from Chen et al. (2018) (see section A.4) to exemplify questions a clinical scientist could be interested in [99]. For a better understanding, we verbalize a question, show its deconstruction into a hypothesis type, describe preparation of input data and in-app configuration necessities in Eatomics and related tools.

Baseline: Two-group comparison

The baseline experiment of a two group comparison as exemplified by the question of *How does protein abundance in failing hearts differ from non-failing hearts?* can be answered by using the `Failing Heart` parameter in the metadata. Eatomics and Perseus are able to directly use the given input data, whereas the metadata file needs reduction to only include the relevant parameter for LFQ analyst and iMetaShiny. SAM needs the most extensive preparation as the protein abundance and the metadata file need to be merged and re-coded to represent the desired design. All samples are assigned to either group of failing or non-failing hearts, thus a removal of samples from both files in not needed in either case. Within all applications the configuration of the experimental design is simple and can be achieved by very few clicks, e.g., selecting the `Failing Heart` parameter and the two subgroups within Eatomics and running `Analyze`.

Complex designs

What are the differences in protein abundance of hearts with a high (preserved) and low (reduced) left ventricular ejection fraction (LVEF)? reads as simple as the first example, however, when the clinical parameters are taken into account, the question is not deterministic yet. Thus, we show how to answer it thoroughly in Eatomics:

1. **Continuous response variable:** `LVEF (%)` was measured for all but one subject (including normal hearts). Within Eatomics, without input preparation, one can simply select `LVEF (%)` from the parameter list and select to use the continuous response instead of grouping. Results show proteins that differ with regard to `LVEF (%)`, i.e., which proteins show higher abundance with higher `LVEF (%)` or lower abundance respectively. The sample for which no `LVEF (%)` value is available is excluded automatically. While the continuous response can not be calculated in Perseus, iMetaShiny, and LFQ analyst, SAM would require input data manipulation: removal of protein abundance data from the file and introduction of the continuous value into the abundance file.

2. **Dichotomized response variable:** A common procedure when the continuous response option is not available is to discretize the continuous parameter. The researcher/clinician needs to set a threshold for discretization, e.g., 40 % to separate patients with reduced ejection fraction (EF) from those with a preserved EF.

3. **Multi-group comparison:** Within the Group parameter, the groups `HCMrEF` and `HCMpEF` denote samples from patients with hypertrophic cardiomyopathy (HCM) with reduced (r) or preserved (p) EF. These two groups can be compared directly in Eatomics, while not excluding the information from all other samples in the statistical model. A similar approach can be configured in Perseus or by manipulating input data in SAM and LFQ analyst. iMetaShiny does not provide an option for multi-group comparisons.

4. **Optional – filter or stratify:** The two previous solutions do include the non-failing heart samples, as `LVEF (%)` is given for them. However, it might be more interesting to further specify the question to find the differences in only failing hearts. In the continuous case, a filter can be used to only include the failing hearts. Similarly, in the dichotomized case, the stratification would give rise to the comparison of Failing heart with `LVEF (%)` < 40 versus Failing heart with `LVEF (%)` > 40.

   In-app usage of discretization, stratification and filtering is only usable in Perseus and Eatomics. All other tools would require manual input manipulation.

5. **Optional – covariates:** The consideration of covariates, such as sex and age, are crucial especially in clinical observational settings as they may have a major influence on the protein profile of a tissue. Covariates can be added to the model or their influence can be calculated by using them as the response variable directly. Covariates can only be included in LFQ analyst/DEP, SAM and Eatomics, as they use linear regression based statistics.

In conclusion, all tools can manage the calculation of a simple two group comparison, but will not suffice when further stratification and/or the inclusion of covariates is needed. Furthermore, especially when Eatomics is used, users require less time for input data preparation.

3.6.3 User studies

User interviews are driven by the question of how clinical scientists and computational biologists perceive the DEAME research application. Thus, user interviews are structured to characterize the study population, test if a user is enabled to perform a given and a self-defined task by using the application, and in the last part assess the user's performance and effort expectancy as well as the further intention to use.

Characterization of study population

Eight testers (n = 8) completed the full interview procedure and provided complete questionnaires. Among them there was an equal proportion of males and females, age 35 years on average (+/- 12.7 standard deviation). Two testers categorized themselves as being computational biologists, two as medical experts/clinicians and three specified their profession as being a biostatistician, biologist, biochemist or scientist. Seven out of eight testers never had performed DE analysis themselves, however all of them qualified for being a tester by showing minimal understanding of gene expression from university experiences or by being involved in the interpretation or description of results of DE analysis.

Perform a given task

In total, 15 items within the questionnaire and the testing notes filled in by the interviewers assessed the user's ability to understand and implement a given task with the help of the DEAME application. While the questionnaire yielded at assessing how well the user understood what they did and where to find the information asked for, the testing notes depict plain functions of the application and if they were found and used or not. The results are visualized in Figure 3.12A. Six out of eight testers translated the given hypothesis into the correct design matrix. Seven or all testers used all other functions correctly for the created hypothesis and thus correctly completed follow-up questions even when a wrong design was created. All testers successfully translated a given design matrix back to a research hypothesis correctly. Only one tester managed to zoom into the list of genes.

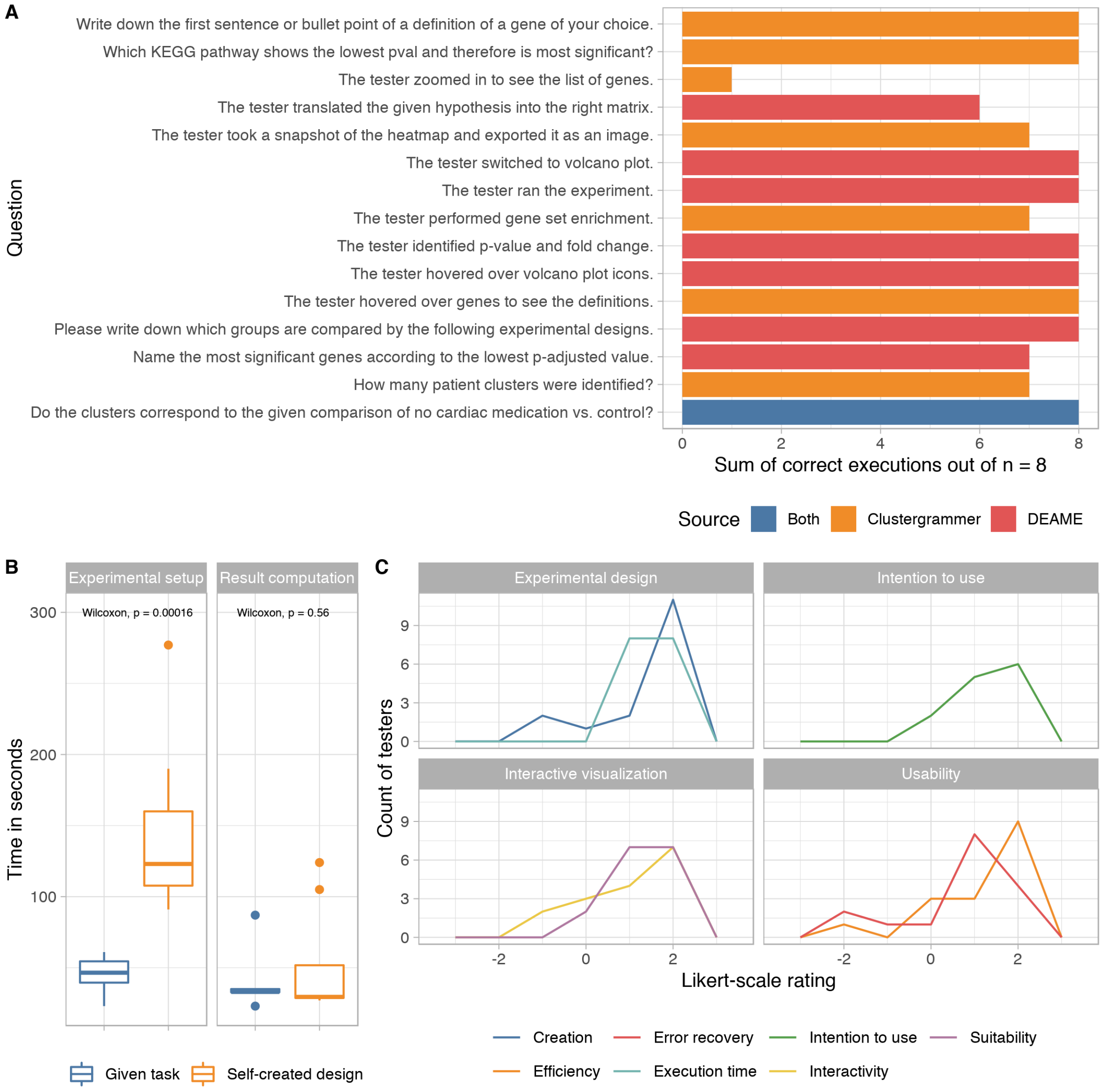


Figure 3.12: Results from User Interviews. (A) Test performance on a given task – the sum of correct executions of a function – is shown with colors denoting the original library being a custom implementation of DEAME, Clustergrammer or a mixture of both. (B) Execution times of experimental setup and calculation time with regard to the given task and the self-created design. The Wilcoxon rank test was used to test differences in setup times. (C) Frequency polygon plot showing the count of testers who rated two items on each functional requirement and for usability and one item on intention to use using a Likert-scale rating (-2 = strongly disagree, -1 = disagree, 0 = neutral, 1 = agree, 2 = strongly agree). Colors denote items. Ratings to negatively formulated items are transformed to represent the same scale as positive items and summarized be represented by one line. All results are compiled from n = 8 interviews.

With the only exception of zooming into the gene list, all other functions provided by the Clustergrammer library appear useful within the DEAME application. In general, testers understood and executed the given task almost error-free and established a good understanding of functionality and its interpretation, provided that testers had at least a minimum understanding of gene expression. The functional requirements were met with regard to a given task.

Performance and user expectation

The applications main aim is to enable users to translate their hypothesis on influences on gene expression as they may have gathered throughout their research or clinical professional experience. Thus, it is crucial that testers manage to translate their own hypothesis into a valid design. The user's intention to use the application can, according to UTAUT, be used as a proxy for their actual use.

Seven out of eight testers were able to create their own design matrix and experiment correctly after verbalizing their hypothesis. The translation of the given hypothesis took 45 (SD +/- 12) seconds on average. The result computation needed 39 (SD +/- 20) seconds and is not significantly different when the design was self created (Figure 3.12B). The three outliers in computation time can partly be explained by network connection or database problems, as the given task for the experimental setup was always identical and always correctly configured. Additionally, there is no correlation between the computation times and the complexity of the design (two-group vs. four-group design) or group sizes. The translation of an own hypothesis took 145 (SD +/- 62) seconds and thus significantly longer than when a hypothesis was given (Wilcoxon rank test, $p < 0.005$). A self-created design includes much more exploration of the available list of parameters and thus a longer setup time is expected.

Figure 3.12C displays the results to evaluate effort and performance expectancy according to UTAUT in one frequency polygon plot for rapid design creation (**R11**), interactive visualization of results (**R13**)), usability, and intention to use each. The majority of testers agree with the statement that experimental design creation was easy for them. Three votes were neutral or negative, two of them coming from the same tester, who managed to correctly set up a more complex four group design, which had not been introduced specifically in the intro or the previous task. Testers were not disturbed by execution times, even when it took more than twice as long as the mean.

Interactivity of diagrams in general seems to be sufficient, however two testers would appreciate further interaction possibilities. However, existing visualizations are suitable for the application. Usability is divided into a subcategory that relates to error recovery, in which three votes asked for a more exhaustive tutorial or documentation, and efficiency, in which one vote stated dissatisfaction with how DEAME works when compared to scientific software they usually use. Needs for further interaction possibilities and elaborate documentation are in line with the only partially fulfilled requirements (**R13** and **R19**) as stated in section 3.6.1. As such, only a short tutorial was available at time of testing for trouble shooting instead of an explanation of the numeric method running in the background and its configuration options.

Despite some shortcomings, 79.5% of all ratings were in the positive range of the scale, demonstrating reassuring positive attitude towards the usage of the application. Accordingly, the intention to use DEAME is positive in six out of eight testers. In the qualitative assessment reasons to not use DEAME can be summarized to the already mentioned need for more documentation of internal methods and fine-tuning options and current lack of a use case for such a tool. DEAME aims at enabling clinical scientists and computational biologists to perform rapid testing of a given hypothesis, potentially also as a communication platform when both users work together.

3.7 Discussion

In Systems Medicine, one approach is to transfer methods from Systems Biology towards their use in the medical setting. As such, many research consortia assemble detailed clinical phenotypes, as well as multiple molecular signatures. The relationship of clinical phenotype and molecular differences are of particular interest. A common analysis strategy in Systems Biology is differential expression/abundance analysis; and a plethora of software platforms to aid and automate the analysis have been proposed. However, at the start of the SMART project, no existing platform had the capability to handle extensive phenotype data and to enable the user to formulate more complex designs for DE analysis with it. Therefore, we utilized a hybrid approach of Design Thinking, scientific requirements engineering and user testing to define a new, viable solution to automated DE analysis software tailored to Systems Medicine.

21 requirements shape our solution to an automated DE/DA analysis software for Systems Medicine

We define 21 requirements, of which 12 detail the crucial requirement spaces of bioinformatics pre-processing, DE design, and annotation and interpretation as defined by the general process of DE/DA analysis and the SCRM/DRUMS model. Another nine requirements specify the system setup and functions, which are shaped by the specific needs of the clinical scientist as the target stakeholder.

Special attention is given to the DE/DA experimental design feature, which represents the heart and the novelty of our approach. The concept and the corresponding requirements detail how existing mathematical models and numerical solutions from Systems Biology may be exploited to readily accept complex designs and how metadata needs to be handled to assemble the design. Furthermore, the user and software interfaces are specified.

Requirements engineering for scientific software is rarely conducted by following dedicated and appropriate methodologies [100]. The complexity of the scientific domain problem hampers the proper understanding by professional developers. Additionally, the scientists who often take over the part of the developer are rarely trained in software development.

Furthermore, requirements may change drastically over time of development. Li et al. (2011) [77] established a sophisticated, yet easy to comprehend modeling technique to capture requirement spaces as features based on the underlying scientific problem with their SCRM tool. In conjunction with modern methods of Design Thinking for iterative user feedback and thorough literature search for a deep domain knowledge, we have added a systematically crafted list of requirements to define automated DE/DA analysis software in general and in handling the metadata and complex design formulation in particular. Although we aimed for a systematic approach, the methodology is still underdeveloped when compared to approaches in commercial software development. This limitation may challenge the results as, e.g., usability and other non-functional requirements are not explicitly mentioned or modelled in SCRM/DRUMS. However, user and expert feedback, as well as the review by Poplawski et al. (2016) [4] helped to fill the gap.

Notably, the requirements may be of practical use in the implementation of HTQ differential analysis software in general as it is not necessarily bound to a specific source of molecular data. Based on the here demonstrated use cases of proteomic and transcriptomic data, the requirements can be deployed to other applications, such as methylation data. For example, the Champ pipeline [48] reuses the Limma model to calculate differential methylation and differential hydroxymethylation, however does not yet allow for complex designs, yet. Future work should address more user research on the specification of the requirement on “actionable information” for further interpretation of results. Especially clinical scientists rely on proper interpretation help and were lost quickly when presented with a large list of results. Manual search for appropriate literature and gene/protein function in the context under investigation is very cumbersome and may be improved by the use of information retrieval software. Approaches for natural language processing and retrieval of meaningful scientific text or production of summaries are manifold, however their usage is less common and could be promoted via a direct interface from the DE/DA analysis software.

Two research applications successfully implement our requirements

To put the presented work into practice, we present two research applications – DEAME and Eatomics – that were implemented on the basis of the stated requirements.

DEAME fulfils most requirements at least partially, with the exception of the report generation and easy installation procedure. These requirements had less priority in the limited scope of the SMART project, which ended in March 2019. While we do provide instructions for deployment within the code repository, it was not a primary goal to open the application to public use. Instead, authorized users are granted direct access to the web application hosted at our development facility. Chipster, the only related tool providing an interface to a more complex design formulation, has been developed in parallel by the CSC, a professional IT Center for Science, jointly run and funded by the Finnish government and universities. While not being mature enough to be used for our purpose in 2015, Chipster has evolved to a versatile platform for various NGS analysis. Chipster lacks a proper handling of phenotype data, as data needs to be entered and encoded manually. Furthermore, complex design formulation is focused on usage by bioinformaticians, which is not suitable for the clinical scientist. However, as Chipster is still under active development by IT specialists, the tool will most probably evolve further to be very valuable to the scientific community.

Within most of the algorithms utilized for our applications, there are many options to fine-tune the analysis. We purposely do not use many of these options as we believe they will confuse the clinical scientist as a user. We expect the results set of regulated genes or proteins to be smaller than within a fine-tuned environment. While this is a drawback in a detailed analysis of a computational biologist, the clinical scientists we spoke to are interested primarily in the strong signals and are pleased with a shorter list of candidate genes/proteins. Apart from this, achieved functionality in DEAME is sufficient to conduct user experiments on perception and intention to use such an application by clinical scientists and computational biologists.

Eatomics is even more mature with regard to the 21 requirements and so are many of the related tools. The reason for their maturity is most probably their head start from developments established in the analysis of microarray data. Gene expression quantification was established much earlier following the development of microarrays and RNAseq as early as 1995 [101]. Label-free MS data started to become feasible ten years later due to the high complexity of pre-processing and cost of MS measurements [3]. However, DA software benefited from expression analysis, as the mathematical model is very similar and thus sophisticated numerical solutions could be transferred successfully from DE to DA analysis [45, 95]. As a result, methods were ready to be wrapped into comprehensive user interfaces and could be extended and published for the new use case. In 2012, the Rstudio developers introduced the R Shiny framework, which is tailored to provide a reactive user interface to R analysis scripts. The framework quickly became popular and was utilized for all kinds of analysis.

With the exception of Perseus, all considered related tools are Shiny applications demonstrating the frameworks suitability. In choosing a reactive programming model as the one implemented in Shiny Eatomics might not be applicable studies with many samples as data is held in the main memory. As we aim at biopsy samples with extensive (clinical) annotations, we expect the HTQ data to be limited and metadata to include sensitive information. Therefore, a local installation may be favourable and can be easily accomplished in a Shiny application. In the need of scaling to many samples, the limitation may be overcome by implementing the application on a professional R Shiny server instance that can utilize multiple cores for calculations. Furthermore, the Shiny server can then be exploited to balance load and handle a larger amount of client requests. This is also true for other Shiny applications, however the code base for own or institutional instances are not available for LFQ analyst and iMetaShiny.

When compared to the DE applications all DA provide an option to upload data instead of manual entry or to accept more metadata than needed for one design. This trend may also reflect the growing need in also handling metadata.

Some functionality is not well supported by neither DEAME, Eatomics or related tools. Adaptations to the experimental design algorithm and implementations should be easily adaptable to retrieve interactions and to extend the possibilities of discretization of numeric beyond two bins. Implementing a possibility for time-series analysis is less straightforward. Although Limma can be used for time-series analysis by utilizing splines, the required input metadata needs to include the time dimension and therefore needs an additional input data specification. Most likely, a time-series analysis would necessitate another tab panel in Eatomics and for now was postponed to a later release.

For Eatomics' functional comparison, we took a closer look at details of R9 and R12, in which the additional functionality of Eatomics becomes more apparent. Here, Perseus' statistical approach of using multiple pairwise comparisons instead of GLM's proves to be not as versatile. While the method was shown to be less sensitive [45, 95], it also does not come with the direct possibility to also model continuous explanatory variables, covariates, interactions, and stratification. Nevertheless, Perseus does provide a versatile software suite valuable to the Systems Biology community, despite its platform dependency. In Eatomics, we implemented the design feature slightly different from DEAME: DEAME lets the user define the contrast, i.e., the two groups to compare, by clicking on the space between the groups. The testers were not expected to use this option within the experiment. However, if they did while exploring the application, they were mainly confused and were not able to comprehend the data anymore as they were expecting to test only one hypothesis instead of all possible combinations of comparisons. For Eatomics we changed the selection of contrast to a drop down list, making the choices more obvious. Furthermore, the merge in DEAME passively results in a filter of samples before DESeq2 is run and thus hindering the method to retrieve all information needed for shrinkage estimation. For a start, we dismissed the merge in Eatomics, however the function itself is useful and should be re-implemented in future versions without a filter.

Also note, that the DA panel is modularized to be reused in the enrichment analysis panel. By using this setup we further exploit the similarity of the different HTQ data sets and prove the generalizability of the approach. In the ssGSEA setup, it is possible to perform enrichment analysis independent from DA analysis. By first transforming the data towards a gene set representation, it is also possible to merge data sets from different institutions more easily as the transformation reduces noise. As a result, one can include samples across batches and thus achieve more power in a statistical analysis of a specific comparison.

Conclusively, Eatomics, which was developed with a clinical scientist as a user in mind, does expand the range of functionality towards the Systems Medicine setting by providing metadata handling and complex design formulation.

User interviews reveal a positive intention to use

Thus, interviews were performed to test if a user is qualified to perform a given and a self-defined task by using the DEAME, and to assess the user's performance and effort expectancy as well as the intention to use. A mean computation time of half a minute is fairly long; especially as the R routine itself only takes a couple of seconds. However, assembly and transfer of data between database, R server, and front-end display need to be considered. Result computation and thus waiting time for the user may impose a threat to the usability of the application. Interestingly, testers were satisfied with computation time and demonstrated patience.

Although there is no baseline time to relate to, we interpret the setup time of the design for a given task and also for the self-created design as fairly low. We argue that when considering that it was the first time to use the application. In addition to the task of configuration, they also had to get familiar with which clinical data is available and how it is displayed and interacted with. Especially for the self-created design, testers needed to find the parameters they were interested in. Furthermore, the users were satisfied with the time they needed to complete the design and felt that it was an easy task. Of note, testers were very inexperienced; some of which learned the basic concept of the calculation only at the time of testing. Incomplete documentation on internal methods poses a major threat to actual usage, which is in line with our crafted requirements and understandable in a community of users that is engaged in producing reasonable results. Additional usability errors are unlikely as the optimal number of testers to find usability errors is met [84].

Overall, the testers' intention to use was positive. In this regard, we need to reiterate on the usage of the UTAUT framework as its scope of application is not specifically within testing of scientific software. Therefore, we needed to reduce the model to be applicable, potentially threatening its validity. However, as the field of scientific software engineering is rather unexplored, it was the most sensible solution to gather insights on the user's perspective. A general threat to the validity of results especially in user interviews with personal contact is the influence of the researcher him-/herself on the results. The possibility of testers wanting to do the researcher a favour cannot be dismissed thoroughly in any such setup. In our sessions, we emphasized that honest feedback is crucial and encouraged reporting of flaws and concerns. This procedure is a vivid element of Design Thinking and agile software development in general and we believe that it reduces the possibility of a positive intention towards the researchers/developers to confound the intention to use the application.

Eatomics was rudimentary validated in an informal testing session, which was not eligible to be included in the research results. Here, we asked one member of the SMART consortium to configure the research questions as introduced in section 3.1 and listed in section A.1 with the help of Eatomics. It quickly became apparent that many questions were not readily defined to be tested. As such, the member had to specify and reiterate on the questions for them to be eligible for testing. While this is only one observation, it could in the worst case lead to early frustration and quitting the tool before any analysis is performed or in the best case to a learning process of how to approach the problem of DA analysis and deepen the understanding. Still, the tester successfully configured a valid design to all questions, which represented a large proportion of possible simple and complex designs.

After taking a short glimpse on results before configuring the next design, the clinical scientist could decide quickly on one particularly interesting hypothesis for follow up, yet reported to be surprised how prominent differences observed in the clinical phenotype did not lead to significant results on the protein level. Further interview sessions to come to a firm conclusion would be desirable, however could not be fitted to the given time frame.

By offering flexibility in design setup, we do touch general concerns of p hacking, i.e., corrupting the calculated p-value by not correcting for the many design setups that may occur throughout data exploration. We therefore strongly advise the user to stay aware of the observational nature of the applications, which as all observational studies helps in exploration and hypothesis generation, while specific findings need validation experiments.

4 Molecular and clinical characterization of AS and MR

Chapter 3 demonstrates the Systems Medicine software solutions inspired by clinical scientists' questions about the SMART and EurValve project's molecular data. Although we considered a multi-omics analysis for the SMART data and found a strong influence of sex, we also let the clinical scientists explore the full data set using Eatomics. The clinical scientists observed that the most changes in protein abundance are found between the three conditions of AS, MR, and control.

As a result, this chapter derives biomedical insights from a joint examination of proteome and clinicome data as assessed in the SMART and EurValve projects. We provide details on how data from both studies are acquired, harmonized, processed, and analyzed. We describe our elaborate study design to compare AS, MR, and healthy controls and disseminate female and male signatures within a disease group. As such, we shed light on how differences in mechanical load in heart valve diseases shape the proteomic and clinical phenotype in a disease- and sex-specific manner.

4.1 Motivation

The SMART project gave rise to a thorough characterization of AS and control (CON) subjects. However, it was not possible to derive all data sources for all subjects for reasons beyond my control. For example, for some subjects there was only enough biopsy material for proteome analysis resulting in their exclusion for RNAseq or exclusion was necessary after thorough quality control of measured data. An overview of available data and their overlap for subjects is given in Figure 4.1. The following advantages undermine our decision to focus on clinicome and proteome analysis:

1. Closeness to phenotype: While the genome and transcriptome provide the instructions, the cells' proteins directly confer function and shape the observed phenotype of the molecular function. Similarly, the clinicome describes the overall organ and whole body phenotype best.

2. Value of proteomic insight: Deep proteome data is still sparse as hardware and algorithms for measurement and processing reached maturity approximately ten years after transcriptome data did.

3. Increase in sample size: Firstly, we arrive at 58 subjects with available clinicome and proteome data just from the SMART study. Secondly, we gain access to a second cohort from the EurValve study. Here, 17 subjects are characterized in proteome and clinicome as well.

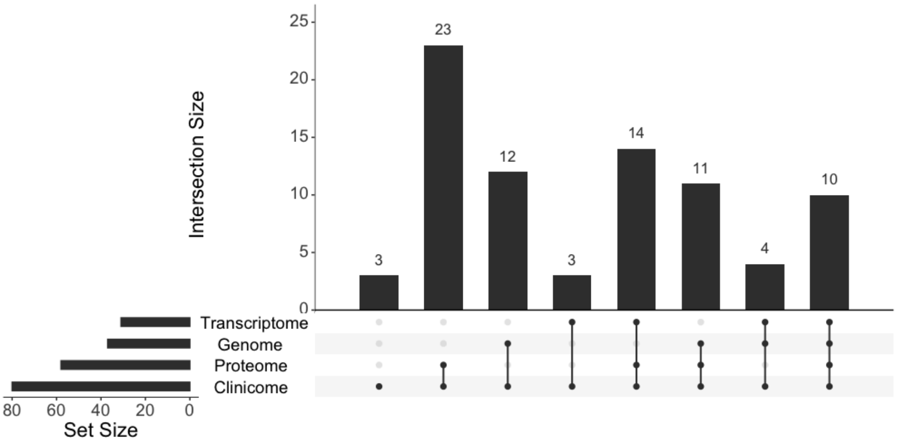


Figure 4.1: Intersection plot of data set combinations for all subjects of the SMART cohort including AS and CON subjects. Bars denote the number of samples having the exact combination of data sets as annotated by the connected dots.

4. Validation data: In cases where validation is needed, we can fall back to find further evidence in the other data sources.

A multi-omics analysis was performed on the SMART data set, i.e., only AS subjects, and results showed a clear separation of female and male subjects. However, the data set proved to be too small to yield robust results (Appendix C). Furthermore, the clinical scientists explored the data and observed that the most protein abundance changes are found between the two conditions of heart valve diseases and controls.

Heart valve diseases may eventually lead to heart failure when left untreated. In aging populations, the incidence is increasing drastically and is becoming a serious health burden. AS and MR are the most frequent types of valve disease and have reached an incidence of more than 12% of the population >65 years for AS and 9% for MR [31, 32]. AS and MR cause chronic cardiac pressure or volume overload, which triggers distinctive forms of cardiac remodeling. One very prominent adaption mechanism is left ventricular hypertrophy, typically concentric in pressure and eccentric in volume overload (Figure 2.11).

In an adapted compensated state, patients can remain asymptomatic for years; however, once there is a transition into heart failure and patients become symptomatic, the prognosis is poor in both patient groups if they remain untreated [33]. Cardiac hypertrophy can be treated by lowering high blood pressure through different medication mechanisms. For example, via angiotensin-converting enzyme (ACE) inhibition or angiotensin II receptor blockage, cardiac remodeling can be slowed down and therefore resembles a first choice before invasive surgical replacement of the malfunctioning heart valve needs to be considered. However, it has been shown that ACE inhibition reduces pressure overload-induced hypertrophy but not volume overload-induced hypertrophy [102]. Furthermore, significant differences in the characteristics of left ventricular hypertrophy and HF in females and males are apparent [34, 35].

Currently, most knowledge about cardiac adaptation mechanisms in valve disease is available at the organ scale, where parameters like ventricular function or myocardial mass, and fibrosis can be investigated with non-invasive imaging methods [36]. Much less is known about cardiac adaptation mechanisms at the cellular or protein expression level. It is important to understand similarities and differences in the adaptation mechanisms of pressure versus volume load to adjust treatment accordingly. Similarly, sex differences need to be elucidated to find personalized treatment options eventually.

4.2 Related Work

Because of the challenges to obtaining human left ventricular myocardial tissue samples, only a few studies and data are available for human heart tissue. If available, transcriptomic descriptions are common; however, its interpretation is limited because transcript abundance is an imperfect proxy for abundance and dynamics of the encoded protein [51].

Doll et al. (2017) have developed a deep proteomic map of 16 different anatomical regions of healthy male human cardiac tissue (n = 3) [103]. Similarly, the Human Protein Atlas provides a novel resource of an antibody-based healthy cardiac proteome [104, 105]. Li et al. (2020) dissect large-scale proteomic and metabolic changes in ischaemic and non-ischaemic heart failure and consider sex-specific effects in 44 individual hearts [106]. Chen et al. (2018) describe distinct changes in cytoskeletal proteins in 21 failing and 13 non-failing human hearts by utilizing mass spectrometry proteomics. They focus on the effect of increased microtubule network density in failing hearts and pharmacologic restoration of contractile function [99]. These recent studies of the deep cardiac proteome serve the particular research question but are limited to healthy organ donor tissue or failing hearts.

Proteomic measurements derived from living individuals are characterized by small-scale quantification or small sample size: Coats et al. (2018) show a targeted analysis of 32 proteins from cardiac biopsies of seven aortic stenosis subjects [49]. Linscheid et al. (2020) compare cardiac protein expression in both atria and the left ventricle of male subjects with MR. By providing high-throughput protein quantification of seven biopsies in total, the authors describe the most extensive large-scale proteomic quantification of *in vivo* collected tissue from patients to date [51].

A direct comparison of volume and pressure overload is described in a mouse model by You et al. (2018) only [50]. Here, six key transcripts and 16 key proteins are measured in a targeted analysis. The authors conclude pressure overload to have stronger maladaptive effects than volume overload. In contrast to previous studies using an aortic fistula, the authors develop an improved model to induce volume overload through aortic regurgitation. Although animal models are crucial and valuable to gain mechanistic insights, they lack age and risk factor-associated components of degenerative aortic valve stenosis of the elderly human being. For example, in a meta-analysis comparing the transcriptomic profiles in human and murine pressure load-induced hypertrophy, the concordance of changes in both organisms was surprisingly low [107]. Furthermore, the common practice of using left ventricular volume load induced by surgical shunts may not fully mimic the pathophysiology of MR [50].

In summary, there are few large-scale proteomic quantifications available on tissue derived from healthy donor hearts or from the severe phenotype of explanted failing hearts. In these studies, sex-specific effects are evident but considered only when sample sizes allow to do so. The most extensive study in living human subjects includes seven subjects with mitral valve regurgitation. The effects of volume and pressure overload in a direct comparison have only been studied in a targeted approach in likely confined animal models. A comparative proteomic exploration of pressure and volume-overloaded left ventricular human myocardium and a direct association to clinical parameters has not been published yet. Therefore, the differences in disease and sex are not well understood and potentially more subtle than when comparing a healthy vs. failing heart. Additionally, the data set would provide a valuable resource of protein expression to other research areas.

Accordingly, the present study aims to obtain deeper insight into heart valve disease-driven protein expression changes and relate the proteomic data to clinical parameters in a well-powered study of human tissue. Additionally, we aim to provide a database of isoform-specific protein quantification data for the research community to explore, form hypotheses and validate own findings.

4.3 Methods and Study Setup

An overview of the study setup is given in Figure 4.2. Proteins are measured from biopsies of the left ventricle of 41 patients with AS (female n=21, male n=20), 17 patients with MR (female n=6, male n=11), and 17 healthy control hearts (CON) (female n=8, male n=9). Peptides from protein extracts are analyzed by high-resolution tandem mass spectrometry (LC-MS/MS). A deep reference proteome is used for MS1 matching. Raw data are processed with MaxQuant and LFQ intensities are used for disease- and sex-specific protein abundance analyses. Joint analyses are performed on proteomic data and clinical imaging (cardiac magnetic resonance imaging) phenotypes. A subset of biopsy samples from AS (n=17) and CON (n=6) are additionally used for quantification in RNA sequencing. The results from disease-specific differential expression analysis from the transcriptome are used to validate proteomic findings where necessary.

4.3.1 Biopsy acquisition and distribution

Left ventricle biopsies are extracted at the time of valve replacement surgery, frozen directly in liquid nitrogen and kept at -80° C. For controls, the cardiac surgeon perfuses the heart in situ while the organ donor is still on the operating table, with sterile Custodiol® solution for cardioplegia and multiorgan protection. Then the heart is removed from the donor and it is placed into ice-cold Custodiol® solution, moved to a separate room where the samples are then placed into 2 mL cryotubes and into liquid nitrogen. Frozen samples are split and

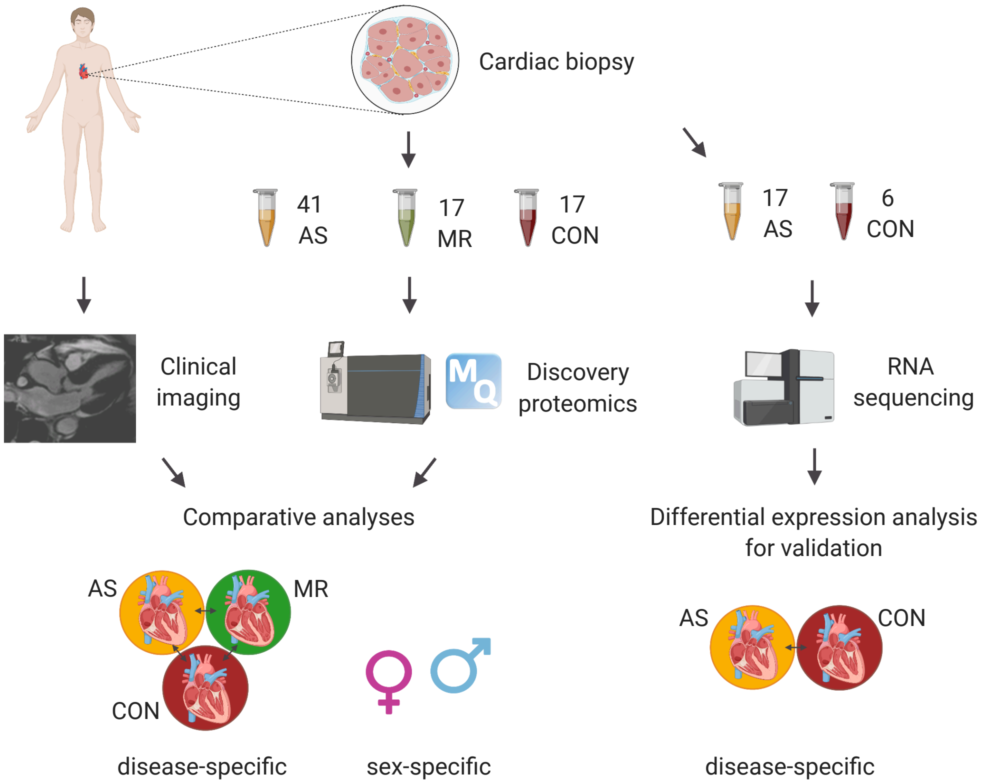


Figure 4.2: Study design. Proteins are extracted from biopsies of the left ventricle of 41 patients with AS (female n=21, male n=20), 17 patients with MR (female n=6, male n=11) and 17 healthy control hearts (CON) (female n=8, male n=9). Protein extracts are digested to peptides and analyzed by high-resolution tandem mass spectrometry (LC-MS/MS). A deep reference proteome is used for MS1 matching. RAW data are processed with MaxQuant and LFQ intensities are used for disease- and sex-specific protein expression analyses. Joint analyses are performed on proteomic data and clinical imaging (cardiac magnetic resonance imaging) phenotypes. A subset of biopsy samples from AS (n=17) and CON (n=6) are additionally used for quantification in RNA sequencing. The results from disease-specific differential expression analysis are used for validation of proteomic findings where necessary.

distributed to the Max Delbrück Center for Molecular Medicine for proteome measurements and to Berlin Institute of Health Genomics Core Facility for RNA sequencing.

4.3.2 Assessment of the clinicome

Similar to SMART, the EurValve study was partially conducted at the German Heart Center Berlin (DHZB) following similar protocols in terms of imaging and proteomic data assessment rendering them well comparable.

Cardiovascular magnetic resonance imaging and post-processing

All cardiovascular MRI examinations are performed using a whole-body 1.5 Tesla MR system (Achieva R 3.2.2.0, Philips Medical Systems, Best, The Netherlands) using a five-element cardiac phased-array coil. Post-processing is performed using View Forum (Philips Medical Systems Nederland B.V; View Forum R6.3V1L7 SP1). Gapless balanced Turbo Field Echo (bTFE) cine 2-dimensional short axis sequences are obtained using a previously applied MRI protocol for left ventricular mass, volume, and function [108]. Calculation of extracellular volume (ECV) follows the procedure as described in Doltra et al. (2014) [109].

$$ECV = (1 - hematocrit) * \frac{1/T_{myopost} - 1/T_{myopre}}{1/T_{bloodpost} - 1/T_{bloodpre}}$$

with *myo* = LV midwall myocardial T1 value and *blood* = LV blood pool T1 value. *Pre* and *post* refer to the measurement before and after contrast administration. Myocardial fibrous tissue content, i.e., absolute ECV (aECV), was calculated using the following equations:

$$aECV = LV_{myovol} * ECV$$

$$LV_{myovol} = LV_{mass}/1.05$$

with 1.05 being the constant of myocardial density given in g/ml.

Quality control and data set preparation

Briefly, data, which was entered into the SMART IT platform for the SMART and EurValve studies is downloaded in key-value format and merged into one file. Key-value format is transformed into a matrix representation of all available information, only omitting redundant and detailed textual information. Further data cleansing includes removing irrelevant details or a summary of values into groups. For example, information on medication, such as prescription name, dose, or interval, is summarized into binary values for seven medication groups: anticoagulation, beta-blockers, diuretics, ACE inhibitors, statins, calcium channel blockers, and angiotensin II receptor blockers, while details are omitted. For certain parameters, rules of missing data deduction are applied. For example, if a subject belonging to the SMART cohort did not have a value for the "Dislipidemia" parameter, the patient did not have the diagnosis and the value could be filled with `FALSE`, whereas for the EurValve cohort, the diagnosis status is not known, leaving the value to stay NA. Furthermore, some parameters are added or calculated, such as the body mass index (BMI) and binary representation of the heart being hypertrophic or dilated as defined by [110].

Statistical analysis

For the study population's clinical characteristics, we compare the differences with a two-tailed Student's T-test in case of normally distributed numerical values or with a chi-squared test in case of discrete categorical values both from the R stats package (v 3.5.1).

4.3.3 Transcriptome analysis

The Berlin Institute of Health Genomics Core Facility provided us with the protocol they followed to prepare samples and conduct sequencing to supply us with three technical replicates of raw FASTQ RNA sequencing data per sample. Parts of the analysis pipeline (steps from initial merging to gene expression quantification) are implemented and executed through Layal abo Khayal. I performed all other steps and summarizing quality control with MultiQC.

Sample preparation and sequencing

Total RNA is extracted from cardiac tissue biopsies utilizing the RNAqueous®-Micro Kit (ThermoFisher Scientific). RNA integrity is assessed with the High Sensitivity RNA assay for TapeStation 4200 (Agilent Technologies). All samples are assorted according to their RNA integrity number values in subgroups for library preparation. Due to low sample concentrations, library preparation is performed with the SMARTer® Stranded Total RNA-Seq Kit v2 - Pico Input Mammalian kit (TaKaRa Clontech) according to the manufacturer's instructions. In brief, 500 pg of total RNA is used for first-strand synthesis with random priming. Due to its terminal transferase activity, the reverse transcriptase adds a few non-template nucleotides to the 3' end of the complementary DNA, which in turn serves as an annealing site for the template switching oligo mix.Thereby an extended template is created, enabling the reverse transcriptase to generate the second strand. After the first round of PCR that attaches the full-length Illumina adapter sample barcode, the ribosomal complementary DNA was depleted utilizing ZapR and mammalian-specific R-probes. ZapR specifically cleaves ribosomal complementary DNA-R-probe hybrids. Non-degraded fragments are enriched by a second polymerase chain reaction with universal primers. Purified libraries are quantified by Qubit® 3.0 fluorometer with the Qubit® High Sensitivity double-strand DNA assay (ThermoFisher Scientific) and analyzed on a TapeStation 4200 system with the High Sensitivity D1000 ScreenTape® assay (Agilent Technologies). All libraries are sequenced on an Illumina HiSeq 4000 platform with 100 base pair paired-end reads.

Raw data processing

RNAseq data is used to quantify gene expression and to extract genomic variation.RNA sequencing reads are processed to extract long non-coding, short non-coding, and protein-coding messenger RNA abundance from RNA sequencing raw reads. The following section documents all details of the processing steps.

Merging, quality control, and trimming

Prior to all analysis, each sample's technical replicates are merged to yield one FASTQ file for forward and one file for reverse reads. Quality control is done by fastqc (v 0.11.5) before and after trimming. Low-quality reads and adapters are removed by Trimmomatic (v 0.36) in palindrome mode using TruSeq3-PE adapter sequences (2 seed mismatches, 30 palindrome clip threshold, 10 simple clip threshold), a sliding window of 4 bases with minimum quality of 15, maxinfo target length 45 and strictness 0.5, a min length of 45, leading and trailing bases with minimum quality of 5, crop to a maximum of 98 bases and headcrop of 3 bases at the beginning. Only paired reads are considered for further analysis. Only paired reads are considered for further analysis. FASTQC reports are summarized using MultiQC (v 1.6), resulting in overall quality reports.

Gene expression quantification

Reads are aligned to the reference genome (Homo-sapiens GRCh38, downloaded from Ensemble on 23 Jan 2018) using the gene library .gtf file (GRCh38.91 from Ensemble on 23 Jan 2018) by TopHat2 (v2.1.1.Linux_x86_64). The same library file is used for the final raw count calculation by featureCounts (subread package v 1.5.1) with paired-end read settings.

The Ensemble BioMart a table (Homo Sapiens, GRCh38.P10) containing the following attributes: Gene name, Gene description (full gene name), Gene type, Gene ID is used to assign a biotype to each quantified transcript. The biotype can be one of the four categories of protein-coding, pseudogenes, long non-coding, and short non-coding [1].

Differential expression analysis

All analyses are performed using R (v 3.5.3). For differential expression analysis, we removed all samples without concurrent proteome data (n: AS = 17, CON = 6). Counts are filtered using Limma's `filterByExpr()` function with min.count set to 10. Linear models for the full sample set are calculated using the Empirical Bayes procedures for residual variance estimation and mean-variance trend correction from Limma (v3.38.3). The contrast is stated to represent the AS vs. CON design. P-values are multiple-testing corrected by BH methodology.

[1] https://www.ensembl.org/Help/Faq?id=468,(accessedon05.02.2019)

4.3.4 Proteome analysis

Biopsy samples and raw data are processed at the Max Delbrück Center for Molecular Medicine. For the sake of completeness, we summarize the procedure here. Quality control, data set preparation, differential abundance analysis, and visualization was planned and executed by me and discussed in consortia meetings.

Sample preparation

For protein extraction, biopsies are lysed in 200 μl lysis buffer containing: 2% SDS, 50 mM ammonium bicarbonate buffer, and EDTA-free Protease Inhibitor Cocktail (Complete, Roche). Samples are homogenized at room temperature using FastPrep-24^{TM} 5G Homogenizer (MP Biomedicals) with 10 cycles of 20 s and 5 s pause between cycles. After heating the samples for 5m̃in at 95° C, 5 freeze-thaw cycles are applied. 25 U of Benzonase (Merck) is added to each sample and after incubation for 30 min the lysates are clarified by centrifugation at 16,000 g for 40 min at 4° C. Protein concentration is measured (Bio-Rad DC Protein assay) and 100 μg of each sample is further processed using the SP3 clean-up and digestion protocol as previously described [111]. Briefly, each sample is reduced with dithiothreitol (10 mM final, Sigma) for 30 min, followed by alkylation with chloroacetamide (40 mM final, Sigma) for 45 min and quenching with dithiothreitol (20 mM final, Sigma). Beads (1 mg) and acetonitrile (70% final concentration) are added to each sample and after 20 min of incubation on an over-head rotor bead-bound protein are washed with 70% ethanol and 100% acetonitrile. 2 μg sequence-grade Trypsin (Promega) and 2 μg Lysyl Endopeptidase LysC (Wako) in 50 mM HEPES (pH 8) are added and after overnight incubation at 37° C peptides are collected, acidified with trifluoroacetic acid and cleaned up using StageTips protocol [112].

Heart reference sample for matching library

A peptide mix for each experimental group (CON, AS, and MR) are generated by collecting 10 μg peptides from each sample belonging to the corresponding group. Equal peptide amounts from each group mixture are combined, desalted using a C18 SepPak column (Waters, 100 mg), and dried down using a SpeedVac instrument. Peptides are reconstituted in 20 mM ammonium formate (pH 10) and 2% acetonitrile, loaded on an XBridge C18 4.6 mm x 250 mm column (Waters, 3.5 μm bead size) and separated on an Agilent 1290 High-Performance Liquid Chromatography (HPLC) instrument by basic reversed-phase chromatography, using a 90 min gradient with a flow rate of 1 ml/min, starting with solvent A (2% acetonitrile, 5 mM ammonium formate, pH 10) followed by increasing concentration of solvent B (90% acetonitrile, 5 mM ammonium formate, pH 10). The 96 fractions are collected and concatenated by pooling equal interval fractions. The final 26 fractions are dried down and resuspended in 3% acetonitrile/0.1% formic acid for LC-MS/MS analyses.

LC-MS/MS analyses

Peptide samples are eluted from stage tips (80% acetonitrile, 0.1% formic acid), and after evaporating organic solvent peptides are resolved in sample buffer (3% acetonitrile/ 0.1% formic acid). Peptide separation is performed on a 20 cm reversed-phase column (75 μm inner diameter, packed with ReproSil-Pur C18-AQ; 1.9 μm, Dr. Maisch GmbH) using a 200 min gradient with a 250 nl/min flow rate of increasing Buffer B concentration (from 2% to 60%) on an HPLC system (ThermoScientific). Peptides are measured on an Orbitrap Fusion (individual samples) and Q Exactive HF-X Orbitrap instrument (reference sample) (ThermoScientific). On the Orbitrap Fusion instrument, peptide precursor survey scans are performed at 120K resolution with a 2×105 ion count target. MS2 scans are performed by isolation at 1.6 m/z with the quadrupole, HCD fragmentation with normalized collision energy of 32, and rapid scan analysis in the ion trap. The MS2 ion count target is set to 2x103 and the max injection time is 300 ms. The instrument is operated in Top speed mode with 3 s cycle time, meaning the instrument would continuously perform MS2 scans until the list of non-excluded precursors diminishes to zero or 3 s. On the Q Exactive HF-X Orbitrap instrument, full scans are performed at 60K resolution using 3x106 ion count target and maximum injection time of 10 ms as settings. MS2 scans are acquired in Top 20 mode at 15K resolution with 1x105 ion count target, 1.6 m/z isolation window, and maximum injection time of 22 ms as settings. Each sample is measured twice, and these two technical replicates are combined in subsequent data analyses.

RAW data processing

Data are analyzed using the MaxQuant software package (v1.6.2.6) [5]. The internal Andromeda search engine is used to search MS2 spectra against a decoy human UniProt database (HUMAN.2019-01, including isoform annotations) containing forward and reverse sequences. The search included variable modifications of oxidation (M), N-terminal acetylation, deamidation (N and Q), and fixed modification of carbamidomethyl cysteine. Minimal peptide length is set to six amino acids and a maximum of three missed cleavages is allowed. The FDR is set to 1% for peptide and protein identifications. Unique and razor peptides are considered for quantification. Retention times are recalibrated based on the built-in nonlinear time-rescaling algorithm. MS2 identifications are transferred between runs with the “Match between runs” option, in which the maximal retention time window is set to 0.7 min. The integrated LFQ quantification algorithm is applied. Gene Symbols assigned by MaxQuant are substituted with gene symbols of the reported UniProt IDs from the used FASTA file.

Quality control and data set preparation

LFQ intensities are extracted from the MaxQuant results and filtered to exclude reverse database hits, potential contaminants, and proteins only identified by site, i.e., proteins identified only by modified peptides. LFQ intensities are log_2 transformed. Homogeneous sample-wise intensity distribution is checked.

Total measured proteins, as well as sample-wise coverage, are assessed as counts. The main axes of variation are calculated in a PCA. Both methods serve as indicators to exclude samples from the analysis: two samples with particularly low coverage need exclusion as well as six samples that skew the PCA analysis because of heavy blood contamination. In human tissue samples, blood contamination is a common artifact, which leads to the detection of mainly large blood proteins instead of proteins of interest. Samples for which the percentage of intensity attributed to blood proteins exceeds 40% of summed LFQ intensity are excluded. Blood proteins are defined by a list from Doll et al. (2017) [103]. Furthermore, we exclude duplicate samples by ensuring equal variance and selecting the sample with higher coverage.

Missing values are imputed by random draws from a Gaussian distribution with 0.3*standard deviation and a downshift of 1.8*standard deviation of the observed values per sample. Furthermore, we employ a filter to exclude proteins detected in less than 50% of samples of at least one group in the respective comparison.

Differential abundance analysis

Analyses are performed using R (v 3.5.3). All proteins with less than 50% valid values in at least one compared group are excluded. A moderated t-test with intensity-trend correction and corresponding Bayesian models for continuous variables are calculated by the Limma package (v 3.38.3). The model formula represents the condition comparisons AS vs. CON, MR vs. CON, and AS vs. MR, and sex-stratified comparisons ASmale vs. CONmale, ASfemale vs. CONfemale, MRmale vs. CONmale, and MRfemale vs. CONfemale. P-values are multiple-testing corrected by BH methodology. Adjusted p-values of <0.05 are considered significant.

Table 4.1: Overview on the definition of result subsets from the comparison of conditions. Table entries denote the direction of effect for significant changes, i.e., ↑ = positive fold change, ↓ = negative fold change or the effect being n.s. = not significant. AS = aortic valve stenosis, MR = mitral valve regurgitation, CON = controls.

	AS vs CON	AS vs MR	MR vs CON
condition-specific	↑	↑	n.s.
	↓	↓	n.s.
	n.s.	↓	↑
	n.s.	↑	↓
shared	↑	-	↑
	↓	-	↓
divergent	↑	↑	↓
	↓	↓	↑
differential in AS vs MR	n.s.	↑/↓	n.s.

We define effects of particular interest in the comparison of conditions as described in Table 4.1. Condition-specific effects are changes found significant in a condition versus control and versus the respective other condition, while the direction, i.e., negative (down) or positive (up) fold change, needs to be conserved. The effect in the other condition vs. CON has to be non-significant (see example in Figure 4.3A). Please note that the direction of effect in AS vs. MR can be switched to MR vs. AS by converting signs. Shared effects are mainly defined through the same direction of effect in both conditions when compared to controls. The effect may, but does not have to be stronger in one condition, i.e., the effect in AS vs. MR is not considered (example in Figure 4.3B). Differential and diverging effects consider the AS vs. MR effect, while the effect may be evident only in that comparison (differential) or show the opposing direction of effects also when compared to control samples (divergent). An example is shown in Figure 4.3C.

In the sex-stratified analysis, we explore proteins found significant in one sex only within a condition. We assume that significant changes found in both are also represented in the condition-specific analysis. The direction of effect is considered in the enrichment analysis, as further described in section 4.3.4.

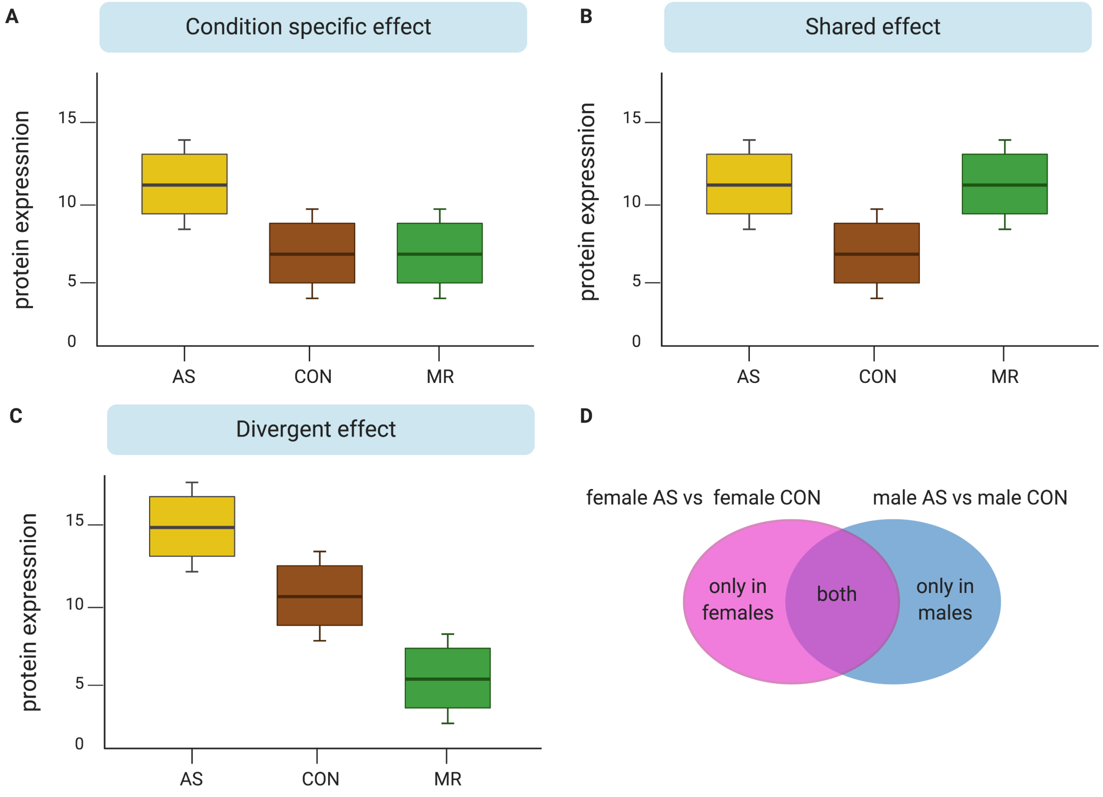


Figure 4.3: Examples to illustrate the definition of result subsets for effects in the comparison of conditions (A-C) and in the sex stratified analysis (D). Conditions AS = aortic stenosis = yellow, MR = mitral regurgitation = green, CON = controls = brown, female = pink, male = blue. Non-overlapping boxes of protein expression in A)-C) resemble significant changes. Circles in D) resemble the set of significant changes in the sex-stratified comparisons.

Enrichment analysis and annotation

Enrichment analysis is performed by gprofiler2 R package (v 0.1.8) with a background set of all detected proteins and two query sets of all up- or down-regulated proteins respectively per condition comparison and sex-stratified comparison in order of largest to smallest absolute fold change against the GO branches biological process (BP), cellular compartment (CC), and molecular function (MF). P-values are controlled by 5% FDR and significant terms are filtered for sets in which the intersection size is less than 5% of the measured proteins of a set and sets of minimum size three. Multiple entries with identical matched gene lists within a GO branch are reduced to the one with the lowest p-value. Further reduction of terms for pie charts is achieved via REVIGO using the following settings: medium reduction, against the homo sapiens database, SimRel similarity measure, and without an order of terms [113]. Manual category assignments for GO terms are given in the appendix (Table D.1). Organelle assignments are adopted from Doll et al. (2017) [103] by mapping UniProt IDs.

4.3.5 Visualization

Schematic drawings are created using the BioRender software. Heatmaps are drawn using the pheatmap R package (version 1.0.12). Proteins to include in a heatmap are combined from gene names enriched in GO terms within a category. Condition group means of log_2(LFQ intensities) are centered and scaled protein-wise and clustered with default values. All other plots are created using ggplot2 (v 3.2.1), ggpubr (v 0.2.5) and cowplot (v 1.0.0) R packages. Box plots show the median and upper and lower hinges representing the 75th and 25th percentile. The whiskers extend from one hinge to the largest/smallest value no further than 1.5 times the interquartile range away. Data outside this range are shown as outlier points. In many boxplots, individual measurements are jittered within the category on the x-axis for a better impression of the actual data.

Intersections are visualized using the UpSetR package (v 1.4.0). We select the most frequent representative of a cluster of redundant terms from the REVIGO annotation for pie charts. Frequency is the percentage of human proteins in UniProt which are annotated with a GO term in the GOA database [114], i.e., a higher frequency denotes a more general term. PCA's are calculated on unscaled, centered matrices, i.e., the filtered data set for the transcriptome and the imputed data set for the proteome.

4.4 Results

This section describes our study cohort concerning their clinical characteristics and summarizes quality control of the transcriptomic and proteomic data. We compare subgroups of the cohort with regard to differences in their clinical and molecular phenotype and report condition and sex-specific effects. For better comprehension, we arrange relevant results to represent biological entities, rather than analytical procedures.

4.4.1 Clinicome - Study cohort

We included 75 human left ventricular myocardial samples in the present study. Samples were taken from 41 patients with AS and 17 patients with MR during valve replacement surgery, and 17 healthy cardiac organ donors (CON) without cardiovascular diseases, whose hearts were not used for transplantation due to non-medical reasons. Patient characteristics are described in Table 4.2. All condition and sex-specific differences in clinical measurements are visualized in Appendix B (Figure B.1 and Figure B.2), whereas relevant parameters are selected to be shown in the respective sections. The pressure gradient from within the left ventricle across the aortic valve to the aorta is increased in AS, whereas it is normal in MR subjects. Similarly, the grade of mitral valve regurgitation is moderate to severe in MR subjects and mild or non-existing in AS subjects.

Aortic valve insufficiency is mild to moderate in both cohorts, such that no further hemodynamic load contributes to cardiac remodeling. AS and MR show a similar degree of hypertrophy as denoted by the indexed myocardial mass and compared to the reference given by Doltra et al. (2014) [109]. The proportion of female and male subjects within groups is balanced. The left ventricular ejection fraction is reduced slightly in both conditions. Co-morbidities are sparse in both cohorts and, if present, balanced across conditions. The age between conditions is different. However, within conditions, the age is similar between males and females (Appendix B Figure B.3).

Table 4.2: Clinical cohort description. Data are presented as counts (%) or mean ± standard deviation. Grades are coded in none/mild, moderate, severe. Statistical comparison between AS and MR is tested by two-sample Wilcoxon-rank test in case of numeric data and χ^2 test in case of categorical data. A reference value is given whenever values for the CON group are not available. ACE-inhibitor = Angiotensin Converting Enzyme-inhibitor, AS = aortic valve stenosis, BMI = body mass index, MR = mitral valve regurgitation;Mean pressure gradient aortic valve describes severity of AS, while the mitral valve regurgitation describes severity of mitral valve insufficiency and are given in bold.

Preoperative Parameters	AS n = 41	MR n = 17	p	CON n = 17 or reference
Age, years	68 ± 9	60 ± 14	0.03	44 ± 15

BMI, kg/m2	28 ± 4	27 ± 3	0.34	25 ± 5
Sex (female), n (%)	21 (51)	6 (35)	0.41	8 (47)
Systolic blood pressure, mmHg	140 ± 19	131 ± 16	0.12	117 ± 21
Diastolic blood pressure, mmHg	74 ± 11	76 ± 14	0.67	74 ± 17
Hypertension, n (%)	27 (69)	11 (65)	1	0
Dyslipidemia	8 (21)	3 (18)	1	0
Diabetes mellitus, n (%)	7 (17)	2 (12)	1	0
Coronary artery disease, n (%)	2 (5)	2 (12)	1	0
Atrial fibrillation paroxysmal	2 (5)	2 (12)	0.71	0
Atrial fibrillation permanent	0 (0)	2 (12)	0.15	0
Left ventricular enddiastolic volume, ml/m2	73 ± 17	108 ± 35	<0.001	74 ± 11 [109]
Left ventricular myocardial mass, g/m2	71 ± 20	67 ± 15	0.385	56 ± 9 [109]
Mean pressure gradient aortic valve, mmHg	**56 ± 15**	**4 ± 8**	**<0.001**	<5 [115]
Mitral valve regurgitation, frequency of grade	**41, 0, 0**	**0, 10, 7**	**<0.001**	0
Aortic valve insufficiency, frequency of grade	36, 5, 0	17, 0, 0	0.321	0
Left ventricular ejection fraction, %	60 ± 7.4	64 ± 6.2	0.13	70 ± 6 [109]
Medication: ACE inhibitor, n (%)	15 (37)	5 (29)	1	0
Medication: Beta blocker, n (%)	20 (49)	10 (59)	0.358	0
Medication: Diuretics, n (%)	12 (29)	5 (29)	1	0

4.4.2 Overview on the myocardial transcriptome

Quality control shows that adapter sequences and technical contamination are removed. Even after trimming, duplication rates are relatively high, which is a less concerning warning in RNAseq experiments in contrast to DNA, where no single sequence should cover more than 0.1% of all sequences. The RNAseq library usually is far less diverse. While checking the duplicated sequences in a manual nucleotide BLAST, we found that some are common across samples and that they map to mitochondrial and long non-coding RNA sequences or they do not map at all. Sequences that do not map are dropped in the alignment process, whereas long non-coding and mitochondrial RNA are potential targets of interest in our analysis and thus should be kept. Quality control plots on RNAseq data after trimming are given in Figure B.4. The density of log_2 counts per million before and after filtering of low-count transcripts for AS and CON is plotted in Figure 4.4A and shows the removal of a large portion of very low-count transcripts, especially in (female) AS. The filtering step results in homogeneous distributions of log_2-transformed counts per million per sample (Figure 4.4B). The PCA reveals a clear separation of AS and CON samples (Figure 4.4C) and large within-group variability more prominent in AS. In the differential expression analysis, we find 824 genes down-regulated in AS vs. CON and 344 to be up-regulated (Figure 4.4D).

Please note that we use the transcriptomic analysis results only for proteomic result validation, where appropriate.

4.4.3 Overview on the myocardial proteome

Four samples are excluded from the main analysis because the percentage of LFQ intensity, which could be assigned to stem from blood particles as defined by Doll et al. (2017) [103], exceeded 40%. Furthermore, nine samples (proteome measurement ID: C02, C03, C19, C21, C22, C15, C24, EV11, EV25) are measured from two distinct specimens taken from the same subject. This duplicate layer of replicates cannot be adequately modeled in differential abundance analysis, and thus we select the samples with the highest coverage. Quality control plots before exclusion of samples can be found in (Appendix B Figure B.9).

To facilitate deep proteome analysis for each sample, we generate an ultra-deep heart reference proteome data set from an equally mixed reference sample consisting of equal parts of AS, MR, and control samples. Using two-dimensional liquid chromatography prior to tandem mass spectrometry analysis, we identify a total 8,365 distinct protein groups. This deep reference proteome is used to match MS2 identification to peptide precursors across individual runs.

Filtering results in uniform coverage across all samples with an average of 3,561 (+/- 187) proteins quantified per individual sample (Figure 4.5A).

Overall, Myosin Heavy Chain 7 (MYH7), Titin (TTN (major isoform)), and Actin Alpha Cardiac Muscle 1 (ACTC1) represent the first quartile of total cumulative protein intensity. Together with Actinin Alpha 2 (ACTN2), these proteins are in line with the top abundant proteins described in human heart tissue as found by Doll et al. (2017) [103]. Other proteins typically found in heart tissue like collagens or heat shock proteins are less abundant, however quantified robustly to be considered for analysis (Figure 4.5B). The PCA analysis (Figure 4.5C) shows a good separation of AS subjects from MR and CON along the first PC, while the second PC separates MR from CON. The AS cohort shows larger differences to the controls and overall higher within-group variability. The intensity distribution after log_2 transformation is uniform over all samples (Figure 4.5D). A significant relationship between age and protein expression within conditions is not evident, although the covered range of age is large (Appendix B Figure B.3). The density of values after missing value imputation shows a slight local maximum at about 24, which can be attributed to imputed values. The local maximum is highest for female and male AS samples, whereas it is very similar in MR and CON males and lowest for female MR and CON.

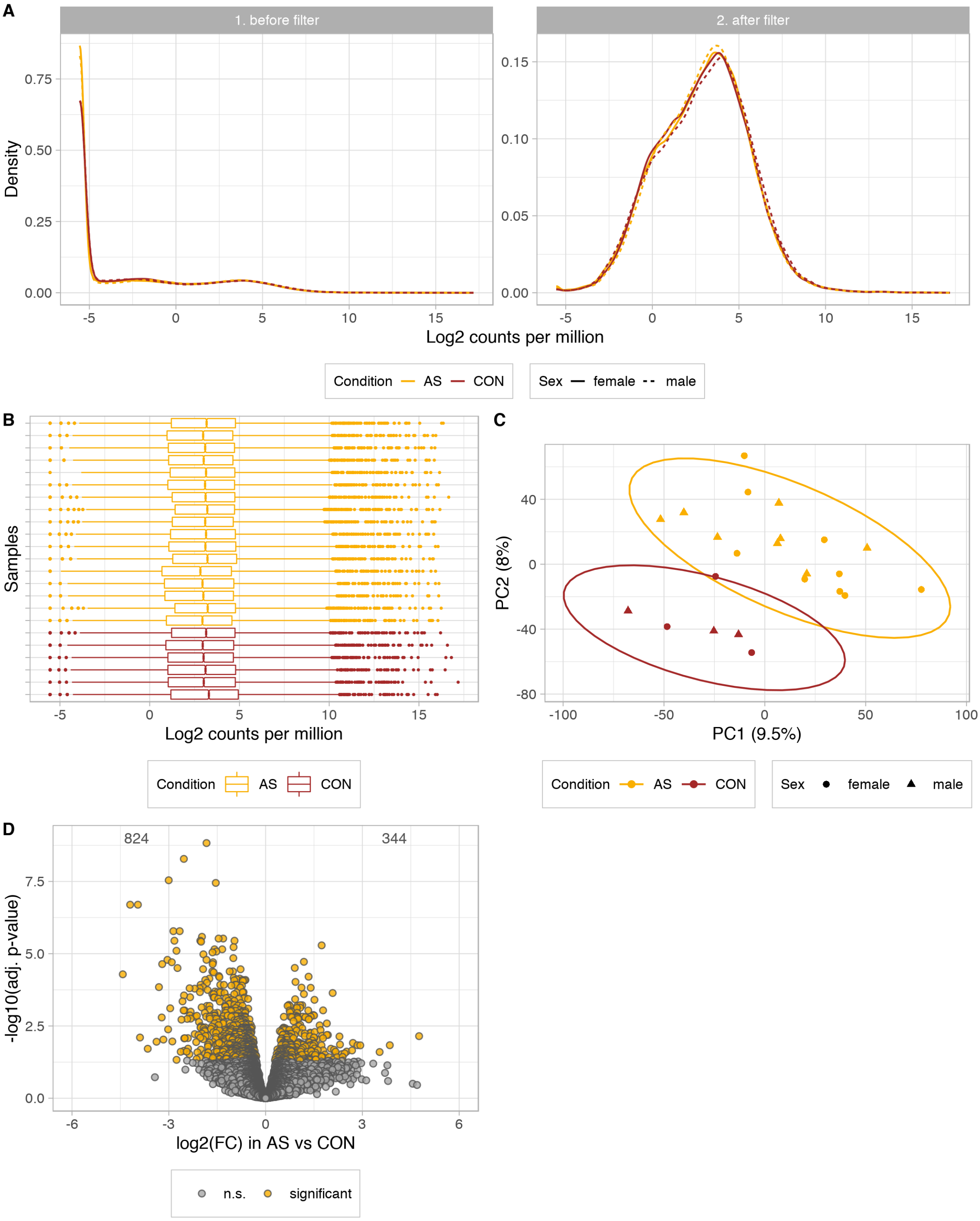


Figure 4.4: Cardiac transcriptome quantification and analysis summary. A) Density of log_2 counts per million before and after filtering of low-count transcripts for AS and CON. B) Counts per million distribution per sample after log_2 transformation and filtering. C) Principal component analysis (PCA) of filtered log_2 counts per million for AS and CON and both sexes. D) Volcano plot denoting the fold change (FC) and p-values for the comparisons of AS vs. CON. Significance is given by BH adjusted p-values < 0.05. AS = aortic valve stenosis, CON = control.

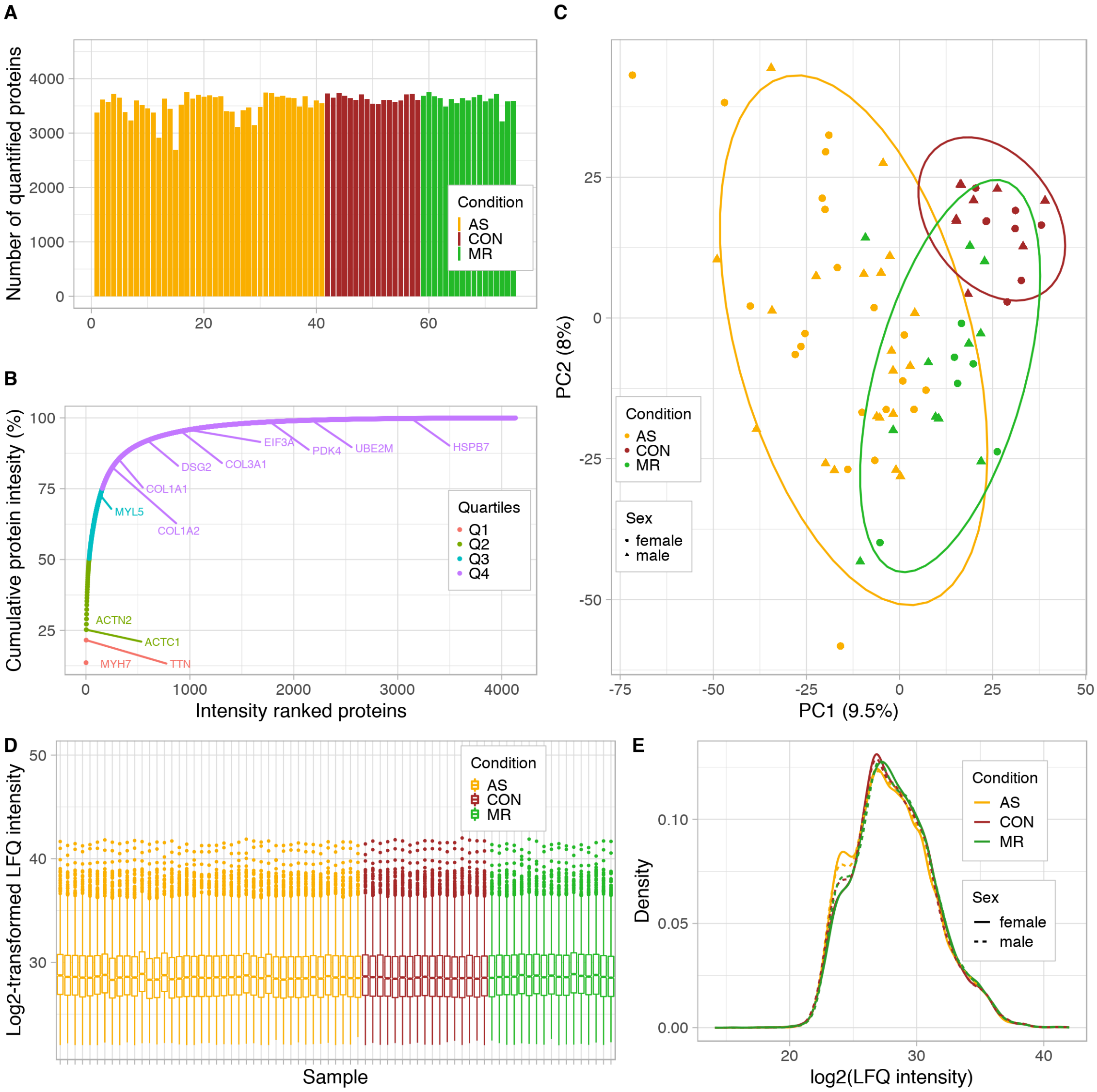


Figure 4.5: Cardiac proteome coverage and distribution. A) Count of quantified proteins per sample B) Cumulative protein intensities in % ordered across the respective protein rank C) Principal component analysis (PCA) of protein measurements displaying all three conditions and sex assignment. Note: In total four points for male CON and two points for female CON overlap giving the impression of only 7 male and 7 female CON. However, all points (9 male CON, 8 female CON) are in the diagram and contribute to the calculation of the ellipse around the group. D) Intensity distribution per sample after log_2 transformation E) Density of values after missing value imputation in condition and sex. AS = aortic valve stenosis, CON = control, MR = mitral valve regurgitation.

4.4.4 Disease- and sex-specific effects in AS and MR

Differential abundance analysis gives rise to significant differences in protein abundance between each condition compared to a healthy state (AS vs. CON (Figure 4.6A), MR vs. CON (Figure 4.6B)), as well as directly between the two different heart valve diseases (AS vs. MR (Figure 4.6C)). By applying a cutoff of BH adjusted p-value of 0.05, our comparisons result in 1332 (AS vs. CON), 400 (MR vs. CON), and 903 (AS vs. MR) differentially expressed proteins. Notably, more than two-thirds of changes show a decrease in protein expression specifically in AS samples, when comparing them to CON and to MR samples. The higher amount of significant hits and more down-regulation than up-regulation remain evident in AS even when a downsampled (n = 17) analysis is performed (see Appendix B Figure B.8). The three-group comparison allows us to define proteins being

- shared in both conditions
- divergent between conditions and
- specific for one condition

(Figure 4.6D). Shared effects show the same direction of change (270 proteins), while diverging effects show opposing directions of regulation between conditions and against control (five proteins). Condition-specific effects are those with a change in protein abundance in one condition when compared to control and when compared to the other condition (refer to section 4.3.4 for details). As such, we find 518 changes specific for AS and 79 changes specific for MR (Figure 4.6D).

We also investigate changes in females and males separately to elucidate sex-specific effects within both pathologies. When comparing to the sex-matched control, we find 462 proteins with significant differences in female AS samples only (97 up, 365 down) and 235 proteins specific for male AS samples only (116 up, 119 down). In MR, we find 70 proteins to be regulated only in females (31 up, 39 down) and 82 only in males (67 up, 15 down) (Figure 4.7A and B).

GO enrichment analysis resulted in 138 GO terms enriched in AS vs. CON, 106 terms enriched in MR vs. CON, and 25 terms enriched in AS vs. MR. REVIGO summary of all up and down-regulated GO terms and subsequent manual assignment to six categories reveal changes mainly in five distinctive categories, which are represented in both pathologies, however in differing proportions (Figure 4.6E and F):

1. extracellular matrix composition,
2. energy metabolism and mitochondria,
3. proteostasis,
4. cytoskeleton and muscle contraction and

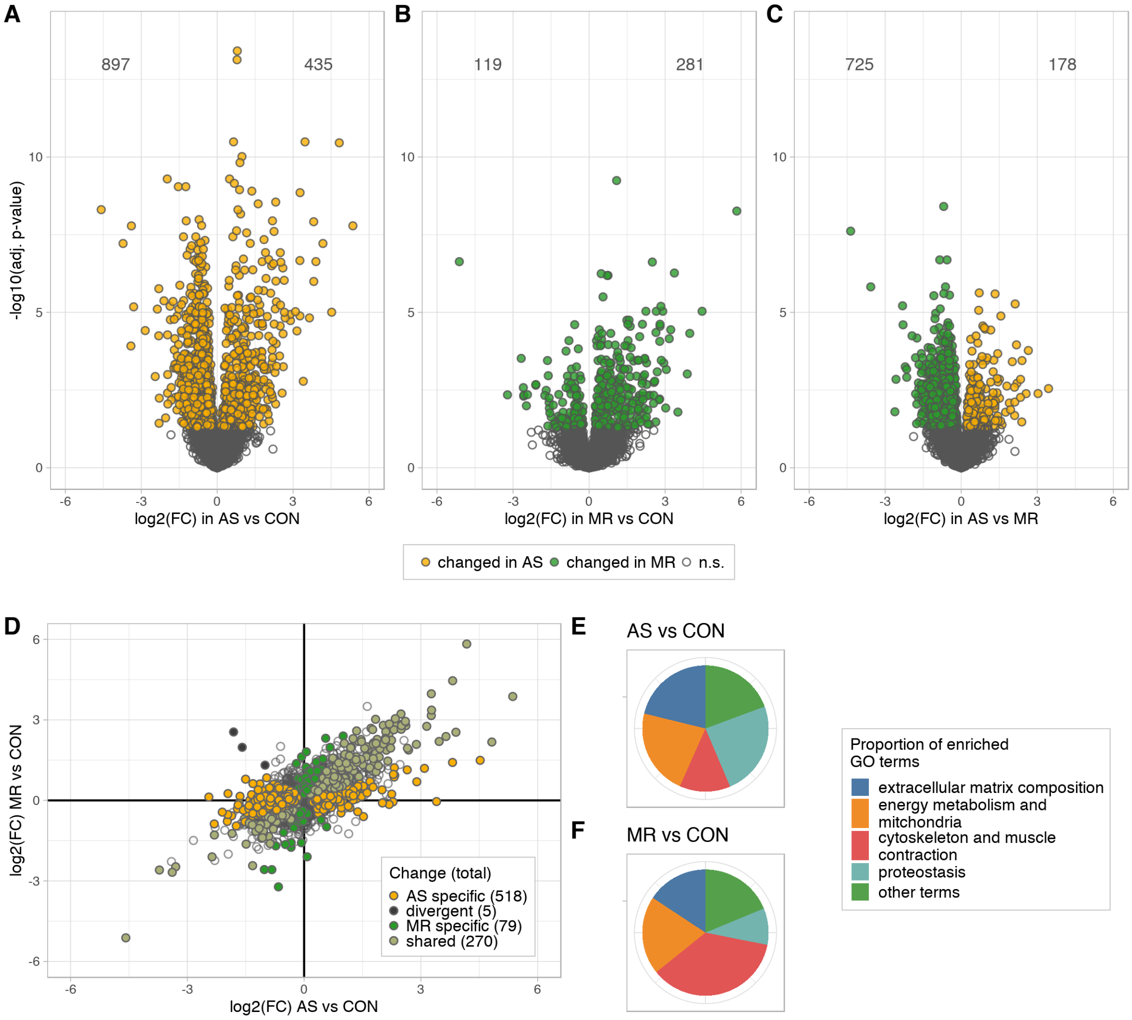


Figure 4.6: Quantitative analyses of disease- and sex-specific differences in protein abundance. Significance is given by BH adjusted p-values < 0.05. A-C) Volcano plots denoting the fold change (FC) and p-values for the comparisons of AS vs. CON (A), MR vs. CON (B) and AS vs. MR (C); D) Scatter plot of fold changes in AS and MR vs. CON. Colors denote changes that are shared, divergent, or condition-specific. E + F) Summary of GO term enrichment analysis from merged up and down-regulation. Proportions are based on REVIGO [113] summary and the GOA frequency [114] of the clusters' most general term, assigned to five categories for AS vs. CON (E) and MR vs. CON (F).

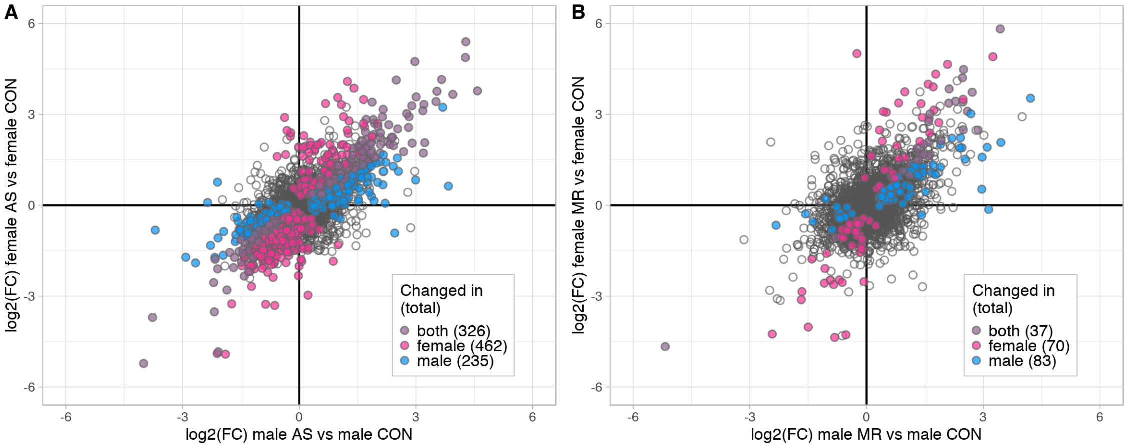


Figure 4.7: Quantitative analyses of sex-specific differences in protein abundance. Significance is given by BH adjusted p-values < 0.05. A) + B) Scatter plots of fold changes in AS and MR stratified to sexes. Colors denote changes only significant in either female, male or in both vs. their sex-matched CON.

5. other terms.

Enriched GO terms and their category assignments are available in Appendix D. Enrichments belonging to the category of other terms are based on 88 proteins with higher abundance in the diseased groups. Of these, 84% are typical body fluid components. When considering the different biopsy collection procedures for the sample groups, blood contamination becomes the most probable source of the signal and impedes any interpretation concerning physiological differences in immune response between the condition and control group. For a more elaborate explanation and visualization, please refer to Appendix B Figure B.5. Therefore we do not further examine the biological relevance of the other terms and focus on disseminating disease and sex-specific effects considering the categories extracellular matrix composition, energy metabolism and mitochondria, proteostasis and cytoskeleton, and muscle contraction.

Extracellular matrix composition

ECM confers structural and mechanical support to the tissue and its composition plays a crucial role in heart disease, especially in cardiac hypertrophy [73]. Proteins related to ECM are higher in abundance in AS and MR when compared to controls (Figure 4.8A). The majority of changes are shared in both conditions; however, disease-specific changes are evident as well. Additionally, more changes are significant in male samples only. A detailed list of GO terms related to the extracellular matrix is shown in Figure 4.8B. All proteins mentioned in the following text are labeled within the heat map shown in Figure 4.8A.

Terms like extracellular matrix, glycosaminoglycan binding, and proteoglycan binding are enriched in both conditions. Within the scope of these broad terms, we find non-fibrillar collagens (collagen type 6, 12, 14, 18) and matricellular proteins (POSTN, TGFBI) to be more abundant in both conditions. In addition, we observe concurrent effects in ECM

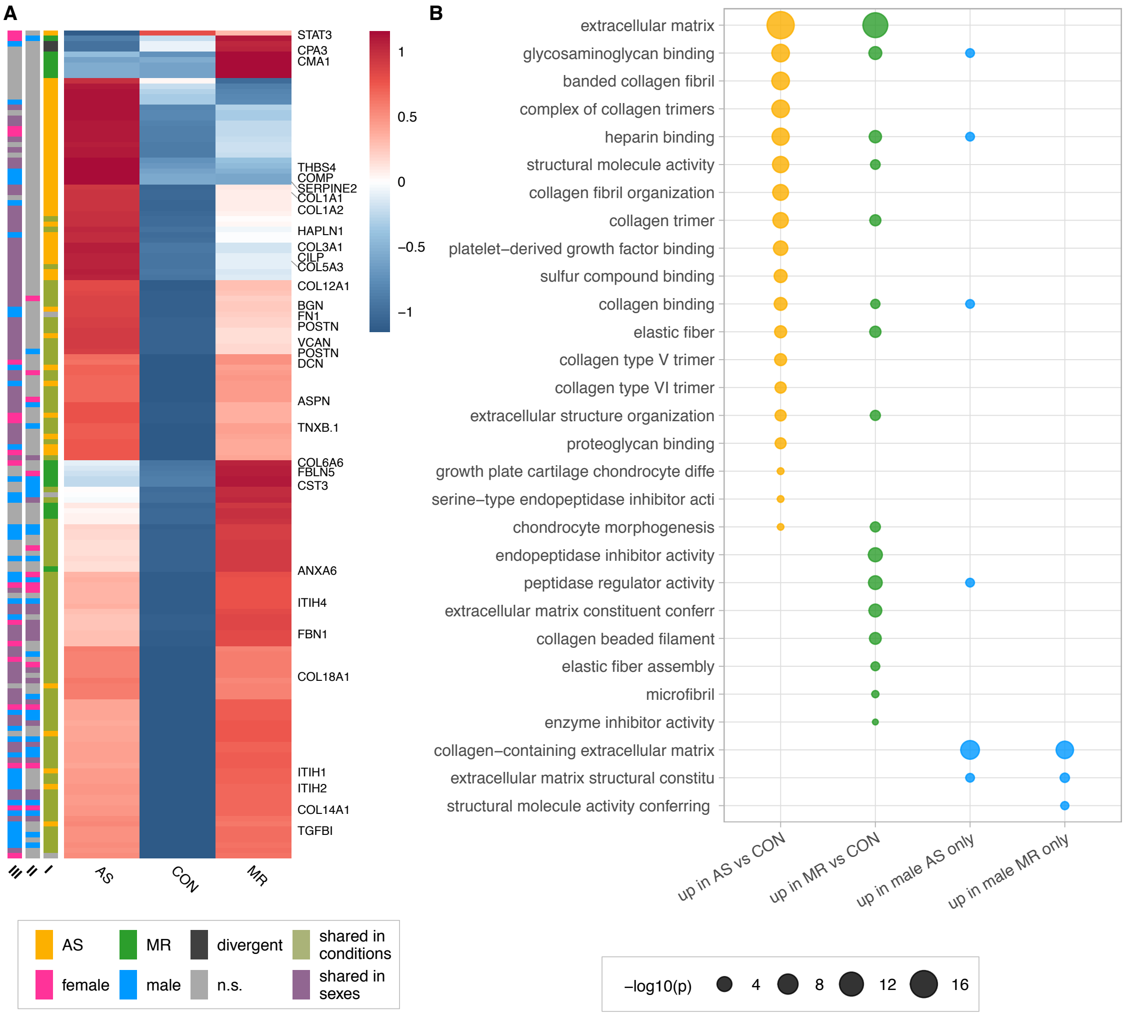


Figure 4.8: Disease- and sex-specific differences in abundance of proteins related to extracellular matrix composition. A) Clustered heatmap showing the condition's mean abundance of proteins belonging to ECM related GO terms. Annotation bars denote significant changes in condition (I) and in sex (II – effect in sex MR, III – effect in sex AS). B) Combined results of metabolic GO term enrichment analysis on up-regulated proteins for all three condition comparisons and on proteins found only in one sex of a condition. GO term names are reduced to a maximum of 40 characters, but are non-ambiguous.

glycoproteins (TNXB, FBN1), proteoglycans, such as VCAN and HAPLN1, and four members of the SLRP (small leucine-rich proteoglycans) class, namely Biglycan (BGN), Decorin (DCN), and Asporin (ASPN). BGN and DCN are believed to regulate the amount of collagen and fibrillogenesis in the heart [116].

In AS, we observe a higher amount of fibrillar collagens (e.g., GO:0098643, GO:0098644), namely type I (COL1A1, COL1A2), which forms thicker and stiffer fibers, as well as type III (COL3A1), which forms more compliant and elastic fibers. The increase in collagen types I and III mRNA is already described in AS in human tissue [117]. Here, we detect an additional increase in collagen type V in AS, which has only been described in animal studies so far [118–120]. In contrast to the literature [34, 35], in which male patients with AS show a stronger increase in collagens, we find a similar increase of collagen type I, III, and V in both female and male AS. In line with findings in pressure-overloaded hearts of mice, thrombospondin-4 (THBS4) is specifically up-regulated in AS [121]. Thrombospondin-5 (COMP) is also specifically increased in AS and may belong to the expression signature of matrifibrocytes, which have been shown to form stiff scar tissue in infarcted mouse hearts [122]. Similarly, CILP1 (Cartilage intermediate layer protein 1) is higher in abundance only in AS and is a mediator of cardiac ECM remodeling and a marker for cardiac fibrosis [123, 124].

In MR, we see specific increase of enzymatic proteins like CPA3, CPB2, and CMA1 and proteins expressed in developing arteries and epithelial cells like FBLN5 and COL6A6. Furthermore, there is an increase in Annexin 6 (ANXA6), which has been described to critically regulate the transition of chronic hypertrophied cardiomyocytes to apoptosis in cultured cardiomyocytes [125]. Interestingly, Cystatin C (CST3), a common serum marker for chronic heart failure, is increased significantly in MR only.

Interestingly, the only two matrix-metalloproteinases detected are too low in abundance to compare between groups. Matrix-metalloproteinases are well-described markers of the progression towards heart failure; however, in our samples, cardiac remodeling does not exhibit these changes (yet), very similar to results reported by Polyakova et al. (2004) [126] in human AS samples and Spinale et al. (1998) in pigs [127]. GO term enrichment of proteins increased in male AS and MR only reveals significant hits related to ECM composition. In contrast, there are no significant enrichments in females of both conditions (Figure 4.8B). Among the proteins found increased in males only are members of the ITIH family, known as ECM stabilizers, e.g., ITIH1, ITIH2, ITIH4 in male AS and ITIH4 in male MR. Serpin E2 is elevated only in male AS, which is in accordance with a previous study, where pressure-overload hypertrophy in mice led to up-regulation of Serpin E2 and accumulation of collagens, thus contributing to cardiac fibrosis [128]. Additionally, fibronectin (FN1), which has been shown to play a prominent role regarding fibrosis and cardiac function in a heart failure animal model [129], is increased in male AS and MR only. In line with a more pronounced elevation of fibrosis-associated proteins in male AS, we find lower levels of STAT3 in female AS, which as a transcription factor is discussed to be an important contributor to collagen synthesis and cardiac fibrosis [130].

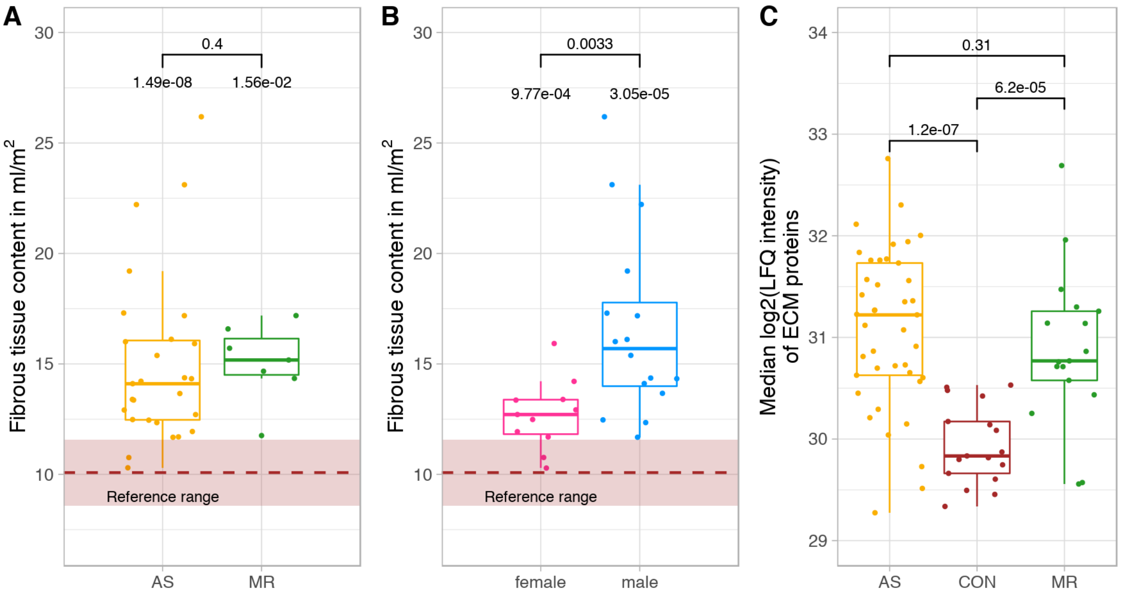


Figure 4.9: Fibrous tissue content and ECM protein intensity. Comparison of fibrous tissue content as measured by MRI in AS and MR (A) and only in AS stratified to sex (B). C) Median log_2 of LFQ intensity of all proteins belonging to the ECM. P-values are calculated via Wilcoxon rank test with two samples between groups (denoted by bracket) and one sample against the reference mean (no bracket for the normal fibrous tissue content). Dots represent individual subjects/samples.

Clinical quantitative imaging of extracellular matrix can be performed via T1 mapping [131]. In our cohort, a subset of AS and MR patients have are T1 mapped before surgery. We see higher values for fibrous tissue content in AS and MR compared to published data of patients without severe pressure or volume overload [109] without a significant difference between conditions (Figure 4.9A and B). Additionally, male AS present a higher degree of fibrous tissue content when compared to female AS. In Figure 4.9C, we used organelle annotations as published by Doll et al. (2017) [103] to compare the differences in log_2 transformed LFQ intensity of all ECM proteins in AS, MR, and CON. Concomitant with the fibrous tissue content, we find an increase in AS and MR vs. control ECM proteins but no significant difference between conditions.

Energy metabolism and mitochondria

The normal cardiac function relies on a constant high energy supply, which in the healthy heart is mainly provided by oxidative phosphorylation from fatty acid oxidation [132]. In our study, proteins involved in energy metabolism are found decreased in AS and MR compared to healthy samples, with a more pronounced effect in AS and a generally stronger effect in males for both conditions.

In Figure 4.10A, the clustered heatmap shows the mean abundance of proteins assigned to energy metabolism and mitochondria-related GO terms. The effects in condition annotation (I) show that the most significant changes are a decrease in abundance in AS. In MR, the effect is less pronounced. However, a few proteins like PYGB, SLCA2A1, and SLC27A6 are increased in both conditions.

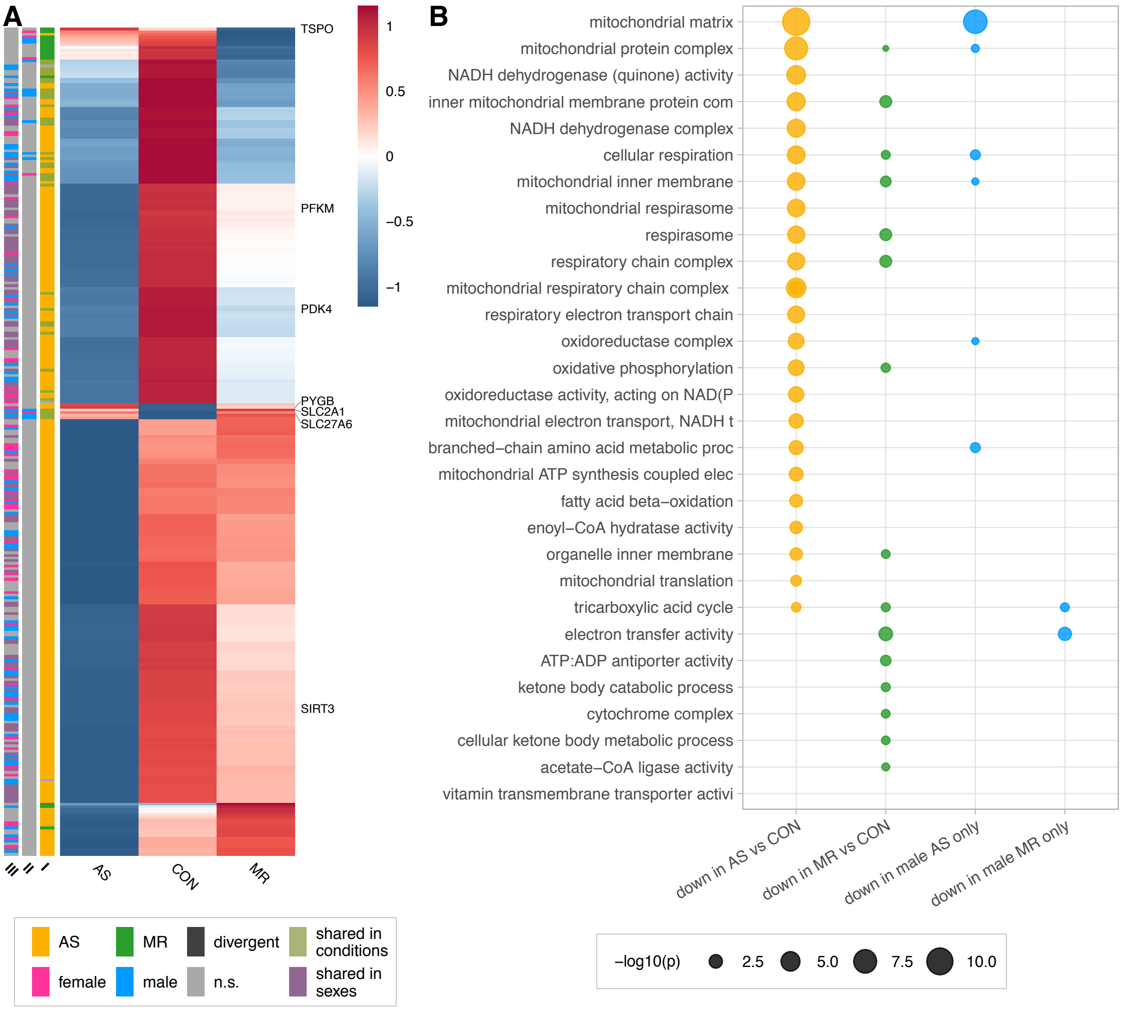


Figure 4.10: Overview on metabolic protein expression and GO enrichment. A) Clustered heatmap showing the condition's mean abundance of proteins belonging to energy metabolism and mitochondria related GO terms. Annotation bars denote significant changes in condition (I) and in sex (II – effect in sex MR, III – effect in sex AS). Proteins described in the text are labeled. B) Combined results of metabolic GO term enrichment analysis on down-regulated proteins for all three condition comparisons and on proteins found only in one sex of a condition. GO term names are reduced to a maximum of 40 characters, but are non-ambiguous.

Figure 4.10B shows GO terms summarized under metabolism and mitochondrial dysfunction for the condition and sex-stratified comparisons. Major aspects of metabolism, especially the tricarboxylic acid (TCA) cycle, respiratory chain, and oxidative phosphorylation, are down-regulated in AS and MR. The lower abundance of proteins involved in fatty acid beta-oxidation and branched-chain amino acid catabolism is mainly seen in AS. Although most effects are shared among sexes, effects found in male AS and MR only yield significant enrichments, while those found in females only do not. Among the enrichments in males are respiratory chain, branched-chain amino acid catabolism and TCA are enriched (Figure 4.10B).

In line with the GO term-based analyses, proteins assigned to mitochondria, the major site of energy generation in cardiac tissue, as published by Doll et al. (2017) [103], show a significant decrease in median log_2 transformed LFQ intensity in AS, but not in MR Figure 4.11.

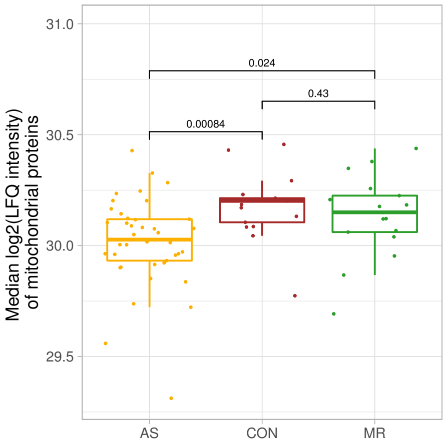


Figure 4.11: Median log_2 of LFQ intensity of all proteins with mitochondrial annotation. P-values are calculated via Wilcoxon-rank test.

We summarize changes in the metabolism in Figure 4.12 for AS and MR. Despite the decrease of proteins involved in fatty acid beta-oxidation, a major transporter for long-chain fatty acids (SLC27A6 or FATP6) is up-regulated in AS and MR. Remarkably, a major glucose transporter in cardiac tissue GLUT1 (SLC2A1) shows a 2.75-fold increase in AS and 4.8-fold in MR, while the main glucose transporter GLUT4 (SLC2A4) is 1.5-fold increased in CON vs. AS. Also, PYGB, a protein responsible for glycogen degradation, is higher in abundance in AS. However, changes in the abundance of proteins involved in glycolysis are seen in only two proteins (PFKM - slight decrease, ENO2, slight increase). Additionally, PDK4 is decreased 13-fold in AS and 6-fold in MR. PDK4 phosphorylates and thus inactivates pyruvate dehydrogenase (PDH). Less inactivation of PDH can contribute to increased glucose oxidation and, as a result, pyruvate utilization for acetyl-CoA generation. Another example of shared up-regulation is SPTLC. The enzyme is crucial in de-novo synthesis of ceramides and overexpression leads to accumulation of ceramides and subsequently to changes in the lipid profile, apoptosis, reduction of oxidative metabolism, and progression of maladaptive remodeling [133].

Down-regulation of sirtuins, i.e., mitochondrial deacetylases, has been implied in mitochondrial dysfunction found in pathological hypertrophy [134]. Sirtuin-3 (SIRT3) is significantly lower in abundance in AS compared to Control and MR samples. TSPO (translocator protein) is down-regulated in MR compared to both control and AS. TSPO belongs to the mitochondrial cholesterol/porphyrin uptake translocator protein family and has been found to be up-regulated in pressure-overloaded hearts in mice. Preventing TSPO

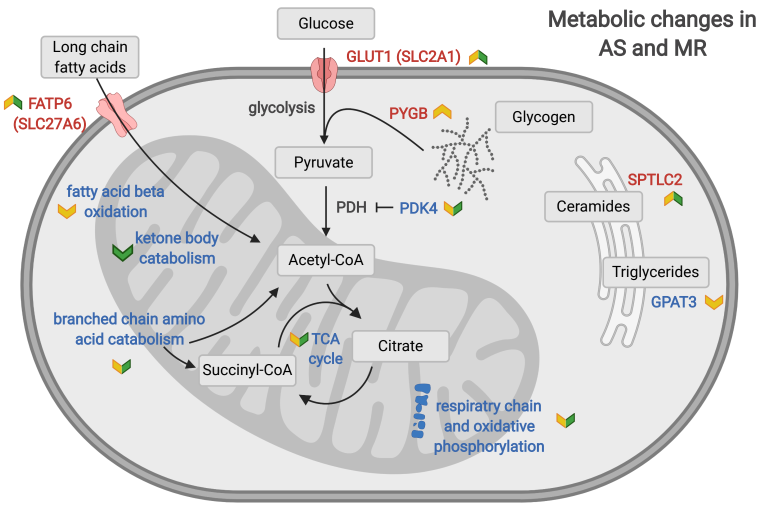


Figure 4.12: Summary of metabolic processes derived from enrichment results and selected key proteins. Colored arrows denote direction of change and the condition the change is found in: yellow = AS, green = MR.

increase limits the progression of heart failure, preserves ATP production, and decreases oxidative stress, thereby preventing metabolic failure [135]. TSPO in AS has a positive fold change, which matches observations in pressure overload by Thai et al.(2018) [135]; however, it misses the significance threshold (adj. p-value = 0.07) in our data. In contrast, significantly less TSPO in MR might point to an opposing mechanism in volume overload.

Changes in myocardial energy supply affect cardiac function, measured reliably as ejection fraction in cardiac MRI [36]. In our cohort, we see a lower ejection fraction in both conditions when compared to published reference values [109], however less severe in MR (Figure 4.13A). In line with the more pronounced decrease of proteins involved in male patients' energy metabolism, cardiac function is better preserved in females, both in AS and MR (Figure 4.13B).

Proteostasis

Proteostasis describes a balance of biological pathways including protein synthesis, folding, quality control, trafficking, and clearance, ensuring proper cell function [136]. In our comparison, proteins related to proteostasis show a major decrease mainly in AS patients and in particular in female AS samples (Figure 4.14).

Down-regulation in AS samples yields enrichments for translation including ribosomes and their subunits, protein folding, and quality control, i.e., chaperonin containing T-complex protein Ring Complex (TRiC, GO:0005832), trafficking such as protein localization to the endoplasmic reticulum (GO:0070972) (see Figure 4.15). Notably, all subunits of the TRiC are lower in abundance in AS compared to CON and MR, pointing to an AS-specific effect. The complex aids in the folding of actin and tubulin, i.e., major parts of the cytoskeleton. Almost all terms found in AS vs. MR, but not when comparing against

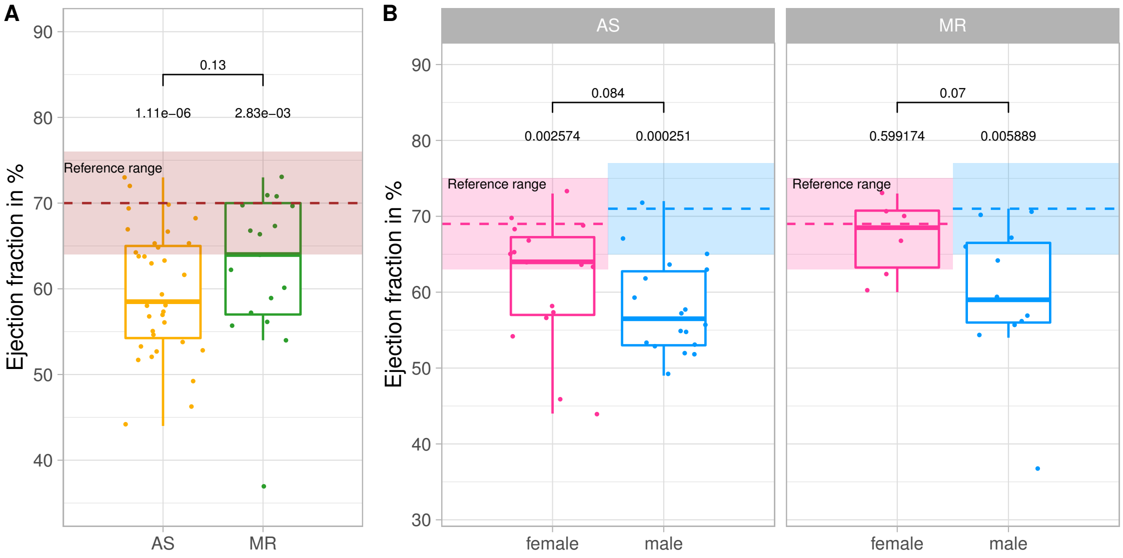


Figure 4.13: Comparison of ejection fraction in % as measured by MRI in AS and MR (A) and stratified to sex (B). P-values are calculated via Wilcoxon rank test with two samples (AS vs. MR, female vs. male, denoted by bracket) and one sample against the reference mean (no bracket). The reference range consists of the mean (dotted line) +/- one standard deviation. For A), we show averaged values from female and male ejection fraction. Dots represent individual subjects.

the controls, are based on proteasome subunits (GO:0000502), i.e., protein degradation (Figure 4.15). As such, the changes in proteasomal subunits are subtle (also see Appendix B Figure B.6), but it is the only concept across the whole analysis that is neither shared nor condition-specific, however not strong enough to be fully divergent between conditions.

Heat shock proteins have been described to play a role in cardiac hypertrophy [137]. The most abundant small heat shock protein in cardiomyocytes αB-Crystallin (CRYAB) has been shown to suppress pressure overload cardiac hypertrophy in mice [138]. In our cohort, αB-Crystallin is down-regulated in AS and MR samples. Heat shock protein beta-7 (HSPB7) is a cardioprotective stabilizer for large sarcomere proteins, whereas a loss leads to autophagic compensation to degrade accumulated protein aggregates [139]. We detect all three isoforms of HSPB7 robustly in control samples, but significantly less HSPB7 (isoform 1 and 2) abundance in AS and MR samples. Furthermore, Hsp70 (HSPA1B) is found to be down-regulated in both disease groups, AS and MR, and Hsp70 knockdown has been described to induce cardiac dysfunction and development of cardiac hypertrophy [140]. Finally, TRAP1/HSP75, known to protect the heart from hypertrophy, was found down-regulated in AS samples only [141]. Few proteins show up-regulation; among them is UCHL1, a deubiquitinase just recently described to stabilize epidermal growth factor receptor, subsequently leading to increased hypertrophy [142].

In the sex-stratified analysis, we recover most translation effects in female AS only (Figure 4.15). More specifically, effects split into cytosolic translation being decreased in females and mitochondrial translation being lower in males. For example, 13 out of

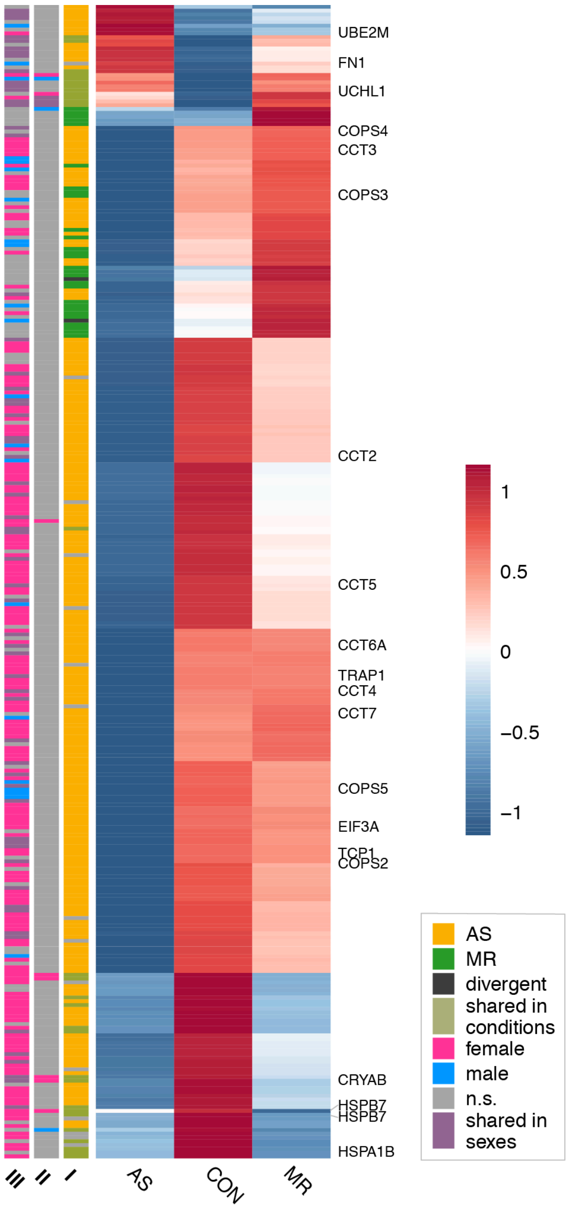


Figure 4.14: Overview on proteostasis protein expression. Clustered heatmap showing the condition's mean abundance of proteins belonging to ECM related GO terms. Annotation bars denote significant changes in condition (I) and in sex (II – effect in sex MR, III – effect in sex AS). Proteins described in text are labeled.

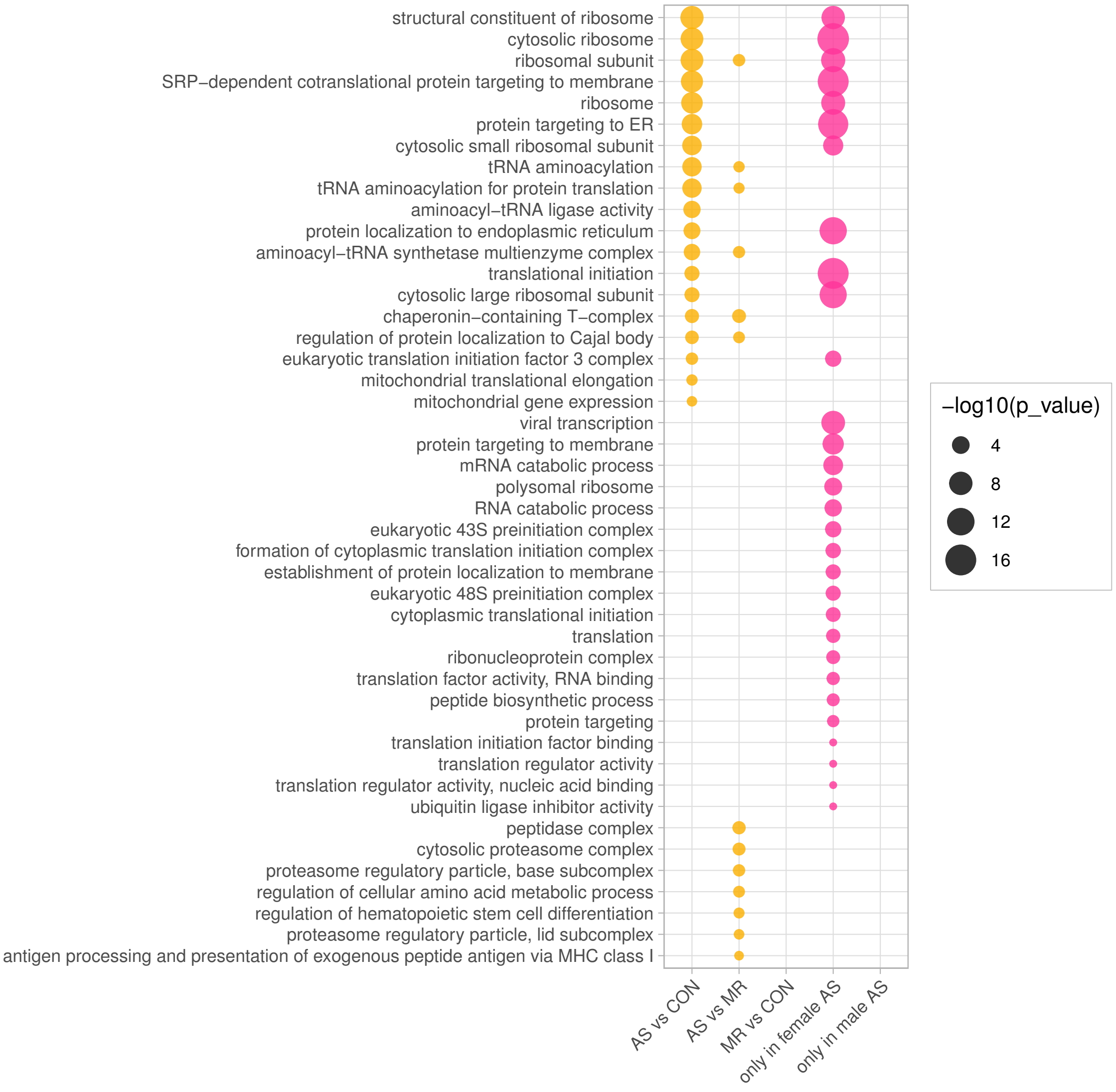


Figure 4.15: Combined results of GO term enrichment analysis on down-regulated proteostasis proteins for all three condition comparisons and on proteins found only in one sex of a condition.

31 detected 40S and 27 out of 42 60S ribosomal subunits are changed in females only; another 9 ribosomal subunits are down-regulated in both. Additionally, we found 16 out of 39 detected subunits spanning all cytosolic translation initiation factor complexes to be less abundant in females compared to none regulated in males (Figure 4.16). Despite the loss of power in the sex-stratified analysis, i.e., only half the sample sizes, nine initiation factor (IF) and ribosomal subunits are only significant in the stratified analysis, not in the overall condition comparison.

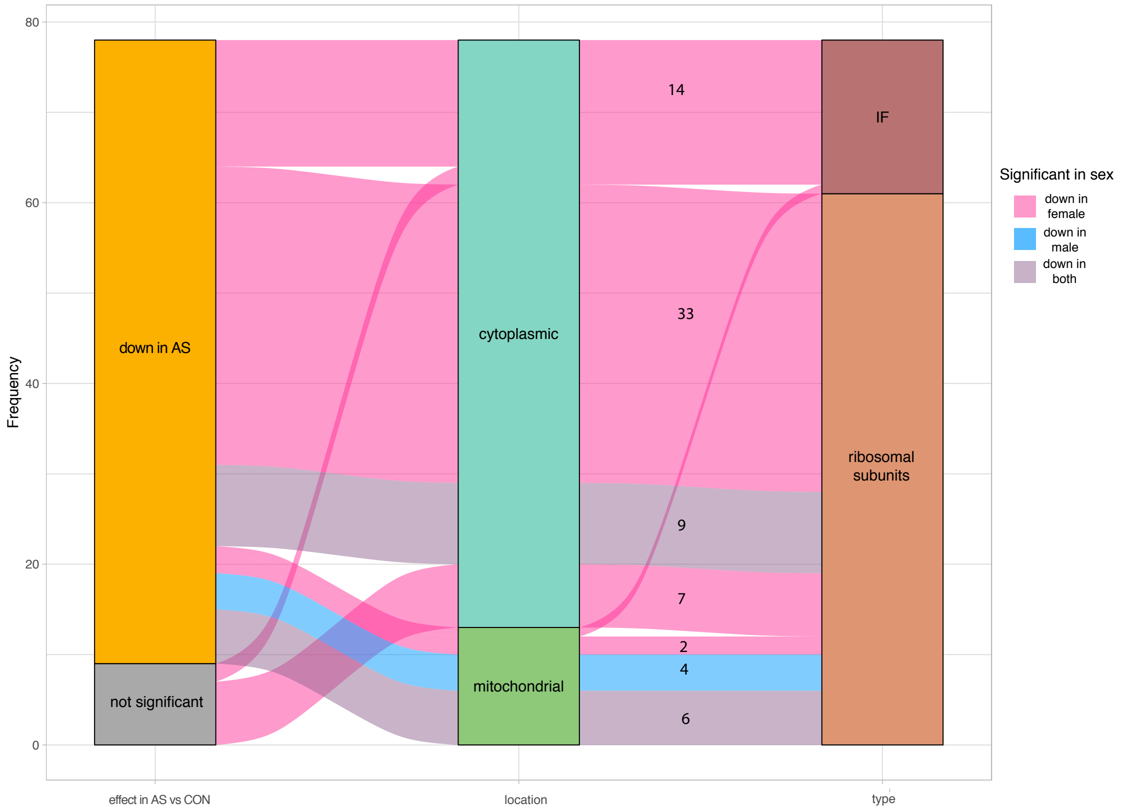


Figure 4.16: Sex-stratified analysis of ribosomal subunits and translation initiation factors (IF) in AS. Frequency of cytoplasmic and mitochondrial ribosomal subunits and translation initiation factors coded through the significant changes found in the sex-stratified analysis in AS. IF = translation initiation factors.

One particular IF, EIF3A, has already been shown to ameliorate cardiac fibrosis [143] and shows a 1.8-fold down-regulation in female samples. Lower tRNA aminoacylation (e.g., GO:0043039) is enriched in both sexes and thus represented in the AS vs. CON comparison; however the proteins behind the enrichment reveal cytosolic enzymes in females, whereas in males the mitochondrial tRNA aminoacylases are less abundant. Protein degradation is regulated via the ubiquitin-proteasome system and lysosomal autophagy. We find some evidence of disturbed degradation and neddylation processes in AS samples, as one major neddylase UBE2M is up-regulated. UBE2M activates the cullin scaffold proteins, which form a potent ubiquitin ligase complex and are crucial for the degradation of many target proteins. However, cullins (1,4 and 5) are lower in abundance in AS patients. Similarly, several COP9 signalosome subunits, which regulate protein homeostasis in the heart, are down-regulated in AS. The disturbance of protein homeostasis may lead to cardiac proteotoxicity.

In female AS patients of our cohort, we see less left ventricular mass, i.e., less hypertrophy, than in male AS patients despite comparable left ventricular pressure load due to severe aortic valve stenosis (Figure 4.17).

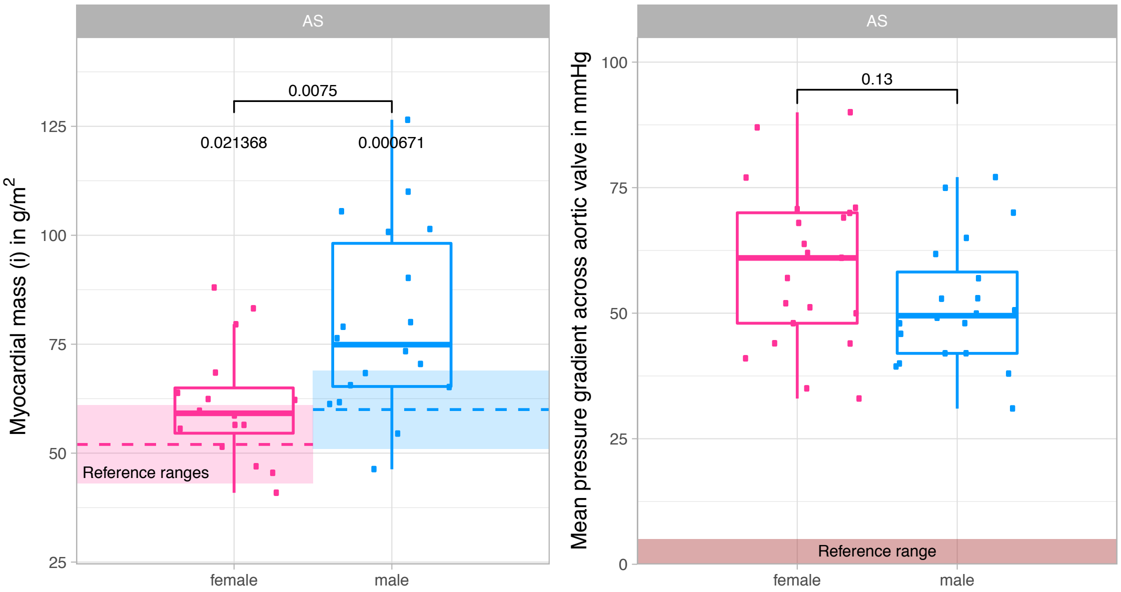


Figure 4.17: Comparison of indexed myocardial mass and the mean pressure gradient across the aortic valve as measured by cardiac MRI in AS stratified to sex. P-values are calculated via Wilcoxon rank test with two samples (female vs. male, denoted by bracket) and one sample against the reference mean (no bracket, only for myocardial mass). Reference ranges for myocardial mass are given as mean (dotted line) ± one standard deviation [109], and the pressure gradient being $<$ 5 mmHG in healthy subjects [144]

Furthermore, proteome measurements suggest a reduced cytosolic protein synthesis capacity in female AS patients, which might be why this difference is seen in myocardial hypertrophy between male and female patients. Proteasome-related proteins are needed for cardiac hypertrophy progression and have been discussed as therapeutic targets to prevent or reduce cardiac hypertrophy [145, 146].

Cytoskeletal, adhesion, and contractile proteins

The cytoskeleton plays a crucial role in maintaining cellular stability and reacting to mechanical stressors through signal transmission and subsequent remodeling. Myofibrils represent the cardiomyocytes' contractile entities and are connected to the ECM and adjacent cells through the cytoskeletal network and adhesion proteins. In Figure 4.18A, GO terms belonging to the cytoskeleton, adhesion, and muscle contraction enriched in AS and MR vs. CON are shown. All other comparisons did not yield any significant enrichment results. We find terms like actomyosin, contractile fiber, and spectrin-associated cytoskeleton to be enriched in both conditions. Similarly, the median expression of proteins belonging to the myofibril and actin-binding proteins as assigned by Doll et al. (2017) [103] is increased significantly in both conditions when compared to CON (Figure 4.18B).

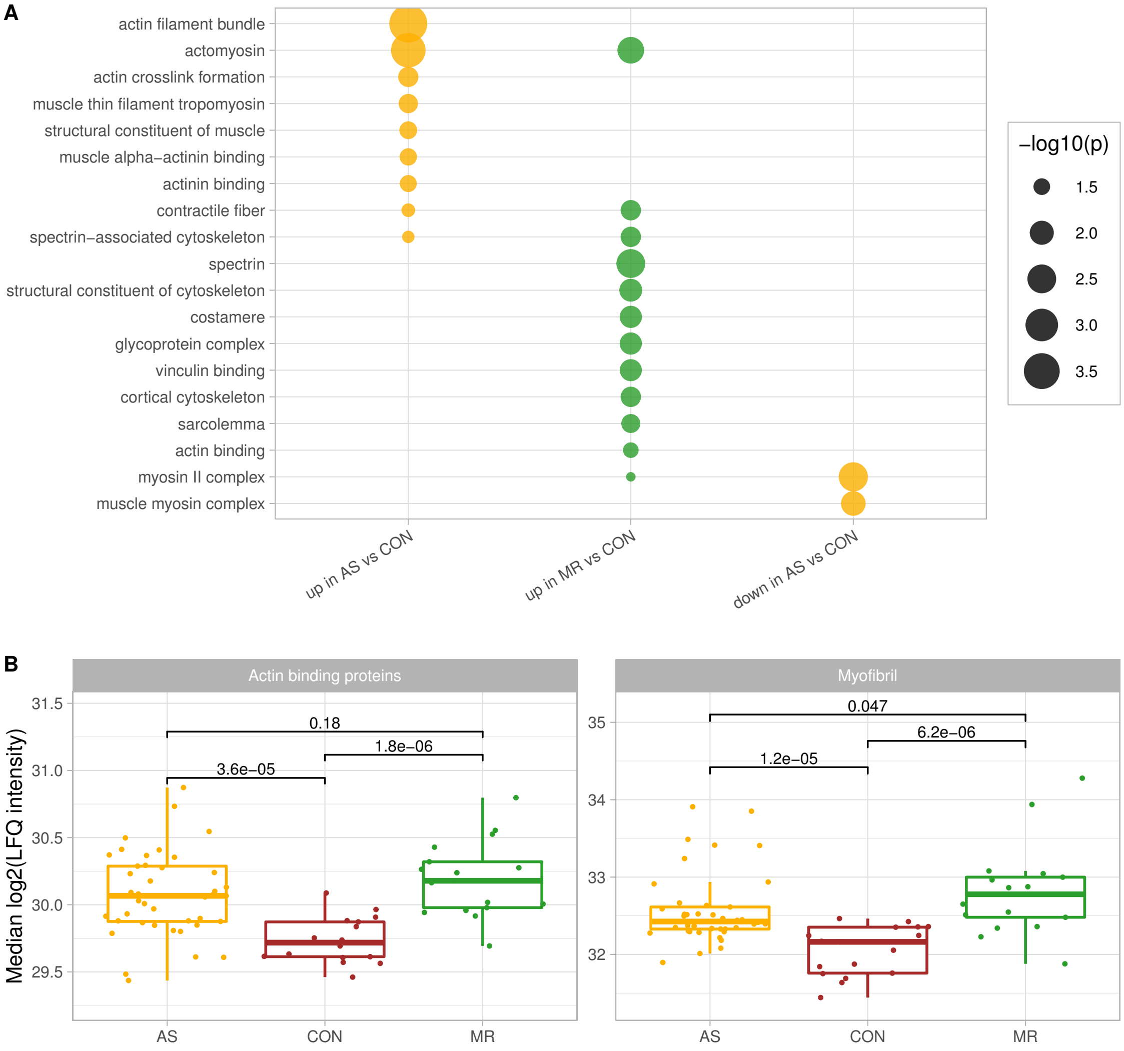


Figure 4.18: Functional and organellar assignments of preoteins in the muscle and cytoskeleton category. A) Summary of GO term enrichment analysis in condition and sex-specific analyses. The negative log10 transformed p-value defines the size of circles for terms enriched in AS vs. CON in yellow and MR vs. CON in green. B) Sample-wise median log_2 transformed LFQ intensities in for all proteins belonging to the actin-binding proteins and myofibrils in AS, MR and CON. P-values are calculated via Wilcoxon-rank test.

Shared effects in actin-binding proteins are found, e.g., in protein 4.1 (EPB41, EPB42) and Filamin A (FLNA), which are both involved in anchoring actin filaments to the membrane. Alpha-actinin 1, another actin-binding protein, was long considered to be expressed in endothelium only, however higher abundance in AS and MR would also support deposition of ACTN1 (see Figure 7F) in cardiomyocytes from patients with aortic stenosis and dilated cardiomyopathy [147]. Intermediate filaments, such as vimentin (VIM), synemin (SYNM), and the nuclear lamin A/C (LMNA), are up-regulated in both conditions. They act as bridges between cell organelles and the sarcolemma. Furthermore, we detect the up-regulation of SYNPO2 (Synaptopodin2). Members of the SYNPO family regulate actin filament assembly, and, e.g., SYNPO2 is known to organize actin bundles in parallel along the long axis of the cell. Many contraction-associated proteins like tropomyosin (TPM1, TPM3) and troponins (TNNI1, TNNT2) are higher in abundance, whereas others are down-regulated (MYH7, MYL5 and 12B, SMPX). Interestingly, many changes in actomyosin proteins may well be assigned to non-sarcomeric structures such as smooth muscle cells (MYL12B, MYH11, CNN3) or to non-muscle myosin 2B (MYH10), for which cardiac remodeling has been described when it is increasingly deposited at costameres in rats and mice [148].

The abundance of proteins involved in calcium handling is shown in Figure 4.19. CASQ2 (Calsequestrin) is higher in abundance and PLN (Phospholamban) levels are lower in AS and MR. Ryanodine receptor 2 (RYR2) is decreased in AS only. RYR2 is a crucial receptor in cardiac calcium-dependent excitation [149]. Additionally, ATP2A2 (SERCA2a), a calcium pump, is lower in abundance in AS, while HRC (Sarcoplasmic reticulum histidine-rich calcium-binding protein) is up-regulated. In combination with the changes in myosin heavy and light chains, these alterations may lead to contractile dysfunction [150].

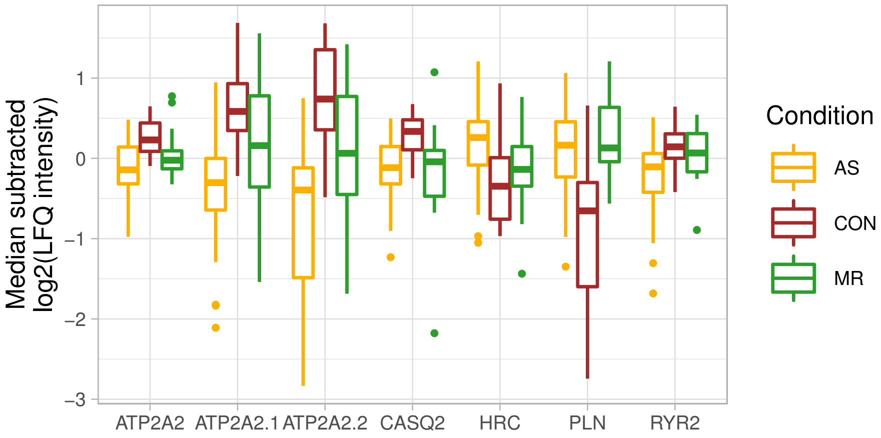


Figure 4.19: Median subtracted log_2 of LFQ intensity for calcium handling proteins. CASQ2 - Calsequestrin, PLN - Phospholamban, RYR2 - Ryanodine receptor 2, ATP2A2 - Sarcoplasmic reticulum histidine-rich calcium-binding protein (several isoforms). P-values are not shown, but for all displayed proteins we find at least one significant (adjusted p-value of < 0.05) result in one of the condition comparisons performed with Limma.

Further disparities are evident in the composition of desmosomes and adherens junctions, i.e., within the intercalated discs over which cardiomyocytes are connected. For example, PKP2 and alphaE-catenin show higher abundance only in AS. Concerning cell-cell adhesion in MR, we find desmoglein-2 (DSG2), desmoplakin (DSP), and alphaT-catenin to be up-regulated specifically.

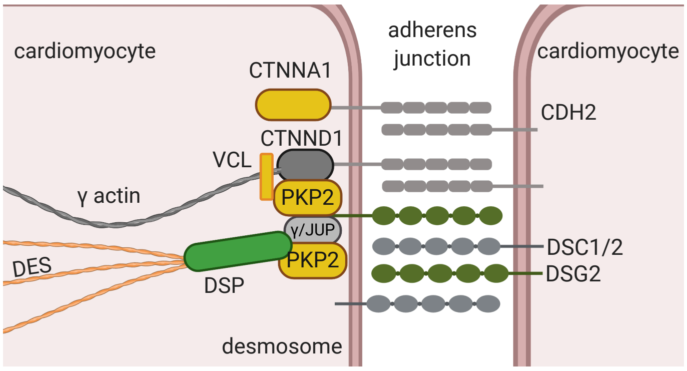


Figure 4.20: Up-regulation in protein abundance at the desomosome, i.e., the connection between two cardiomyocytes, specific for MR in green and AS in yellow. DSG2 - desmoglein-2, DSP - desmoplakin, PKP2 - Plakophilin-2, CTNNA1 - alpha-catenin, VCL - vinculin, DES - desmin, DSC1/2 - desmocollin-1 and -2, CTNND1 - Catenin delta 1, CHD2 - N-cadherin, JUP - Junction plakoglobin. Grey: Detected, but not regulated.

We find several AS-specific effects (Figure 4.21), such as a decrease of alpha-integrins (ITGA1, ITGA5, ITGA6, ITGAV) and melusin (ITGB1BP2), which interacts with integrin beta-1 and which was found to have a protective effect in response to chronic pressure overload is also down-regulated in AS samples [151].

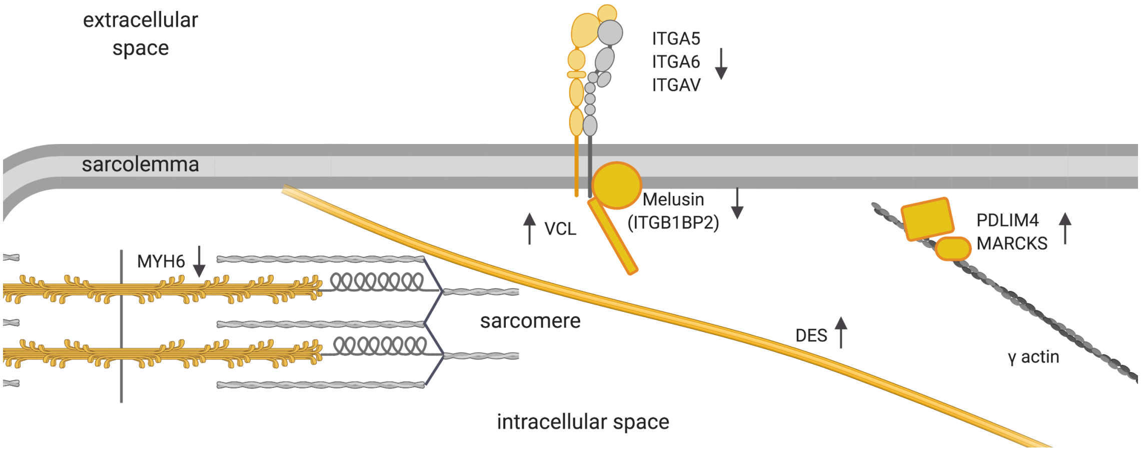


Figure 4.21: AS-specific changes in in the cytoskeleton and sarcomere. Colors denote a significant change in the respective protein and arrows indicate on the direction of change.

Vinculin (VCL) is higher in abundance in AS and is believed to enhance the stiffening of cells in response to strain, and therefore, cells become less susceptible to further deformations [152]. With regard to intermediate filaments, we found Nestin (NES) and Desmin (DES) to be of higher abundance only in AS. Nestin is expressed only during early heart development [153]. Desmin is the major connector of costameres, desmosomes, myofibrils, nucleus, and other organelles within the cardiomyocyte and was found up-regulated in heart failure [99]. Further inspection of the term actin filament bundle

(GO:0032432), revealed a multitude of proteins involved in actin bundles and stress fiber formation (FSCN1, PDLIM1, MARCKS) and in transducing mechanical stress signals towards the nucleus (TRIP6, ZYX, LPP, ABLIM1, SEPT7).

AS-specific alterations in muscle contraction are found, e.g., in MYH3, MYH13, and MYL6B, which contribute to the enrichment of contractile fibers in AS in general. Enrichment results among down-regulated proteins in AS are based on major cardiac myosin heavy chain isoforms MYH6, MYH7 in addition to heavy and light chains previously not believed to be expressed in cardiac tissue (MYL5, MYH4, MYH8). Altered levels of MYH6 and MYH7 manifest in a significantly lower ratio of the two sarcomeric heavy chains (MYH6/MYH7) only in AS (Figure 4.22), which is compliant with current literature [154–156].

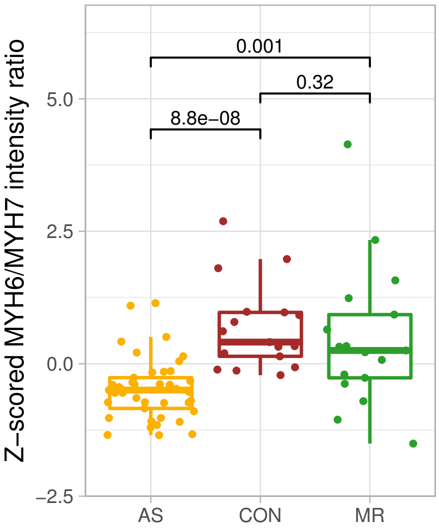


Figure 4.22: Ratio of myosin heavy chains MYH6 and MYH7 in AS, MR and CON. P-values are calculated via Wilcoxon-rank test.

Changes specific in MR (summarized in Figure 4.23) concentrate towards the structures involved in cell-matrix adhesion (costamere, sarcolemma, glycoprotein complex) and cytoskeletal proteins just beneath the sarcolemma (cortical cytoskeleton, GO:0030863). Within the dystrophin-glycoprotein complex, dystrophin (DMD), a dystrophin binding protein (SNTB2), and a membrane-spanning protein (SGCE) are higher in abundance only in MR. Actin-binding proteins with an MR-specific increase are mainly anchoring proteins such as SPTBN1, SPTB, TLN2, EBP41L2, and ADD3. Furthermore, up-regulation of the actin-binding non-muscle alpha-actinin ACTN4 provides evidence on the reactivation of fetal actinin forms described in failing hearts [157]. Proteins involved in muscle contraction show higher levels in MR for MYH2, MYH14 and lower levels for MYH7B, which were previously considered less important in adult hearts. Up-regulation of MYLK3, a cardiac-specific myosin light chain kinase, may positively affect contractility in MR [158].

While we did not find changes in the primary cardiac muscle ankyrin repeat proteins, ANKRD29 (Ankyrin repeat domain-containing protein 29) shows a strong divergent effect

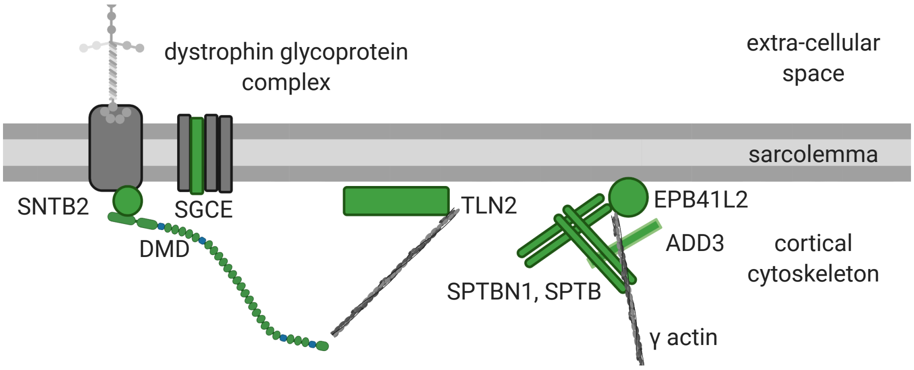


Figure 4.23: Graphical illustration of regulated proteins belonging to the cortical cytoskelelon. The color indicates an MR-specific (green) increase.

between AS and MR, favoring the direction of increase in anchor proteins in MR. However, little is known about the protein despite its ubiquitous expression.

GO term enrichment of proteins enriched exclusively in sex resulted in a single hit of the term actin-binding (GO:0003779) in male MR (not shown). However, except for GC, DST, and PFN2, the enrichment may result from the slight difference in statistical power.

Observations on ACE2

Human ACE2 (angiotensin-converting enzyme 2) is recognized as the main receptor for SARS-CoV-2. It is expressed in many organs, including the respiratory tract, kidney, and heart. Thus, some reports suggest that ACE2 plays a role in cardiac SARS-CoV-2 infection [159, 160]. ACE2 is the rate-limiting enzyme in the degradation of the fibrogenic and proinflammatory AngII (angiotensin II) peptide and therefore a major player in the pathophysiology of heart disease. In patients with AS, ACE2 protein is 4.76-fold up-regulated compared to controls (adj. P<0.0001) and 4.04-fold compared to MR (adj. P<0.001). In contrast, in patients with mitral valve regurgitation, ACE2 abundance does not show any significant differences when compared to controls (Figure 4.24A and C). To confirm the validity of these results, protein abundance is compared with available cardiac transcriptomic data of 17 patients with AS and 6 controls from the same cohort. Equivalent to the proteomic results, ACE2 is significantly (adj. P<0.05) up-regulated in AS compared to controls (Figure 4.24C). Moreover, there is a significant correlation between ACE2 protein abundance and messenger ribonucleic acid (mRNA) expression levels (R=0.6, P<0.01), suggesting a direct link between cardiac ACE2 transcription levels and the amount of generated ACE2 (Figure 4.24D).

In addition to a relevant pressure gradient across the valve, patients with AS do not have higher blood pressures than MR. The intensity of ACE2 abundance positively correlates with the pressure gradient in AS (Pearson correlation coefficient 0.36; p=0.036). Despite covering a large range of proteins, the abundance of ACE, another major angiotensin-converting enzyme, is below the detection limit and thus could not be quantified robustly (Appendix B Figure B.7). ACE mRNA is detected, and expression levels increase in AS

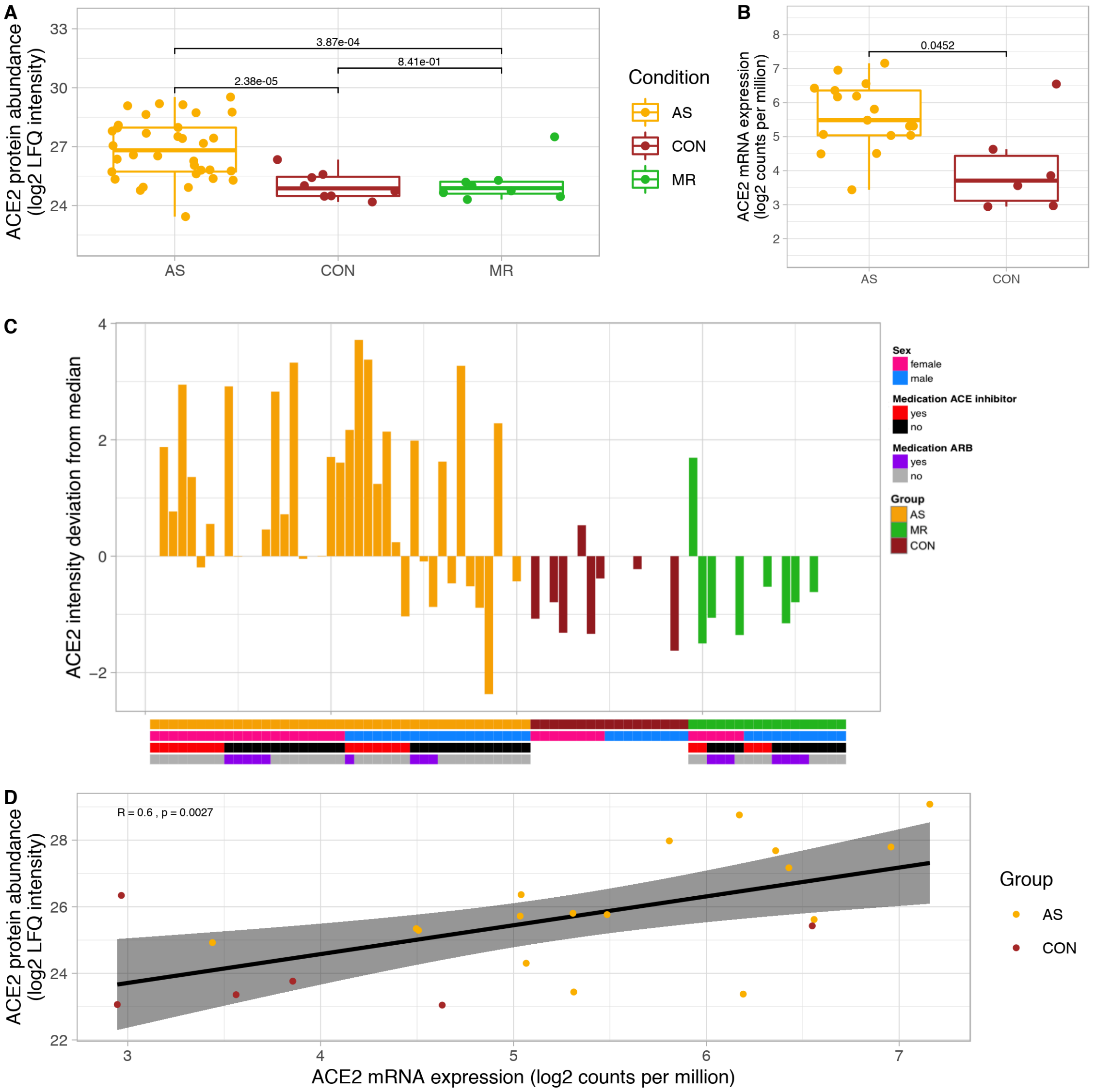


Figure 4.24: Distinct expression of ACE2, the putative severe acute respiratory syndrome coronavirus 2 receptor in heart disease. ACE2 expression values from (A) protein measurements as log_2 LFQ intensities (n = 41 AS, MR, 17 CON), and (B) RNA sequencing as log_2 counts per million (n = 17 AS, 6 CON). Individual values per patient are plotted as dots. Missing values are not shown but were down-shift imputed for statistical testing. (C) Sample-wise derivation of cardiac ACE2 protein abundance from the median log_2 expression value (set to 0). Each bar represents one sample, while annotation columns below denote selected baseline characteristics. P-values stem from the differential analysis results and are BH- adjusted. ARB – angiotensin II receptor blocker, ACE – angiotensin converting enzyme. (D) Correlation between ACE2 protein abundance and ACE2 mRNA expression levels with R being the Pearson correlation coefficient.

when compared to controls, however not significantly (adj. p = 0.06).

4.5 Discussion

Our curated data set from the SMART and EurValve studies comprises a clinical characterization of 124 subjects in total. Proteomic, transcriptomic, and genomic data further describe subsets of these individuals. Exploratory analysis motivated us to focus on the proteome and clinicome.

After overall quality control and data set description, we used differential expression and enrichment analysis to depict condition- and sex-specific alterations in the proteome together with corresponding clinical imaging data of human left ventricular myocardial samples of patients with severe aortic valve stenosis or severe mitral valve regurgitation and healthy donors. Transcriptomic data was used for validation purposes.

The largest set of deep molecular data from living individuals

Quality control of transcriptomic data showed successful removal of measurement artefacts and resulted in good quality reads. Filtering and transformation of counts yielded homogeneous distributions of read counts across all samples. The preparation of a deep proteomic reference of more than 8300 protein groups enabled a high mean coverage of more than 3500 cardiac proteins and more than 80 isoforms comparable to or more extensive than related publications [51, 103]. By providing proteomic data on 58 living subjects, we outgrow previous studies by a factor of eight [51]. For a subset of 23 subjects, there is additional data on RNA expression available.

Larger variability in AS subjects

Despite uniform intensity distributions per sample in the transcriptomic and proteomic data, we observed larger within-group variability in AS proteome and transcriptome and in MR proteome when compared to controls. Relating to this, Doll et al. (2017) report stronger per-subject variability in the proteome of atrial fibrillation samples, whereas control samples are homogeneous. The authors suspect an underlying sub-classification of atrial fibrillation subjects [103]. The differences in within-group variability can also be interpreted as greater susceptibility to personal influences in certain pathologies, in our case in AS and to a lesser extent in MR. These observations can only be ascertained in larger cohorts and need further evaluation, but if confirmed, they make the need for a personalized approach to disease even more pressing.

Low count/abundance in (female) AS

The "imputation bumps" that are strongest for AS point to more imputed values in both female and male AS, comparable to the many filtered low-count transcripts described in the transcriptomic analysis. However, in contrast to the transcriptomic data, for which all present transcripts should be "caught", but low counts are less reliable and thus filtered, it is common practice to impute missing values in proteomic data (e.g., as implemented in Tyanova, Stefka et al. (2016) [43]). Here, the reference sample provides evidence for all protein groups found in the studied tissue. However, the quantity in a sample or sample groups may lie below the detection limit of the mass spectrometer. Therefore, imputed values come from a distribution mimicking intensities at the lower detection limit. As such, we find a higher amount of low abundance proteins, i.e., missing values and low-count transcripts in AS, especially in females.

Compared to control samples, we find more than twice as many genes and proteins to be down-regulated in AS. This imbalance is in cohesion with the observation of low counts/abundance; however, it is rather unexpected with respect to literature, e.g., Kararigas et al. (2014) [161]. Interestingly, the effect vanishes in male AS samples but remains in female AS in our sex-stratified analysis. Several biologic and technical reasons are probable and thus discussed in the following together with other biologic interpretations and as part of limitations in section 4.5.

AS and MR represent equal cardiac hypertrophy of differing etiology

Our AS subjects suffer from high systolic left ventricular pressure, whereas patients with MR are subject to high diastolic volume load. Both cohorts develop a comparable increase in total myocardial mass, i.e., hypertrophy, but are different in left ventricular end-diastolic volume and myocardial wall thickness. As such, the cohort represents the two pathologies well with regard to organ and hemodynamic characteristics [162]. Co-morbidity is fairly low in general and comparable across conditions. Similarly, relevant medication is taken by both cohorts to a similar extent. As such, the cohort is well suited to study the proteomic effects of the differences in mechanical load.

A massive increase in ECM in AS and MR

The massive increase in cardiac ECM in AS is consistent with previous studies, where chronic pressure overload triggers profibrotic activation and subsequent increase in ECM and myocardial stiffness [73]. The effect of volume overload on ECM remodelling, as in MR, is mainly described in animal models so far, but in general not understood at an equivalent level of detail [73]. Compared to pressure overload, volume overload was previously associated with ECM degradation and less fibrosis [50, 102]. However, fibrosis is not defined sharply – in most publications, the amount of fibrillar collagen, measured through staining, is used as a proxy for fibrosis. As such, our finding of an increase of fibrillar collagens like Collagen I, III, and V specifically in AS and not MR is in accordance with published findings [163]. Additionally, MR-specific increase in enzymatic proteins and proteins related to vessel formation is in line with the described activation of proteases, increased vessel formation, and ventricular dilation in volume-overloaded ventricles [50, 73]. In summary, we find a massive shared increase in ECM in pressure and volume overload in the proteome and cardiac imaging. However, the major difference lies between proteins increasing myocardial stiffness to adapt to the increased pressure load in AS and a slight increase of ECM-degrading proteins in MR.

Decrease in metabolic proteins strong in MR and stronger in AS

Proper cardiac function is dependent on a high energy supply. In cardiac hypertrophy and heart failure, energy metabolism changes [164]. We show a decrease of proteins involved in energy metabolism in AS and MR subjects. The effect is more pronounced in AS – here, especially the mitochondrial proteins are reduced. A key enzyme, Sirtuin3, inhibits processes that lead to mitochondrial dysfunction and pathological hypertrophy [134]; however, it is down-regulated specifically in AS.

Additionally, we detect a stronger decrease in male AS subjects. This is also in line with a former study in which transcriptome characterization detected down-regulation of oxidative phosphorylation pathway in male, but not in female overloaded ventricles [161]. The disturbance of mitochondrial translation found in male AS offers an explanation for the strong decrease.

Interestingly, in our cohort, overall male subjects tend towards a higher degree of cardiac hypertrophy and reduced cardiac function than female patients. Similar sex differences are seen in a transcriptional profile in a murine model of pressure overload, which showed a pronounced increase in myocardial hypertrophy and fibrosis in male animals. In contrast, female animals exhibited less down-regulation of genes related to mitochondrial function and respiration [165, 166].

Proteostasis is reduced in female AS

The maintenance of healthy protein homeostasis, i.e., protein synthesis, folding, quality control, trafficking, and clearance, ensures proper cell function and is of major importance in cardiac tissue due to its limited regenerative potential [136]. At the same time, cardiac hypertrophy relies on increased protein synthesis, which depends among others on translation initiation factors and ribosomes [143, 167].

Notably, AS proteostasis changes are driven by strong downregulation in AS females. We find exclusive down-regulation of many ribosomal subunits and translation initiation factors in female AS subjects, which can ultimately decrease protein translation capacity.

Down-regulation of proteostasis proteins in AS women offers a link to lower cardiac hypertrophy and better cardiac function of female AS than male AS, thus, possibly providing a molecular explanation for the phenotypic clinical observation. In this context, we need to discern that an increase in the abundance of a protein, despite the heavy reduction of transcriptional capacity, points to a strong need for these proteins. All the more intriguing, we do not find enriched terms among the more than 100 proteins with increased abundance only in female AS. Even when raising the threshold of significance towards more proteins for enrichment analysis, there are no hits (not shown). For comparison: we do find significant GO terms in less than 50 significant proteins up-regulated in male MR. Conversely, no enrichment means a seemingly random assembly of up-regulation in female AS with regard to the GO ontology. These proteins require an extensive further manual inspection to understand female pressure overload hypertrophy mechanisms.

Druggable targets for the treatment of hypertrophy

A prominent example of a druggable target could be TSPO. Thai et al. (2019) show a recent approach to regulating the expression of TSPO in mice and re-establishing the branched-chain amino acid catabolism [135, 168]. TSPO is increased in heart failure; however, prevention of its increase is associated with better cardiac function and outcome. Based on our data, inhibition of TSPO may prove to have an effect in pressure overload, where we find TSPO up-regulated but not in volume overload hypertrophy where we find no differential regulation.

Bi et al. (2020) identified UCHL1, a deubiquitinase, as a target for hypertrophic therapy through LDN-57444 in mice undergoing transverse aortic constriction [142]. The latter is a model for pressure load. As we find UCHL1 increase in both AS and MR, further efforts in the clinical applicability of LDN-57444 could prove beneficial in both conditions.

Furthermore, proteasomal subunits are not changed compared to controls but increased in MR vs. AS. The change is subtle but contributes to the debate on the role of the ubiquitin/proteasome pathway in cardiac hypertrophy. Depre et al. (2006) describe the proteasome's activation to promote hypertrophy in mice and highlight the potential to inhibit the mechanism through epoxomicin [169]. An activated ubiquitin pathway is associated with a continued decrease in left ventricular function in volume load even after mitral valve repair [170].

The reduction of hypertrophy is an important goal in drug-based therapy of heart valve diseases. Therefore, it is essential to understand the underlying mechanisms leading to a difference in applicability, similar to ACE inhibitors' ineffectiveness in volume overload hypertrophy [102].

There are distinct cytoskeletal changes in pressure and volume load

Cytoskeletal changes can cause or be triggered by cardiac dysfunction [164]. The pressure overloaded myocardium in patients with AS has to cope with pressures of approximately 200-250 mmHg, whereas intraventricular pressures in patients with MR are between 120-150 mmHg. In contrast, end-diastolic volume, which describes the volume load on the heart, is increased only in MR patients.

In general, we find higher levels of cytoskeletal and contractile proteins to cope with the increase in mechanical load in both conditions. However, a closer look reveals a major increase in the cortical cytoskeleton in MR samples. Proteins found exclusively increased in MR are known to contribute to anchoring the cytoskeletal actin to the sarcolemma and cross-linking between cytoskeletal entities, thus providing structural integrity to the cell. Furthermore, alterations in the glycoprotein complex and desmosomal changes point to an increased interconnectedness towards the ECM and neighboring cardiomyocytes, which may be an adaptation to increased stretch caused by volume load.

In AS, among the actin-binding proteins, we find many LIM-domain-containing proteins, of which many display a role in transducing mechanical stress towards the nucleus. Additionally, proteins promoting actin bundles and stress fibers are more prominent in AS. Pressure load in the heart may have a stronger effect on mechanosensing than volume load.

Changes in contractile proteins do not necessarily stem from the typical adult cardiac sarcomere. Instead, we found many changes associated with the non-muscle, skeletal, smooth muscle, and fetal expression profiles. Non-muscle and smooth muscle contractile proteins may originate from endothelial cells or (myo-)fibroblasts. However, deposition of increased amounts of, e.g., MYH10, has been detected in cardiomyocytes in animal models before [148]. A reconsideration of the role of myosin light and heavy chains proclaimed to be non-cardiac is a reasonable goal for the future. Here, the expression of, e.g., skeletal myosin heavy chains as found in our data, could also be reproduced by searching in available databases for normal and failing hearts [99, 104].

Towards the fetal gene programme?

A switch back to the fetal gene expression has been observed and discussed as an adaptation mechanism to various stressors of cardiac tissue, including mechanical load [171, 172]. It is mainly defined by an increase of glucose utilization for ATP generation, as it is common in the pre-natal heart, instead of predominant reliance on fatty acids, which is common in the post-natal heart [171].

While proteins involved in fatty acid oxidation are clearly reduced in both conditions, the evidence for increased glucose utilization is sparse. The expression of the main glucose transporters GLUT1 and GLUT4 are changed. Contrary to results in mice [173], the rate-limiting enzyme PFKM shows a slight decrease in abundance in our data. The exact mechanism and succession of a switch towards glucose utilization are still under current debate [174]. With our data, we add information on non-failing human hearts. We show that an increase in glucose transporters, and presumably higher glucose availability, is not encompassed by a direct increase in glucose metabolic proteins.

Adding to findings related to the fetal gene program, we find evidence on increased levels of key enzymes of ceramide synthesis, which subsequently would lead to changes in the lipid profile, apoptosis, reduction of oxidative metabolism, and progression of maladaptive remodeling [133]. However, for a proper judgement of the actual metabolic processes and substrate usage, further measurements of, e.g., metabolic fluxes would be needed.

Changes in MYH6/MYH7 ratio are a common marker for a switch to fetal gene expression as a response to aortic stenosis [154]. The fetal isoform of Troponin I - TNNI1 - shows a strong increase in AS and MR but has been claimed not to be reactivated regardless of cardiac condition before [175]. In contrast, Asp et al. (2017) found TNNI1 mRNA to be highly expressed in one HF patient combined with down-regulation of mRNA of mitochondrial proteins [176].

The exact mechanism and succession towards the fetal gene program in failing hearts are not fully understood [172]. In our non-failing hearts, we do see evidence that can be attributed to a switch to the fetal gene programme.

Putative SARS-Cov-2 receptor up-regulated only in pressure-overload hypertrophy

In light of the current COVID-19 pandemic, we set out to show the additional benefit of our data's availability to the research community. Although not related to our main research aims, we noticed that Lindner et al. (2020) reported the myocardium affected by SARS-CoV-2, often [177]. Up-regulation of ACE2 has been reported previously on the transcriptomic level and in Western Blots in left ventricle tissue collected from obstructive hypertrophic cardiomyopathy [178] and in five AS samples of single nucleus RNAseq [160]. Here, we provide further evidence of higher ACE2 expression in pressure-overload hypertrophy and additionally no increase in expression in volume-load hypertrophy. The pathophysiological reason for an increased ACE2 expression in pressure-overload hearts might be a compensatory mechanism that mediates the well-described antihypertrophic and antifibrotic actions of ACE2 in the heart [179]. Further research is needed to investigate whether other pressure load conditions such as arterial hypertension also lead to increased ACE2 expression and whether such conditions can enhance ACE2 expression also in tissues of initial SARS-CoV-2 infection, such as the upper airways.

Limitations

The difference in age between controls and patients is a major limitation of the study. With an increase in age, the risk factors and *in vivo* perturbations of the cardiac tissue accumulate. As such, a healthy control group matching in age would have been the appropriate choice but is hardly available. A derailment of proteostasis proteins and its interaction with the induction of senescence in cardiomyocytes has been reported [136]. However, we are positive that the described effects are also, if not mainly, the result of the underlying pathology as the difference in age between AS and MR is significant but much lower than when compared to controls. Additionally, within conditions, the age is homogeneous between sexes and a significant continuous relationship between protein expression and age within conditions was not found (not shown). As such, a sole impact of age is unlikely as the proteostasis effects are strongest in female AS. However, when interpreting the data, the effects cannot be completely distinguished.

Blood contamination is an additional confounding factor with implications on interpretation. A biopsy taken at the valve replacement surgery will most probably not be free from blood and cannot be washed as efficiently as a sample taken from an explanted heart. Many proteins are of high molecular mass and may sensitively disturb the detection and quantification of other proteins in a sample. As such, we exclude heavily contaminated samples after careful consideration and do not interpret enrichments that pointed towards body fluids. However, as the immune system depends on body fluids, we are impeded in interpreting the effects of inflammation, which have been identified as a hallmark of ventricular hypertrophy and are also under differential regulation in sexes [161, 180].

Our coverage of protein measurement is high when compared to similar studies. Still, good coverage of low abundance transcription factors is hard to achieve even with deep proteomic reference samples. As such, we do not have the potential to uncover full mechanisms of a hypertrophic response as described, e.g., by Haque et al. (2017) and Nakamura et al. (2018) [72, 162]. While we do find expression signatures of downstream effectors of certain transcription factors, we cannot show the changes in the transcription factor's abundance itself. A similar issue is a mechanistic action resulting from post-transcriptional modifications, such as phosphorylation events, which denote actual activity or inhibition of a protein's function.

Additionally, we need to consider that all observations relate to the full tissue, i.e., a mixture of cells from the biopsy. The exact shares of cellular composition are unknown and, due to the multiplicity of nuclei inside a single cardiomyocyte, also hard to infer [103]. As such, all changes in protein abundance may not only be due to a regulation of expression but also due to composition effects, e.g., a relative increase in fibroblasts.

Lastly, one assumption of the methodological framework in DE/DA analysis, or more specific in pre-processing of the data, is an equal amount of total RNA or protein in a sample. A common approach also used by us is to normalize to a uniform distribution and size of library or cumulative intensities. Normalization methods for HTQ measurements and their impact are a topic of ongoing debate [181–183]. The procedure most probably results in a balanced amount of negative and positive fold changes in DE/DA analysis. The strong deviation from this balance in the transcriptomic and proteomic analysis in the comparison against AS fits the biologic explanation of reduced transcriptional and translational capacity. Compositional effects are another reasonable explanation, especially because the effect is consistent across both information levels. As the measurement techniques for both artefacts are based on completely different concepts, they are not prone to show the same bias to the same contamination sources. Additionally, the effect is not seen in MR, which precludes a bias in sampling. However, an impact of the normalization strategy is hard to dismiss and may require further attention.

5 Conclusion and Outlook

In this section, we revisit and answer the two main research questions separately and put our findings into the overall aim of progressing towards Systems Medicine regarding software solutions and by providing a proteomic landscape for heart valve diseases.

The first part of the thesis aimed at rethinking and adapting automated DE/DA calculation software to the wealth of medical data and the exploratory nature of analysis in a Systems Medicine setting.

We used a hybrid approach of Design Thinking, scientific software engineering, and literature research to define 21 requirements for DE/DA software in general and for Systems Medicine in particular. As such, the DE/DA software must cover crucial analysis steps, such as automated pre-processing and quality control, DE/DA design setup, various visualization options, annotation, and interaction possibilities. Automated pre-processing is ensured by providing default configuration for full-fledged pipelines of acknowledged, independent tools and algorithms. Furthermore, the researcher as a user needs functions to assure reproducible results like reports and data downloads. Overall, the software needs to be well documented, example and test data need to be available, and in the best case, no IT skills are needed for usage. A platform-independent straightforward installation procedure and data security are further requirements to the overall system.

To progress towards use in Systems medicine, we focused on the unique challenges in medical research, i.e., handling and using the wealth of clinicome data and other molecular *omes within our SMART IT platform as well as in DE/DA analysis software in general. We established a specific feature space and algorithmic concept, in which the clinicome is the basis for defining complex designs for DE/DA, like multi-group comparisons and continuous variables. Several options for adding covariates, filtering and stratification are also considered. A design formula is created automatically based on the selected options, which serves as input for the widely adopted GLMs. Using GLMs, we exploit the specific extensions tailored to the different kinds of molecular HTQ data. Furthermore, automated pre-processing and DE/DA analysis are independent of each other, and as a result, redundant computations can be avoided.

DEAME and Eatomics are implementations of the general requirements and include instances of the flexible design setup module. Both research applications cover or outnumber the majority of functionalities of existing tools. While it is a standard requirement to prepare one experimental setup prior to using a DE/DA analysis tool, DEAME and Eatomics can handle full phenotypic annotation as it may be available in clinical investigation of biopsy samples or other observational settings. The utility of the working prototype for exploratory

DE analysis was shown as it is applied to the need of clinical scientists and computational biologists. We utilized user testing to evaluate the DE/DA software's specific features with regard to the user's perception and intention to use. We show that novice users, like clinical scientists, are qualified to configure given and own hypotheses into valid designs without any prior computational biology knowledge. The intention to use DE/DA software is high, especially when documentation and help pages are readily available.

The general requirements and our novel experimental design module provide a blueprint for further development efforts in similar software as it is applicable to a large variety of HTQ data sets. Since such data sets' availability rises, the potential user group will expand likewise. Our user testing insights may be useful for further research and practice when developing scientific software. In the meantime, Eatomics is freely available to the research community as an easy-to-use R Shiny application. Therefore, Eatomics may be used in settings of simple designs, as they are common in molecular biology/Systems Biology, but also in medical research settings, such as in Systems Medicine. Furthermore, anyone may reuse the experimental design module or adapt the software to other HTQ data sets.

We believe that our applications may also provide a platform for communication on DE/DA results between the clinical scientist and the computational biologist.

Because of the difficulty to obtain human biopsy samples and the only recently emerging large-scale proteomics measurements, the changes in protein abundance in the myocardium of AS and MR are based on animal models or established in small-scale studies [49]. Furthermore, sex-differences play a role in cardiac disease [34,35] but are seldom considered in molecular studies.

In the second part of the thesis, we aimed at exploring the human myocardial proteome in heart valve diseases. We wanted to obtain a deeper insight into condition- and sex-specific differences in human heart valve disease and to relate the extensive proteomic data to clinical parameters in a well-powered study of human tissue. Our main findings can be summarized as follows:

- AS and MR show many shared mechanisms, of which the most prominent are an increase of ECM and a decrease in metabolism. Both effects are stronger in AS. Additionally, AS shows a larger variability among subjects in general.

- In muscle and cytoskeletal adaptations, we see a strong increase in mechanotransduction in AS and an increase in the cortical cytoskeleton in MR. The adaptations may result from the differences in mechanical stress to the ventricle.

- A strong decrease in proteostasis was revealed to be driven by changes in female AS. A reduced translational capacity may explain less cardiac hypertrophy and better clinical outcomes in females.

- We confirm the expression of several proteins currently under investigation as druggable targets to reduce hypertrophy and add a distinction to their beneficial effects with regard to etiology.

Although an imbalance of more down-regulation in AS was unexpected, the evidence on both transcriptomic and proteomic levels preclude potential methodological flaws, such as the proteomics measurement being sensitive to blood contaminations. Furthermore, age disparities among groups may influence proteostasis. However, the distinctiveness in female AS and only slight changes in all other groups make age effects unlikely. Our depiction of the results highlights mechanisms and proteins that confirm findings from animal models. Validation resembles a first step in translating findings into clinical care and is a major objective in Systems Medicine [26]. Furthermore, we elaborate on changes contradicting current literature, e.g., no increase of ECM in MR, or give hints on blank spots of unknown courses of events, e.g., the changes towards a fetal gene program in non-failing hearts.

Apart from the general landscape, we show the added value of the data set in providing evidence of increased levels of the putative SARS-CoV-2 virus receptor (ACE2) in pressure, but not volume loaded myocardial tissue in the proteome and transcriptome. Furthermore, the data lead to the development of a novel cardiac metabolism model named Cardiokin1. Cardiokin1 can unravel differences in the myocardium's energetic state and help gain deeper insight into metabolic alterations in different types of heart valve diseases [184].

The full set of proteomic quantification will be available to the research community within the publication of Nordmeyer, Kraus, Ziehm, and Kirchner et al. [185]. By providing proteomic data on 58 living subjects, we outgrow previous studies of the human myocardial proteome by a factor of eight [51].

Our description is by far not exhaustive, as it is out of scope to discuss all significant effects. However, the data is continuously used to provide evidence on more hypotheses, e.g., the correlations of sex hormones, protein abundance, and the clinical presentation. As mentioned in Appendix C, specific patients received a hormonal treatment that may have changed their clinical and protein expression profile towards the respective other sex. These specific analysis results are not robust, but further evidence on these mechanisms is currently generated. The findings could ultimately lead to a re-purposing of existing hormonal substances to establish the female phenotype and, as a result, a favourable outcome in the treatment of cardiac hypertrophy. Similarly, other described effects need a further examination of the underlying mechanisms.

Our addition of extensive human evidence extends our general knowledge base on heart valve valve disease. It may help guide new targeted therapy approaches and avoid interventions that are not efficient in a specific condition or sex.

We anticipate that this exciting resource of information about heart valve disease-driven human myocardial proteome changes will further be exploited in future studies to understand differences in cardiac remodeling better and, thus, improve disease- and sex-specific therapy concepts in the future.

This thesis provides insights into a Systems Medicine approach for heart valve diseases. The wealth of HTQ data combined with a rich clinical phenotype as a prominent feature in Systems Medicine was approached in a dedicated IT platform, an extension of DE/DA analysis software, and an exploration of the myocardial proteomic landscape. Although the overall background is rooted in Systems Medicine, our work is valid in other research areas as well and extends the basic knowledge base on heart valve diseases.

The founders of Systems Medicine as a research area struggled to find a definition to grasp its overarching scope. Instead, they provided a road map to follow along for the whole expedition or just for a couple of milestones of a short trip. We believe that in this thesis we have taken a short trip and important steps along this road in advancing a Systems Biology approach to cope with the new challenges unique to Systems Medicine and in an addition of knowledge on the molecular basis of heart valve diseases.

Author Contributions and Credits

A considerable portion of this thesis originates from close collaborative efforts as they are crucial to Systems Medicine approaches. Citations and references are not sufficient to create a clear picture. In this section, I provide a dedicated and detailed attribution of credits to account for the efforts of others and sharpen my own contribution.

All sections not accounted for here are solely written by me - content from others is properly referenced adhering to scientific standards.

Regarding Chapter 2, the implementation of parts of the SMART IT platform was subject of a Master's project led by me and Dr. Matthieu-P. Schapranow. The students enrolled in the project were Lars Rückert, Friedrich Horschig, Benjamin Reißaus and Markus Dücker. Several parts may overlap with details given in Kraus et al. (2017) [53].

Major parts of Chapter 3 are subject of other publications, of which I am the first author, e.g. [54, 55]. Especially the sections on related work and implementation details are drawn from these articles and were expanded to provide more depth within my thesis. The conceptual approach, a major part of the methods and the generalized requirements, as well as the results of user testing are unique to this thesis and are my work. User testing was planned and executed by me, with student assistance from Tamara Slosarek.

My efforts in development and implementation of DEAME was supported by student assistants Tamara Slosarek, Marius Danner, Ajay Kesar and Akshay Bhushan. The major implementation work for Eatomics was done by myself with support of my intern Mariet Mathew Stephens. Whenever needed, we approached Dr. Schapranow for his advice as supervisor.

Chapter 4 is based on the data acquired in the SMART and EurValve studies. A simplified version of the process and stakeholders are depicted in Figure 2.3. The Heart Center refers to the German Heart Center Berlin with Prof. Titus Kühne as principal investigator, Dr. Sarah Nordmeyer and Dr. Marcus Kelm as responsible clinician scientists, who enrolled patients and performed or overlooked clinical data assessment. Patient biopsy samples were collected by Dr. Christoph Knosalla and prepared and distributed for further analysis by Daniel Lehmann under Prof. Vera Regitz-Zagrosek's supervision. RNA sequencing was performed by the Berlin Institute of Health Genomics Core Facility, whereas proteomic measurements were conducted by Dr. Marieluise Kirchner under supervision of Prof. Philipp Mertins at the Berlin Institute of Health Proteomics Core facility. Dr. Matthias Ziehm performed raw data processing, in-depth initial quality control and provided protein abundance levels for further analysis. Overview analyses as shown in Figure 4.5 were performed simultaneously, but independently by me and Dr. Ziehm (in rare cases also Dr. Kirchner) and results were compared and discussed. However, all results shown in this thesis (if not stated otherwise in the respective caption) are based on my own analyses. The analysis strategy was discussed with the whole team. Result interpretation is based on an initial draft of main biological signals prepared by me, e.g., the main categories for enrichment terms. Subsequently, Dr. Nordmeyer and me worked in close collaboration to generate a detailed proteomic landscape. Especially, details of the molecular findings were written by me and were then reworked by Dr. Nordmeyer. Dr. Nordmeyer provided the clinical motivation, discussion and conclusions for the study, which were then reworked by me. Additionally, all co-authors listed on the respective publication [185] hold their legitimate contribution.

Transcriptomic data was in part processed by Dr. Layal abo Khayal to the point of generating gene counts. Choice of tools to do so were jointly agreed upon. Overall quality control and all further analyses, visualization and interpretation of gene expression was done by me.

Dr. Johannes Stegbauer contributed the clinical motivation for our findings on ACE2 receptor expression and wrote large parts of the publication together with Prof. Kühne [56] - I conceived and performed the analyses to support the hypothesis and created the figures. Excerpts from the publication were in part reworked for presentation in this thesis.

In general and if not stated in the respective caption, all figures and tables were conceived and created by me. I generated most diagrams using R's ggplot2 and accessory libraries. I used the BioRender app to create schematic figures - for this purpose a personal academic license was purchased by the Digital Health Center.

List of Acronyms

ACE	angiotensin-converting enzyme
AS	aortic valve stenosis
BH	Benjamini Hochberg
BMI	body mass index
BP	biological process
BPMN	Business Process Modeling Notation
CC	cellular compartment
CON	control
DA	differential abundance
DE	differential expression
DEAME	Differential Expression Analysis Made Easy
DNA	deoxyribonucleic acid
DRUMS	Domain specific ReqUirements Modeling for Scientists
ECG	electro cardiogram
ECM	extracellular matrix
ECV	extracellular volume
ES	enrichment score
FC	fold change
FCS	functional class scoring
FDR	false discovery rate
G-DOC	Georgetown Database of Cancer
GLM	generalized linear models
GO	Gene Ontology

GSEA	gene set enrichment analysis
HF	heart failure
HPLC	High-Performance Liquid Chromatography
HTQ	high-throughput quantification
iBAQ	intensity Based Absolute Quantification
IF	initiation factor
IT	information technology
LFQ	label-free quantification
LV	left ventricle
MF	molecular function
MR	mitral valve regurgitation
MRI	magnetic resonance imaging
mRNA	messenger ribonucleic acid
MS	mass spectrometry
NGS	next-generation sequencing
ORA	over-representation analysis
PA	pathway analysis
PCA	principal component analysis
PO	pressure overload
QC	quality control
RNA	ribonucleic acid
RNAseq	RNA sequencing
SCRM	Scientific Computing Requirements Model
SMART	Systems Medicine Approach for Heart Failure
SR-NCD	San Raffaele Systems Medicine Platform for Non-Communicable Diseases
ssES	single sample enrichment score
ssGSEA	single sample gene set enrichment analysis

TCA	tricarboxylic acid
UI	user interface
UTAUT	The Unified Theory of Acceptance and Use of Technology

Bibliography

[1] Health Directorate European Commission. Workshop report: from systems biology to systems medicine, 2010. `http://ec.europa.eu/research/health/pdf/systems-medicine-workshop-report_en.pdf` accessed on 11.03.2021.

[2] Marc Kirschner. Systems medicine: sketching the landscape. In *Systems Medicine*, pages 3–15. Springer, 2016.

[3] Johan Malmström, Hookeun Lee, and Ruedi Aebersold. Advances in proteomic workflows for systems biology. *Current opinion in biotechnology*, 18(4):378–384, 2007.

[4] Alicia Poplawski et al. Systematically evaluating interfaces for RNA-seq analysis from a life scientist perspective. *Briefings in Bioinformatics*, 17(2):213–223, 2016.

[5] Stefka Tyanova, Tikira Temu, and Juergen Cox. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nature Protocols*, 11(12):2301, 2016.

[6] Ana Conesa et al. A survey of best practices for RNA-seq data analysis. *Genome biology*, 17(1):13, 2016.

[7] Yidong Chen, Edward R Dougherty, and Michael L Bittner. Ratio-based decisions and the quantitative analysis of cdna microarray images. *Journal of Biomedical optics*, 2(4):364–374, 1997.

[8] Matthew E Ritchie et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research*, 43:e47–e47, 2015.

[9] Michael Love, Simon Anders, and Wolfgang Huber. Differential analysis of count data – the DESeq2 package. *Genome Biology*, 15:550, 2014.

[10] Matthias Gietzelt et al. Models and data sources used in systems medicine. *Methods of Information in Medicine*, 55(02):107–113, 2016.

[11] Matthias Gietzelt et al. The Use of Tools, Modelling Methods, Data Types, and Endpoints in Systems Medicine: A Survey on Projects of the German e: Med-Programme. *Studies in health technology and informatics*, 228:670–674, 2016.

[12] Matthias Ganzinger, Matthias Gietzelt, Christian Karmen, Blanca Flores, and Petra Knaup. Implementing systems medicine: A medical informatics perspective. In *MIE*, pages 875–879, 2018.

[13] Nicholas R. Anderson et al. Issues in biomedical research data management and analysis: needs and barriers. *Journal of the American Medical Informatics Association*, 14(4):478–488, 2007.

[14] The CASyM Consortium. The CASyM Roadmap, 2017. `https://www.casym.eu/lw_resource/datapool/_items/item_325/roadmap_1.0.pdf` accessed on 11.03.2021.

[15] EASyM. About EASyM. `https://easym.eu/about-easym/` accessed on 11.03.2021.

[16] International Network and Systems Medicine Association e.V. `https://www.insma.net/` accessed on 11.03.2021.

[17] Department of Human Systems Medicine. `http://hmsysmd.snu.ac.kr/` accessed on 11.03.2021.

[18] Sona Vasudevan. First International Conference in Systems and Network Medicine: Applications of Systems Science and Thinking to Biomedicine, 2019. `https://sites.google.com/georgetown.edu/sysmedconf/home?authuser=0` accessed on 11.03.2021.

[19] Ursula Klingmüller. 8th Conference on Systems Biology of Mammalian Cells, 2021. `https://sbmc2020.bioquant.uni-heidelberg.de/index.php/prgm` accessed on 11.03.2021.

[20] Harald H H W Schmidt and Jan Baumbach. `https://home.liebertpub.com/publications/systems-medicine/643/overview` accessed on 11.03.2021.

[21] Damjana Rozman et al. Workshop 01: CASyM: Modeling Tools for Pharmacokinetics and Systems Medicine, 2014. `http://videolectures.net/mdo2014_stuttgart/` accessed on 11.03.2021.

[22] AMC Graduate School. Systems Medicine (AMC PhD Program), 2020. `https://www.amc.nl/web/leren/graduate-school/phd-1/systems-medicine-amc-phd-program.htm` accessed on 11.03.2021.

[23] Paola Leon-Mimila, Jessica Wang, and Adriana Huertas-Vazquez. Relevance of multi-omics studies in cardiovascular diseases. *Frontiers in cardiovascular medicine*, 6:91, 2019.

[24] Florian Schlotter, Arda Halu, Shinji Goto, Mark C Blaser, Simon C Body, Lang H Lee, Hideyuki Higashi, Daniel M DeLaughter, Joshua D Hutcheson, Payal Vyas, et al. Spatiotemporal multi-omics mapping generates a molecular atlas of the aortic valve and reveals networks driving disease. *Circulation*, 138(4):377–393, 2018.

[25] Linda Pattini, Roberto Sassi, and Sergio Cerutti. Dissecting heart failure through the multiscale approach of systems medicine. *IEEE Transactions on Biomedical Engineering*, 61(5):1593–1603, 2014.

[26] Frank Kramer, Steffen Just, and Tanja Zeller. New perspectives: systems medicine in cardiovascular disease. *BMC Systems Biology*, 12(1):57, 2018.

[27] Kalliopi Trachana, Rhishikesh Bargaje, Gustavo Glusman, Nathan D Price, Sui Huang, and Leroy E Hood. Taking systems medicine to heart. *Circulation research*, 122(9):1276–1289, 2018.

[28] The University of Sheffield. Personalised Decision Support for Heart Valve Disease, 20146. `https://cordis.europa.eu/project/id/689617/de` accessed on 11.03.2021.

[29] Howard E Morgan and Kenneth M Baker. Cardiac hypertrophy. Mechanical, neural, and endocrine dependence. *Circulation*, 83(1):13–25, 1991.

[30] Warren J Manning. Asymptomatic aortic stenosis in the elderly: a clinical review. *Jama*, 310(14):1490–1497, 2013.

[31] Vuyisile T Nkomo et al. Burden of valvular heart diseases: a population-based study. *The Lancet*, 368(9540):1005–1011, 2006.

[32] BJ Bouma et al. To operate or not on elderly patients with aortic stenosis: the decision and its consequences. *Heart*, 82(2):143–148, 1999.

[33] Thomas A Kelly et al. Comparison of outcome of asymptomatic to symptomatic patients older than 20 years of age with valvular aortic stenosis. *The American journal of cardiology*, 61(1):123–130, 1988.

[34] George Petrov et al. Regression of myocardial hypertrophy after aortic valve replacement: faster in women? *Circulation*, 122(11_suppl_1):S23–S28, 2010.

[35] George Petrov et al. Maladaptive remodeling is associated with impaired survival in women but not in men after aortic valve replacement. *JACC: Cardiovascular Imaging*, 7(11):1073–1080, 2014.

[36] Pamela K Woodard et al. ACR practice guideline for the performance and interpretation of cardiac magnetic resonance imaging (MRI). *Journal of the American College of Radiology*, 3(9):665–676, 2006.

[37] Mattia D'Antonio et al. RAP: RNA-Seq Analysis Pipeline, a new cloud-based NGS web application. *BMC genomics*, 16(6):S3, 2015.

[38] Jilong Li et al. From gigabyte to kilobyte: a bioinformatics protocol for mining large RNA-Seq transcriptomics data. *PloS one*, 10(4):e0125000, 2015.

[39] Prerana Wagle, Miloš Nikolić, and Peter Frommolt. QuickNGS elevates Next-Generation Sequencing data analysis to a new level of automation. *BMC genomics*, 16(1):487, 2015.

[40] Ilenia Boria et al. NGS-trex: next generation sequencing transcriptome profile explorer. *BMC bioinformatics*, 14(S7):S10, 2013.

[41] Kyuri Jo, Hawk-Bin Kwon, and Sun Kim. Time-series RNA-seq analysis package (TRAP) and its application to the analysis of rice, Oryza sativa L. ssp. Japonica, upon drought stress. *Methods*, 67(3):364–372, 2014.

[42] Markus Wolfien et al. TRAPLINE: a standardized and automated pipeline for RNA sequencing data analysis, evaluation and annotation. *BMC bioinformatics*, 17(1):21, 2016.

[43] Stefka Tyanova et al. The Perseus computational platform for comprehensive analysis of (prote) omics data. *Nature Methods*, 13(9):731, 2016.

[44] Xiaofei Zhang et al. Proteome-wide identification of ubiquitin interactions using UbIA-MS. *Nature Protocols*, 13:530, 2018.

[45] Anup D Shah et al. LFQ-Analyst: An Easy-To-Use Interactive Web Platform To Analyze and Visualize Label-Free Proteomics Data Preprocessed with MaxQuant. *Journal of Proteome Research*, 19(1):204–211, 2020.

[46] Rob Tibshirani. *samr: SAM: Significance Analysis of Microarrays*, 2018. R package version 3.0.

[47] Bo Liao et al. iMetaLab 1.0: a web platform for metaproteomics data analysis. *Bioinformatics*, 34(22):3954–3956, 2018.

[48] Yuan Tian et al. Champ: updated methylation analysis pipeline for illumina beadchips. *Bioinformatics*, 33(24):3982–3984, 2017.

[49] Caroline J Coats et al. Proteomic analysis of the myocardium in hypertrophic obstructive cardiomyopathy. *Circulation: Genomic and Precision Medicine*, 11(12):e001974, 2018.

[50] Jieyun You et al. Differential cardiac hypertrophy and signaling pathways in pressure versus volume overload. *American Journal of Physiology-Heart and Circulatory Physiology*, 314(3):H552–H562, 2018.

[51] Nora Linscheid et al. Quantitative proteomics of human heart samples collected in vivo reveal the remodeled protein landscape of dilated left atrium without atrial fibrillation. *Molecular & Cellular Proteomics*, 2020.

[52] Tamara Slosarek, Milena Kraus, Matthieu-P Schapranow, and Erwin Boettinger. Qualitative comparison of selected indel detection methods for rna-seq data. In *International Work-Conference on Bioinformatics and Biomedical Engineering*, pages 166–177. Springer, 2019.

[53] Milena Kraus and Matthieu-P. Schapranow. An In-Memory Database Platform for Systems Medicine. In *Proceedings of the 9th Int'l Conf. on Bioinformatics and Computational Biology*. ISCA, 2017.

[54] Milena Kraus et al. DEAME – Differential Expression Analysis Made Easy. In *44th International Conference on Very Large Data Bases, Workshop on Heterogeneous Data Management, Polystores, and Analytics for Healthcare*, pages 162–174. Springer, 2018.

[55] Milena Kraus, Mariet Mathew Stephen, and Matthieu-P Schapranow. Eatomics: Shiny Exploration of Quantitative Proteomics Data. *Journal of Proteome Research*, 20(1):1070–1078, 2020.

[56] Johannes Stegbauer, Milena Kraus, Sarah Nordmeyer, et al. Proteomic analysis reveals upregulation of ACE2, the putative SARS-CoV-2 receptor in pressure-but not volume-overloaded human hearts. *Hypertension*, 76(6):e41–e43, 2020.

[57] Henry Han and Xiaoqian Jiang. Disease biomarker query from RNA-seq data. *Cancer informatics*, (Suppl. 1):81, 2014.

[58] Michael I Love et al. RNA-Seq workflow: gene-level exploratory analysis and differential expression. *F1000Research*, 4, 2015.

[59] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[60] Purvesh Khatri, Marina Sirota, and Atul J Butte. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Computional Biology*, 8(2):e1002375, 2012.

[61] Aravind Subramanian et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102:15545–15550, 2005.

[62] Ravi Mathur et al. Gene set analysis methods: a systematic comparison. *BioData mining*, 11(1):1–19, 2018.

[63] Karsten Krug et al. A curated resource for phosphosite-specific signature analysis. *Molecular & Cellular Proteomics*, 18(3):576–593, 2019.

[64] e:Med Management Office. SMART: Systems Medicine of Heart Failure. http://www.sys-med.de/en/demonstrators/smart/, April 2013.

[65] Subha Madhavan et al. G-DOC: a systems medicine platform for personalized oncology. *Neoplasia*, 13(9):771–783, 2011.

[66] Alfredo Cesario et al. A systems medicine clinical platform for understanding and managing non-communicable diseases. *Current pharmaceutical design*, 20(38):5945–5956, 2014.

[67] Brian D Athey et al. tranSMART: an open source and community-driven informatics and data sharing platform for clinical and translational research. *AMIA Summits on Translational Science Proceedings*, 2013:6, 2013.

[68] Matthieu-P Schapranow et al. The Medical Knowledge Cockpit: Real-time analysis of big medical data enabling precision medicine. In *Proceedings of the 2015 IEEE International conference on Bioinformatics and Biomedicine (BIBM)*, pages 770–775. IEEE, 2015.

[69] Matthieu-P. Schapranow, Franziska Häger, and Hasso Plattner. High-Performance In-Memory Genome Project: A Platform for Integrated Real-Time Genome Data Analysis. In *Proceedings of the 2nd Int'l Conf on Global Health Challenges*, pages 5–10. IARIA, Nov 2013.

[70] Cole Trapnell et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, 7(3):562–578, 2012.

[71] Peter JA Cock et al. The Sanger FASTQ File Format for Sequences with Quality Scores, and the Solexa/Illumina FASTQ Variants. *Nucleic Acids Research*, 38(6):1767–1771, 2009.

[72] Michinari Nakamura and Junichi Sadoshima. Mechanisms of physiological and pathological cardiac hypertrophy. *Nature Reviews Cardiology*, 15(7):387–407, 2018.

[73] Nikolaos G Frangogiannis. The extracellular matrix in ischemic and nonischemic heart failure. *Circulation research*, 125(1):117–146, 2019.

[74] Matthew J Czarny and Jon R Resar. Diagnosis and management of valvular aortic stenosis. *Clinical Medicine Insights: Cardiology*, 8:CMC–S15716, 2014.

[75] Pallavi Gaur and Anoop Chaturvedi. A Survey of Bioinformatics-Based Tools in RNA-Sequencing (RNA-Seq) Data Analysis. In *Translational Bioinformatics and Its Application*, pages 223–248. Springer, 2017.

[76] Vahid Jalili et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. *Nucleic acids research*, 48(W1):W395–W402, 2020.

[77] Yang Li et al. Requirements engineering for scientific computing: A model-based approach. In *2011 IEEE Seventh International Conference on e-Science Workshops*, pages 128–134. IEEE, 2011.

[78] Simon Anders et al. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nature Protocols*, 8(9):1765, 2013.

[79] Eystein Oveland et al. Viewing the proteome: How to visualize proteomics data? *Proteomics*, 15(8):1341–1355, 2015.

[80] Chanchal Kumar and Matthias Mann. Bioinformatics analysis of mass spectrometry-based proteomics data sets. *FEBS letters*, 583(11):1703–1712, 2009.

[81] Hasso Plattner et al. *Design Thinking Research.* Understanding Innovation. Springer, 2012.

[82] Viswanath Venkatesh et al. User acceptance of information technology: Toward a unified view. *MIS quarterly*, pages 425–478, 2003.

[83] Claire Anderson. Presenting and evaluating qualitative research. *American journal of pharmaceutical education*, 74(8), 2010.

[84] Jakob Nielsen and Thomas K Landauer. A mathematical model of the finding of usability problems. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, pages 206–213, 1993.

[85] Simon Andrews. FASTQC, 2017. `https://www.bioinformatics.babraham.ac.uk/projects/fastqc/` accessed on 11.03.2021.

[86] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.

[87] Cole Trapnell, Lior Pachter, and Steven L Salzberg. TopHat: Discovering Splice Junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111, 2009.

[88] Alexander Dobin et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[89] Yang Liao, Gordon K Smyth, and Wei Shi. FeatureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features. *Bioinformatics*, 30(7):923–930, 2014.

[90] Nicolas F Fernandez. *Clustergrammer's Documentation.* `http://clustergrammer.readthedocs.io/index.html` accessed on 10.03.2018.

[91] Milena Kraus et al. Olelo: a web application for intuitive exploration of biomedical literature. *Nucleic Acids Research*, 45(W1):W478–W483, 2017.

[92] National Library of Medicine. `https://pubmed.ncbi.nlm.nih.gov/` accessed on 11.03.2021.

[93] Núria Queralt-Rosinach et al. DisGeNET-RDF: harnessing the innovative power of the Semantic Web to explore the genetic basis of diseases. *Bioinformatics*, 32(14):2236–2238, 2016.

[94] David A Barbie et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*, 462:108, 2009.

[95] Marek Gierlinski, Francesco Gastaldello, and Geoffrey J Barton. Proteus: an R package for downstream analysis of MaxQuant output.

[96] Anna Pursiheimo et al. Optimization of statistical methods impact on quantitative proteomics data. *Journal of Proteome Research*, 14(10):4118–4126, 2015.

[97] Cosmin Lazar. *imputeLCMD: A collection of methods for left-censored missing data imputation*, 2015. R package version 2.0.

[98] M Aleksi Kallio et al. Chipster: user-friendly analysis software for microarray and other high-throughput data. *BMC genomics*, 12(1):507, 2011.

[99] Christina Yingxian Chen et al. Suppression of detyrosinated microtubules improves cardiomyocyte function in human heart failure. *Nature Medicine*, 24(8):1225–1233, 2018.

[100] Judith Segal and Chris Morris. Developing scientific software. *IEEE software*, 25(4):18–20, 2008.

[101] Mark Schena et al. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470, 1995.

[102] Thomas D Ryan et al. Left ventricular eccentric remodeling and matrix loss are mediated by bradykinin and precede cardiomyocyte elongation in rats with volume overload. *Journal of the American College of Cardiology*, 49(7):811–821, 2007.

[103] Sophia Doll et al. Region and cell-type resolved quantitative proteomic map of the human heart. *Nature Communications*, 8(1):1–13, 2017.

[104] Mathias Uhlén et al. Tissue-based map of the human proteome. *Science*, 347(6220), 2015.

[105] Cecilia Lindskog et al. The human cardiac and skeletal muscle proteomes defined by transcriptomics and antibody-based profiling. *BMC genomics*, 16(1):1–15, 2015.

[106] Mengbo Li et al. Core functional nodes and sex-specific pathways in human ischaemic and dilated cardiomyopathy. *Nature Communications*, 11(1):1–12, 2020.

[107] Peng Yu et al. Transcriptome analysis of hypertrophic heart tissues from murine transverse aortic constriction and human aortic stenosis reveals key genes and

transcription factors involved in cardiac remodeling induced by mechanical stress. *Disease markers*, 2019.

[108] Joao Filipe Fernandes et al. Beyond pressure gradients: the effects of intervention on heart power in aortic coarctation. *PloS one*, 12(1):e0168487, 2017.

[109] Adelina Doltra et al. Potential reduction of interstitial myocardial fibrosis with renal denervation. *Journal of the American Heart Association*, 3(6):e001353, 2014.

[110] Lucy E Hudsmith et al. Normal human left and right ventricular and left atrial dimensions using steady state free precession magnetic resonance imaging. *Journal of cardiovascular magnetic resonance*, 7(5):775–782, 2005.

[111] Christopher S Hughes et al. Single-pot, solid-phase-enhanced sample preparation for proteomics experiments. *Nature Protocols*, 14(1):68–85, 2019.

[112] Juri Rappsilber, Yasushi Ishihama, and Matthias Mann. Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Analytical chemistry*, 75(3):663–670, 2003.

[113] Fran Supek et al. REVIGO summarizes and visualizes long lists of gene ontology terms. *PloS one*, 6(7):e21800, 2011.

[114] Daniel Barrell et al. The GOA database in 2009 – an integrated Gene Ontology Annotation resource. *Nucleic acids research*, 37:D396–D403, 2009.

[115] Douglas P Zipes et al. *Braunwald's Heart Disease E-Book: A Textbook of Cardiovascular Medicine*. Elsevier Health Sciences, 2018.

[116] Ariane Melchior-Becker et al. Deficiency of biglycan causes cardiac fibroblasts to differentiate into a myofibroblast phenotype. *Journal of Biological Chemistry*, 286(19):17365–17375, 2011.

[117] Jens Fielitz et al. Activation of the cardiac renin-angiotensin system and increased myocardial collagen expression in human aortic valve disease. *Journal of the American College of Cardiology*, 37(5):1443–1449, 2001.

[118] Masaaki Honda et al. Biochemical remodeling of collagen in the heart of spontaneously hypertensive rats: prominent increase in type V collagen. *Japanese circulation journal*, 57(5):434–441, 1993.

[119] Paula Grippa Sant'Ana et al. Heart remodeling produced by aortic stenosis promotes cardiomyocyte apoptosis mediated by collagen V imbalance. *Pathophysiology*, 25(4):373–379, 2018.

[120] Javier Barallobre-Barreiro et al. Proteomics analysis of cardiac extracellular matrix remodeling in a porcine model of ischemia/reperfusion injury. *Circulation*, 125(6):789–802, 2012.

[121] Erja Mustonen et al. Thrombospondin-4 expression is rapidly upregulated by cardiac overload. *Biochemical and biophysical research communications*, 373(2):186–191, 2008.

[122] Xing Fu et al. Specialized fibroblast differentiated states underlie scar formation in the infarcted mouse heart. *The Journal of clinical investigation*, 128(5):2127–2143, 2018.

[123] Frans A van Nieuwenhoven et al. Cartilage intermediate layer protein 1 (CILP1): a novel mediator of cardiac extracellular matrix remodelling. *Scientific Reports*, 7(1):1–9, 2017.

[124] Shuin Park et al. Cardiac fibrosis is associated with decreased circulating levels of full-length CILP in heart failure. *JACC: Basic to Translational Science*, 5(5):432–443, 2020.

[125] Priyam Banerjee, V Chander, and A Bandyopadhyay. Balancing functions of annexin A6 maintain equilibrium between hypertrophy and apoptosis in cardiomyocytes. *Cell death & disease*, 6(9):e1873–e1873, 2015.

[126] Victoria Polyakova et al. Matrix metalloproteinases and their tissue inhibitors in pressure-overloaded human myocardium during heart failure progression. *Journal of the American College of Cardiology*, 44(8):1609–1618, 2004.

[127] Francis G Spinale et al. Time-dependent changes in matrix metalloproteinase activity and expression during the progression of congestive heart failure: relation to ventricular and myocyte function. *Circulation research*, 82(4):482–495, 1998.

[128] Xuelian Li et al. Overexpression of SerpinE2/protease nexin-1 contribute to pathological cardiac fibrosis via increasing collagen deposition. *Scientific Reports*, 6:37635, 2016.

[129] Valiente-Alandi et al. Inhibiting fibronectin attenuates fibrosis and improves cardiac function in a model of heart failure. *Circulation*, 138(12):1236–1252, 2018.

[130] Saiful Anam Mir et al. Inhibition of signal transducer and activator of transcription 3 (STAT3) attenuates interleukin-6 (IL-6)-induced collagen synthesis and resultant hypertrophy in rat heart. *Journal of Biological Chemistry*, 287(4):2666–2677, 2012.

[131] Dina Radenkovic et al. T1 mapping in cardiac MRI. *Heart failure reviews*, 22(4):415–430, 2017.

[132] William C Stanley and Margaret P Chandler. Energy metabolism in the normal and failing heart: potential for therapeutic interventions. *Heart failure reviews*, 7(2):115–130, 2002.

[133] Ruiping Ji et al. Increased de novo ceramide synthesis and accumulation in failing myocardium. *JCI insight*, 2(9), 2017.

[134] Shouji Matsushima and Junichi Sadoshima. The role of sirtuins in cardiac disease. *American Journal of Physiology-Heart and Circulatory Physiology*, 309(9):H1375–H1389, 2015.

[135] Phung N Thai et al. Cardiac-specific conditional knockout of the 18-kDa mitochondrial translocator protein protects from pressure overload induced heart failure. *Scientific Reports*, 8(1):1–17, 2018.

[136] Robert H Henning and Bianca JJM Brundel. Proteostasis in cardiac health and disease. *Nature Reviews Cardiology*, 14(11):637, 2017.

[137] Hae Jin Kee et al. Activation of histone deacetylase 2 by inducible heat shock protein 70 in cardiac hypertrophy. *Circulation research*, 103(11):1259–1269, 2008.

[138] Asangi RK Kumarapeli et al. αB-crystallin suppresses pressure overload cardiac hypertrophy. *Circulation research*, 103(12):1473–1482, 2008.

[139] Emily J Mercer et al. Hspb7 is a cardioprotective chaperone facilitating sarcomeric proteostasis. *Developmental biology*, 435(1):41–55, 2018.

[140] Yun-Kyung Kim et al. Deletion of the inducible 70-kDa heat shock protein genes in mice impairs cardiac contractile function and calcium handling associated with hypertrophy. *Circulation*, 113(22):2589–2597, 2006.

[141] Yan Zhang et al. HSP75 protects against cardiac hypertrophy and fibrosis. *Journal of cellular biochemistry*, 112(7):1787–1794, 2011.

[142] Hai-Lian Bi et al. The deubiquitinase UCHL1 regulates cardiac hypertrophy by stabilizing epidermal growth factor receptor. *Science Advances*, 6(16):eaax4826, 2020.

[143] B Li et al. Knockdown of eIF3a ameliorates cardiac fibrosis by inhibiting the TGF-β1/Smad3 signaling pathway. *Cellular and Molecular Biology*, 62(7):97–101, 2016.

[144] Pamela S Douglas et al. Gender differences in left ventricle geometry and function in patients undergoing balloon dilatation of the aortic valve for isolated aortic stenosis. NHLBI Balloon Valvuloplasty Registry. *Heart*, 73(6):548–554, 1995.

[145] Anita YM Chan et al. Activation of AMP-activated protein kinase inhibits protein synthesis associated with hypertrophy in the cardiac myocyte. *Journal of Biological Chemistry*, 279(31):32771–32779, 2004.

[146] Erik A Blackwood et al. Designing Novel Therapies to Mend Broken Hearts: ATF6 and Cardiac Proteostasis. *Cells*, 9(3):602, 2020.

[147] Stefan Hein et al. Deposition of nonsarcomeric alpha-actinin in cardiomyocytes from patients with dilated cardiomyopathy or chronic pressure overload. *Experimental & Clinical Cardiology*, 14(3):e68, 2009.

[148] Pragati Pandey et al. Cardiomyocytes sense matrix rigidity through a combination of muscle and non-muscle myosin contractions. *Developmental cell*, 44(3):326–336, 2018.

[149] Sakima A Smith et al. Dysfunction of the β2-spectrin-based pathway in human heart failure. *American Journal of Physiology-Heart and Circulatory Physiology*, 310(11):H1583–H1591, 2016.

[150] Masashi Arai et al. Sarcoplasmic reticulum genes are selectively down-regulated in cardiomyopathy produced by doxorubicin in rabbits. *Journal of molecular and cellular cardiology*, 30(2):243–254, 1998.

[151] Guido Tarone and Mara Brancaccio. The muscle-specific chaperone protein melusin is a potent cardioprotective agent. *Basic research in cardiology*, 110(2):10, 2015.

[152] Kathryn A Rosowski et al. Vinculin and the mechanical response of adherent fibroblasts to matrix deformation. *Scientific Reports*, 8(1):1–10, 2018.

[153] Thomas Sejersen and Urban Lendahl. Transient expression of the intermediate filament nestin during skeletal muscle development. *Journal of Cell Science*, 106(4):1291–1300, 1993.

[154] Peter J Reiser et al. Human cardiac myosin heavy chain isoforms in fetal and failing adult atria and ventricles. *American Journal of Physiology-Heart and Circulatory Physiology*, 280(4):H1814–H1820, 2001.

[155] Brian D Lowes et al. Changes in gene expression in the intact human heart. downregulation of alpha-myosin heavy chain in hypertrophied, failing ventricular myocardium. *The Journal of clinical investigation*, 100(9):2315–2324, 1997.

[156] Setsuya Miyata et al. Myosin heavy chain isoform expression in the failing and nonfailing human heart. *Circulation research*, 86(4):386–390, 2000.

[157] Ayse Cetinkaya et al. Radixin relocalization and nonmuscle α-actinin expression are features of remodeling cardiomyocytes in adult patients with dilated cardiomyopathy. *Disease Markers*, 2020, 2020.

[158] Jason Y Chan et al. Identification of cardiac-specific myosin light chain kinase. *Circulation research*, 102(5):571–580, 2008.

[159] Mahmoud Gheblawi et al. Angiotensin-converting enzyme 2: SARS-CoV-2 receptor and regulator of the renin-angiotensin system: celebrating the 20th anniversary of the discovery of ACE2. *Circulation research*, 126(10):1456–1474, 2020.

[160] Luka Nicin et al. Cell type-specific expression of the putative SARS-CoV-2 receptor ACE2 in human hearts. *European heart journal*, 41(19):1804–1806, 2020.

[161] Georgios Kararigas et al. Sex-dependent regulation of fibrosis and inflammation in human left ventricular remodelling under pressure overload. *European journal of heart failure*, 16(11):1160–1167, 2014.

[162] Zaffar K Haque and Da-Zhi Wang. How cardiomyocytes sense pathophysiological stresses for cardiac remodeling. *Cellular and Molecular Life Sciences*, 74(6):983–1000, 2017.

[163] Dong Fan et al. Cardiac fibroblasts, fibrosis and extracellular matrix remodeling in heart disease. *Fibrogenesis & tissue repair*, 5(1):15, 2012.

[164] Sowndramalingam Sankaralingam and Gary D Lopaschuk. Cardiac energy metabolic alterations in pressure overload–induced left and right heart failure (2013 Grover Conference Series). *Pulmonary circulation*, 5(1):15–28, 2015.

[165] Daniela Fliegner et al. Female sex and estrogen receptor-β attenuate cardiac remodeling and apoptosis in pressure overload. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 298(6):R1597–R1606, 2010.

[166] Henning Witt et al. Sex-specific pathways in early cardiac response to pressure overload in mice. *Journal of Molecular Medicine*, 86(9):1013, 2008.

[167] RD Hannan et al. Cardiac hypertrophy: a matter of translation. *Clinical and Experimental Pharmacology and Physiology*, 30(8):517–527, 2003.

[168] Mengping Chen et al. Therapeutic effect of targeting branched-chain amino acid catabolic flux in pressure-overload induced heart failure. *Journal of the American Heart Association*, 8(11):e011625, 2019.

[169] Christophe Depre et al. Activation of the cardiac proteasome during pressure overload promotes ventricular hypertrophy. *Circulation*, 114(17):1821–1828, 2006.

[170] Feng-Chun Tsai et al. Ubiquitin pathway is associated with worsening left ventricle function after mitral valve repair: A global gene expression study. *International journal of molecular sciences*, 21(14):5073, 2020.

[171] Mitra Rajabi et al. Return to the fetal gene program protects the stressed heart: a strong hypothesis. *Heart failure reviews*, 12(3-4):331–343, 2007.

[172] Heinrich Taegtmeyer, Shiraj Sen, and Deborah Vela. Return to the fetal gene program: a suggested metabolic link to gene expression in the heart. *Annals of the New York Academy of Sciences*, 1188:191, 2010.

[173] Ronglih Liao et al. Cardiac-specific overexpression of glut1 prevents the development of heart failure attributable to pressure overload in mice. *Circulation*, 106(16):2125–2131, 2002.

[174] Teresa Pasqua et al. Cardiometabolism as an interlocking puzzle between the healthy and diseased heart: New frontiers in therapeutic applications. *Journal of Clinical Medicine*, 10(4):721, 2021.

[175] Fikru B Bedada et al. Acquisition of a quantitative, stoichiometrically conserved ratiometric marker of maturation status in stem cell-derived cardiac myocytes. *Stem cell reports*, 3(4):594–605, 2014.

[176] Michaela Asp et al. Spatial detection of fetal marker genes expressed at low level in adult human heart tissue. *Scientific Reports*, 7(1):1–10, 2017.

[177] Diana Lindner et al. Association of cardiac infection with SARS-CoV-2 in confirmed COVID-19 autopsy cases. *JAMA cardiology*, 5(11):1281–1285, 2020.

[178] J Martijn Bos et al. Marked up-regulation of ACE2 in hearts of patients with obstructive hypertrophic cardiomyopathy: Implications for SARS-CoV-2-mediated COVID-19. In *Mayo Clinic Proceedings*. Elsevier, 2020.

[179] Koichi Yamamoto et al. Deletion of angiotensin-converting enzyme 2 accelerates pressure overload-induced cardiac dysfunction by increasing local angiotensin II. *Hypertension*, 47(4):718–726, 2006.

[180] Fanmui Yang et al. Coronary artery remodeling in a model of left ventricular pressure overload is influenced by platelets and inflammatory cells. *Plos one*, 7(8):e40196–e40196, 2012.

[181] Marie-Agnès Dillies et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics*, 14(6):671–683, 2013.

[182] Shintaro Katayama et al. Guide for library design and bias correction for large-scale transcriptome studies using highly multiplexed RNAseq methods. *BMC bioinformatics*, 20(1):1–9, 2019.

[183] Ciaran Evans, Johanna Hardin, and Daniel M Stoebel. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Briefings in Bioinformatics*, 19(5):776–792, 2018.

[184] Nikolaus Berndt et al. Cardiokin1: Computational assessment of myocardial metabolic capability in healthy controls and patients with valve diseases. *submitted at Circulation*, 2021.

[185] Sarah Nordmeyer, Milena Kraus, Matthias Ziehm, Marieluise Kirchner, et al. Myocardial proteome profiling reveals disease- and sex-specific alterations in patients with aortic valve stenosis and mitral valve regurgitation. *Submitted.*

[186] Winston Chang et al. *shiny: Web Application Framework for R*, 2018. R package version 1.2.0.

[187] Hadley Wickham. *modelr: Modelling Functions that Work with the Pipe*, 2019. R package version 0.1.5.

[188] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York, 2016.

[189] JJ Allaire et al. *markdown: 'Markdown' Rendering for R*, 2018. R package version 0.9.

[190] Nicolai J Wewer Albrechtsen et al. Plasma proteome profiling reveals dynamics of inflammatory and lipid homeostasis markers after Roux-en-Y gastric bypass surgery. *Cell systems*, 7(6):601–612, 2018.

[191] Qianxing Mo et al. A fully bayesian latent variable model for integrative clustering analysis of multi-type omics data. *Biostatistics*, 19(1):71–86, 2018.

[192] Bo Wang et al. Similarity network fusion for aggregating data types on a genomic scale. *Nature methods*, 11(3):333, 2014.

[193] Damien McParland and Isobel Claire Gormley. Model based clustering for mixed data: clustMD. *Advances in Data Analysis and Classification*, 10(2):155–169, 2016.

A Supporting information on DE/DA analysis software

A.1 Clinically motivated questions towards assessed molecular data (German)

1. Gibt es Unterschiede zwischen Controls Männern und Frauen?
2. Gibt es Unterschiede zwischen allen controls und allen Patienten mit Aortenstenose?
3. Gibt es Unterschiede zwischen controls Männer und Patienten mit Aortenstenosen Männer?
4. Gibt es Unterschiede zwischen Controls Frauen und Patienten mit Aortenstenosen (AS und SMART) Frauen?
5. Gibt es Unterschiede zwischen Frauen und Männern der Patienten mit Aortenstenosen (AS und SMART)?
6. Gibt es Unterschiede zwischen Patienten mit und ohne Medikamente (beta blocker vs. No cardiac medication)?
7. Nur Smarts: Gibt es Unterschiede zwischen Patienten, die besser oder schlechter belastbar waren vor OP? (NYHA 1-2 vs 3-4 zum Beispiel)
8. Gibt es Unterschiede zwischen Patienten mit und ohne Hypertrophie? (Myocardial mass in g/BSA) - ja/nein oder kontinuierlich? ggf. aufgetrennt in Männer und Frauen
9. Gibt es Unterschiede zwischen Patienten mit guter und schlechter Herzfunktion? (EF in %; - entweder ja/nein, oder kontinuierlich? ggf. aufgetrennt in Männer und Frauen
10. Gibt es Unterschiede zwischen Patienten mit viel und wenig Fibrose? (Extra cellular volume) - entweder ja/nein oder kontinuierlich?ggf. aufgetrennt in Männer und Frauen
11. Gibt es Unterschiede zwischen Patienten mit vergrößertem und nicht vergrößertem linken Herzen? (EDV/BSA) - entweder ja/nein, oder kontinuierlich? ggf. aufgetrennt in Männer und Frauen
12. Gibt es Unterschiede zwischen Patienten mit erhöhtem und nicht erhöhtem NT-proBNP? - entweder ja/nein, oder kontinuierlich? ggf. aufgetrennt in Männer und Frauen

13. Gibt es Unterschiede zwischen Patienten mit niedrigem und hohem Testosteron? (Dehydrotestosterone)- entweder ja/nein, oder kontinuierlich? ggf. aufgetrennt in Männer und Frauen

14. Nur Smarts: Gibt es Unterschiede zwischen Patienten mit niedriger und hoher internal heart power? - entweder ja/nein, oder kontinuierlich? ggf. aufgetrennt in Männer und Frauen

15. Nur Smarts: Gibt es Unterschiede zwischen Patienten mit niedriger und hoher myokardialen efficiency? - entweder ja/nein, oder kontinuierlich? ggf. aufgetrennt in Männer und Frauen

A.2 Model of the preprocessing requirements space

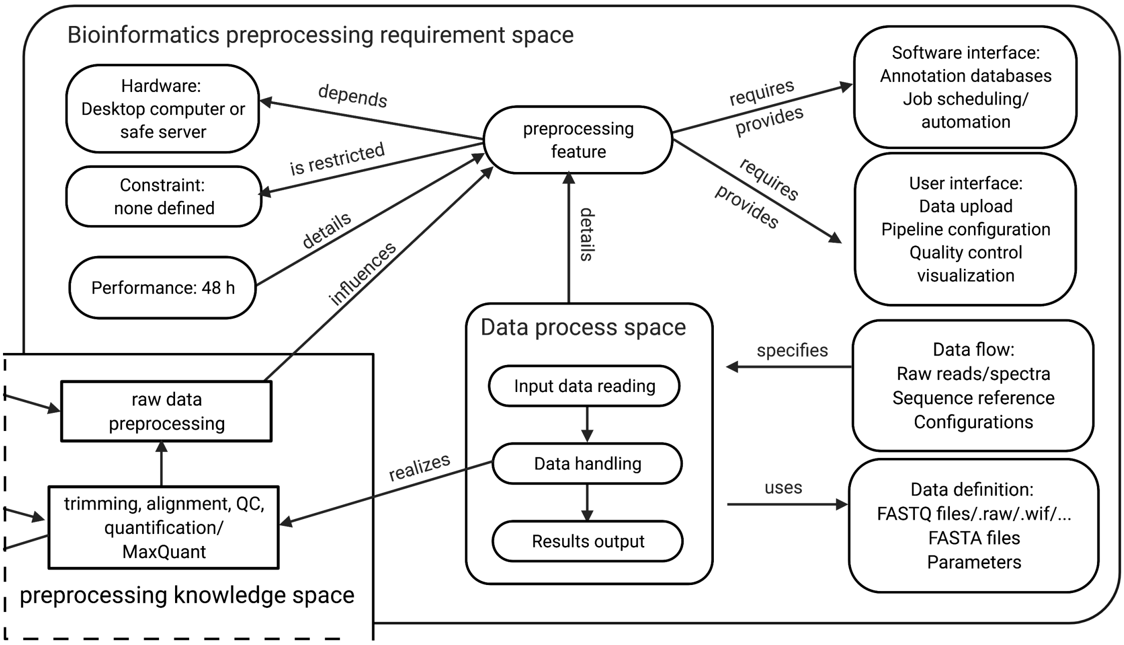

Bioinformatics preprocessing requirement space
Hardware: Desktop computer or safe server
depends
Constraint: none defined
is restricted
Performance: 48 h
details
influences
preprocessing feature
requires
provides
Software interface: Annotation databases Job scheduling/ automation
requires
provides
User interface: Data upload Pipeline configuration Quality control visualization
details
Data process space
Input data reading
Data handling
Results output
specifies
Data flow: Raw reads/spectra Sequence reference Configurations
uses
Data definition: FASTQ files/.raw/.wif/... FASTA files Parameters
realizes
raw data preprocessing
trimming, alignment, QC, quantification/ MaxQuant
preprocessing knowledge space

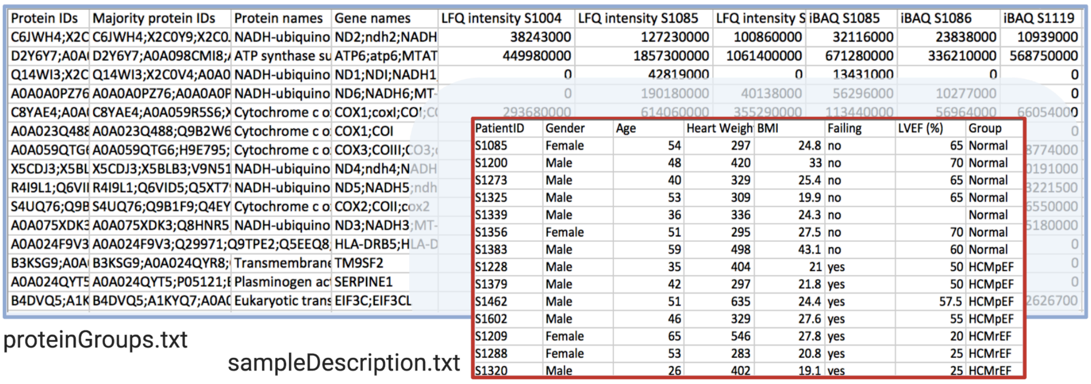
Figure A.1: SCRM diagram of the preprocessing pipeline requirements space.

A.3 Eatomics: Installation and dependencies

The most important packages Eatomics depends on for proper function are: R Shiny [186] and several add-on packages are used as a general framework to transfer R code for interactive analysis and HTML display. imputeLCMD [97] and a custom implementation of the Perseus' sampling from Gaussian distribution are used for missing value imputation. The latter function was provided generously by Matthias Ziehm (Orcid ID: 0000-0001-7074-4054). Limma's linear models with empirical Bayes estimation for accurate results in small sample settings are used for differential protein abundance analysis [8]. Model formulas, i.e., the experimental design is generated with the help of the modelr package [187].

ggplot2 [188], plotly and gridExtra packages are used for plotting and rmarkdown [189] for report generation The tidyverse dogma is used whenever reasonable for tidy data usage. The ssGSEA tab panel wraps the available code (`https://github.com/broadinstitute/ssGSEA2.0`) into a user-friendly shiny interface. Eatomics uses four MSigDB gene sets namely C1, C2, C5 and H to calculate the enrichment score. The installation procedure for the local instance requires R and R studio, which are available for a multitude of different operating systems. Furthermore, the R packages devtools, shiny and janitor need to be installed by the user to then be able to run the application in their R studio and web browser by using the runUrl() function pointing to the R subdirectory in the github repository at `https://github.com/Millchmaedchen/Eatomics/archive/master.zip`. In the repository, we also provide a list of all other package dependencies. For institutional use, Shiny applications are well suited for server installations and scale out to serve more users. The application opens in the users standard browser similar to a web application.

A.4 Examples of input files needed for Eatomics.



Protein IDs	Majority protein IDs	Protein names	Gene names	LFQ intensity S1004	LFQ intensity S1085	LFQ intensity S	iBAQ S1085	iBAQ S1086	iBAQ S1119
C6JWH4;X2C	C6JWH4;X2C0Y9;X2C0	NADH-ubiquino	ND2;ndh2;NADH	38243000	127230000	100860000	32116000	23838000	10939000
D2Y6Y7;A0A	D2Y6Y7;A0A098CMI8;	ATP synthase su	ATP6;atp6;MTAT	449980000	1857300000	1061400000	671280000	336210000	568750000
Q14WI3;X2C	Q14WI3;X2C0V4;A0A0	NADH-ubiquino	ND1;NDI;NADH1	0	42819000	0	13431000	0	0
A0A0A0PZ76	A0A0A0PZ76;A0A0A0P	NADH-ubiquino	ND6;NADH6;MT-	0	190180000	40138000	56296000	10277000	0
C8YAE4;A0A	C8YAE4;A0A059R5S6;	Cytochrome c o	COX1;coxI;COI;C	293680000	614060000	355290000	113440000	56964000	66054000
A0A023Q488	A0A023Q488;Q9B2W6	Cytochrome c o	COX1;COI						0
A0A059QTG6	A0A059QTG6;H9E795;	Cytochrome c o	COX3;COIII;CO3;						8774000
X5CDJ3;X5BL	X5CDJ3;X5BLB3;V9N51	NADH-ubiquino	ND4;ndh4;NADH						0191000
R4I9L1;Q6VII	R4I9L1;Q6VID5;Q5XT7	NADH-ubiquino	ND5;NADH5;ndh						3221500
S4UQ76;Q9B	S4UQ76;Q9B1F9;Q4EY	Cytochrome c o	COX2;COII;cox2						6550000
A0A075XDK3	A0A075XDK3;Q8HNR5	NADH-ubiquino	ND3;NADH3;MT-						5180000
A0A024F9V3	A0A024F9V3;Q29971;Q9TPE2;Q5EEQ8;		HLA-DRB5;HLA-D						0
B3KSG9;A0A	B3KSG9;A0A024QYR8;	Transmembrane	TM9SF2						0
A0A024QYT5	A0A024QYT5;P05121;	Plasminogen ac	SERPINE1						0
B4DVQ5;A1K	B4DVQ5;A1KYQ7;A0A0	Eukaryotic trans	EIF3C;EIF3CL						2626700

proteinGroups.txt

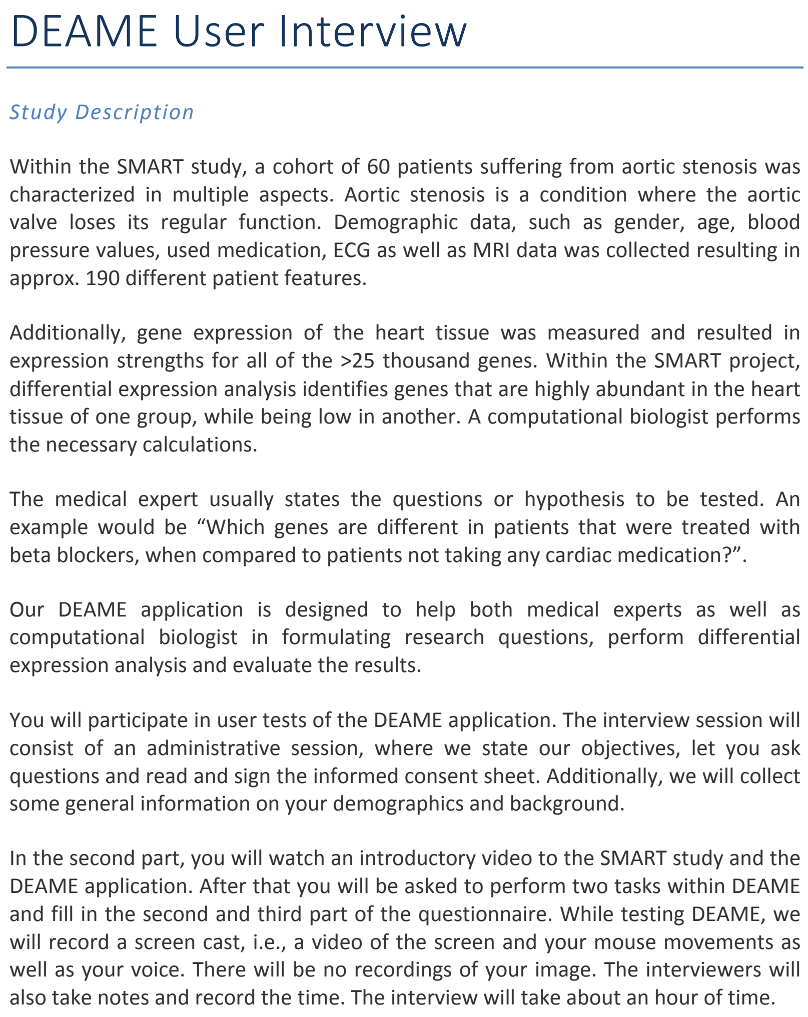
PatientID	Gender	Age	Heart Weigh	BMI	Failing	LVEF (%)	Group
S1085	Female	54	297	24.8	no	65	Normal
S1200	Male	48	420	33	no	70	Normal
S1273	Male	40	329	25.4	no	65	Normal
S1325	Male	53	309	19.9	no	65	Normal
S1339	Male	36	336	24.3	no		Normal
S1356	Female	51	295	27.5	no	70	Normal
S1383	Male	59	498	43.1	no	60	Normal
S1228	Male	35	404	21	yes	50	HCMpEF
S1379	Male	42	297	21.8	yes	50	HCMpEF
S1462	Male	51	635	24.4	yes	57.5	HCMpEF
S1602	Male	46	329	27.6	yes	55	HCMpEF
S1209	Female	65	546	27.8	yes	20	HCMrEF
S1288	Female	53	283	20.8	yes	25	HCMrEF
S1320	Male	26	402	19.1	yes	25	HCMrEF

sampleDescription.txt

Figure A.2: Examples of input files needed for Eatomics are an evidence file as produced by the MaxQuant algorithm (left) and a sample description file which may contain as many parameters as available (right).

A.5 User Interview Testing Artefacts

Short study description handed to all participants of the DEAME user interviews.



DEAME User Interview

Study Description

Within the SMART study, a cohort of 60 patients suffering from aortic stenosis was characterized in multiple aspects. Aortic stenosis is a condition where the aortic valve loses its regular function. Demographic data, such as gender, age, blood pressure values, used medication, ECG as well as MRI data was collected resulting in approx. 190 different patient features.

Additionally, gene expression of the heart tissue was measured and resulted in expression strengths for all of the >25 thousand genes. Within the SMART project, differential expression analysis identifies genes that are highly abundant in the heart tissue of one group, while being low in another. A computational biologist performs the necessary calculations.

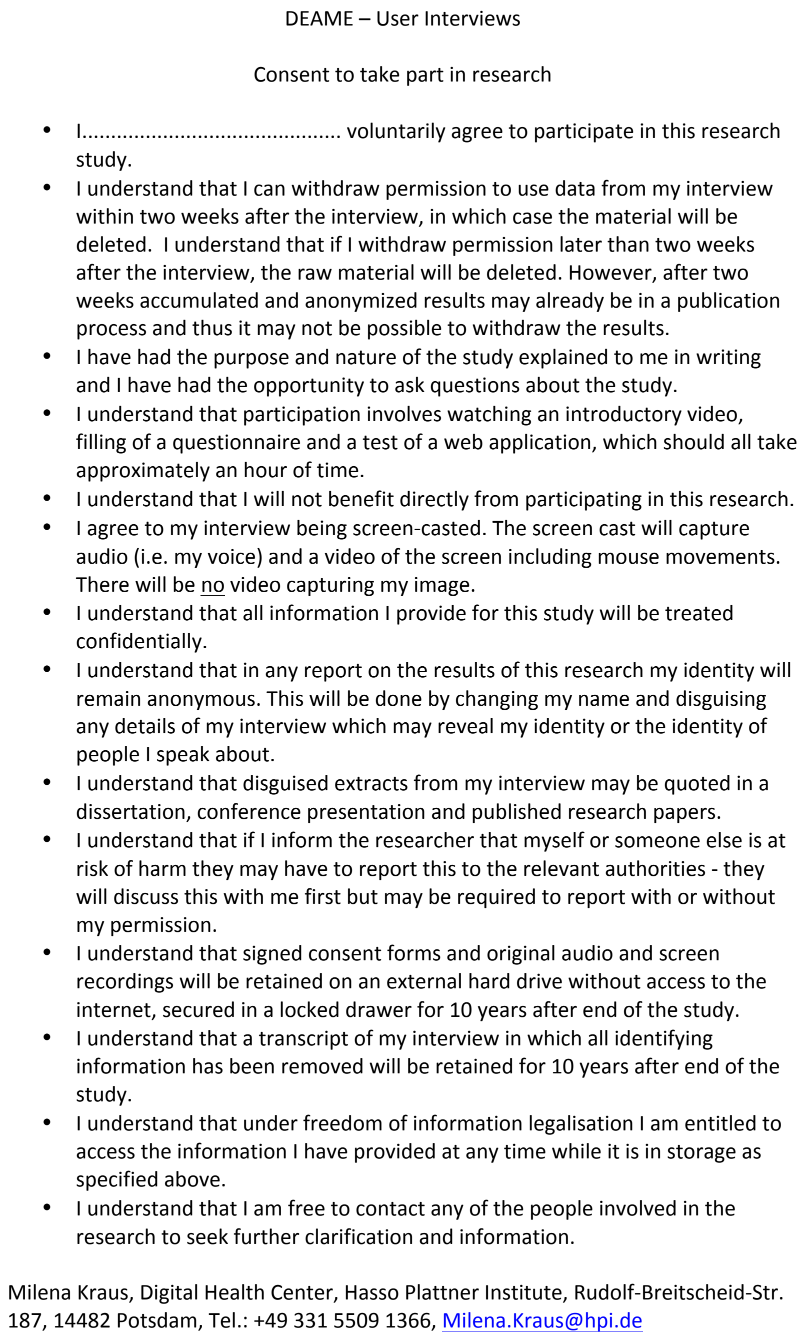
The medical expert usually states the questions or hypothesis to be tested. An example would be “Which genes are different in patients that were treated with beta blockers, when compared to patients not taking any cardiac medication?”.

Our DEAME application is designed to help both medical experts as well as computational biologist in formulating research questions, perform differential expression analysis and evaluate the results.

You will participate in user tests of the DEAME application. The interview session will consist of an administrative session, where we state our objectives, let you ask questions and read and sign the informed consent sheet. Additionally, we will collect some general information on your demographics and background.

In the second part, you will watch an introductory video to the SMART study and the DEAME application. After that you will be asked to perform two tasks within DEAME and fill in the second and third part of the questionnaire. While testing DEAME, we will record a screen cast, i.e., a video of the screen and your mouse movements as well as your voice. There will be no recordings of your image. The interviewers will also take notes and record the time. The interview will take about an hour of time.

Consent form supplied and signed by all participants of the DEAME user interviews.



DEAME – User Interviews

Consent to take part in research

- I... voluntarily agree to participate in this research study.
- I understand that I can withdraw permission to use data from my interview within two weeks after the interview, in which case the material will be deleted. I understand that if I withdraw permission later than two weeks after the interview, the raw material will be deleted. However, after two weeks accumulated and anonymized results may already be in a publication process and thus it may not be possible to withdraw the results.
- I have had the purpose and nature of the study explained to me in writing and I have had the opportunity to ask questions about the study.
- I understand that participation involves watching an introductory video, filling of a questionnaire and a test of a web application, which should all take approximately an hour of time.
- I understand that I will not benefit directly from participating in this research.
- I agree to my interview being screen-casted. The screen cast will capture audio (i.e. my voice) and a video of the screen including mouse movements. There will be no video capturing my image.
- I understand that all information I provide for this study will be treated confidentially.
- I understand that in any report on the results of this research my identity will remain anonymous. This will be done by changing my name and disguising any details of my interview which may reveal my identity or the identity of people I speak about.
- I understand that disguised extracts from my interview may be quoted in a dissertation, conference presentation and published research papers.
- I understand that if I inform the researcher that myself or someone else is at risk of harm they may have to report this to the relevant authorities - they will discuss this with me first but may be required to report with or without my permission.
- I understand that signed consent forms and original audio and screen recordings will be retained on an external hard drive without access to the internet, secured in a locked drawer for 10 years after end of the study.
- I understand that a transcript of my interview in which all identifying information has been removed will be retained for 10 years after end of the study.
- I understand that under freedom of information legalisation I am entitled to access the information I have provided at any time while it is in storage as specified above.
- I understand that I am free to contact any of the people involved in the research to seek further clarification and information.

Milena Kraus, Digital Health Center, Hasso Plattner Institute, Rudolf-Breitscheid-Str. 187, 14482 Potsdam, Tel.: +49 331 5509 1366, Milena.Kraus@hpi.de

User questionnaire filled in by all participants of the DEAME user interviews.

DEAME Questionnaire

Name: ____

Gender: □ female □ male □ other

Age: ______ □ prefer not to say

Part I – General Information

1. What is your profession?

   □ Clinician /Medical expert
   □ Computational biologist
   □ Other: _________________

2. Have you ever performed differential expression analysis (the whole computational pipeline starting from raw sequencing reads to a list of differentially expressed genes) yourself?

   □ Yes
   □ No

3. Have you ever interpreted the results of differential expression analysis?

   □ Yes
   □ No

4. Have you been an author/a co-author of a research article that included results of differential expression analysis?

   □ Yes
   □ No

5. Did your university or college studies cover the topic of gene expression?

   □ Yes
   □ No

Part II – Complete a given task in the DEAME app

This part is about using the DEAME app while performing a pre-defined task. Keep in mind that we are testing the app (not you), so in case you need any further explanation or help please first consult the information given when clicking the small question mark in the upper right corner. Additionally, we would appreciate it if you think aloud and tell us why you are performing which action.

For this task, we hypothesize that male patients that took no cardiac medication have differentially expressed genes when compared to the control study group. Execute an experiment that compares the study group of no cardiac medication against the control group while restricting the subjects to be only male.

1. How many patient clusters were identified?

____ □ I don't know

2. Do the clusters correspond to the given comparison of no cardiac medication vs. control?

____ □ I don't know

3. Write down the first sentence or bullet point of a definition of a gene of your choice.

Gene name: ____

Definition: ____

____ □ I don't know

4. Name the most significant genes according to the lowest p-adjusted value. Please reduce the values to include only 3 digits after the dot.

   Downregulated gene: ______________________________ □ I don't know

   Fold change: ______________________________ □ I don't know

   P-adjusted: ______________________________ □ I don't know

   Upregulated gene: ______________________________ □ I don't know

   Fold change: ______________________________ □ I don't know

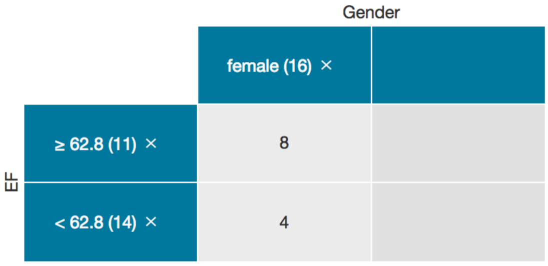
   P-adjusted: ______________________________ □ I don't know

Additionally, perform a gene set enrichment analysis based on the KEGG (2016) pathways and export the heatmap as an image.

5. Which KEGG pathway shows the lowest pval and therefore is most significant?

____ ☐ I don't know

6. Please write down which groups are compared by the following experimental designs.

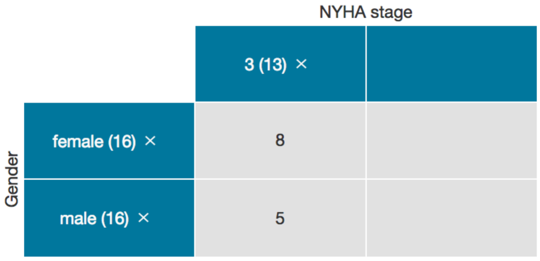


		Gender	
		female (16) ×	
EF	≥ 62.8 (11) ×	8	
	< 62.8 (14) ×	4	

________________________________ compared to ________________________________

filtered to include only ____

____ ☐ I don't know



		NYHA stage	
		3 (13) ×	
Gender	female (16) ×	8	
	male (16) ×	5	

________________________________ compared to ________________________________

filtered to include only ____

____ ☐ I don't know

Part III – Explore the DEAME app

Please give us a short note as soon as you start this part of the questionnaire.

In this part you create and test your own hypothesis. First, think of a hypothesis you would like to test based on the clinical parameters given within the app and write it down below.

Hypothesis:
_________________________________ compared to _________________________________

filtered to include only (optional) ____

____

Create and execute an experiment to test your hypothesis in the DEAME app. What do you conclude after an exploration of the results, are genes differentially expressed? Do you have any concerns regarding the significance of results?

Conclusion: ____

____

____ □ I don't know

Please answer the questions below.

	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
1. It was no effort to translate my hypothesis into a valid design matrix.	□	□	□	□	□
2. The app provides an interactive representation of the results.	□	□	□	□	□
3. The visualization of the results is suitable.	□	□	□	□	□
4. The app is easy to use when compared to other scientific software.	□	□	□	□	□
5. It took a long time until I managed to create a valid experiment design.	□	□	□	□	□
6. I would appreciate additional interaction possibilities with diagrams.	□	□	□	□	□
7. The app lacks diagrams for a usable visualization.	□	□	□	□	□

	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
8. It was difficult to use the app when compared to the software I normally use for my research.	□	□	□	□	□
9. It is clear and understandable how the app works.	□	□	□	□	□
10. The calculation time is acceptable.	□	□	□	□	□
11. As a clinician, I can imagine to use the app in my work.	□	□	□	□	□
12. A more exhaustive tutorial would help me to better understand the app.	□	□	□	□	□
13. The calculation time renders the app to be unusable.	□	□	□	□	□
14. I would prefer a computational biologist to test my hypotheses over using DEAME.	□	□	□	□	□

After completing the questionnaire until this point you may explore DEAME further. The open exploration is optional and we encourage you to do so. Apart from that you may as well complete the questions below and finish the interview.

15. Would you keep using the DEAME app after you tried it today?

□ Yes, because ____

____

□ Maybe, if ____

____

□ No, because ____

____

16. Do you think the app is going to help you with your research?

□ Yes, because ____

____

□ Maybe, if ____

____

☐ No, because ____

____

17. Based on your previous expertise in gene expression analysis, have you encountered any content-related errors in the result representation? If yes, please describe, which.

____

____

____

____

____

18. How likely is it that you recommend the DEAME app to a colleague, on a scale from 1 (not likely) to 5 (very likely)?

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

19. Have you already tried out similar tools or applications?

☐ Yes
☐ No

20. If yes, please write down which tools:

____

____

____

21. What rate would you give our app, on a scale from 1 (worst rate) to 5 (best rate)?

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

Thank you for your participation!

Testing notes filled in by the two moderators during the DEAME user interviews.

DEAME User Testing – Testing Notes

Tester: ____

Participant: ____

Part I – Welcome, General Information and Video Tutorial

1. Prior to interview, send out consent forms an study descriptions so time is not wasted within interview
2. Prepare setting including to open the app, the video, plug in charger and mouse, open quicktime player, check for app availability
3. Welcome, introduce ourselves as testers, clarify preferred language of communication
4. Give an overview on the setup of the experiment.
5. Let them read the short intro and sign consent form – while reading the form the user should formulate any questions.
6. Let user fill in the general information part of the questionnaire
7. Watch the video
8. After the video, ask if there are any remaining questions.

Part II – Given Task

Time needed to run working experiment (start building the design to clicking the run button): ______

Time needed by app for result computation: ____

The tester…	Completed	Completed after asking	Not completed
…translated the given hypothesis into the right matrix.	□	□	□
…ran the experiment.	□	□	□
…zoomed in to see the list of genes.	□	□	□
…hovered over genes to see the definitions.	□	□	□
…performed gene set enrichment.	□	□	□
…took a snapshot of the heatmap and exported it as an image.	□	□	□
…switched to volcano plot.	□	□	□
… hovered over volcano plot icons.	□	□	□
…identified p-value and fold change.	□	□	□

Notes (at least the order of the groups in the design):

____

____

____

____

____

____

____

Part III – App Exploration

Time needed to run working experiment (start building the design to clicking the run button): ______

Time needed by app for result computation: ____

The tester…	Completed	Not completed
…created a valid design.	□	□

Created Design:

Notes:

____

____

____

B Quality control reports and and additional analyses

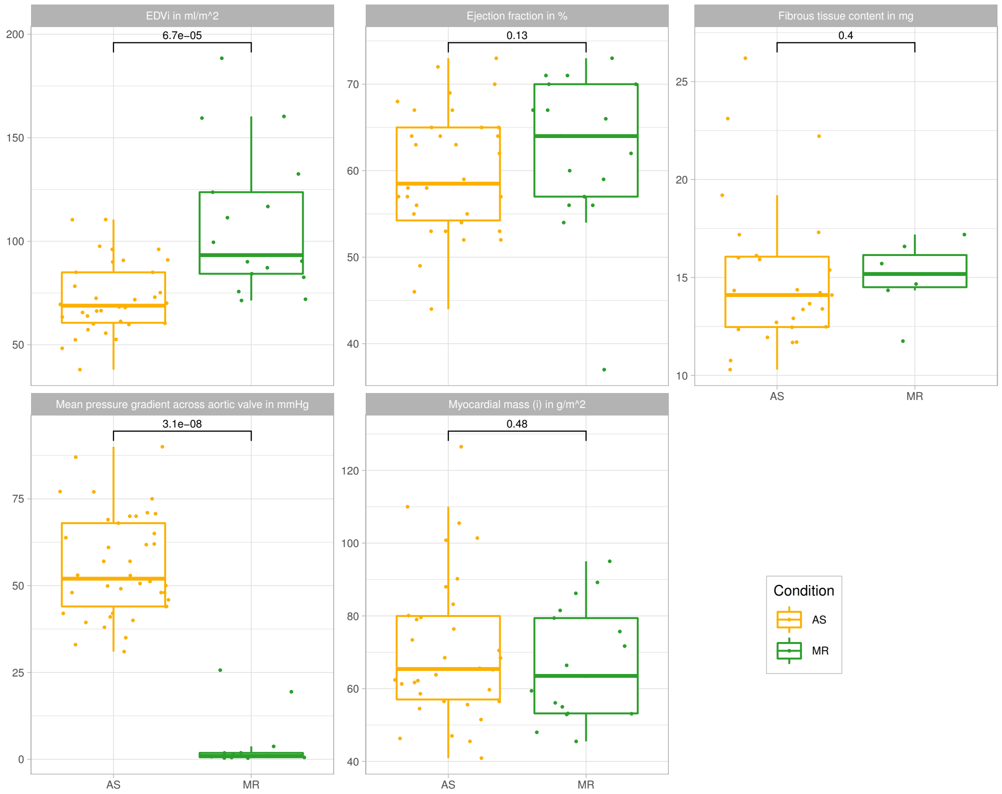


Figure B.1: Overview on clinical parameters from magnetic resonance imaging stratified to conditions. End-diastolic volume, ejection fraction, fibrous tissue content and myocardial mass of the left ventricle are described and the mean pressure gradient across the aortic valve. AS = Aortic valve stenosis, EDVi - end-diastolic volume indexed to body surface area, MR = mitral valve regurgitation. Statistical comparison was performed via two-sided, two-sample Wilcoxon-rank test. Missing values are not shown and not included in the statistical test.

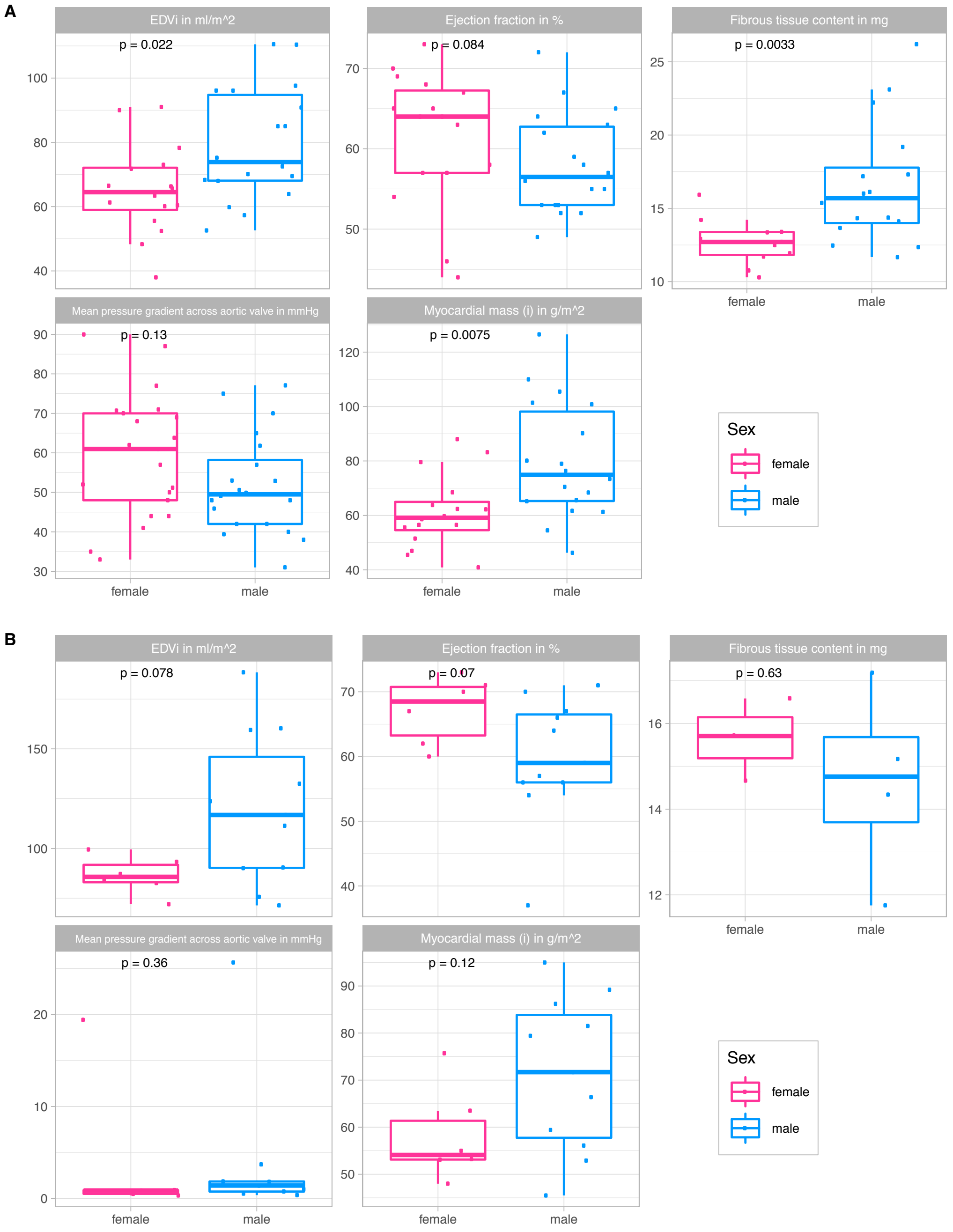


Figure B.2: Overview on clinical parameters from magnetic resonance imaging stratified to sex in AS (A) and MR (B). End-diastolic volume, ejection fraction, fibrous tissue content and myocardial mass of the left ventricle are described and the mean pressure gradient across the aortic valve. AS = Aortic valve stenosis, EDVi - end-diastolic volume indexed to body surface area, MR = mitral valve regurgitation. Group-wise statistical comparison was performed via two-sided, two-sample Wilcoxon-rank test. Missing values are not shown and not included in the statistical test.

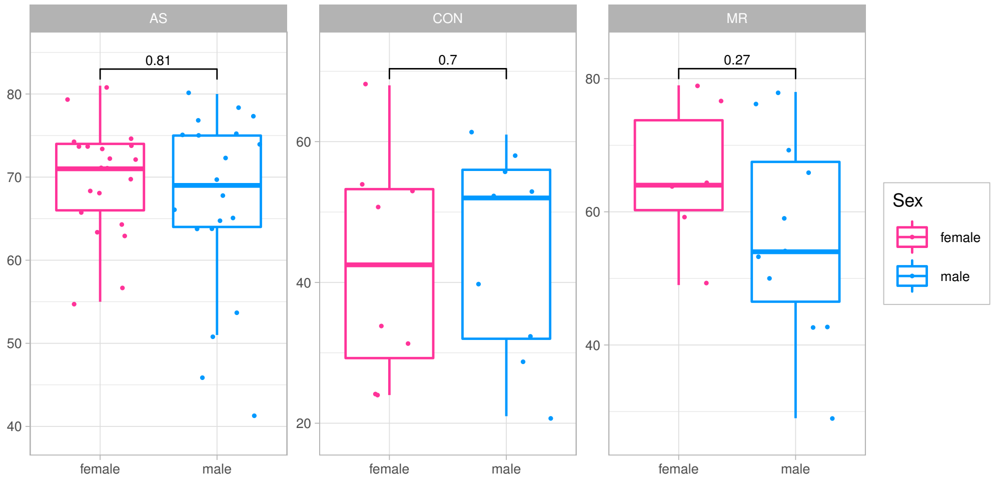


Figure B.3: Overview on the age at surgery stratified to sex in AS (left), CON (middle) and MR (right). AS = Aortic valve stenosis, MR = mitral valve regurgitation, CON = controls. Group-wise statistical comparison was performed via two-sided, two-sample Wilcoxon-rank test. Missing values are not shown and not included in the statistical test.

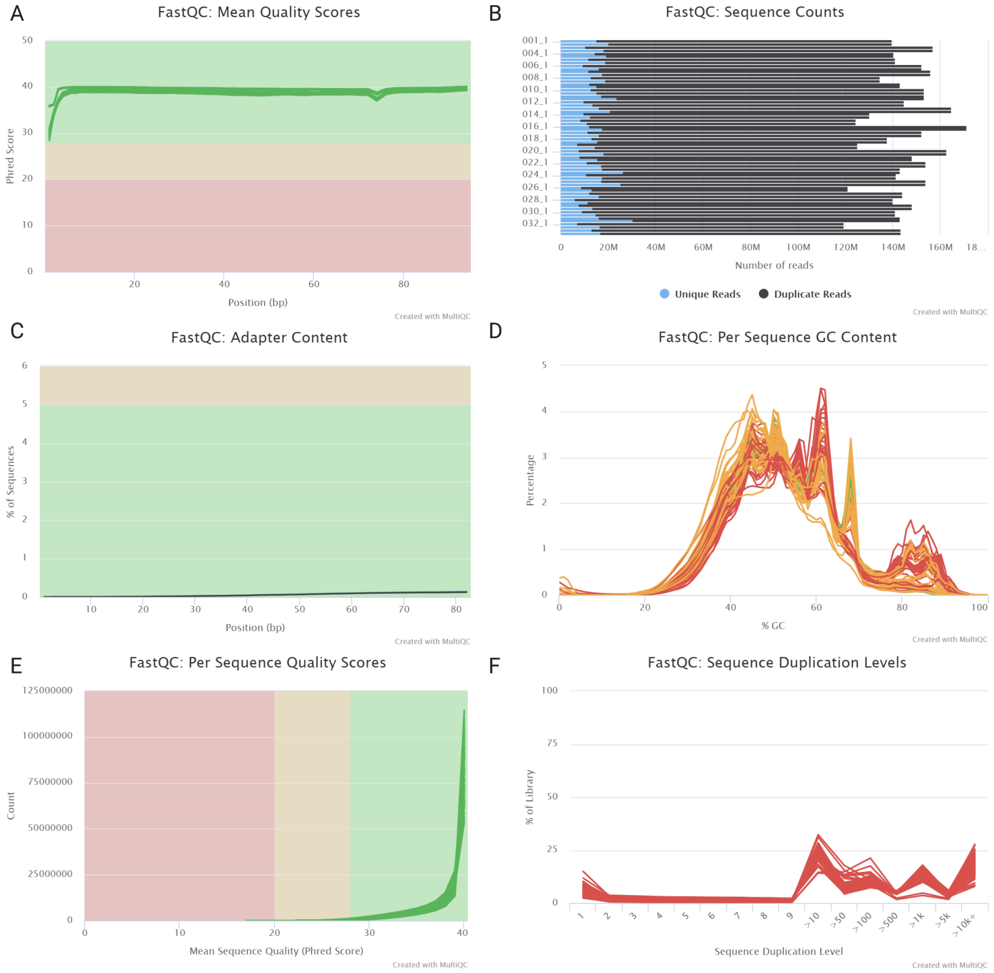
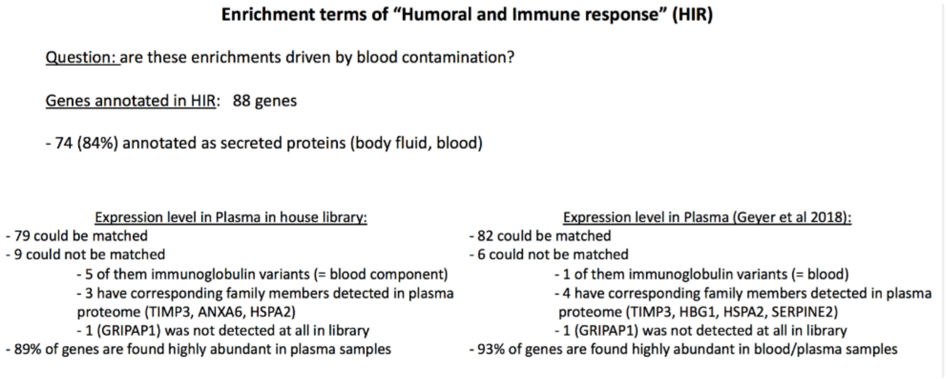


Figure B.4: Summary reports created by MultiQC for all 33 paired end RNA sequencing FASTQ files after trimming. A) Mean quality histogram showing the mean quality value across each base position in the read. B) Sequence counts for each sample showing unique and duplicate read estimation. C) Adapter content denoting the cumulative percentage count of the proportion of the library which has seen each of the adapter sequences at each position. D) Per sequence GC content shows the average GC content of reads. No organism information was given as a reference to fastqc. E) Per sequence quality scores shows the number of reads with average quality scores. F) The relative level of duplication found for every sequence.



Enrichment terms of "Humoral and Immune response" (HIR)

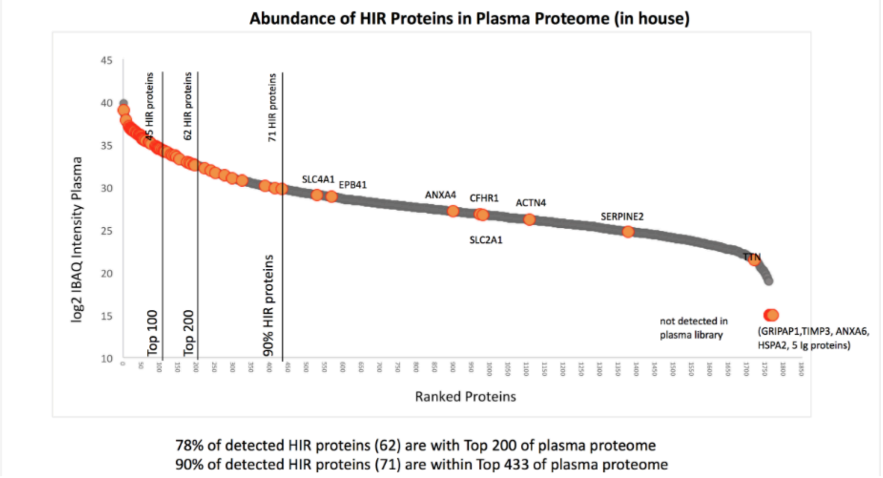
Question: are these enrichments driven by blood contamination?

Genes annotated in HIR: 88 genes

- 74 (84%) annotated as secreted proteins (body fluid, blood)

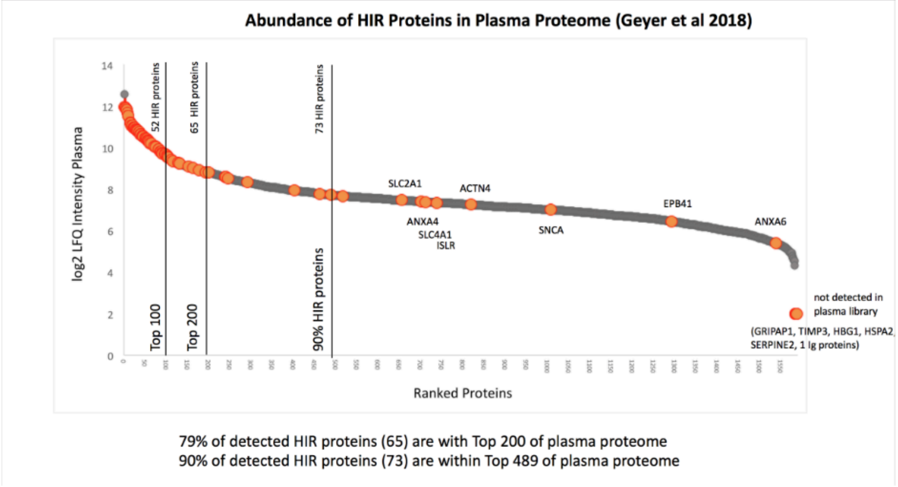
Expression level in Plasma in house library:
- 79 could be matched
- 9 could not be matched
  - 5 of them immunoglobulin variants (= blood component)
  - 3 have corresponding family members detected in plasma proteome (TIMP3, ANXA6, HSPA2)
  - 1 (GRIPAP1) was not detected at all in library
- 89% of genes are found highly abundant in plasma samples

Expression level in Plasma (Geyer et al 2018):
- 82 could be matched
- 6 could not be matched
  - 1 of them immunoglobulin variants (= blood)
  - 4 have corresponding family members detected in plasma proteome (TIMP3, HBG1, HSPA2, SERPINE2)
  - 1 (GRIPAP1) was not detected at all in library
- 93% of genes are found highly abundant in blood/plasma samples



Abundance of HIR Proteins in Plasma Proteome (in house)

78% of detected HIR proteins (62) are with Top 200 of plasma proteome
90% of detected HIR proteins (71) are within Top 433 of plasma proteome



Abundance of HIR Proteins in Plasma Proteome (Geyer et al 2018)

79% of detected HIR proteins (65) are with Top 200 of plasma proteome
90% of detected HIR proteins (73) are within Top 489 of plasma proteome

Figure B.5: Critical examination of proteins belonging to GO terms which were assigned to the other terms category (abbreviated as HIR for this figure). Both disease groups showed significant GO term enrichment of these other terms compared to control. The enrichments are based on 88 proteins with higher abundance in the disease groups. Of these, 84% are typical body fluid components (secreted, erythrocyte, hemoglobin, immunoglobins, complement factor etc.). In addition, matching these genes to an in-house and a published plasma proteome data [190] revealed, that 98% are detected in plasma and the majority (80%) is found to be highly abundant in plasma (top 200). Considering the different biopsy collection procedures for the samples groups, blood contamination becomes the most probable source of signal and impedes any interpretation with regard to physiological differences in immune response between the disease and control groups. The analysis was performed and diagrams were created by Dr. Marieluise Kirchner from Berlin Institute of Health, Berlin, DE (orcid ID 0000-0002-7049-534X)

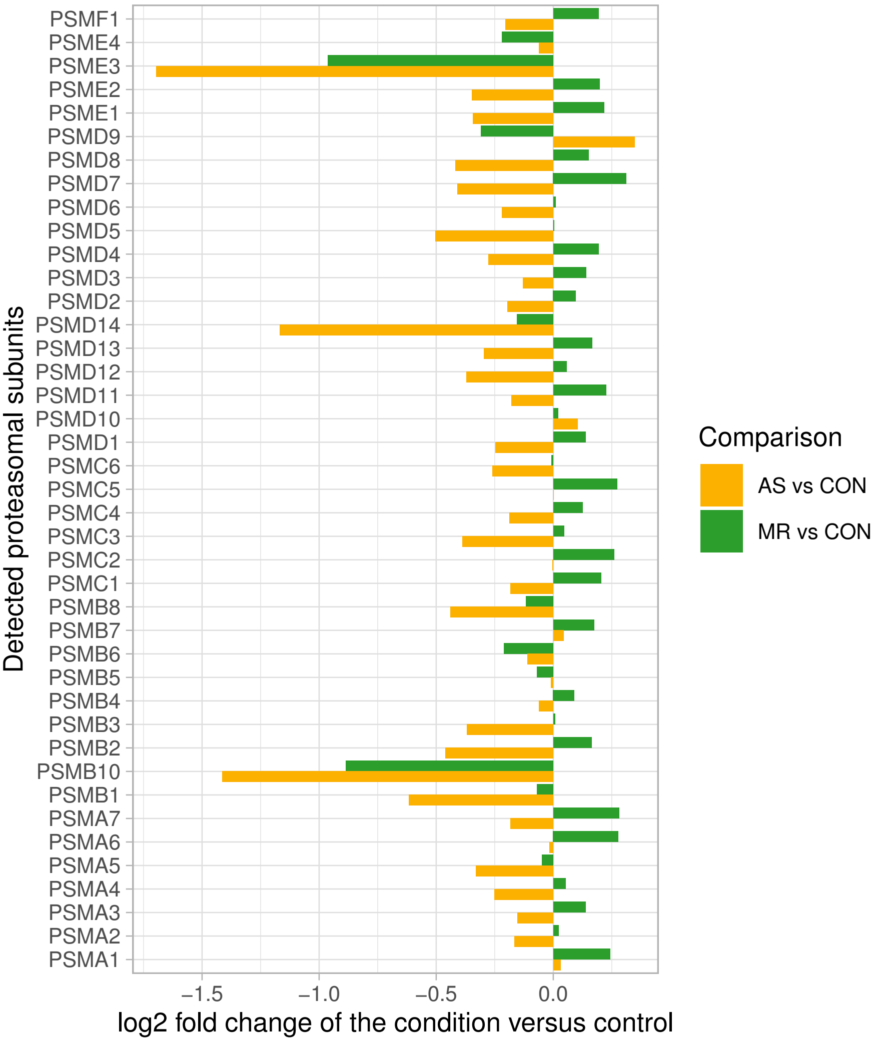


Figure B.6: Log2 fold changes of detected proteasomal subunits in condition (AS = aortic stenosis, MR = mitral regurgitation) versus control comparisons.

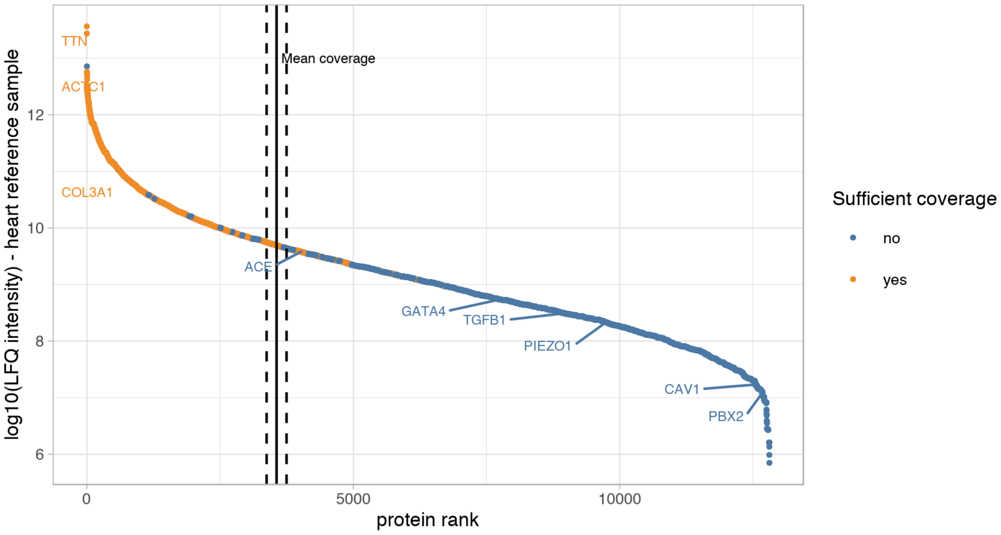


Figure B.7: LFQ intensity as measured in the reference sample (a mixture of all available samples) is plotted for every detected protein based on the overall rank of the protein. Yellow denotes proteins, which are detected in the samples sufficiently to base analysis on them. Blue dots and labels denote proteins that were not detected or not sufficiently covered in the sample measurements. Mean coverage and standard deviation (dashed) are denoted by black lines. The mean sample coverage approximates the detection limit of our measurements. ACE intensity in the reference sample ranks very close to the limit (ACE rank = 3582) which is an indicator of a decreased probability of detecting and quantifying ACE in a sufficient amount of samples robustly.

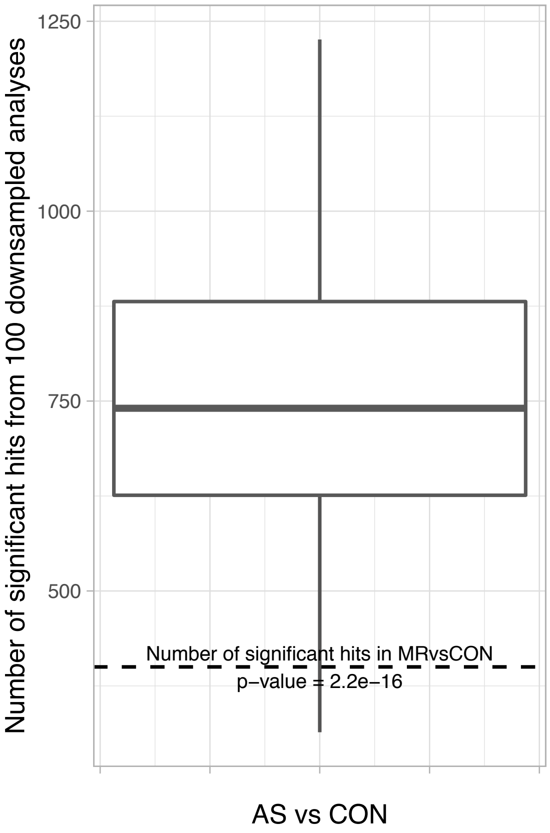


Figure B.8: Down-sampled analyses to exclude bias of having more samples in AS. In total 100 times, we randomly select 17 AS samples from the pool pf 41 AS samples and compare them against the 17 CON samples in an otherwise identical fashion as described in section section 4.3.4. The average number of proteins significant in the comparison is 746 ± 179, which is significantly more than in MR (n = 17) vs CON (n = 17), for which we found 400 significant hits. Additionally, approximately two thirds (64%) of all significant hits show a decrease in fold change.

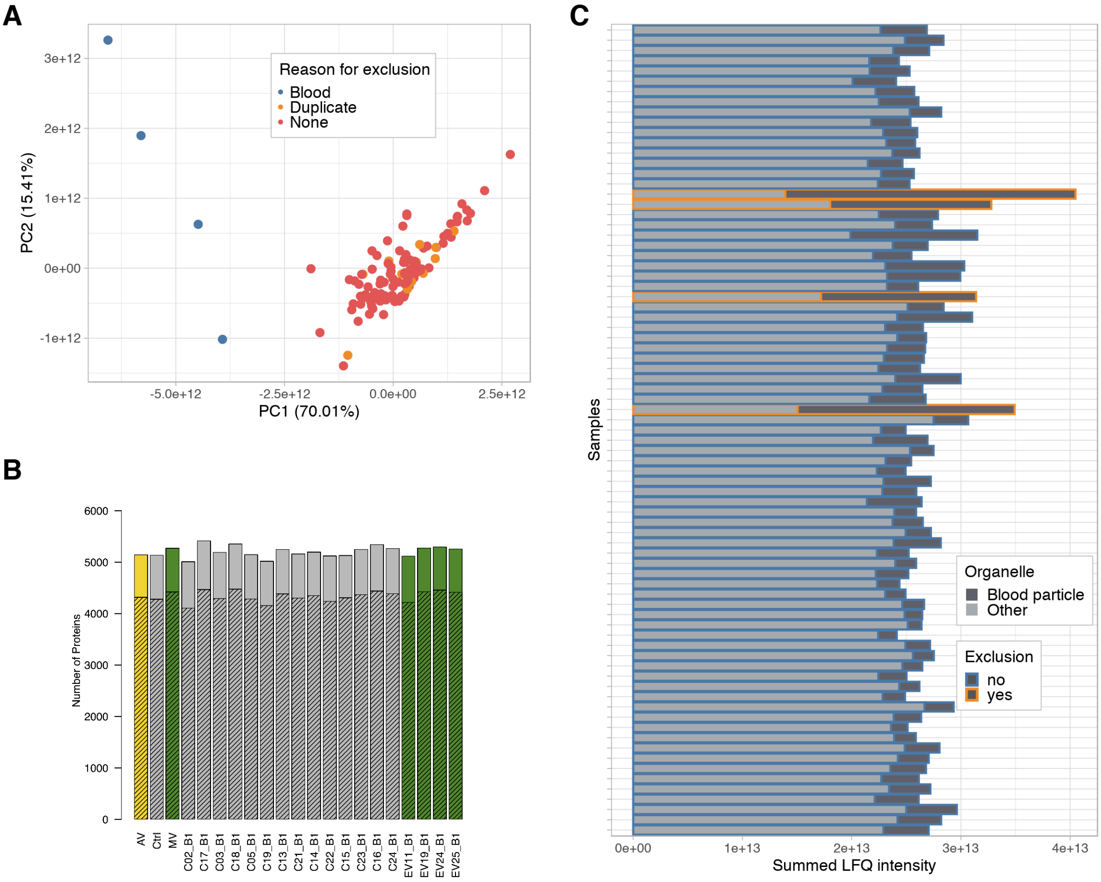


Figure B.9: Summary of sample exclusion because of duplicate measurements and strong blood contamination. A) Principal component analysis before exclusion of samples. Colors denote the reason for exclusion. B) Coverage among duplicate biopsy measurements and in relation to overall coverage of AS in yellow, MR in green and CON in grey. Samples with proteome measurement ID: C02, C03, C19, C21, C22, C15, C24, EV11, EV25 were excluded bcause of lower coverage when compared to their counter parts. Image created by Matthias Ziehm (Orcid ID: 0000-0001-7074-4054). C) Summed LFQ intensity of samples before exclusion of blood contaminated samples. LFQ intensity assigned to blood particles as defined by Doll et al. [103] are shown in darker grey and samples to be excluded from the main analysis are rendered in orange.

C Exploratory comparison of clustering strategies on a Systems Medicine data set

Unsupervised subgroup detection (USD) is a common approach to stratify tumours into subtypes. Algorithms utilize the full spectrum of available multi-omics information and thus aim to provide a holistic picture of the disease. We hypothesize that USD can be readily applied in other complex diseases as well. Genetic and lifestyle factors may influence the onset and progression of disease, while gene and protein expression in the heart muscle, as well as imaging data contain information on the status of the disease. We utilized the SMART study data to explore

1. the applicability of three established USD algorithms in a complex disease data set and
2. if the found subgroups are associated to clinical outcome and relevant biological terms.

As the data set is very small, we do not expect to find robust subtypes. However, small sample sizes are very common as multi-omics analysis in the medical setting are still mainly used as demonstrators with small sample sizes. However, means to make sense of the data are desperately needed and it is of particular interest to evaluate existing solutions to USD detection on these data sets. Additionally, we hope that USD algorithms might still be helpful to find interesting patterns and features among patients that lead to new hypotheses with regard to mechanisms in the disease.

Methods

Available methods are selected when eligible according to the following characteristics:

- Need to perform clustering/unsupervised subgroup detection
- Need to handle mixed-type data
- Generally suitable/tested for biomedical data

Out of a variety of publicly available cluster algorithms, we selected iClusterBayes (iCB) [191], SNF [192], and clustMD [193] for a comparison and qualitative assessment of advantages and drawbacks in the application of these algorithms on the SMART data set of 21 aortic stenosis patients.

The genome, proteome and clinicome (binary, numeric, categorical) are first analyzed by every algorithm separately (single-omics experiments) and additionally in the respective tool's combined mode (multi-omics experiments) for a range of $k = 2$ to $k = 6$ possible clusters. As a result, there are 6 datasets x 5 cluster possibilities = 30 solutions per algorithm in the single-omics mode and 5 solutions per algorithm in the multi-omics setting. All solutions, i.e., the assignment of samples to clusters, are tested for association to binary clinical outcome (CO) estimators. Genes and proteins meeting a tool-specific significance threshold are included in the enrichment analysis against the Gene Ontology database. Technical evaluation metrics are the completion of process, the run time in seconds and the mean suggested number of clusters. Biological evaluation metrics are the number of significant GO terms found through enrichment analysis and the number pf significant associations to clinical outcome estimators. Additionally, we take a close look on the actual clusters and important features of the best multi-omics solution. An overview of the procedure is given in Figure C.1.

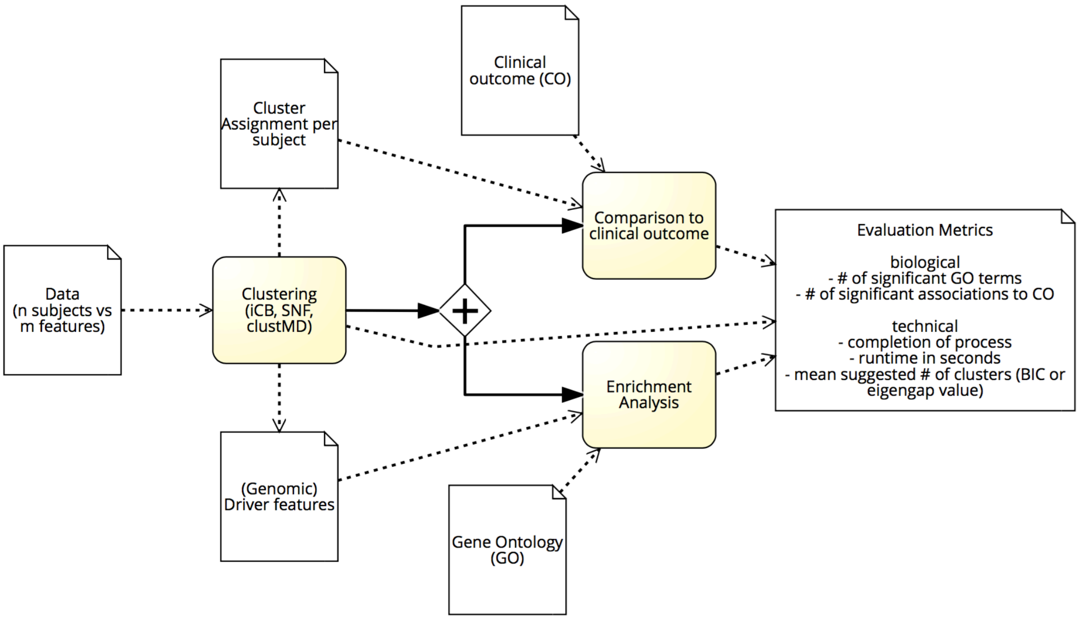

Clinical outcome (CO)
Cluster Assignment per subject
Comparison to clinical outcome
Evaluation Metrics
biological
- # of significant GO terms
- # of significant associations to CO
technical
- completion of process
- runtime in seconds
- mean suggested # of clusters (BIC or eigengap value)
Data (n subjects vs m features)
Clustering (iCB, SNF, clustMD)
Enrichment Analysis
(Genomic) Driver features
Gene Ontology (GO)

Figure C.1: Experimental setup for multi-omics experiments. Every data set is analyzed by every algorithm separately (single-omics experiments) and once in the respective tool‘s combined mode (multi-omics experiments) for a range of $k = 2$ to $k = 6$ possible clusters. All cluster assignments are tested for association to the binary clinical outcome (CO) estimators. Genes and proteins meeting a tool-specific significance threshold are included in the enrichment analysis against the Gene Ontology database. Technical and biological evaluation metrics are used for comparison.
# – number

Data preprocessing: Whole genome sequencing data is processed according to GATK best practices and annotated via the variant effect predictor (VEP) plugin dbNSFP. Deleterious variants (REVEL score < 0.05) are selected and summarized on the gene level yielding a binary feature of harbouring no deleterious variant (=0) or at least one (=1). LFQ intensities for protein expression were extracted from the MaxQuant output file according to [4]. Proteins detected in at least half of the patients are selected, $log2$-transformed and missing values are imputed based on a normal distribution of width $= 0.3$ and downshift $= 1.8$. Clinical parameters with $> 90\%$ valid values are split into being of numeric, binary or categorical type. The final data sets consists of 3812 numerical features for protein expression, 116 genes harbouring at least one deleterious variant for one subject (i.e., a binary feature), and 46 numerical, 28 binary and 3 categorical clinical features.

Algorithm parameters: Settings are kept close to the suggested default values, but needed the following adjustments:

- SNF's K is set to 10.
- iClusterBayes settings: n.burnin = 18000, n.draw = 20000, prior.gamma = 0.5, sdev = 0.05, thin = 3, pp.cutoff = 0.5, gaussian and binomial priors for respective numerical and binary data.
- clustMD settings: all six available models are tested for best fit according on scaled data to estimated Bayesian Information Criterion (BIC), the best model is chosen to calculate the final result. Kmeans clustering is used to initiate the clustering. Furthermore, Nnorms=50000 and MaxIter=500.

The suggested number of clusters is calculated based on BIC for iCB, estimated BIC for clustMD and the eigengap value for SNF.

Clinical outcome: A favourable clinical outcome is defined by reduced hypertrophy, decrease in NYHA stage or in nt-proBNP or an increase in left ventricular ejection fraction (LVEF) with regard to a post-surgery examination. All cluster assignments are tested for association to the binary outcome estimators (favourable outcome = 1) using Fisher‘s exact test and Benjamini-Hochberg multiple testing correction. An adjusted p-value of $<$ 0.05 is considered to be significant.

Enrichment analysis: Enrichment analysis is performed via enrichR (2.1) against all Gene Ontology (GO) terms (2018). Only genes and proteins with a posterior probability of > 0.7 for iCB or a normalized mutual information (NMI) > 0.3 for SNF are included in the analysis.

Experiments were conducted on a MacBook Pro (8 GB 1867 MHz DDR3, 2.9 GHz Intel Core i5, 2 cores, OS X El Capitan 10.11.6), R version 3.5.1.

Results

First, we performed a comparison of SNF, iCB and clustMD in a single-omics and a multi-omics setting and assessed technical metrics such as run time per feature and biological metrics. The results are summarized in Table C.1.

SNF stays below 1 second of computation time per feature in single omic and multi-omics mode. Within all tested single omics data sets and all possible k, SNF finds two solutions that show a significant association to clinical outcome variables after multiple testing correction: the numeric clinicome data is associated to a decrease in hypertrophy and a decrease in nt-proBNP at visit 2 when k = 6. There are no significant associations found in the multi-omics approach. Furthermore, enrichment analysis of relevant genomic and proteomic features results in six significant GO terms in the single omic and the multi-omics mode.

iCB also stays below a second of computation time per feature as long as the feature number is low as in the single-omics experiment. However, it computation time per feature triples in the multi-omics setting. The mean number of suggested clusters for a best solution is two in both experimental modes. Similar to SNF, there are two significant associations to clinical outcome in the single-omics solutions. We find an association of the numeric clinicome, when k = 3 to a decrease in hypertrophy and an association of the proteome (k =4) to a decrease in NYHA class. In the multi-omics mode we again find no associations to clinical outcome, however there are 75 terms enriched in the GO analysis.

ClustMD requires almost a minute of computation time per feature in the single-omics experiments, as iCB suggests two clusters to give rise to the best solution, however there are no significant hits in the clinical outcome and enrichment analysis. clustMD finds no solution in the multi-omics setting by the process not showing any further activity for at least a day while not finishing or aborting.

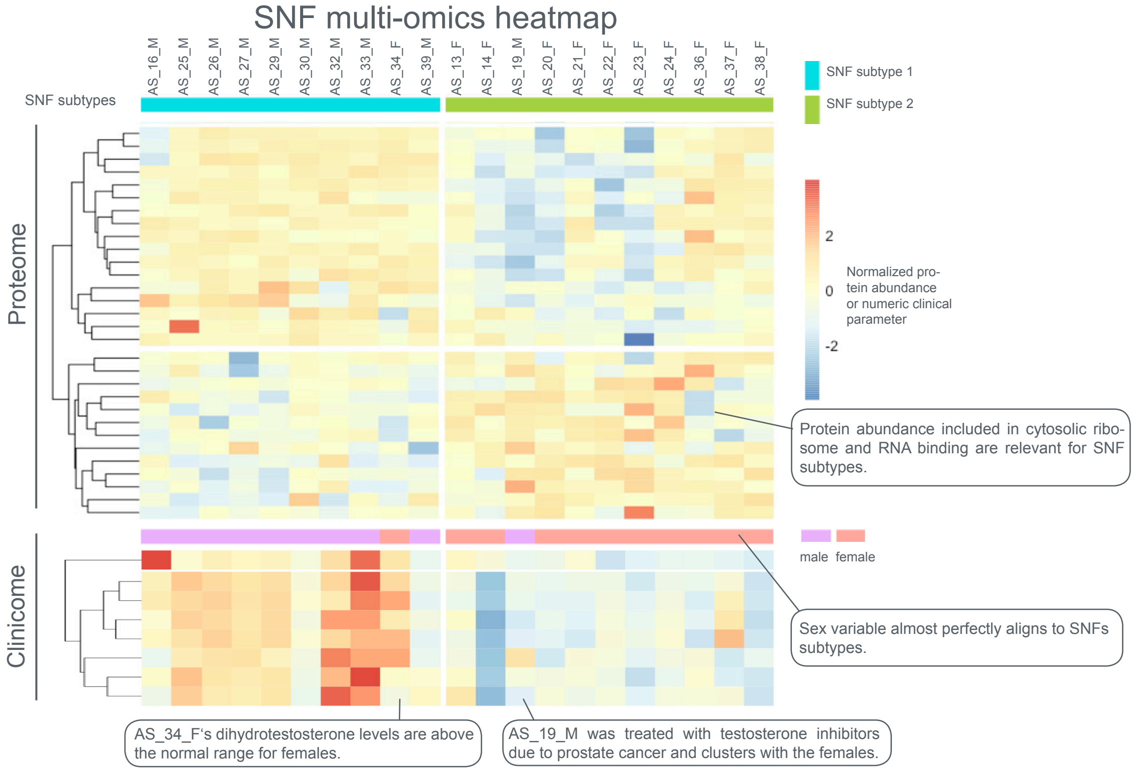
In summary, SNF is fastest and in our setup shows no obvious dependency of feature number to computation time. iCB's solutions show the most associations in the biological evaluation metrics. clustMD is very limited in feature capacity and the implementation is unstable.

Table C.1: Summary of quantitative metrics for our comparison of SNF, iCB and clustMD. Results are given for the single and multi-omics experiments. Stars on numbers denote and association of * the numeric clinicome, when k = 6 to a decrease in hypertrophy and a decrease in nt-proBNP at visit 2; ** the numeric clinicome, when k = 3 to a decrease in hypertrophy and an association of the proteome (k =4) to a decrease in NYHA class.

	single *ome			multi omics		
	SNF [192]	iCB [191]	clustMD [193]	SNF [192]	iCB [191]	clustMD [193]
Computation time per feature in sec	<1	0.084	58.8	<1	0.299	no solution
Mean suggested number of clusters	3	2	2	2	2	no solution
Associations to clinical outcome	2*	2**	0	0	0	no solution
Associations to Gene Ontology	6	6	0	6	73	no solution

In a second step, we take a closer look at the best solutions found in the multi-omics setting. There is no solution for clustMD, and a solution both identifying two clusters of samples by SNF and iCB. Features and cluster assignments of samples are visualized in a heatmap for SNF in Figure C.2 and for iCB in Figure C.3.

Although SNF and iCB both find two clusters, the assignment of samples to these clusters does not correspond across algorithms (adjusted Rand index = 0.032) and is based on different features. Relevant features for SNF are defined by a normalized mutual information of > 0.3. Only features from the clinicome (numeric and binary) and proteome meet the criterion. The sex of patients is the only relevant binary feature and almost perfectly devides the SNF subtypes one and two. Only two samples deviate from this rule: `AS_34_F`, a female, is assigned to the "male" subtype 1 with dihydrotestosterone levels being above the normal range of females. Similarly, `AS_19_M` is assigned to the "female" subtype 2. The patient was treated with testosterone inhibitors due to a history of prostate cancer. Enrichment analysis of proteomic features driving the separation resulted in terms for RNA binding, cytosolic ribosomal subunits and focal adhesion.



Protein abundance included in cytosolic ribosome and RNA binding are relevant for SNF subtypes.

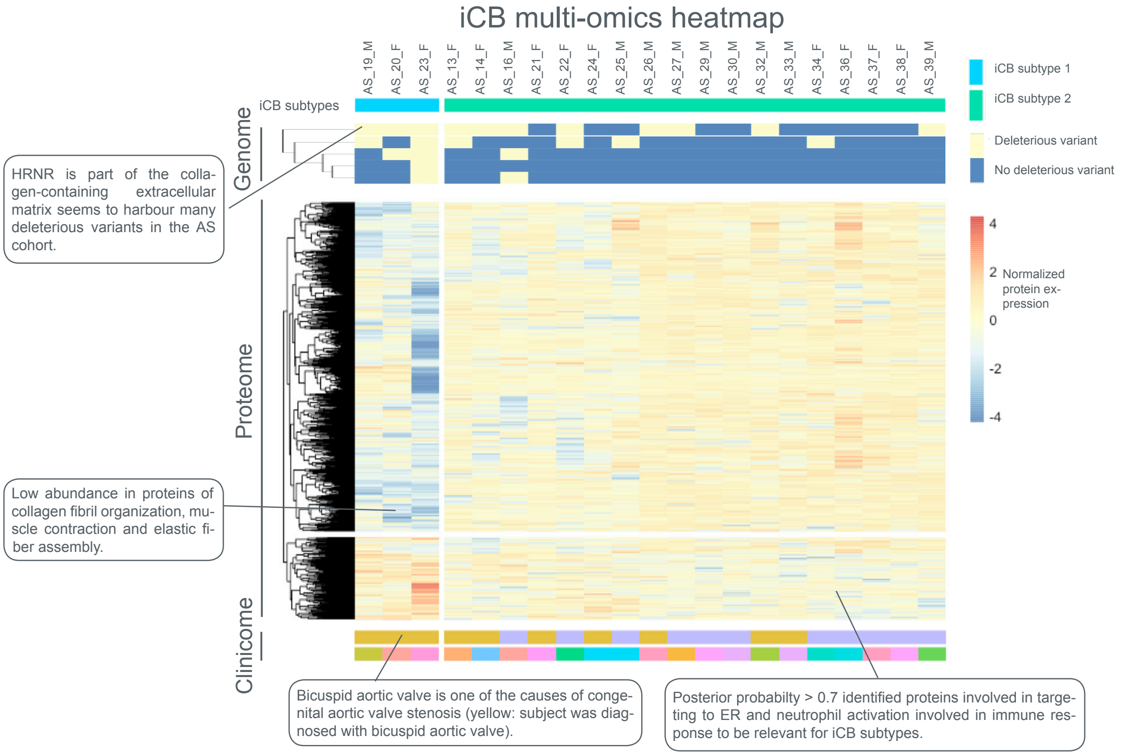
Sex variable almost perfectly aligns to SNFs subtypes.

AS_34_F's dihydrotestosterone levels are above the normal range for females.

AS_19_M was treated with testosterone inhibitors due to prostate cancer and clusters with the females.

Figure C.2: Multi-omics heatmap representation of SNF's results for the two suggested disease subtypes. Only relevant features (normalized mutual information >0.3) are displayed. Column clusters are defined by SNF's subtype assignment, row clusters are pruned to show two clusters per dataset, based on Euclidean distance. If only a single feature is selected from a data set (e.g. from binary clinical data) it is displayed as a single row.

Relevant features for iCB are defined by a posterior probability of > 0.7 and are found among the genomic, proteomic and clinicome data (see Figure C.3). Among the genomic variation data, we find deleterious variants for many samples in HRNR, a gene, which encodes for a protein in the collagen-containing extracellular matrix. According to the enrichment analysis, iCB subtype 1 shows a low abundance of proteins involved in collagen fibril organization, muscle contraction and elastic fiber assembly and an increase in proteins involved in targeting of proteins to the endoplasmic reticulum (ER) and neutrophil activation involved in immune response. Among the clinical features we find the diagnosis of bicuspid aortic vales to be relevant. Bicuspid aortic valve is a congenital defect, which is a risk factor for aortic valve stenosis.



HRNR is part of the collagen-containing extracellular matrix seems to harbour many deleterious variants in the AS cohort.

Low abundance in proteins of collagen fibril organization, muscle contraction and elastic fiber assembly.

Bicuspid aortic valve is one of the causes of congenital aortic valve stenosis (yellow: subject was diagnosed with bicuspid aortic valve).

Posterior probabilty > 0.7 identified proteins involved in targeting to ER and neutrophil activation involved in immune response to be relevant for iCB subtypes.

Figure C.3: Multi-omics heatmap representation of iCB's results for the two suggested disease subtypes. Only relevant features (posterior probability of >0.7 for iCB) are displayed. Column clusters are defined by iCB‘s subtype assignment, row clusters are pruned to show two clusters per dataset, based on Euclidean distance. If only a single feature is selected from a data set (e.g. from binary clinical data) it is displayed as a single row. ER: endoplasmic reticulum

Discussion and Conclusion

SNF, iCB and clustMD were tested on single and multi-omics data of 21 aortic stenosis patients. Our results show that SNF and iCB found solutions to any given combination that in part showed associations to clinical outcome and GO terms. In addition to genomic and proteomic features, the clinicome showed relevance in both successful multi-omics clusterings. The interpretation of results is limited by the extremely small sample size, while the number of features is high (n << p problem). Resulting subtypes are not robust across algorithms and multi-omics subtypes do not necessarily confer more associations to clinical outcome than single-omics USD. However, SNF was the fastest algorithm and shows no obvious dependency of feature number to computation time in our setup. iCB's solutions show the most associations in the biological evaluation metrics. clustMD is very limited in feature capacity, most probably because it models all distributions of available features with a very elaborate mathematical modes that requires a lot of computational power, especially in the case of categorical data. Despite the limited possibility of interpretation, the SNF subtypes remarkably capture differences in sexes also found in the in-depth proteome analysis in chapter 4. As such, we can recommend to at least try one of the algorithms for means of hypothesis generation or in larger cohorts, as SNF and iCB are well suited for all tested data types and for a large number of features.

D Data tables

Table D.1: Assignments of GO terms to five categories. ECM = extracellular matrix composition, other terms, MET = metabolism and mitochondrial functions, MUSC_SKEL = cytoskeleton and muscle contraction, PROT = proteostasis

ECM	other terms	MET	MUSC_SKEL	PROT
GO:0003418	GO:0002455	GO:0006099	GO:0005862	GO:0005788
GO:0030199	GO:0002576	GO:0006119	GO:0016460	GO:0031983
GO:0043062	GO:0002920	GO:0006120	GO:0030016	GO:0000184
GO:0048251	GO:0006956	GO:0006635	GO:0030016	GO:0006413
GO:0090171	GO:0006957	GO:0009060	GO:0042383	GO:0006418
GO:0001527	GO:0006958	GO:0009081	GO:0042641	GO:0006613
GO:0005581	GO:0006959	GO:0022904	GO:0043292	GO:0006614
GO:0005588	GO:0007597	GO:0032543	GO:0043292	GO:0043039
GO:0005589	GO:0010873	GO:0032981	GO:0097517	GO:0045047
GO:0005593	GO:0010903	GO:0033108	GO:0003779	GO:0070125
GO:0005604	GO:0010951	GO:0042775	GO:0008307	GO:0070972
GO:0031012	GO:0015671	GO:0045333	GO:0051371	GO:0140053
GO:0062023	GO:0015701	GO:0046950	GO:0005859	GO:1904869
GO:0071953	GO:0030193	GO:0046952	GO:0016460	GO:0005832
GO:0098643	GO:0030195	GO:0005743	GO:0001725	GO:0005840
GO:0098644	GO:0030300	GO:0005746	GO:0005859	GO:0005852
GO:0098647	GO:0030449	GO:0005759	GO:0097513	GO:0015934
GO:0004857	GO:0031639	GO:0019866	GO:0051764	GO:0015935
GO:0004866	GO:0034371	GO:0030964	GO:0008091	GO:0017101
GO:0004867	GO:0034372	GO:0070069	GO:0014731	GO:0022625
GO:0005198	GO:0042730	GO:0070469	GO:0030863	GO:0022626
GO:0005201	GO:0042744	GO:0098798	GO:0032432	GO:0022627
GO:0005518	GO:0048821	GO:0098800	GO:0043034	GO:0044391
GO:0005539	GO:0050818	GO:0098803	GO:0090665	GO:0003735
GO:0008201	GO:0050819	GO:1990204	GO:0005200	GO:0004812
GO:0030020	GO:0051917	GO:0003987	GO:0017166	GO:0002479
GO:0030021	GO:0051918	GO:0004300	GO:0042805	GO:0006521
GO:0030023	GO:0061045	GO:0005471	GO:0016010	GO:0042590
GO:0043394	GO:0072376	GO:0009055	GO:0090665	GO:0070125
GO:0048407	GO:0072378	GO:0016655	GO:0003779	GO:0140053
GO:0061134	GO:1900046	GO:0050136	GO:1900025	GO:1902036
GO:0097493	GO:1900047	GO:0050136	GO:0002102	GO:1904869
GO:1901681	GO:1904478	GO:0006084	GO:0003779	GO:0008540
GO:0005583	GO:0005577	GO:0006104	GO:0035374	GO:0008541
GO:0098643	GO:0005579	GO:0006551	GO:0030055	GO:0031597

GO:0038063	GO:0031089	GO:0006573	GO:0044307	GO:1905368
GO:0005587	GO:0031091	GO:0009063		GO:0004812
GO:0098643	GO:0031093	GO:0010257		GO:0002479
	GO:0031838	GO:0010510		GO:0006521
	GO:0034363	GO:0016054		GO:0042590
	GO:0034364	GO:0035383		GO:1902036
	GO:0034365	GO:0042773		GO:1904869
	GO:0034366	GO:0046395		GO:0008540
	GO:0034774	GO:0046459		GO:0008541
	GO:0042627	GO:0071616		GO:0031597
	GO:0044217	GO:1904182		GO:1905368
	GO:0071682	GO:1904183		GO:0000956
	GO:0072562	GO:0005750		GO:0001732
	GO:0001848	GO:0030062		GO:0002183
	GO:0004064	GO:0045261		GO:0006401
	GO:0004089	GO:0045271		GO:0006402
	GO:0030492	GO:0003954		GO:0006412
	GO:0031210	GO:0003985		GO:0006605
	GO:0031721	GO:0004473		GO:0006612
	GO:0035473	GO:0004774		GO:0019083
	GO:0060228	GO:0008948		GO:0043043
	GO:0070325	GO:0016408		GO:0090150
	GO:0070653	GO:0016615		GO:0140053
	GO:0120020	GO:0016878		GO:1904869
	GO:0008228	GO:0050136		GO:0005844
	GO:0001849	GO:0051287		GO:0016282
	GO:0004064	GO:0090482		GO:0033290
	GO:0004089			GO:0042788
	GO:0002002			GO:0071541
	GO:0008228			GO:1990904
	GO:0001849			GO:0003743
	GO:0002003			GO:0008135
	GO:0010951			GO:0016875
	GO:0015701			GO:0031369
	GO:0044217			GO:0045182
	GO:0001848			GO:0090079
	GO:0004064			GO:1990948
	GO:0004089			
	GO:0031210			
	GO:0001849			
	GO:0002697			
	GO:0007596			
	GO:0009611			
	GO:0010951			
	GO:0015701			

GO:0015914
GO:0019433
GO:0019835
GO:0034378
GO:0042060
GO:0043687
GO:0043691
GO:0045916
GO:0050817
GO:0051006
GO:0051346
GO:0051919
GO:0061041
GO:0070527
GO:0071830
GO:0090303
GO:0098869
GO:0098883
GO:1904729
GO:2000427
GO:0001652
GO:0030669
GO:0031232
GO:0044216
GO:0046930
GO:0001848
GO:0004064
GO:0004089
GO:0016209
GO:0031210
GO:0043395
GO:0050750
GO:0061135