Inaugural dissertation
for
obtaining the doctoral degree
of the
Combined Faculty of Mathematics, Engineering and Natural Sciences
of the
Ruprecht - Karls - University
Heidelberg

Presented by

Dr. med. Jan David Lanzer
born in Berlin, Germany
Oral examination: 13.07.2023

From Mouse Models to Patients: A Comparative Bioinformatic Analysis of HFpEF and HFrEF

Referees:
Prof. Dr. Marc Freichel
Prof. Dr. Julio Saez-Rodriguez

"... how false the most profound book turns out to be when applied to life."

W. Faulkner

Abstract

Heart failure (HF) represents an immense health burden with currently no curative therapeutic strategies. Study of HF patient heterogeneity has led to the recognition of HF with preserved (HFpEF) and reduced ejection fraction (HFrEF) as distinct syndromes regarding molecular characteristics and clinical presentation. Until the recent past, HFrEF represented the focus of research, reflected in the development of a number of therapeutic strategies. However, the pathophysiological concepts applicable to HFrEF may not be necessarily applicable to HFpEF. HF induces a series of ventricular modeling processes that involve, among others, hallmarks of hypertrophy, fibrosis, inflammation, all of which can be observed to some extent in HFpEF and HFrEF. Thus, by direct comparative analysis between HFpEF and HFrEF, distinctive features can be uncovered, possibly leading to improved pathophysiological understanding and opportunities for therapeutic intervention. Moreover, recent advances in biotechnologies, animal models, and digital infrastructure have enabled large-scale collection of molecular and clinical data, making it possible to conduct a bioinformatic comparative analysis of HFpEF and HFrEF.

Here, I first evaluated the field of HF transcriptome research by revisiting published studies and data sets to provide a consensus gene expression reference. I discussed the patient clientele that was captured, revealing that HFpEF patients were not represented.

Thus, I applied alternative approaches to study HFpEF. I utilized a mouse surrogate model of HFpEF and analyzed single cell transcriptomics to gain insights into the interstitial tissue remodeling. I contrasted this analysis by comparison of fibroblast activation patterns found in mouse models resembling HFrEF. The human reference was used to further demonstrate similarities between models and patients and a novel possible biomarker for HFpEF was introduced.

Mouse models only capture selected aspects of HFpEF but largely fail to imitate the complex multi-factor and multi-organ syndrome present in humans. To account for this complexity, I performed a top-down analysis in HF patients by analyzing phenome-wide comorbidity patterns. I derived clinical insights by contrasting HFpEF and HFrEF patients and their comorbidity profiles. These profiles were then used to predict associated genetic profiles, which could be also recovered in the HFpEF mouse model, providing hypotheses about the molecular links of comorbidity profiles.

My work provided novel insights into HFpEF and HFrEF syndromes and exemplified an interdisciplinary bioinformatic approach for a comparative analysis of both syndromes using different data modalities.

Zusammenfassung

Herzinsuffizienz (HF) stellt eine immense gesundheitliche Belastung dar, für die es derzeit keine heilenden therapeutischen Strategien gibt. Die Untersuchung der Heterogenität von HF-Patienten hat dazu geführt, dass HF mit erhaltener (HFpEF) und verminderter Ejektionsfraktion (HFrEF) als unterschiedliche Syndrome hinsichtlich molekularer Merkmale und klinischer Präsentation anerkannt wurden. Die HF-Forschung konzentrierte sich im vergangenen Jahrhundert in der Regel auf HFrEF, was sich im Arsenal der Pharmakotherapie widerspiegelt, die auf kompensatorische Mechanismen abzielt, die unser Modell des Verständnisses von HFrEF erklären, aber nicht vollständig auf HFpEF zutreffen. HF induziert eine Reihe von ventrikulären Modellierungsprozessen, zu denen unter anderem Merkmale der Hypertrophie, Fibrose und Entzündung gehören, die alle in gewissem Umfang bei HFpEF und HFrEF beobachtet werden können. Durch eine direkte vergleichende Analyse von HFpEF und HFrEF lassen sich also Unterscheidungsmerkmale aufdecken, die möglicherweise zu einem besseren pathophysiologischen Verständnis und zu Möglichkeiten der therapeutischen Intervention führen. Darüber hinaus haben die Fortschritte in der Biotechnologie, bei Tiermodellen und in der digitalen Infrastruktur die Erhebung hochdimensionaler molekularer und klinischer Daten ermöglicht, welche eine vergleichende bioinformatische Analyse von HFpEF und HFrEF ermöglichen.

Hier bewertete ich zunächst den Bereich der HF-Transkriptomforschung, indem ich veröffentlichte Studien und Datensätze überprüfte, um eine konsensfähige Genexpressionsreferenz zu erstellen. Ich diskutierte die erfasste Patientenklientel und stellte fest, dass HFpEF-Patienten nicht vertreten waren.

Daher wandte ich alternative Ansätze zur Untersuchung von HFpEF an. Ich nutzte ein HFpEF-Mausmodell und analysierte die Genexpression einzelner Zellen, um Einblicke in den Umbau des interstitiellen Gewebes zu gewinnen. Ich verglich die Aktivierungsmuster der Fibroblasten in HFpEF mit denen in Mausmodellen, welche HFrEF ähneln. Die menschliche Genexpressionsreferenz wurde verwendet, um weitere Gemeinsamkeiten zwischen Mausmodellen und Patienten aufzuzeigen, und ein neuer möglicher Biomarker für HFpEF wurde vorgestellt.

Mausmodelle erfassen nur ausgewählte Aspekte der HFpEF, können aber das komplexe Multifaktor- und Multiorgan-Syndrom des Menschen kaum nachahmen. Um dieser Komplexität Rechnung zu tragen, führte ich bei HF Patienten eine Top-down-Analyse durch, indem ich Phänomweite Komorbiditätsmuster analysierte. Durch die Gegenüberstellung von HFpEF- und HFrEF-Patienten, habe ich klinische Erkenntnisse

hinsichtlich ihrer Komorbiditätsprofile gewonnen. Diese Profile wurden dann zur Vorhersage zugehöriger genetischer Profile verwendet, die auch im HFpEF-Mausmodell bestätigt werden konnten und Hypothesen über die molekularen Zusammenhänge der Komorbiditätsprofile lieferten.

Meine Arbeit lieferte neue Erkenntnisse über die HFpEF und HFrEF Syndrome und veranschaulichte einen interdisziplinären, bioinformatischen Ansatz für eine vergleichende Analyse beider Syndrome unter Verwendung verschiedener Datenmodalitäten.

Acknowledgements

This thesis was my effort to incorporate bioinformatics into the study of medicine. To achieve this, a knowledgeable and supportive environment was indispensable. I am very lucky to have found such.

I am indebted to Julio Saez-Rodriguez, who accepted me as a PhD student. He provided me with the time and space to learn and to step-by-step grow accustomed to scientific problem formulation and solution. I received his trust and kindness which were the basis for joyful learning.

The colleagues that populated this scaffold were of fundamental importance to me and I thank all my teachers and companions, especially Ricardo O. Ramirez-Flores, Alberto Valdeolivas, Jovan Tanevski, Rebecca T. Levinson, Attila Gabor, Martin Garrido, Daniel Dimitrov, Sophia Müller-Dott, Olga Ivanova, Robin Fallegger, Lorna Wessel, Charlotte Boys, and Pau Badia i Mompel.

I thank Florian Leuschner, Laura Wienecke and all collaboration partners for their valuable contributions to generating and evaluating data, and for engaging in thoughtful scientific discussions.

I want to thank Informatics for Life funded by the Klaus Tschira foundation together with Jobst H. Schultz and Hans-Christoph Friederich who provided funding for my work and continuous support. The MD/PhD program of the Heidelberg University was indispensable for it provided the necessary financial, administrative and educational framework.

I am indebted to my parents, Renate and Peter Lanzer, who provided the right mixture of critique and trust for me to grow. I cherish my siblings, Anne, Nina and Claudia Lanzer; you are a great team to have by my side.

I am grateful to Sophia Frangos who accompanied me, listening to all sorrows without growing tired, at least not showing it. You were part of any success I had.

Table of contents

List of Figures and Tables

Introduction

A tale of the failing heart

The cardiovascular system supplies the mammalian body with oxygenic and nutritious blood. It consists of the heart and the vasculature who build a system where the blood circulates while it is enriched or depleted for various metabolites by the different tissues it traverses. The regulation of blood flow is complex and multifactorial and involves many feedback mechanisms that adjust the blood supply to the organs needs. If the heart fails to sufficiently meet the demands of the body, a pathologic state is reached that is termed heart failure (HF).

One possible model of understanding HF is an initial damage to the cardiac tissue, resulting in an impaired pump function. This initial damage launches a series of compensatory mechanisms to ensure blood supply. While these mechanisms can alleviate HF short term, they overstrain the heart long term and cause a chronification of HF and deterioration of the residual pump function. Thus, pharmaceutical treatment strategies aim to block these compensatory mechanisms (e.g. beta-blockers, Angiotensin converting enzyme inhibitors, Mineralocorticoid receptor antagonists). While these treatments can improve patient outcomes, they cannot reverse disease progression.

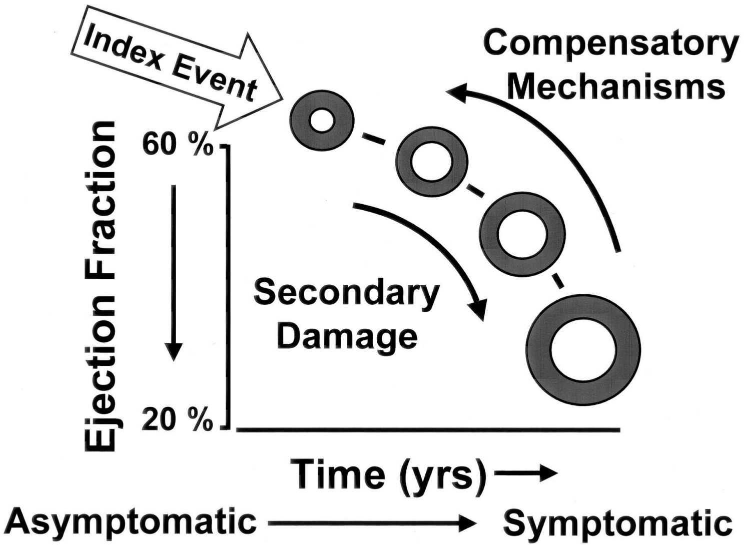


Figure I.1. Possible model of heart failure progression.

An index event causing cardiac damage (primary damage) leads to decreased pump function. Compensatory mechanisms are activated to maintain pump function but overstrain the heart with time leading to decompensated heart failure (secondary damage). Reprinted with written permission from[1].

Thus, to advance clinical care for HF, it is necessary to study the syndrome more closely. One approach is to categorize HF into subgroups to identify possible pathophysiologic subgroup characteristics as novel treatment targets. The categorization of HF today spans multiple dimensions describing functional (forward or backward), clinical (e.g. NYHA, compensated or decompensated, acute or chronic), morphological (e.g. hypertrophic, fibrotic), anatomical (e.g. right, left or global), etiological (eg. ischemic, hypertensive or dilated) aspects of the syndrome. An important classification is the functional distinction between a systolic and a diastolic heart failure. The diastole is the phase of the cardiac cycle where the ventricle relaxes and increases in volume, filling with the blood inflow from the atria. When filled, the systolic phase begins with the ventricle contraction, increasing intraventricular pressure. When the pressure exceeds the arterial diastolic pressure the semilunar valves open and blood is pumped into the arterial system until ventricular pressure falls below arterial pressure and the diastolic phase begins again. Thus, systolic HF describes the inadequate discharge of blood from the ventricle while diastolic HF describes the inadequate filling of the ventricle, both ultimately leading to a reduced cardiac output. This distinction is recognized with increasing importance.

In clinical practice a standard measure to assess LV function is the left ventricular ejection fraction (LVEF). LVEF is a ratio of the LV volume before and after the systolic phase and thus describes the share of the blood in the LV that is ejected during the systolic phase with physiologic values ranging between 50% and 65%. Thus, systolic HF is characterized by reduced LVEF (<40%) while diastolic HF often presents with a preserved LVEF (>50%). Systolic HF has been the major focus of twentieth century HF research and was addressed in most clinical trials while diastolic HF has been recognized only later[2]. It was noted that patients with a preserved LVEF could present with clinically severe HF symptoms, however, at first a transient systolic recovery was assumed. The CHARM-Preserved trial (2003) [3] was the first larger clinical trial that addressed patient outcome for diastolic HF patients (LVEF >40%) receiving angiotensin receptor blocker. The following years increased the attention to diastolic HF. It was noted that patients with a preserved LVEF displayed clinically different profiles and diastolic dysfunction was only one of many characterstics. Over time, this coined the distinction of HF into HF with reduced or preserved ejection fraction (HFrEF and

HFpEF, respectively), which are increasingly used to describe two major forms of chronic HF (Figure I.2).

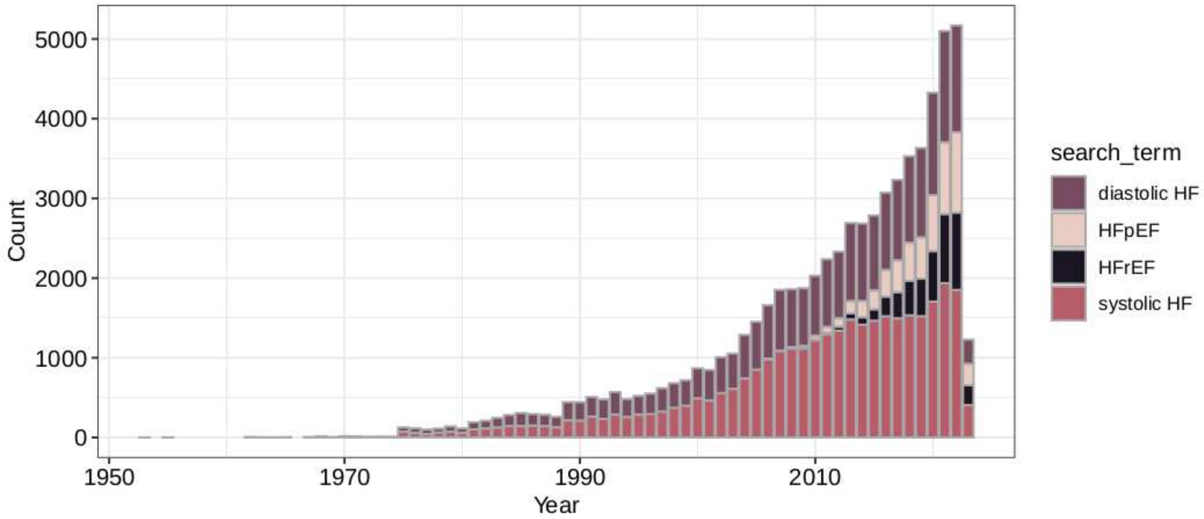


Figure I.2 Pubmed article counts of HF related terms.

Number of articles in the NCBI's Pubmed database mentioning HF related search terms. Emergence of HFpEF and HFrEF literature can be dated around 2010 while the distinction and terminology was probably coined in the early 2000's. Database accessed on March 15th, 2023.

HFpEF as an evolving concept

HFpEF has since been the focus of intense research efforts. While early studies already suggest that epidemiologically HFpEF patients are no curiosity [4], the epidemiological importance of the syndrome that makes up around 50% of HF patients was appreciated only lately[5]. Considering the high prevalence of HF of up to six percent in western countries like Germany[6], HFpEF is responsible for an immense burden for patients and the health care system. However, apart from gliflozins, no effective treatment strategies exist today to reduce the associated diastolic dysfunction, fibrosis, hypertrophy and the resulting pronounced morbidity and mortality. Therapeutic concepts and established drugs for the treatment of heart failure with reduced ejection fraction (HFrEF) failed broadly when tested for beneficial effects in HFpEF, suggesting fundamentally different pathomechanisms that would not follow the HF paradigm of overcompensation (see Figure I.2) [7]. Moreover the recent years of HFpEF research shaped our notion of this syndrome as a systemic disease with a multi-organ involvement. Influential factors were learned from patient cohorts, and include aging, hypertension, obesity and metabolic syndrome and female sex. These factors are thought to induce a series of pathomechanisms which in turn were mainly derived from animal models. HFpEF pathomechanisms include hypertrophy, myocardial fibrosis, excitation–contraction coupling defects, sarcomere dysfunction, cGMP–PKG signaling deficiency, nitrosative–oxidative stress, microvascular insufficiency, inflammation, and mitochondrial and metabolic defects [7]. These pathways presumably cause cardiomyocyte stiffness, hypertrophy and cardiac fibrosis typically associated with concentric cardiac remodeling (Figure I.3).

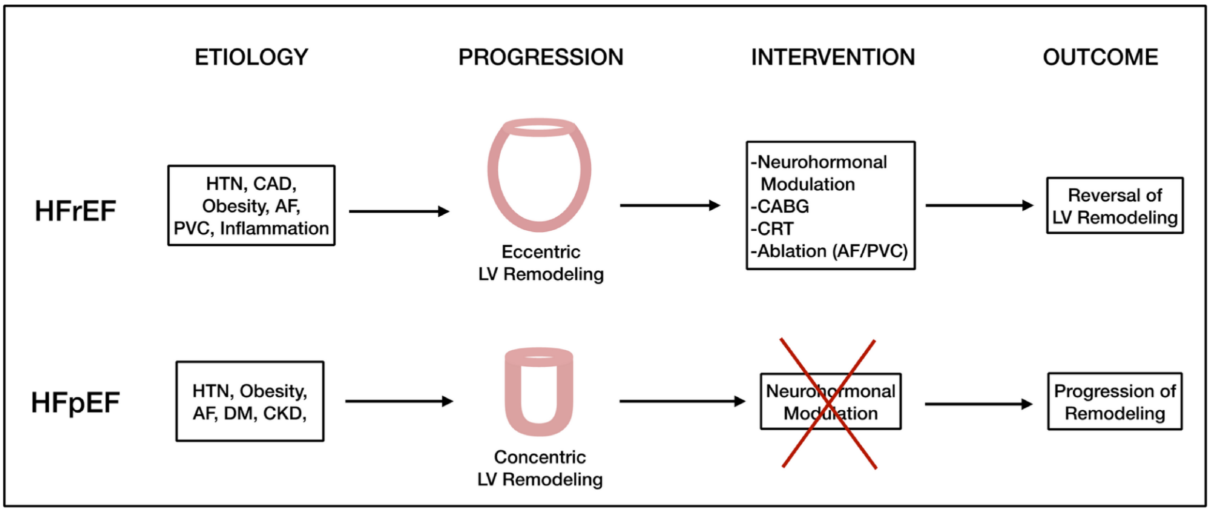


Figure I.3. Contrasting therapeutic outcomes in eccentric and concentric LV remodeling.

AF indicates atrial fibrillation; CABG, coronary artery bypass graft; CAD, coronary artery disease; CKD, chronic kidney disease; CRT, cardiac resynchronization therapy; DM, diabetes mellitus; HFpEF, heart failure with preserved ejection fraction; HFrEF, heart failure with reduced ejection fraction; HTN, hypertension; LV, left ventricle; and PVC, premature ventricular contraction. *Reprinted from* [8]

While the disentanglement of these factors has been difficult, one reason for slow progress is the observed heterogeneity of HFpEF patients. Patient clustering has been previously shown to yield novel subgroups of HFpEF defined by multivariate similarity [9–11]. While these subgroups require external validations, the evaluation on molecular level could address a key question on how patients develop common characteristics of HFpEF via different disease pathways (e.g. hypertensive vs. metabolic vs. aging). NHLBI's HFpEF working group formulated research priorities for HFpEF that included i) improved animal models, ii) human tissue and phenotype data and iii) application of computational approaches and multiscale modeling [12].

Ventricular tissue remodeling

The cardiac ventricles in HF undergo a morphological, cell compositional and functional adaptation process that is called remodeling. The hallmarks of this process include inflammation, hormone and cytokine dysregulation, myocardial fibrosis, cardiac hypertrophy, excitation-contraction coupling defects, oxidative stress, and metabolic and mitochondrial defects. These processes are involved in HF remodeling with different emphasis and character in HFpEF and HFrEF. Here, I will briefly introduce hallmark features of cardiac hypertrophy, fibrosis and oxidative stress while recent literature can provide a more comprehensive overview [7,13,14].

Muscle hypertrophy is the physiologic adaptation to strain and thus the result of muscle training. In the heart, physiologic and pathologic hypertrophy differ fundamentally, the former being reversible, eccentric, the latter being mainly irreversible, concentric and associated with fetal gene program expression. In HF, cardiomyocyte growth leads to the concentric (i.e. inwards) thickening of the ventricle, which is a morphological transformation leading to diastolic dysfunction and thus more typical in HFpEF. With time, eccentric hypertrophy can develop, characterized by rather elongated than thickened cardiomyocytes (CMs) leading to impaired systolic function which is more typical in HFrEF. CM growth is regulated by a complex interplay of biological signals including among others NO, insulin like growth factors, thyroid hormones, mTOR pathway, catecholamines and the renin-angiotensin II system [15].

Cardiac hypertrophy is usually accompanied by cardiac fibrosis which is an evolutionary conserved process to preserve tissue integrity after injury. In the heart, replacement fibrosis can be differentiated from reactive fibrosis. The former is induced by cell death (e.g. after myocardial infarction) and is characterized by the rapid deposition of extracellular matrix (ECM) to preserve tissue integrity, comparable with tissue scarring. The latter is a slower process that leads to histological patterns of interstitial or perivascular ECM deposition without significant loss of CMs [16]. In HFpEF, reactive fibrosis is a typical pattern that has been associated with increased risk of arrhythmias and diastolic dysfunction, myocardial stiffness [17] and mortality [18]. While certain cross-organ preserved fibrotic pathways are well established like TGFbeta pathway and fibroblast to myofibroblast transdifferentiation, mechanistic and cellular differences regarding different fibrosis types and patterns remain unclear [19,20] [19].

Oxidative stress ist a hallmark of HFpEF that is defined as a dysbalance towards reactive oxygen species with damaging effects to cell structure and function. In HFpEF, this stress is linked to metabolic comorbidities such as hypertension and diabetes. Important actors include NADPH oxidases (e.g. NOX2), xanthine oxidase and the inducible nitric oxide synthases [7]. Recent studies showed that endoplasmic reticulum stress via XBP1 is a downstream effector of impaired NO signaling and directly related to the pathophysiology of the studied mouse model [21].

Bioinformatic approaches in HF research

In this section I will introduce previous work in the intersection of data science and HF research. I summarized and reviewed the state of this field in the beginning of my PhD in a peer-reviewed manuscript [22]. This article was written by me, supervision and edits were provided by Rebecca Levinson and Julio Saez-Rodriguez. Extracts and structure of this paragraph are cited from the review here.

In the past 5-10 years data science and bioinformatics have become an integral part of the study of the cardiovascular system. Novel technologies and improved data infrastructure led to a steep increase of voluminous, available data. To extract relevant information from high dimensional data, dimension reduction methods, linear and non linear modeling, enrichment based analysis, clustering algorithms and other machine learning techniques are routinely applied. HF, and especially HFpEF, is a prime target for bioinformatic research due to the complex etiology of the syndrome, the large number of risk factors and involved organs, the high degree of comorbidity in patients and the prolonged and progressive disease course. Data used for the study of HF are derived from a variety of sources, while some are dependent on tissue such as blood or myocardial samples, others are ascertained through clinical care or wearable devices. In my thesis I have worked with both data types and will introduce technology, data analysis principles and previous work in the next section.

Molecular HF profiles

High throughput methods enable researchers to study molecular profiles of tissues at high resolutions. This field is generally referred to with the suffix *-omics*. Omics technologies can be described as non-targeted - those that aim to measure complete molecular profiles in an unbiased manner - and targeted - those that have predefined molecules of interest. In HF, the specimen for omics analysis usually is cardiac tissue or blood (e.g. PBMCs). While myocardial omic analyses can help elucidate disease mechanisms and identify biomarkers and therapeutic targets, the tissue availability for human samples is limited. Blood samples are easier to access and can help survey HF patients at a higher temporal resolution. They are used for biomarker detection and to study genomics as well as the role of circulating cells while the origin and pathophysiology of circulating molecules can be difficult to define.

Different omics technologies pose similar challenges on data analysis and evaluation, including problems concerning accuracy, imputation, integration, replication and interpretation.

Transcriptomics

The transfer from genetic code to cellular function is mediated by the transcription of ribonucleic acid (RNA). RNA can be translated to proteins (coding RNA or messenger RNA), or execute structural (e.g. ribosomal RNA, transfer RNA) or gene regulatory functions (e.g. micro RNA, long non-coding RNA). The quantification of the set of RNA molecules (transcripts) produced by the genome is generally referred to as transcriptomics and provides important understanding of disease mechanisms[23].

This quantification is routinely performed via RNA sequencing technology. Briefly, RNA is extracted from the sample, possibly followed by mRNA enrichment or ribosomal RNA depletion. To stabilize the fragile RNA, it is transcribed into complementary DNA, which then can be amplified via PCR after adaptor ligation [24]. The amplified cDNA is sequenced by various technologies and the *Illumina* workflow is widely used, including all gene expression analysis in this thesis. Here, cDNA molecules are clustered on a flowcell and a complementary DNA strand is synthesized by using fluorescently labeled nucleotides, such that for each nucleotide extension the growing DNA strands are detected via imaging the fluorescence signals. The data that originates from this process is typically stored as FASTQ files, which contain each flowcell strand sequence together with metadata describing data quality. To quantify gene expression, these sequences are mapped on to a reference genome. While this process is computational intensive, pseudo aligners can offer an alternative approach by assessing for each read the compatible transcripts [25]. The alignment results in a count table describing the number of transcripts counted for each genomic region of interest (i.e. often genes). To compare genes between samples, technical factors such as sample sequencing depths are taken into account via normalization procedures. Gene abundance also affects gene variance and is routinely controlled for by variance stabilization approaches.

To relate gene expression patterns to a sample/patient phenotype such as HF, case control study designs are necessary. Contrast analysis is routinely performed via linear gene-level modeling, however, as transcriptome profiles can cover up to ~20.000 coding and ~15.000 non coding genes [26], this is subject to the burden of multiple hypothesis testing. Thus, the functional interpretation is aided by reducing data dimensions by incorporating prior biological knowledge together with overrepresentation analysis (ORA)[27]. Prior biological knowledge is the result of previous studies that generated associations between a biological topic (e.g. gene set name) and its associated features (e.g. gene set members). This knowledge can take many forms, such as a signaling pathway and its members, or a signaling pathway and its gene regulatory targets, or a cell type and its markers, or a disease and its associated genes. When analyzing transcriptomic data, this knowledge can then be used to decipher gene expression patterns by estimating an effect size of a gene set to be consistently up or down regulated together with the probability of observing this pattern by chance. There are many different methods to approach this task. The hypergeometric test relies on estimating the intersection between two gene sets and thus necessitates a cut off decision for the regulatory pattern (i.e. which genes are considered up-regulated). A fundamental alternative was introduced by the development of the gene set enrichment analysis (GSEA) [28]. Here, the effect size of enrichment is calculated as a running sum by adding or subtracting the gene-level statistic for a given gene set, thus increasing if gene sets are located close to the top of the ranking. This approach does not depend on a cut-off and thus considers the full transcriptome ranking. However, hypothesis testing can be difficult and computationally intensive permutation tests were initially suggested [29]. Other approaches include the combinations of gene-level statistics in univariate or multivariate linear models, variance analysis or simple averaging. Method selection should ultimately consider the research question at hand, the format of prior knowledge, the choice and reliability of the gene level statistic, the methods assumptions and characteristics and, if available, benchmarking results[30] .

The first high-throughput transcriptomic study on myocardial human HF was published in 2000 [31]. In the subsequent decades, technological and bioinformatic advances in transcriptomics have improved our comprehension of cardiac hypertrophy [32], reverse remodeling [33], cardiac metabolism [34,35], cardiac fibrosis [36], and immune dysregulation [37] in HF. Several studies made their data sets and protocols publicly available on platforms like NCBI's gene expression omnibus. However, few attempts have been made to compare transcriptomic HF studies [38–40]. The continuing development of sophisticated data analysis methods invites the retrospective re-analysis and integration of published HF studies, although data integration from different platforms, centers and technologies presents many challenges [41].

Single cell RNAseq

With single cell RNAseq, the transcriptome of individual cells can be measured, providing tissue profiling at unprecedented granularity. Bulk RNAseq fails to account for a functional diversity of cell types that might be crucial in understanding the orchestration of myocardial syncytium in health and disease. Single cell expression profiles can inform about cell lineage heterogeneity [42], cell-cell communications [43], individual transcription factor and pathway activity levels [44], or can be integrated within multi-omic approaches [45]. The main challenges in the application of this technology included separation of single, viable cells and subsequent amplification of a minute amount of RNA and the approaches to overcome these hurdles vary in gene coverage and multiplexing ability (i.e. the capacity to process in parallel) [46,47]. A widely used technology developed for this task is *10x Chromium,* which separates individual cells with a microfluidic approach and merges these droplets with gel beads in emulsion which are labeled with various oligonucleotides used for identifying unique transcripts and cells possibly samples (so-called demultiplexing). This system allows high throughput and reduces the need for sorting equipment or workflows that involve large numbers of assay plates and is used in this thesis in chapter II.
The cardiac tissue poses additional challenges. Since cardiomyocytes (CM) are too large for many cell sorting approaches, single nucleus RNAseq can be applied, which involves isolating the nucleus rather than the whole cell prior to sequencing. The transcriptional profile of single cell and single nucleus RNAseq has been reported to be comparable during CM differentiation [48]. As the transcriptional profiles of mono- and polynucleated CM were reported to be similar [49], application of single nucleus RNAseq on cardiac tissue is encouraged.

After the sequencing of transcripts, count matrices are obtained with the dimensions of number of barcodes (which usually represent cells) and number of transcripts (which represent genes or genomic regions of interest). Quality control steps then attempt to identify the barcodes that relate to viable cells. This is done via estimating cut-offs for count depth, feature number (low values are expected to be empty droplets, while high values can origin from doublets), and by assessing evidence for cell lysis (i.e. tissue dissociation patterns, unexpected mitochondrial or ribosomal gene expression). After barcode filtering, the high-dimensional count matrix can be further reduced by feature selection, removing low expressed genes. For many downstream analyses, only highly variable genes are considered which helps to improve signal to noise ratio. In general, these processing steps are context dependent and should consider the data and tissue at hand in an often iterative process to ensure the careful removal of non-plausible data without dispensing biological signal. After these filtering steps normalization procedures are applied, which scale count data to correct for differences in count depths between cells and further provide approximations of normal distributions of gene expression often via log transformation.

Gene expression is subjected to tissue handling and other technical factors which can introduce strong batch effects. To compare multiple samples, species or data sets, uncorrected normalized data is subjected to these batch effects and prevents joint clustering and annotation. Thus, single cell integration methods are used to overcome these challenges with the aim to remove technical and keep biological effects while considering differences in cell type composition, ultimately resulting in a joint feature space that enables meaningful distance calculations[50]. These methods provide different degrees of flexibility for batch correction and should be chosen regarding the expected strength of batch effect (e.g. sample integration < technology integration < species integration) [51].

The integrated space allows for the joint analysis of samples. This typically involves non-linear dimensionality reduction approaches which exchange distance interpretability for visualization purposes. Thus, clustering is not performed in the reduced dimensions but in the high-dimensional gene expression space, often by making use of local similarity in community graphs to identify clusters of similar cells. Then, identification of cluster markers via differential gene expression analysis is used to annotate clusters by assigning cell lineage based on prior biological knowledge or trusted annotated single cell data. Cell lineage is typically a major source of gene expression variation and thus single cell analysis often directs functional analysis by considering and comparing distinct cell lineages. Here, cluster resolution affects granularity of this assignment and sparked lively debate about cell identities and sub-cellular clusters, often called cell states [52,53].

In case control studies, the cellular profiles are typically processed together to achieve an integrated space of all samples. Then, disease associated patterns are often analyzed via i) compositional analysis of cell clusters or ii) by comparing molecular profiles to identify transcriptional shifts [54]. This enables the quantification of cellular contributions to tissue remodeling as well as prioritization of cell lineages. To interpret resulting gene expression patterns, similar strategies as in bulk transcriptomics can be applied [44].

In general, the plethora of information gathered by single cell RNAseq poses new challenges to data analysis that have only partially been met. These include the need to quantify uncertainty in measurements and efficiently handle gene dropout rates; the limited benchmarking possibilities; the need to scale to higher dimensional data, as more cells can be measured; and the integration of multiple levels of single cell omics [55].

Single cell RNAseq has already been applied to study the cardiovascular system (reviewed in [56–60]). To date, studies have focused on the description of cardiac cell lineage heterogeneity and trajectory in mice [58,61–63], as well as on human cardiogenesis [64–68] and more recently on profiling human hearts in health [69] and disease [70,71]. However, the description of HFpEF on single cell level is missing to date.

Clinical HF profiles

Clinical data and omic data can be analyzed by similar methods, however they differ regarding their data structure. While omic data are structured measurements, clinical data is often a combination of unstructured, semi-structured, and structured data with the added complication that free text can be subjective or spurious. Thus, clinical data often requires significant pre-processing prior to analysis, a major hurdle for clinical data analysis on a large scale. Highly promising approaches to this challenge of extracting relevant information from unstructured clinical data include natural language processing [72–74], but even structured clinical data is subject to noise resulting from entry errors. Clinical data is frequently sparse, subject to care utilization and documentation habits, and biased, in that health states outside of clinical encounters are rarely reported. Once preprocessing challenges are overcome, clinical data analysis is often subjected to similar statistical and mathematical modeling as omics data for predictive or inference purposes. In HF, patient outcomes have been associated with the presence of a wide variety of comorbid conditions and ejection fraction sub-group. Despite this, mortality and risk of rehospitalization in HF patients remains high. As a result, two major trends have emerged in the use of clinical data for the study of HF: sub-phenotyping and deep phenotyping.

The emergence of sub-phenotyping has caused a shift from the tendency to view HF patients as a single population (or as two clearly defined populations) towards the tendency to view them as a large heterogeneous supergroup composed of many smaller and potentially unknown subgroups [75,76]. Predicting the outcomes of HF patients, especially within subgroups, is a major area within big data studies using electronic health record (EHR) data or other data relevant to clinical care [77]. Adler et al. were able to divide HF patients into those at high and low risk of death based on clinical variables, and their classifier had a better predictive power than any of the individual classifier components, and better than other comparison markers including NT-proBNP [78]. Ahmad et al. divided a group of HF patients into four clusters which differed in age, sex, clinical measures, and comorbid conditions, before building a classifier to predict survival. They found that cluster membership had a modest predictive ability, but performed better than left ventricular ejection fraction alone as the gold standard measure of cardiac function [79]. Other studies have tested multiple types of algorithms for predicting outcomes including HF hospitalization and mortality amongst HFpEF patients [80], and phenogrouped HF patients who had been randomized to cardiac resynchronization therapy with a regular or implantable cardiac defibrillator prior to evaluation of the effect on HF events and death [81].

Deep phenotyping- the characterization of a phenotype through the comprehensive evaluation of components and intermediate manifestations- has resulted in the use of many diverse types of data. Data including echocardiography [82], electrocardiography[83], cardiac magnetic resonance imaging[84], tissue imaging [85], implantable monitors [81], and other wearable and non-invasive cardiac monitors [86,87] are used in combination with machine learning methods for the event prediction and monitoring of HF patients. Laboratory values and intermediate phenotypes are also widely analyzed. The diversity of data types used for the study of HF is rapidly expanding. Analyzing populations that have multiple data in the same individuals can provide detailed information about the progression of disease as well as insights into clinical characteristics that may indicate negative outcomes.

As a whole, the recent years of large clinical data analysis has provided great insight into the true phenotypic diversity of HF and has begun to provide links between that diversity and patient outcomes. However, despite the increased understanding of phenotypic heterogeneity, there is still a significant amount to be learned about the relationship between sub-phenotypes and outcomes. While this is a rapidly expanding field, questions about the necessary manual curation of certain data types, inconsistencies in imaging between clinical sites, and privacy concerns remain.

Comorbidity studies

The disease profile, i.e. the enumeration of diagnosed diseases of a patient, is an important characteristic that is fundamental for patient phenotyping and cohort selections. Often an index disease is studied with respect to the accompanying disease profiles which can be called comorbidities. The concept of comorbidities is subject to debate and the term is in clinical context often used interchangeably with multimorbidity. One possible conceptual distinction is that comorbidities are co-occurring more frequently than expected by chance and thus dependent while multimorbidity is the co-existence of independent diseases [88]. However, both concepts assume an unambiguous disease classification that is able to meaningfully separate diseases, syndromes and symptoms, which is often challenging considering the continuum of pathologies and the evolving clinical and societal perspective on some disorders [89].

The study of comorbidities is thus often of epidemiological character, where disease co-occurrences are statistically assessed in patient cohorts. These associations are then interpreted and further analyzed for possible reasons of disease dependency that can lead to improved clinical assessment or pathophysiologic understanding of the index disease. In the case of HF, patients typically suffer from a wide range of comorbidities, which are considered important for HF development and progression [90]. In the pathogenesis of HFpEF, comorbidities have been suggested as causal factors [7,13] and could possibly be linked to genetic etiology. Treatment of comorbidity has also been shown to have beneficial effects of cardiac physiology [91], emphasizing the potential to address HF subtypes through their comorbidities.

Systems medicine attempts to model disease in a holistic manner. One facet of this, network medicine, is used to analyze complex systems such as patients, organs or cells via network representation [92,93]. Comorbidity networks represent diseases as nodes, connected via edges based on co-occurrence in patients. If these networks become large, they cannot be visualized or inspected by eye, thus a toolbox of graph theory is usually applied to summarize network topology or to answer specific research questions that have to be translated into the network perspective. E.g. these networks can be used to i) define disease modules i.e. clusters of co-occurring diseases ii) to assess network based distance to describe node relationships or iii) assess the influence of a node in a network which is often estimated by a nodes centrality or capacity to influence network structure. These concepts have been applied to study comorbidity patterns in patients [94–97] including cardiovascular cohorts [98]. However, these network topologies are subject to many possible biases in the data as well as technical factors of the assessment and statistical evaluation of comorbidity [99]. A highly influential factor is the disease ontology which, as discussed above, lies at the very heart of studying comorbidities [100].

Another aspect of studying comorbidity networks is the coupling to multi-layer gene networks. Previous work has shown that disease comorbidity is also often linked to shared disease genes that locate close together in gene-based networks like protein-protein interaction networks [97,101]. This observation is often the basis of network-based gene prediction, where novel disease genes are predicted based on network proximity to known disease genes.

Summary

HF represents a huge health burden with currently no curative therapeutic strategies. Study of HF patient heterogeneity has led to the recognition of HFpEF and HFrEF as distinct syndromes regarding molecular and clinical characteristics. HF research has been typically focused on HFrEF in the past century which is reflected by the arsenal of pharmacotherapy targeting compensatory mechanisms that explain our model of understanding of HFrEF but does not fully apply to HFpEF. Thus, novel approaches are necessary to study the mechanisms driving and distinguishing the HFpEF syndrome.
HF induces a series of ventricular modeling processes that involve hypertrophy, fibrosis, inflammation, conductive coupling, all which can be observed to some extent in HFpEF and HFrEF. Thus by direct comparative analysis between HFpEF and HFrEF, more distinctive features could be uncovered leading to improved pathophysiological understanding and opportunities for therapeutic intervention.

This thesis evaluated HF by considering and comparing HFpEF and HFrEF syndromes on molecular and clinical level (Figure I.4). In **chapter I**, I review existing transcriptomic data sets and compile new state-of-the-art knowledge for consensus transcriptomic changes in HF. I discuss the patient clientele that is captured, revealing misrepresented HF patient collectives ethnically but also phenotypically, which identifies a knowledge gap for the molecular landscape in HFpEF ventricular remodeling. HFpEF patient biopsies are difficult to obtain, thus HFpEF can be addressed via study models. In **chapter II**, I utilized a mouse model that resembles HFpEF and analyzed single cell transcriptome data to gain insights into fibroblast activation. I contrast this analysis by comparison with mouse models resembling HFrEF and use the human reference to demonstrate similarities between models. Furthermore, a possible novel biomarker for HFpEF is introduced. As mouse models only capture selected aspects of the HF syndrome, I next performed a top-down analysis in human HF patients in **chapter III.** By phenotyping HFpEF and HFrEF patients, I extract and describe distinctive comorbidity profiles, yielding novel HFpEF characteristics. These profiles are then used to predict associated recurrent gene candidates, which could be linked to the HFpEF mouse model, providing hypotheses about the molecular links of comorbidity profiles.

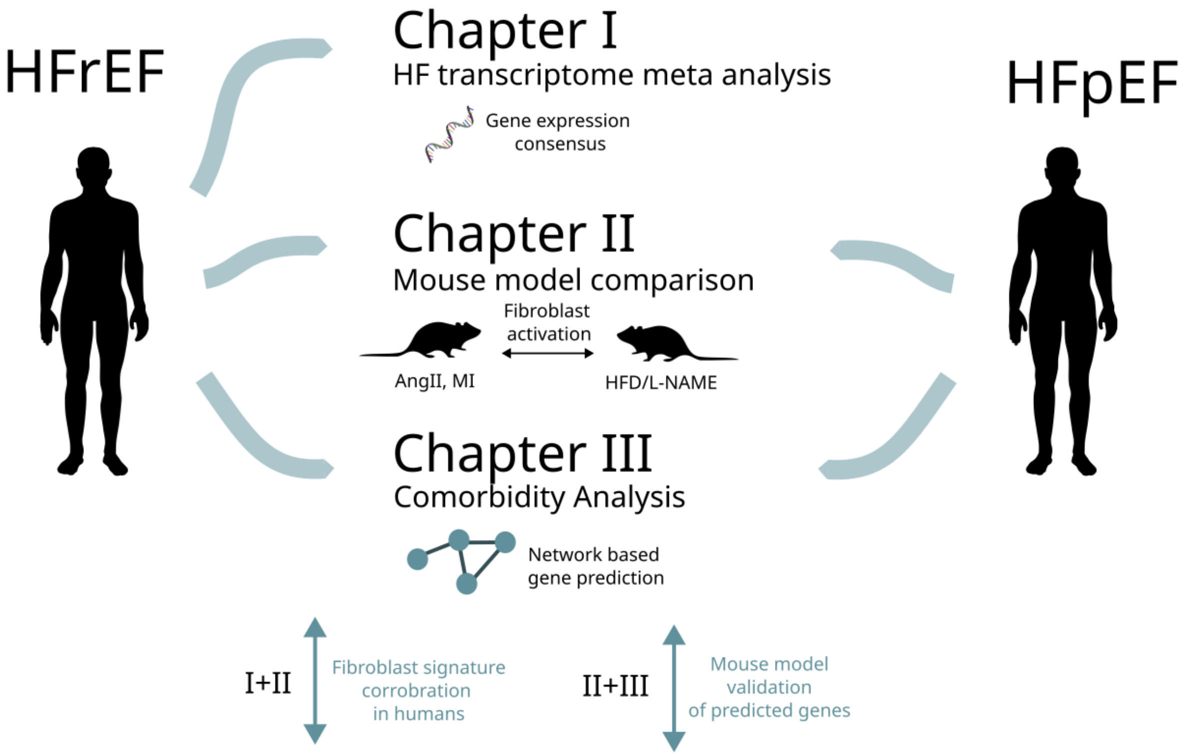

HFrEF
Chapter I
HF transcriptome meta analysis
Gene expression consensus
HFpEF
Chapter II
Mouse model comparison
Fibroblast activation
AngII, MI
HFD/L-NAME
Chapter III
Comorbidity Analysis
Network based gene prediction
I+II
Fibroblast signature corrobration in humans
II+III
Mouse model validation of predicted genes

Figure I.4 Overview of how this thesis addresses HFrEF and HFpEF.

Graphical overview of how this thesis addresses the two HF subtypes in three chapters. Chapter one is reviewing existing HF transcriptome data which turns out is typically HFrEF. Chapter II and III address HFpEF by comparative analysis to HFrEF. Chapter II is a scRNAseq data comparison of mouse models that resemble HFrEF or HFpEF, while Chapter III analyzes comorbidity profiles of HF patients to detect different patterns in comorbidity profiles which are also translated to associated genes. Bottom vertical arrows indicate important cross chapter analyses. AngII, angiotensin II; MI, myocardial infarction.

Chapter I - A consensus transcriptional landscape of human heart failure

1.1 Background

In this chapter, the state of the art of human heart failure bulk transcriptome research is assessed. While targeted gene expression studies have been conducted since the advent of polymerase chain reaction, high throughput approaches allowed for parallel quantification of thousands of transcripts. The first high-throughput transcriptomic study on myocardial human HF was published in 2000, to the best of my knowledge [31]. In the subsequent decades, technological and bioinformatic advances in transcriptomics have improved our comprehension of cardiac hypertrophy [32], reverse remodeling [33], cardiac metabolism [34,35], cardiac fibrosis [36], and immune dysregulation [37] in HF. Several studies made their data sets and protocols publicly available on platforms like NCBI's gene expression omnibus. However, few attempts have been made to compare transcriptomic HF studies [38–40]. The existing studies mainly relied on comparing intersects of differentially expressed genes. They lack an assessment of general comparability of gene expression patterns as well as providing results in an accessible and unified database. Moreover, the continuing development of sophisticated data analysis methods invites the retrospective re-analysis and integration of published HF studies, although data integration from different platforms, centers and technologies presents many challenges [41].

20 years of sampling myocardial gene expression in heart failure patients will be reviewed by comprehensively curating and integrating existing data sets. This work was a joint project with Ricardo Omar Ramirez-Flores. While disentangling collaborative work is a challenge, I will refer specifically to the parts of the project that have been my contribution by using the first person, highlighting my responsibilities of developing a clinical concept of this project, the cohort descriptions and curations, processing of RNAseq data, comparative analysis between studies as well as interpretation of gene expression patterns. Furthermore, in each figure legend a contribution statement is added. This chapter has been published in a peer reviewed journal [102]. The article was jointly written by Ricardo Omar Ramirez Flroes and me with all authors providing minor edits. This work was supervised by Rebecca T. Levinson and Julio Saez-Rodriguez.

1.2 Curation and review of data sets

1.2.1 Inclusion criteria

The first aim of this project was to identify and curate relevant HF transcriptome studies. For this purpose I defined i) a set of key words to query databases and ii) inclusion criteria to ensure technical and clinical comparability.

There are different databases that publicly host gene expression data, often with redundant data entries. I queried three distinct data bases: NCBI's Gene Expression Omnibus database (GEO), the European Nucleotide Archive (ENA) and ArrayExpress. I decided to use the following keywords: "heart failure", "ischemic cardiomyopathy", "dilated cardiomyopathy", "cardiac failure" and "heart disease", to capture a rather broad spectrum of all HF associated studies. Then, I manually reviewed all matching data entries from these studies and further applied the following inclusion criteria. I will spell out the criteria and briefly explain their rationale.

1. **case samples came from biopsies of the left cardiac ventricle**

HF patients were also often sampled from the right heart, especially in catheter based interventional approaches. However, the right and left ventricle display unique gene expression profiles that would not justify a joint analysis [103]. Another frequent biopsy location was found to be the sampling of peripheral blood mononuclear cells, which we did not include for the same reason. This criteria thus ensured a comparability of left ventricular remodeling.

2. **case samples of the human heart of end stage HF patients with either ischemic cardiomyopathy (ICM) or dilated cardiomyopathy (DCM)**

These criteria ensured that the two main etiological branches of chronic heart failure were included, while more rare HF etiologies such as infectious or inherited HF were not considered.

3. **control samples were obtained from patients with non-failing heart**

As I expected strong batch effects to be associated with each study, I only considered studies that included control samples.

4. **data from at least 5 samples were available**

Smaller sample size can complicate statistical reevaluation.

5. **a publication or preprint with a detailed methodology was available**

I relied only on data sets where a detailed methodology section enabled me to comprehend experimental protocol for data generation.

1.2.2 Curated studies and compared metadata

A total of 16 data sets were retained, after manually evaluating studies for the discussed inclusion criteria (Table 1). These 16 data sets contained 263 control, 372 DCM and 281 ICM samples (Figure 1B). The studies were published between 2005 and 2019 and their sizes varied between 5 and 313 samples. The comparison of the country of study origin further revealed a research bias: Ten cohorts were from the USA, while the remaining six cohorts were from Europe (Spain, Germany, Italy).

#	Study ID	GEO ID	Samples (CT + HF)	Technology	Year	Country	Disease	Citation
1	Liu15_M	GSE57345	313	Microarray	2015	USA	ICM, DCM	[35]
2	Hannenhalli06	GSE5406	210	Microarray	2006	USA	ICM, DCM	[104]
3	vanHeesch19	not in GEO	77	RNAseq	2019	Germany	DCM	[105]
4	Sweet18	GSE116250	64	RNAseq	2018	USA	ICM, DCM	[106]
5	Kittleson05	GSE1869	37	Microarray	2005	USA	ICM, DCM	[107]
6	Tarazon14	GSE55296	35	RNAseq	2014	Spain	ICM,DCM	[108]
7	Spurrell19	GSE126573	33	RNAseq	2019	USA	DCM	[109]
8	Kong10	GSE16499	30	Microarray	2010	USA	ICM	[110]
9	Molina-Navarro	GSE42955	29	Microarray	2013	Spain	ICM, DCM	[111]

	13							
10	Greco12	GSE26887	24	Microarray	2012	Italy	DCM	[112]
11	Yang14	GSE46224	24	RNAseq	2014	USA	ICM, DCM	[113]
12	Barth06	GSE3585	12	Microarray	2006	Germany	DCM	[114]
13	Pepin19	GSE123976	9	RNAseq	2019	USA	ICM, DCM	[115]
14	Kim16	GSE76701	8	Microarray	2016	USA	ICM	[116]
15	Schiano17	GSE71613	6	RNAseq	2017	Italy	DCM	[117]
16	Liu15_R	GSE57344	5	RNAseq	2015	USA	DCM	[35]

Table 1.1 Overview of curated studies.

Overview of meta-analyzed studies. This data was curated and represented by me. *Reprinted from* [102].

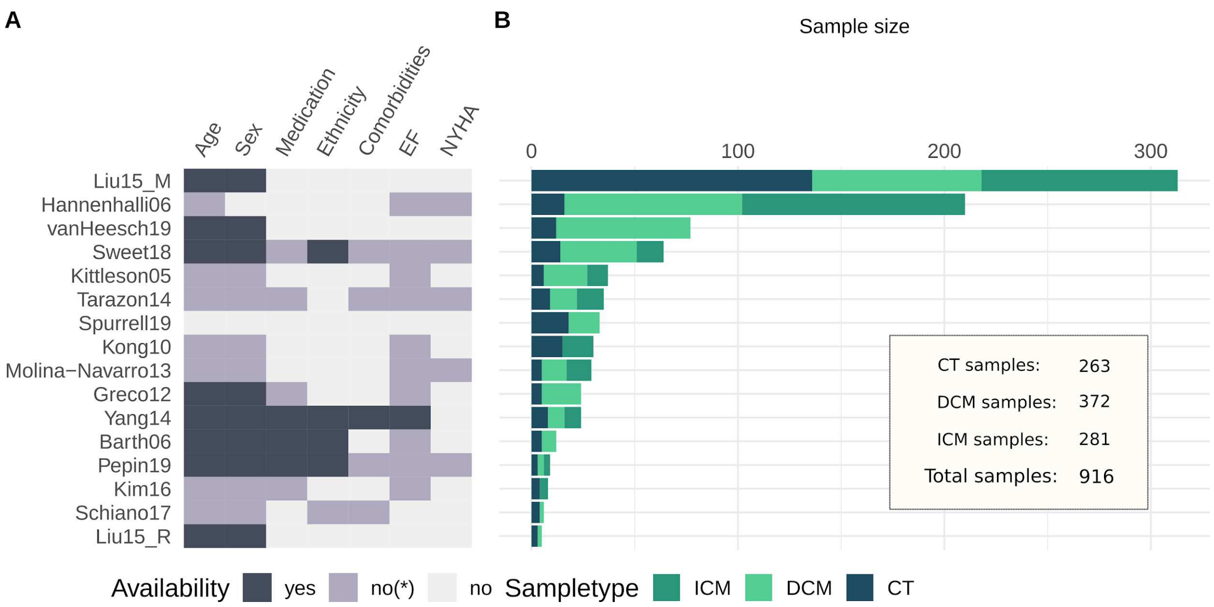


Figure 1.1 Curation results of HF transcriptome studies.

A) Sample information availability per study. yes, information available per sample; no*, incomplete information or only summary statistics; no, no information available. B) Sample size comparison of studies. CT, Control; DCM, dilatative cardiomyopathy; ICM, ischemic cardiomyopathy. This data was curated and represented by me. *Reprinted from* [102].

I next assessed how clinical and technical information was available. The databases that store gene expression data provide meta information stored by the author of the data. However, there is no consensus on i) which sample features should be recorded and ii) how sample features should be formatted. My strategy for gathering as much information as possible was first to define features of interest for our study. This included clinical & demographic patient features as well as technical features of sample processing. Then, I processed and unified all metadata stored in the gene expression databases. To increase coverage, I next assessed for each study whether cohort information was stored within the published manuscript. Here, I often found only summary statistics of cohorts which I extracted. The results regarding availability of demographic (age & sex) as well as clinically relevant features (Comorbidities, EF, Medication, NYHA) is displayed in Figure 1.1A. All studies but one reported at least some information of age and sex, while only eight provided complete sample level information. For clinical features the sparsity of available information rendered it almost useless for patient cohort comparison. This heterogeneity of reporting sample information that I discovered constitutes a major impediment for data sharing and reutilization. Thus a clear need for the standardization for data sharing is evident. Furthermore, I found multiple relevant studies that did not publish their data which prevented its full exploitation by the research community. Since many studies are funded one way or another through government initiatives and taxes, data sharing should become non-optional, if patient privacy rights are not in danger of being violated. Furthermore the standardization for sample information deposition is necessary.

Since age and sex were reported by most studies, I was able to compare patient demographics between studies. Here I found that most patients were male in (Figure 1.2) revealing another bias in the field of HF transcriptome research. Furthermore, patients were younger than expected, as HF incidence increases with age. It can be speculated that the young age in these cohorts is related to sample availability of HF patients: In general, myocardial biopsies of the human heart are difficult to obtain. Usually, a clear clinical indication is necessary to justify the probing of myocardium which causes complications in ~1 % of patients. Ventricular rupture is a feared complication that necessitates further surgical treatment but occurs rarely (0.42%) [118]. Thus, purely scientifically motivated biopsies were not performed and clinically motivated opportunities for sampling included left ventricular device implantation, surgical ventricular restoration or heart transplantation. What is common to these indications is the last line of treatment character. Patients went typically through stepwise escalating heart failure pharmaceutical therapy before being subjected to a surgical or endovascular interventional therapy. For control biopsies, all studies relied on donor hearts that could not be transplanted due to reasons of size disparity or AB0 mismatches. This bottleneck of sample availability prevents randomized patient sampling and thus, results have to be considered with caution.

When comparing technical protocols of sample acquisition and handling, I also found that many studies did not provide complete information. While all studies described sampling location as left ventricle (see inclusion criteria #1), the exact location was not described by six studies. The other studies sampled apex (four studies) or free wall (five studies) or both (one study). Information on tissue storage was often incomplete as well. While all studies snap froze samples to -80C°, intermediate storage was often done at 4°C for often unknown duration. Gene expression is highly sensitive to tissue handling and thus, batch effects due to technical circumstances can have a high impact on analysis results.

In summary, HF patients represented in transcriptome studies were typically male, from western countries, and suffered from severe HF at a relatively young age of 50-60 years justifying clinical interventions and enabling sample acquisition. These are important cohort characteristics that might not generalize well to patients with a different profile.

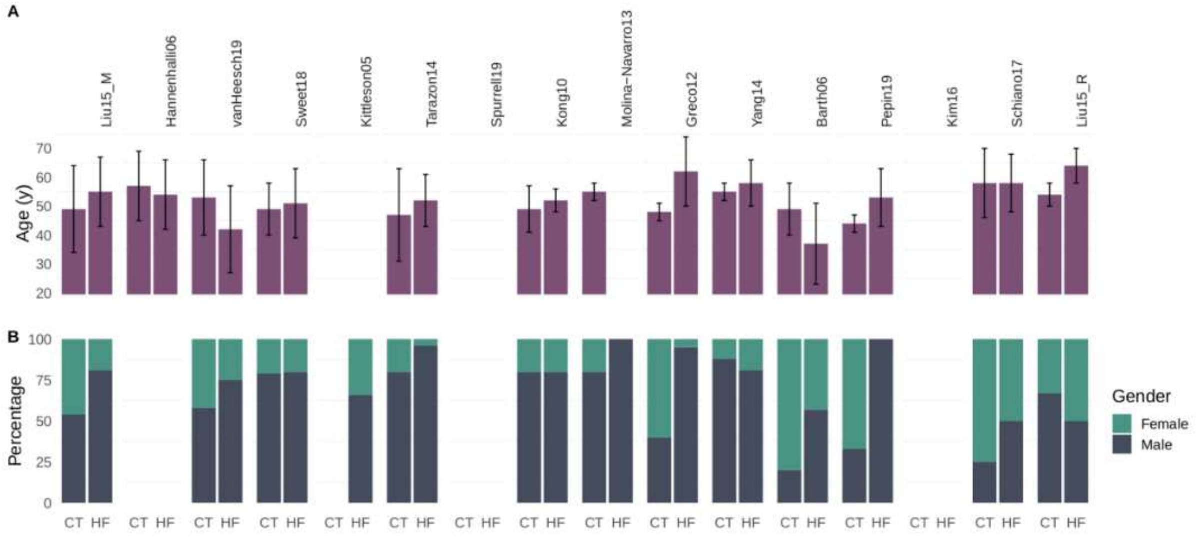


Figure 1.2 Comparison of age and sex distribution.

A) Age distribution in years of control (CT) and heart failure samples (HF) per study. Displayed is mean and standard deviation. B) Sex of patients in % per study. This data was curated and represented by me. *Reprinted from* [102].

1.2.3 Assessing technical variation in gene expression between studies

While differences in experimental study protocols can induce study specific gene expression patterns, different computational pipelines can have similar effects. For this reason, we downloaded raw expression data and processed all studies uniformly. As mentioned in the introduction, there are two different technologies to quantify gene expression, RNA sequencing and microarray.

For microarray data we read CEL files and normalized them using Robust Multi-array Average (RMA). The gene probes were then annotated by using the HUGO Gene Nomenclature Committee (HGNC) gene symbols for each platform. For RNA-Seq studies, I downloaded FASTQ files and aligned sequence reads to the human GRCh38 reference by using Kallisto implemented in the ARCHS4 pipeline. For convenience I utilized the BioJupies platform to run this pipeline. We only regarded protein coding genes to increase the comparability of microarray with RNAseq data. Gene expression counts were then filtered for low expressed genes and multiple measures of the same gene symbol summarized by calculating the mean. I normalized samples using the *Trimmed mean of M-values* (*edgeR* [119]) and subsequent variance-stabilizing transformation with voom (*limma* [120]).

After these processing steps, study quality was assessed by visually comparing the distribution of gene expression values for all samples. Multidimensional scaling was performed for each study separately and satisfying separability of heart failure and control samples was observed. I compared gene coverage and found that ~14k genes were reported by at least 10 studies while the mean gene coverage was ~16k. Notably microarray studies did not yield inferior coverage (Figure 1.3)

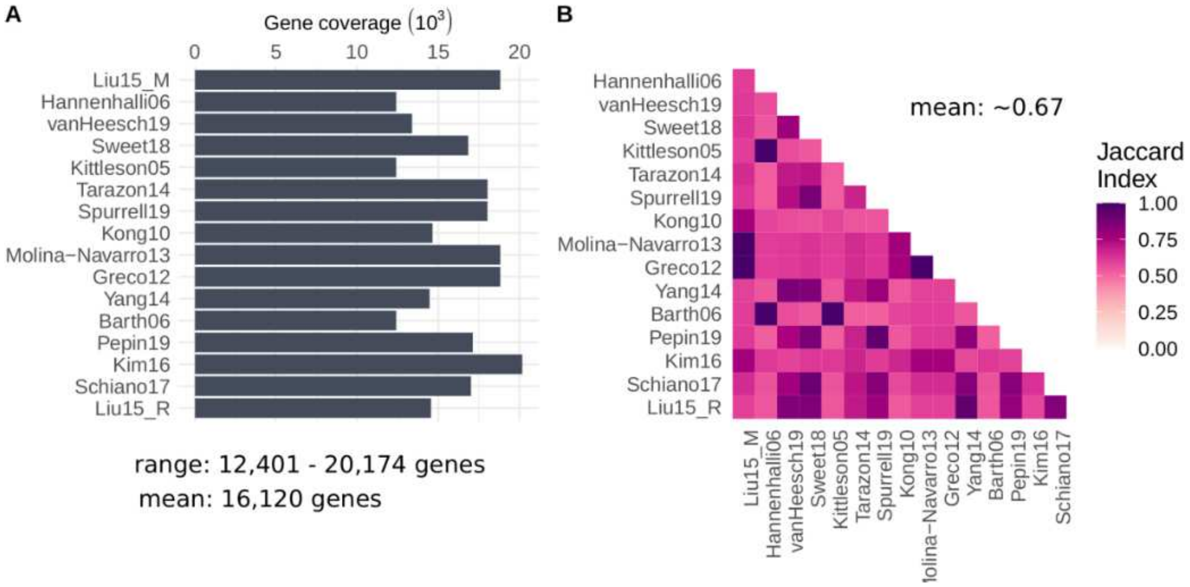


Figure 1.3 Comparison of gene coverage between studies.

A) Absolute gene coverage per study after processing. B) Pairwise comparison of covered genes measured with Jaccard Index. This data was analyzed by me. Reprinted from [102].

Next, we estimate the variance in association to study labels (i.e. study batch effect), HF, and HF etiology. For this we applied Principal Component Analysis (PCA) and Analyses of Variance (ANOVA) to various transformations.

First, we combined all normalized gene expression data and performed PCA (Figure 1.4A). We then summed up the explained variance by PC if they were associated with study labels (ANOVA, $p< 0.05$). As expected, the joint analysis of gene expression data from different studies results in mainly gene expression variance driven by study batch effects (85%), thus preventing this simple analysis approach.

Second, we tested whether simple z-transformation can ameliorate batch effects (Figure 1.4B). For this calculated mean and standard deviations per gene in control samples and transformed gene expression in HF samples to z-values. We repeated the PCA analysis and found that the estimated variance associated to study batch decreased to 74%. I used tsne to demonstrate the persistence of the batch effect (Figure 1.4C)

Third, we performed gene standardization. For this, we standardized gene expression for control and HF samples per study (mean= 0, sd= 1). After performing PCA, the explained variance by study decreased to 0, suggesting that the scale of gene expression is relevant for the study batch effects (data not shown).

After estimating study batch effect, we next addressed the association of the clinical covariates with variance in gene expression. For this we used PCA for each study independently and tested with linear models ($p < 0.05$) for association of each covariate . We performed this analysis for all patients (Figure 1.5A) and for only HF patients (Figure 1.5B). Since the availability of sample level information was poor (see section 1.2.2), we could not draw a general conclusion. However, we observed that HF status together with age and sex explained most of the variance per study. When comparing HF patients, we observed that HF etiology (ICM vs DCM) was associated with very little variance (6.98%). Of note, our approach underestimated the variance in studies with low sample size, since the variance is distributed on few principal components. However, as the sample sizes of studies were balanced we expect this effect to be ameliorated.

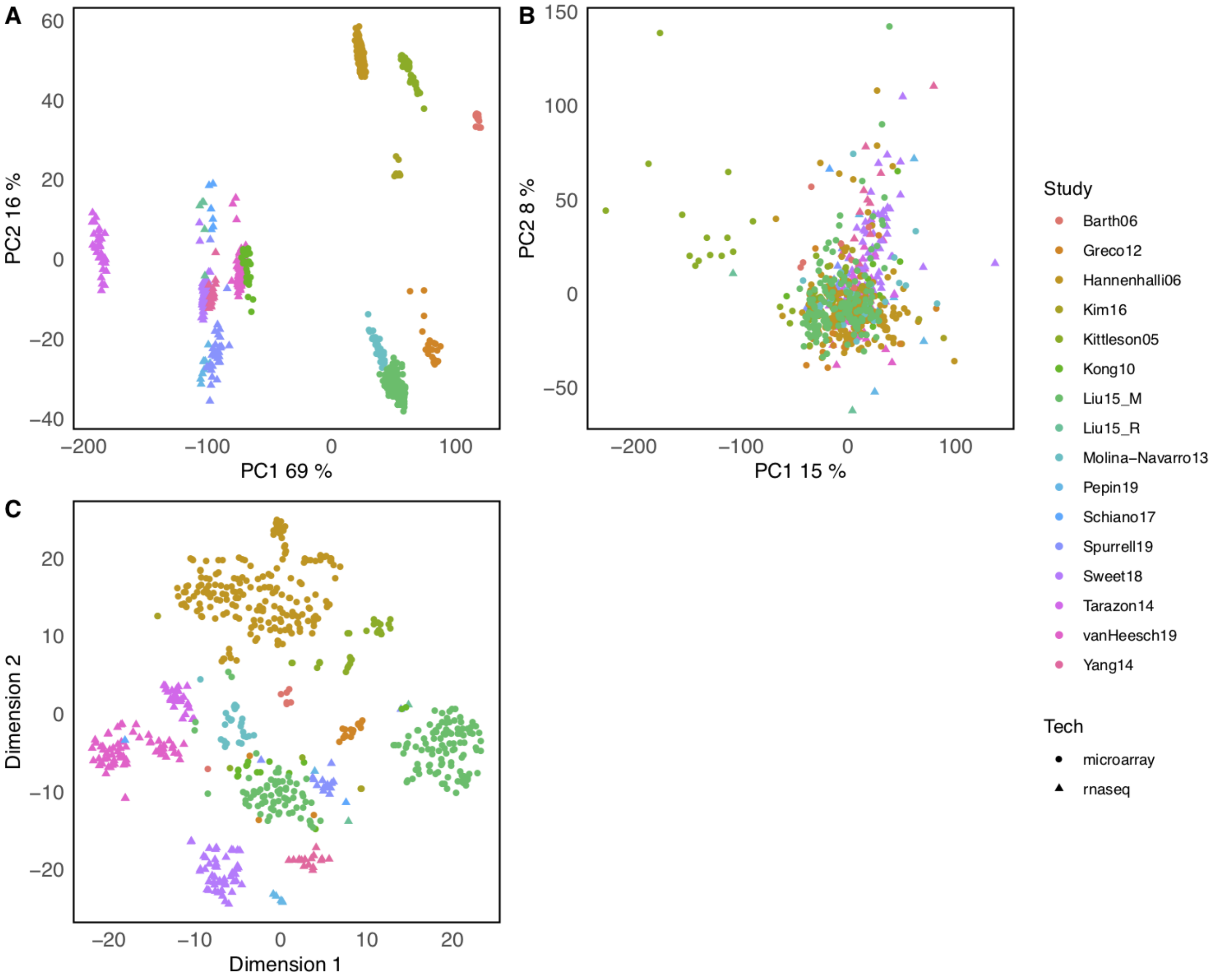


Figure 1.4 Differences in samples included in the study.

A) First two components from a Principal Component Analysis (PCA) done to all samples. B) First two components from a PCA done to all z-transformed heart failure samples. C) t-distributed stochastic neighbor embedding of all z-transformed heart failure samples. This data was analyzed by Ricardo Ramirez Flores (panel A, B) and me (panel C). *Reprinted from* [102].

In conclusion, this series of analysis demonstrated that i) gene expression magnitude is subjected to strong study batch effects and ii) clinical covariates are important drivers of gene expression. Thus the reporting of as many possible observed confounders is necessary to decipher gene expression, especially when patient cohorts number is small and cohorts were not randomized (see section 1.2.2).

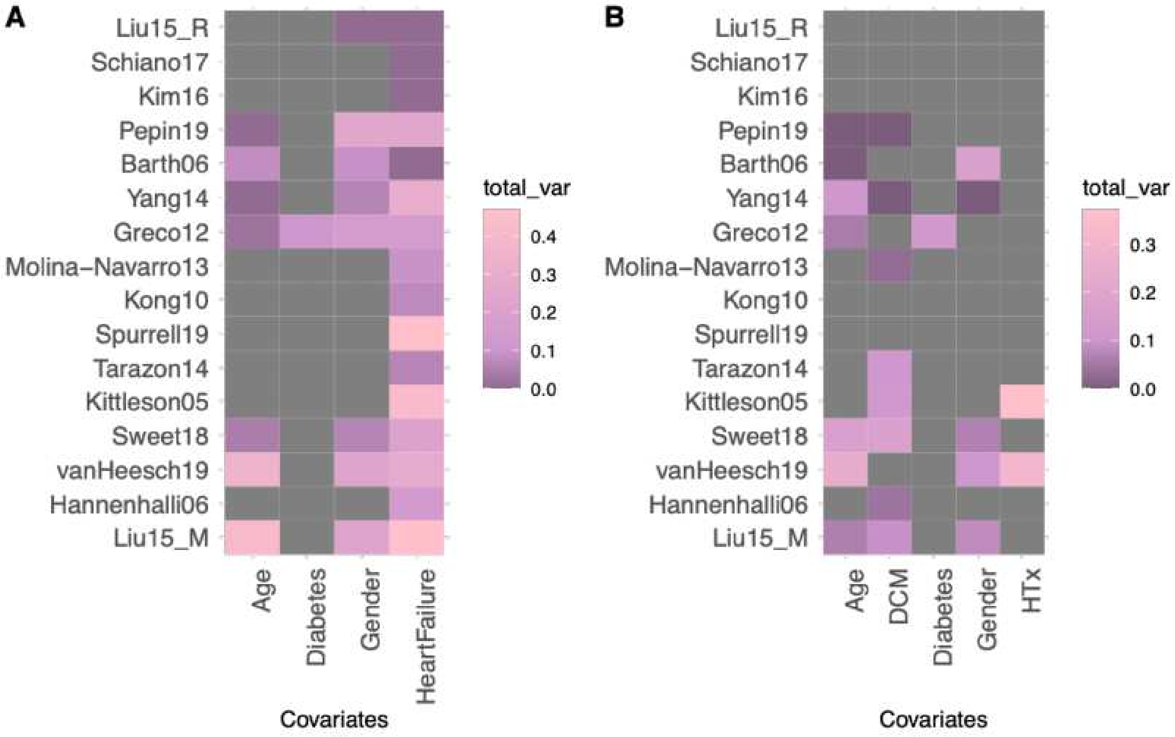


Figure 1.5 Contribution of the covariates to the variability of individual studies.
Estimated proportion of explained variance associated with the different covariates used in the differential expression analysis in A) all patients and B) only heart failure patients. Grey tiles represent missing reported data. HTx, heart transplantation. This data was analyzed by Ricardo Ramirez-Flores. *Reprinted from* [102].

1.2.4 Evaluating consistency of gene expression changes in HF

We have established that HF is associated with variance in gene expression in most studies. Thus we posed the question, whether gene expression changes reported by each study were consistent. To estimate the gene expression changes in HF per study, we applied linear models as implemented in the limma package by controlling for clinical covariates if available. This yielded gene-level statistics of p-value, t-value and effect size (log fold change). We developed a series of different analyses that each yielded a different perspective on the question of consistency in HF gene expression changes.

Results of differential gene expression analysis are often reported by choosing an alpha level and reporting the resulting gene by disregarding those that do not reach significance. However, applying the same alpha level to all studies would yield a very different number of differentially expressed genes (DEGs) for the differences in sample size. Thus we ranked genes by p-value and compared the top 500 genes among studies by calculating Jaccard indices (Figure 1.6A). This yielded an almost null concordance of DEGs (mean Jaccard index = 0.05), suggesting that each study reported different top DEGs. This was in fact the strategy of previous reports of HF transcriptome meta analysis [38].

This analysis might suggest that different HF related expression patterns were reported in each study. We next asked whether the top 500 genes reported from one study can separate HF from control patients in another study. For this we calculated a disease score for a given set of genes by multiplying the t-values reported by one study with the sample level expression values by another study. For each sample we summed up the values for all genes, and then estimated differences between HF and control samples by calculating the area under the receiver operator curve (AUROC) (Figure 1.6B). This yielded high separability in all pairwise comparisons (median AUROC = 0.94). Studies that profiled only patients with ischemic forms of heart failure (eg, Kong10) effectively classified studies that profiled only patients with dilated cardiomyopathy (eg, Spurrell19) (AUROC, 1) and vice versa (AUROC, 0.95). We observed no association between each study's mean AUROC and their technology (Wilcoxon test, p-value = 0.72), sample size, or estimated proportion of variance captured by HF (Pearson correlation, 0.17, 0.18, respectively p-value > 0.4. These results indicate that patterns of coexpression of genes are more stable between cohorts than substantial changes in expression of specific genes. Thus, while each study reported different DEGs, the direction of expression was conserved in other studies. This indicates that trends of transcriptional regulation are more stable than the ranking of top marker genes.

Now that conservation of transcriptional regulation has been confirmed, I asked whether the ranking of DEGs also was conserved. For this I applied fast gene set enrichment analysis (fgsea) [121] as a pairwise study comparison. Here, I selected the top 500 up- and downregulated DEGs from one study and enriched them separately in the t-value gene ranking of another study. This yielded mostly consistent results, Differentially up-regulated and down-regulated genes had a median enrichment score of 0.55 (Figure 1.6C, upper panel) and -0.56 (Figure 1.6C, lower panel), respectively. In GSEA the enrichment score is punished for genes in the top of the ranking that are not in the tested gene set. Thus this analysis provided additional insight as we know that there is little overlap between the top 500 genes, however, that considering the full ranking of genes the DEGs of one study still tend to be in the top of the ranking in the other.

To summarize, these results suggest that the proper way to combine the evidence of the curated studies is by looking at the consistency of deregulation of genes and not at the dimension of the change in expression.

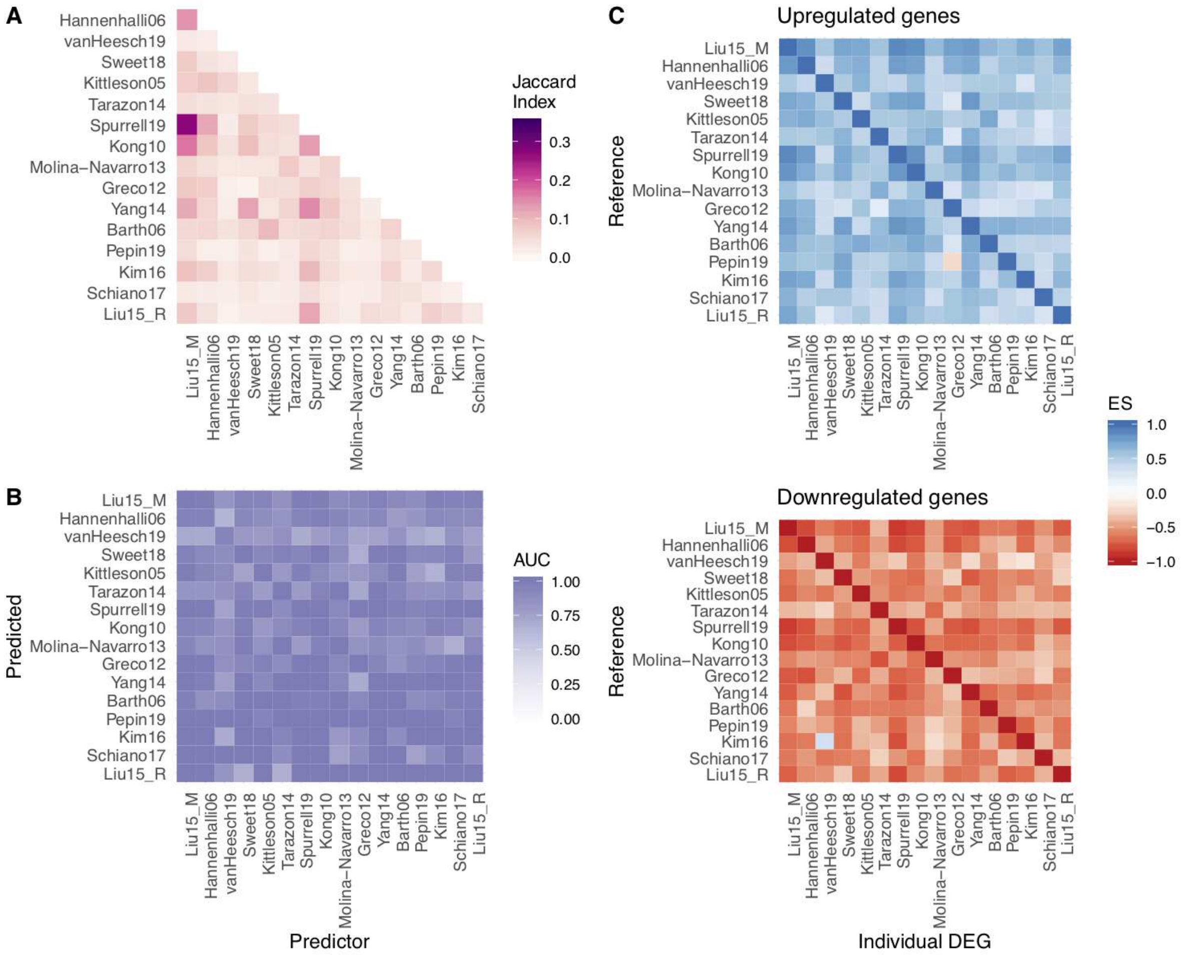


Figure 1.6 Consistency of the transcriptional signal of end-stage HF among studies.

A) Pairwise-comparison of the top 500 differentially expressed genes of each study using the Jaccard index. B) AUROC of pairwise predictions using a disease score with the top 500 differentially expressed genes of each study. C) Enrichment score (ES) of the top 500 differentially expressed of each study in sorted gene level statistics lists. This data was analyzed together with Ricardo Ramirez-Flores. *Reprinted from* [102].

1.3 Meta-analysis of the transcriptional responses in end-stage HF

1.3.1 Building the heart failure consensus signature (HF-CS)

After observing that directionality of gene regulation was conserved between studies, we next sought to collect the consensus of gene regulation in HF. For this analysis we selected the Fisher combined probability test. This test calculates a test statistic based on log transformed p-values of independent studies describing the same null hypothesis. In our case, a given gene was tested whether the difference between mean expression in HF and control patients is non zero for each study separately. The meta-test statistic is large when single p-values tend to be small. The null hypothesis of the meta-test is that all null hypotheses are true, while the alternative hypothesis is that at least one of the alternative hypotheses is true. After applying this test for each gene independently, meta p-values for each gene were corrected (Benjamini and Hochberg, BH). The ranking of genes based on this meta p-value then could be regarded as a ranking of the most consistently regulated genes in HF. We considered only genes which were reported in ten or more studies, resulting in the HF consensus signature (HF-CS) of 14,041 genes (Figure 1.7A). We found that the top ~500 genes displayed an elbow in the meta p-value distribution and thus contain the most conserved genes.

Since the individual study p-values were derived from two sided tests, a gene could receive a low meta p-value by displaying significant but inconsistent regulation within a study. Although we demonstrated earlier that the directionality is highly conserved between studies, we demonstrated the agreement of directionality of the top 500 HF-CS genes (Figure 1.7B).

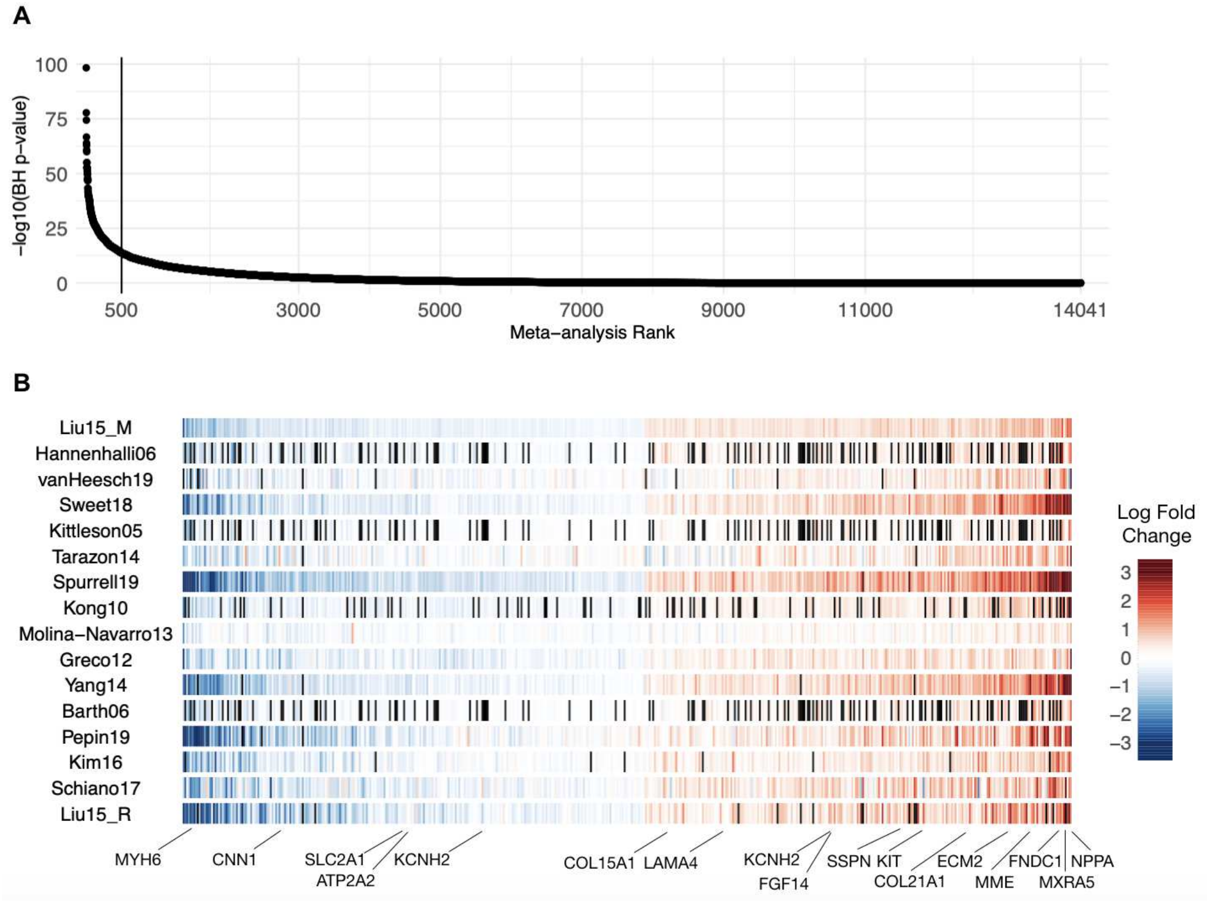


Figure 1.7 Meta-analysis summary.

A) Sorted −log10 (meta analysis BH P values) of the 14 041 genes included in the Fisher combined test, representing the heart failure consensus signature (HF-CS). B) Top 500 genes sorted by their mean log fold change across all studies; black lines represent genes that were not measured in specific studies. A selection of HF marker genes are highlighted. BH indicates Benjamini–Hochberg. This data was analyzed by Ricardo Ramirez-Flores. *Reprinted from* [102].

We found no correlation between the sample size of a study and its contribution to the meta-analysis (Spearman correlation 0.24, p-value = 0.37), suggesting that proper experimental design and representative sampling could compensate for study size [122]. The consensus ranking captured known HF markers such as MYH6, MYH7, MME, CNN1, NPPA, NPPB, KCNH2 and ATP2A2; extracellular associated proteins such as COL21A1, COL15A1, and MXRA5; fibroblast associated protein FGF14; mast cells associated protein KIT; proteins mapped to force transmission defects like FNDC1, LAMA4, SSPN, or related to ion channels like KCNN3. Importantly, the myosin heavy chain isoform switch that is known to be related to the contractile velocity and energy economy the human heart [123] was consistently found in all studies (Fig 1.8).

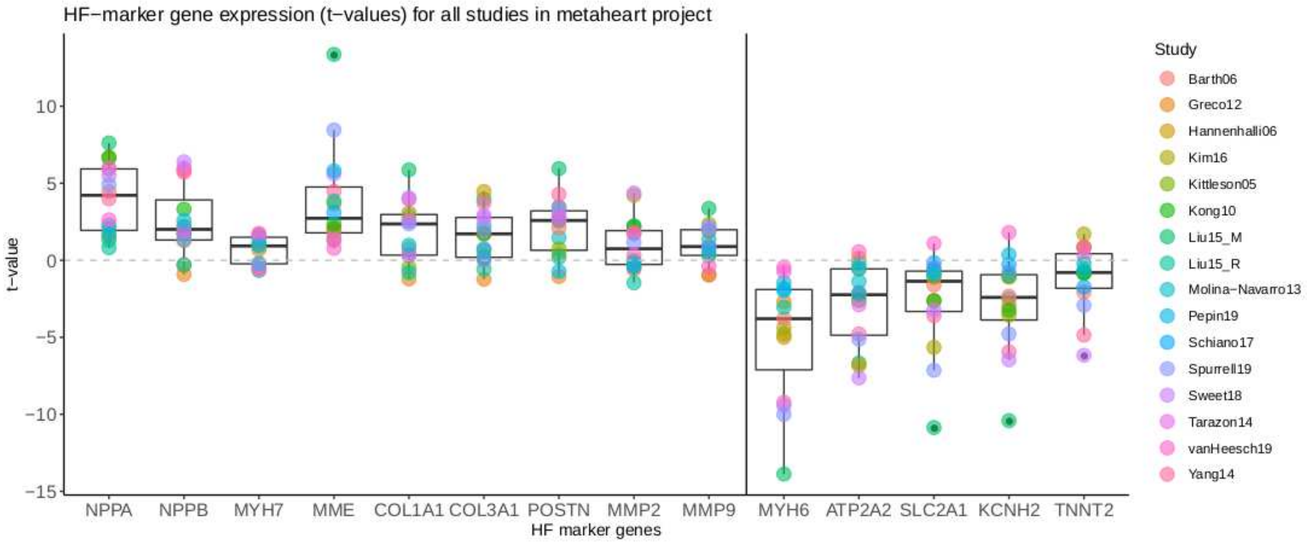


Figure 1.8 t-values from the differential expression analysis of genes that are established as dysregulated in heart failure (HF).

This data was analyzed by me. *Reprinted from* [102].

1.3.2 The added value of the HF-CS

We proposed three approaches to estimate the added value of the HF-CS over individual study reports.

First, for the comparison of studies we built a disease score and classifier approach to assess if reported DEGs in one study could also successfully separate HF from control samples in other studies (see section 1.2.4). We used the same approach now to assess whether the top 500 genes of the HF-CS further improves general classification performance over DEGs derived from individual studies. We found that AUROCs and enrichment scores improved significantly with top 500 genes from the HF-CS compared to individual study DEGs (Wilcoxon paired test, p-value < 1x10e-16).

Second, to assess the added value of the HF-CS was to compare the diversity of reporting top 500 genes between studies (Figure 1.9A). I demonstrated that genes that are highly ranked in the HF-CS were only reported originally by a few studies as genes associated with HF. I propose that these genes could be regarded as an added value as they have been reported by few studies, but the meta-analysis uncovered their highly consistent changes (Figure 1.9B).

Third, I compared the HFCS with a previous meta-analysis of HF transcriptomes. This meta-analysis [38] also reported highly consistent genes in HF. To compare the resulting gene lists, I performed an enrichment of the top 500 reported genes from Alimadadi et al. in the HF-CS (Figure 1.9C) resulting in a positive and significant enrichment. Not surprisingly, as the same data has been used to generate both signatures. However, many genes from Alimadadi et al. have been labeled to be less informative in our signature. I expected that these genes could not be reproduced by other studies and are likely part of the less informative noise of gene regulation and can thus serve as an example for the added value of our analysis.

In conclusion, the HF-CS outperforms individual studies in the classification task, which can be possibly explained by including highly consistent gene candidates that reached significance in few studies due to low effect sizes. Moreover, noisy genes were eliminated by data integration as demonstrated in comparison with a smaller meta-analysis.

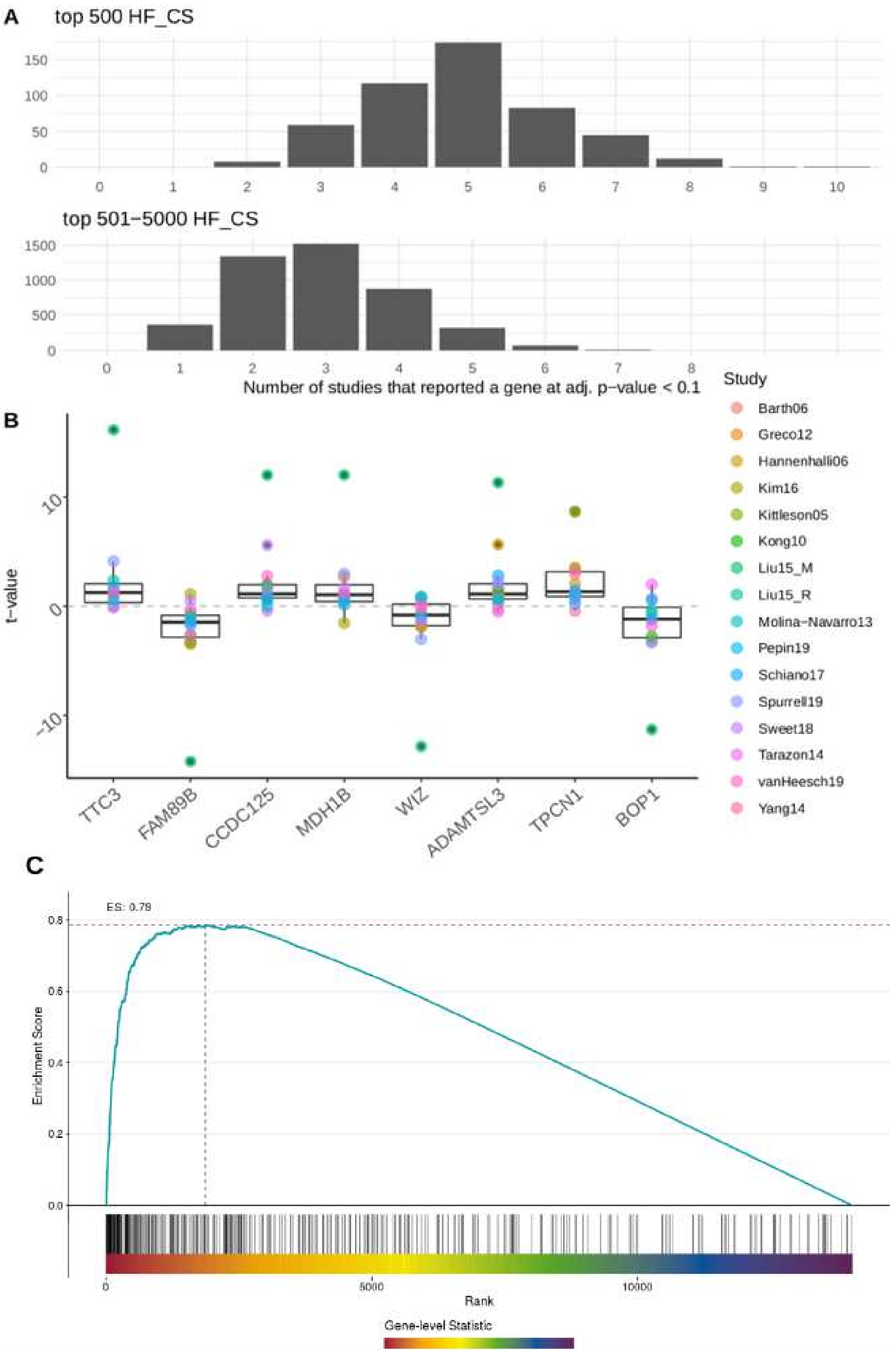


Figure 1.9 Added value of the heart failure consensus signature (HF-CS) on single gene level.

A) Histogram of genes that were reported by single studies (with adj. p-value <0.1), grouped by HF-CS rank < 501 (upper panel) and rank between 501-5000 (lower panel). Distribution of both groups varies significantly (p-value <0.0001, Wilcoxon test). B) Genes that were reported by only 2 individual studies (adj. p-value <0.1) and with a HF-CS rank < 500. Single study t-values are displayed for each gene to visualize consistency in expression. C) Running sum visualization of the top 500 genes from a previous meta-analysis [38] in the HF-CS. This data was analyzed by me. *Reprinted and modified from* [102].

1.4 Functional interpretation of the HF-CS

1.4.1 Comparison of diverse etiologies

Personalized medicine attempts to tailor treatment strategies to the individual patients' needs. The etiology of HF is currently not considered for treatment selection, because it is poorly understood how the diverse pathophysiological stimuli induce ventricular remodeling and whether or not this is a common pathway independent of etiology or whether more specific pathways exist that could offer targets for more personalized treatment options. In our study, we found that ICM and DCM failing hearts display a very similar gene expression program. I now posed the question, whether the HF-CS that characterized both, further characterizes HF of more diverse etiologies. For this reason I curated studies that were initially not included in the meta-analysis because inclusion criteria were not met (i.e. HF etiology, biopsy location or profiling platform) (Table 1.2).

GEO ID	**n (CT)**	**n (HF)**	**HF etiology**	**Reason for exclusion from meta analysis**	**Technology**	**Year**	**Country**	**Citation**
GSE10161	7	20	Aortic stenosis	HF etiology	Microarray	2008	Netherlands	124
GSE4172	4	8	Inflammatory DCM due to PVB19 infection	HF etiology & samples from right ventricle	Microarray	2006	Germany	125
GSE84796	7	10	Chagas disease	HF etiology	Microarray	2016	France	126
GSE9800	11	19	Eosinophilic myocarditis, alcoholic cardiomyopathy, hypertrophic cardiomyopathy, sarcoidosis, peripartal cardiomyopathy, ICM, DCM	HF etiology, no publication	Microarray	2007	Japan	-
GSE52601	8	12	ICM, DCM (additional 4 fetal samples)	Technical	oligonucleotide beads	2013	USA	127
GSE3586	15	13	DCM	Technical	Microarray	2013	German	114

GEO ID	n (CT)	n (HF)	HF etiology	Reason for exclusion from meta analysis	Technology	Year	Country	Citation
							y	
GSE76701	4	4	ICM	Technical	Microarray	2016	USA	[116]

Table 1.2 Additional HF studies with etiological and technical variation.

This data was curated and presented by me.

I used these studies now to investigate how well the HF-CS signature can be recovered within the gene expression changes from each individual study (Figure 1.10). I calculated the mean disease score of each sample of these excluded studies using the top 500 genes of the HF-CS and the gene level statistics of the studies included in the meta-analysis. AUROCs were calculated to estimate classifier success. We found that studies with technical exclusion reasons performed better than studies with etiological exclusion reasons. Nevertheless, the high separability suggested that the HF-CS captures a common pathway independent of etiologies. Thus, I concluded that the HF-CS might serve as reference to delineate subcohort specific gene expression patterns that diverge from the common pathway in HF. As such it might provide a highly valuable resource for future personalized medicine approaches.

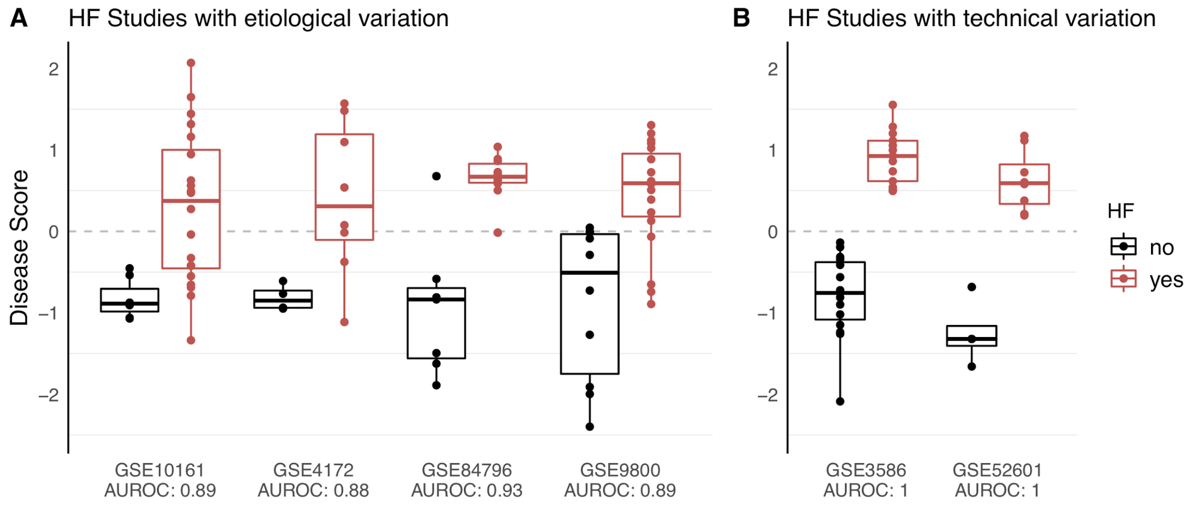


Figure 1.10 Disease score calculation based on the top 500 genes from the consensus signature for diverse HF studies.

A) HF with diverse etiologies: aortic stenosis (GSE10161); PVB19 infection (GSE4172); chagas disease (GSE84796); eosinophilic myocarditis, alcoholic cardiomyopathy, hypertrophic cardiomyopathy, sarcoidosis, peripartum cardiomyopathy, ICM , DCM (GSE84796). B) HF studies with ICM and DCM samples but processed with different bioinformatic pipelines (GSE3586, GSE52601).This data was analyzed by me. *Reprinted from* [102].

1.4.2 Pathway and transcription factor activities within the HF-CS

After confirming that the HF-CS contained a robust ranking of gene expression dysregulation in HF, we sought to functionally characterize the signature. For this we relied on a toolbox of functional genomics tools that incorporate prior biological knowledge. Since the HF-CS ranking contained up and downregulated genes, we added this directional information to the ranking by weighting the log-transformed meta p-values by the mean expression change. This directed ranking now allowed us to interpret gene expression patterns for up or down regulation.

First, I used the MSIG DB database to curate a total of 5,998 gene sets describing GO terms, canonical and hallmark pathways. 579 gene sets yielded an enrichment in the HF-CS (p-value <0.05; Figure 1.11A). Positively enriched gene sets were associated with the matrisome, and thus indicative of cardiac fibrosis while down regulated gene sets associated with inflammatory and metabolic processes.

Second, we used two footprint based tools to estimate activities of transcription factors (TFs) and pathways. We found that the regulons of 65 TFs were consistently up- or down regulated (p-value<0.05). Among active TFs were SOX2, MEF2A, MEF2B, MEF2C, ARNT and MEIS1-2, RUNX2. RUNX2 is involved in osteoblast differentiation and is suspected to drive pro-fibrotic soft tissue remodeling. MEF family members are crucial for cardiac development and have been associated with hypertrophic remodeling in adult HF [128]. ARNT is indicative of metabolic stress as it is involved in fatty acid oxidation together with PPARγ, and the cardiac-specific depletion of ARNT resulted in improved cardiac function in mice [129].
The pathway analysis revealed two main findings. JAK-STAT was the only pathway with a high activity (p-value<0.05). The JAK-STAT pathway is activated by growth factors and cytokines and is an important regulator of cardiac development and inflammation. We found a down-regulation of the targets of the TNFα and NFKB pathways. While TNFα plasma level is associated with the worsening of cardiac function [130], possible cardioprotective effects are [131,132]. The decreased pathway activities of TNFα and NFKB together with the decreased TF activities of RELA and NFKB1 that we found in the HF-CS indicated that TNFα related signaling is decreasing in the end stage Heart. Future

studies and experimental approaches are necessary to mechanistically elucidate the role of TNFα in HF.

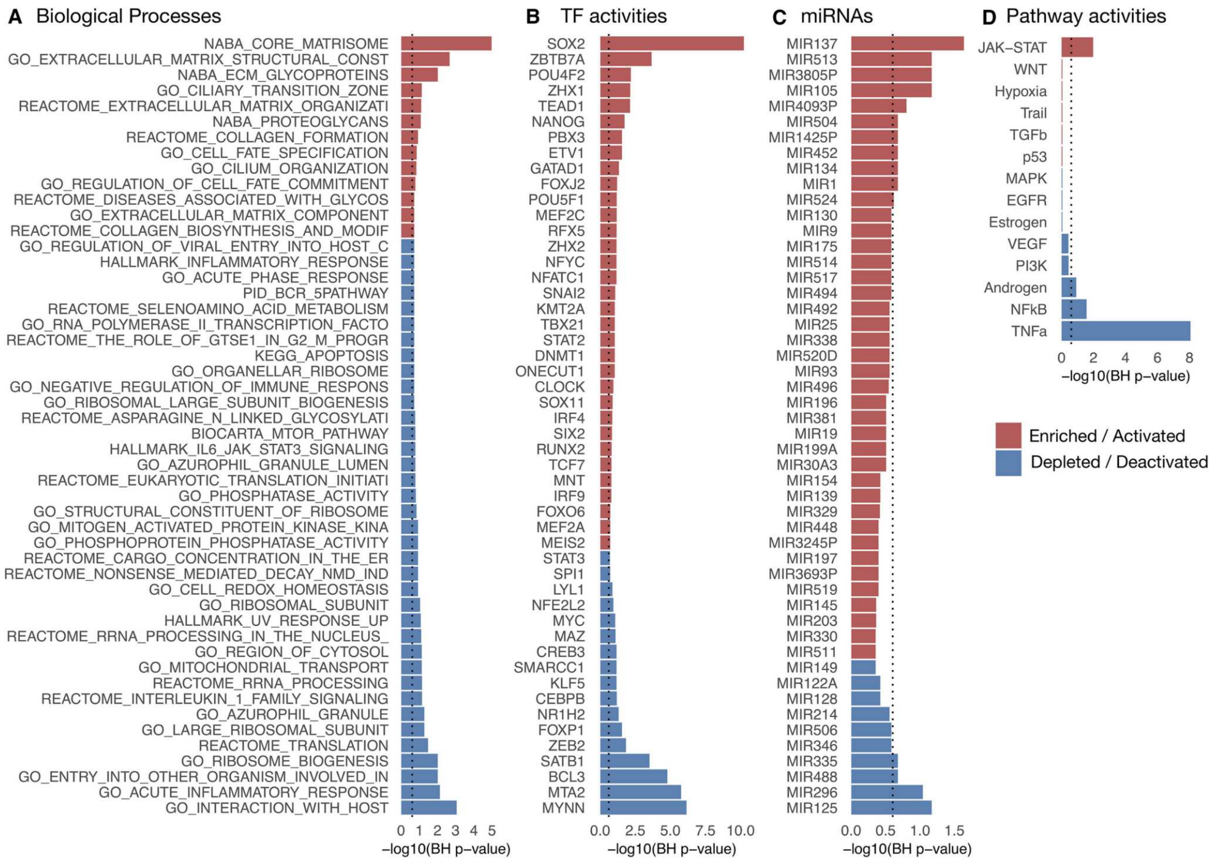


Figure 1.11 Functional characterization of the HF-CS.

−log10 (BH P-values) coloured by direction of enrichment (A and C) or by direction of activation (B and D) of the top 50 (A) most enriched canonical and hallmark gene sets, (B) transcription factor activities, (C) miRNAs' targets, and (D) all signaling pathway activities. Dashed line indicates BH P=0.25. BH indicates Benjamini-Hochberg; HF-CS, heart failure consensus signature; and miRNA, micro RNA.This data was analyzed by Ricardo Ramirez-Flores and me. *Reprinted from* [102].

1.4.3 HF-CS as a resource for biomarker detection and hypothesis building

The HF-CS can be used to learn about consistent patterns of gene expression in a large cohort of HF patients. While these patterns can inform us about underlying biology (see section 1.4.2), it can also be a powerful resource for prioritizing candidates from independent data. In this section, I will demonstrate how an integrative approach using independent data together with the HF-CS can provide a prioritization strategy by combining multiple sources of evidence. First, I will use a plasma proteome study to suggest novel cardiac HF markers and second, I will use fetal transcriptome studies to decipher fetal reprogramming in HF.

The general framework that I proposed is the signature generation in a case-control setting. Next, a test for enrichment of the top signature features is performed in the HF consensus signature using GSEA. Finally, highly consistent features can be filtered by directionality of dysregulation and significance levels. Here, I used a combination of the leading edge of GSEA and the ranking of the HF-CS (Figure1.12A).

The plasma proteome study by Egerstedt *et al* [133], analyzed the plasma proteome in an untargeted fashion of early and manifest HF patients as well as controls. While HF is a systemic disease, changes in plasma proteome could be attributed to multi-organ involvement or comorbidities of HF patients. Thus, the identification of novel biomarkers of cardiac origin is tempting to identify biomarkers indicative of cardiac remodeling. I propose that the HF-CS can be utilized to hypothesize about possible cardiac origin of plasma markers. I observed a clear enrichment of manifest HF proteins (GSEA p-value = 0.0001) and a modest enrichment of early HF proteins (GSEA p-value = 0.13) in the top of the HF-CS (Figure 1.12B), indicating that plasma protein markers also tend to be upregulated on gene expression level in the heart (Figure 1.12C). Further, this association is higher for HF patients in manifest disease stages. While gene transcription often is poorly correlated to protein translation, I sought to filter candidates by considering protein expression information by the human protein atlas [134]. For this I processed protein expression data (Figure 1.13A) to identify genes with recorded protein expression in cardiac tissue. Next, I calculated a cardiac specificity score by considering whether protein expression was recorded in other tissues as well (Figure 1.13B). Resulting Candidate markers included the established HF marker NPPA which might serve as a proof of concept for this approach. Novel potential markers induced CCDC80, BID, MAP2K1, MRC2, JAK2, and LTBP4. To sum up, these markers were suggested because they were i) consistently upregulated in HF on gene expression level, ii) reported on protein level in the heart and iii) upregulated in plasma proteome of HF patients.

The second use case for the HF-CS as a reference could be to address the concept of fetal reprogramming. This reactivation of gene expression patterns typical for fetal cardiac cells and is thought to be an adaptive process to counteract pathophysiological stress. These patterns are involved in metabolism, the contractile apparatus and conduction system [135]. I investigated whether these patterns are part of the HF-CS by analyzing two studies (Spurrell19 & GSE52601, Table 1.2) that compared healthy human hearts with fetal hearts. First I analyzed the separability of the two studies by calculating the disease score per sample based on the top 500 genes of the HF-CS which yielded perfect separability (AUROC = 1) (Figure 1.14A). I performed differential expression analysis and estimation of TF activities of the fetal studies by comparing fetal to healthy adult hearts as described before. Then, I tested the enrichment of top differentially expressed TFs in the HF-CS (Figure 1.11B). In the Spurell19 data, 221 of the top 500 genes from the consensus signature correlated with fetal, indicating that these genes might be part of the fetal reprogramming (Figure 11D). Furthermore, 32 TFs correlated with TFs active in fetal hearts including SOX2, NANOG, MEF2C, MEIS1 (Figure 11E). GSE52601 displayed similar results (Figure 14B,C). This indicated that a great share of the HF-CS can be attributed to changes related to fetal reprograming, and the HF-CS enables us to decipher which patterns are part of the fetal reprogramming and which patterns can be attributed to adult heart disease remodeling.

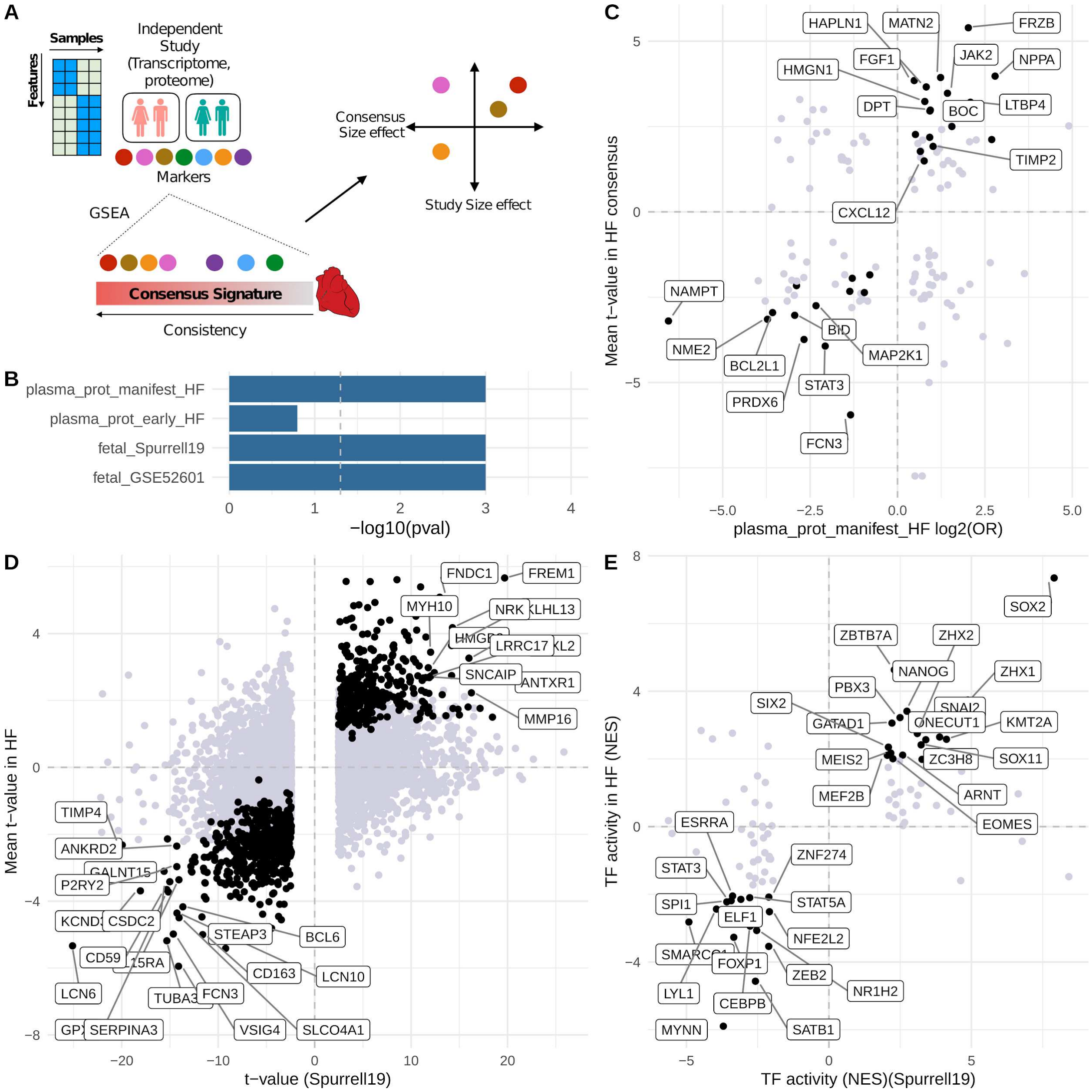


Figure 1.12. HF-CS as a reference that complements independent studies.

A) Schematic of a suggested framework. Marker features from independent studies are enriched in the HF-CS with GSEA. Features that belong to the leading edge are further filtered, e.g. by correlation or ranking in the HF-CS. B) Enrichment results of marker features from four individual studies. C) Plasma proteome of HF patients mapped to the HF-CS. D) Fetal cardiac transcriptome (Spurrell19) mapped to HF-CS on gene level and E) TF level. Black dots in C & D indicate correlated features in the enrichment leading edge; labeled features in C & D indicate genes with a rank < 500 in HF-CS. Black dots in E indicate overlap with significantly dysregulated TFs derived from the HF-CS. This data was analyzed by me. *Reprinted from* [102].

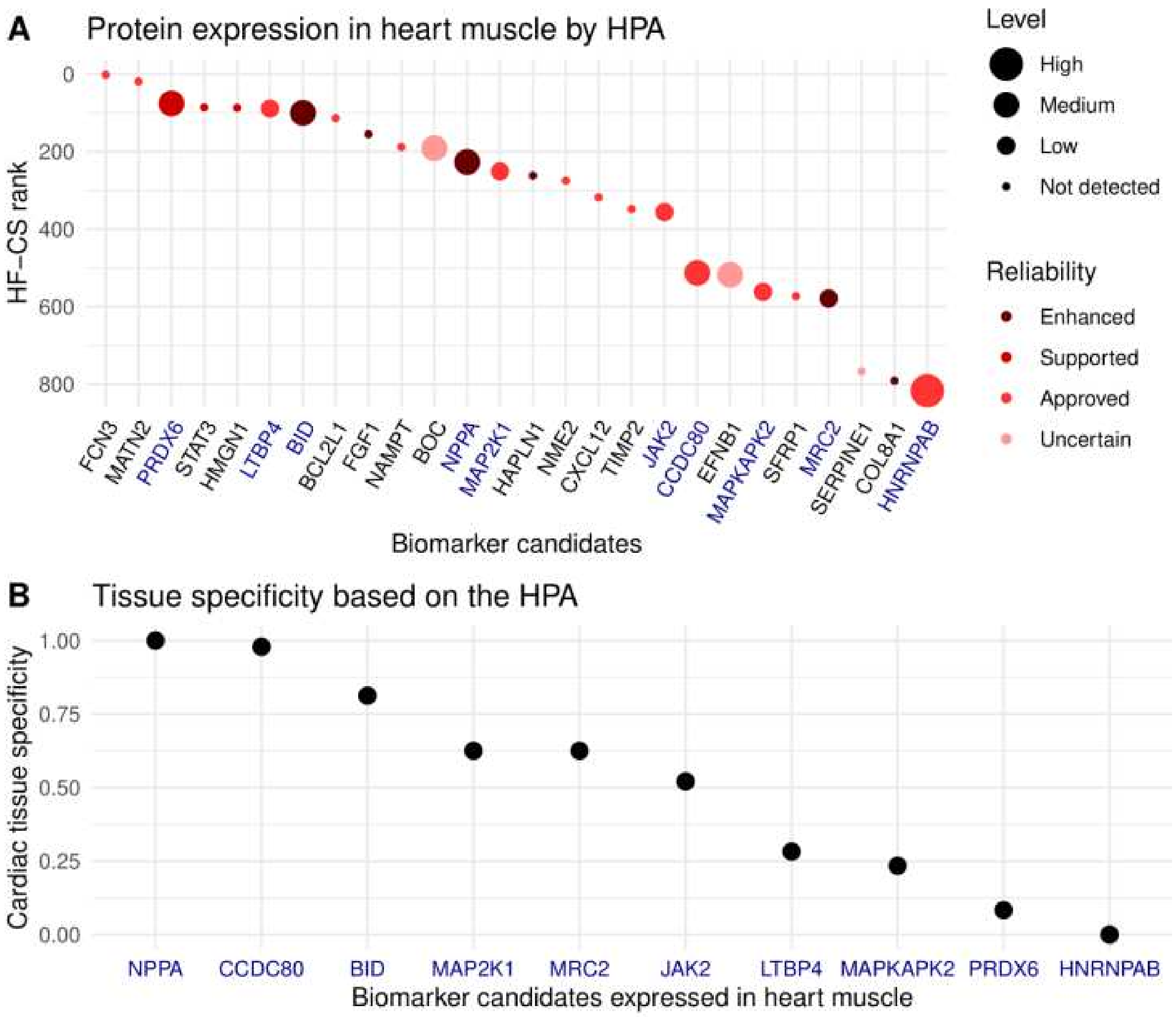


Figure 1.13. Biomarker candidates and their expression in the Human Protein Atlas (HPA).

A) Relevant biomarker candidates taken from figure 5 and analyzed for their reported protein expression in heart muscle tissue in the HPA. Protein expression was reported for genes labeled in red including PRDX6, LTBP4, BID, BOC, NPPA, MAP2K1, JAK2 with a rank in the heart failure consensus signature (HF-CS) < 500 and CCDC80, MAPKAPK2, MRC2, HNRNPAB with rank between 500-1000. Expression of FRZB, TIMP3, F3 and DPT were not assessed by the HPA. B) Assessment of tissue specificity of protein expression using the HPA. The total number of measured non-cardiac tissues in the HPA per candidate ranged between 46 and 48. Tissue specificity was calculated as the ratio of tissues not expressing the protein (Low or Not detected) to the total number of measured tissues. NPPA is not expressed in any non-cardiac tissue. CCDC80 and BID are showing high to moderate specificity while HNRNPAB is suggested to be unsuitable for a cardiac biomarker as it is reported in all non-cardiac tissues. This data was analyzed by me. *Reprinted from* [102].

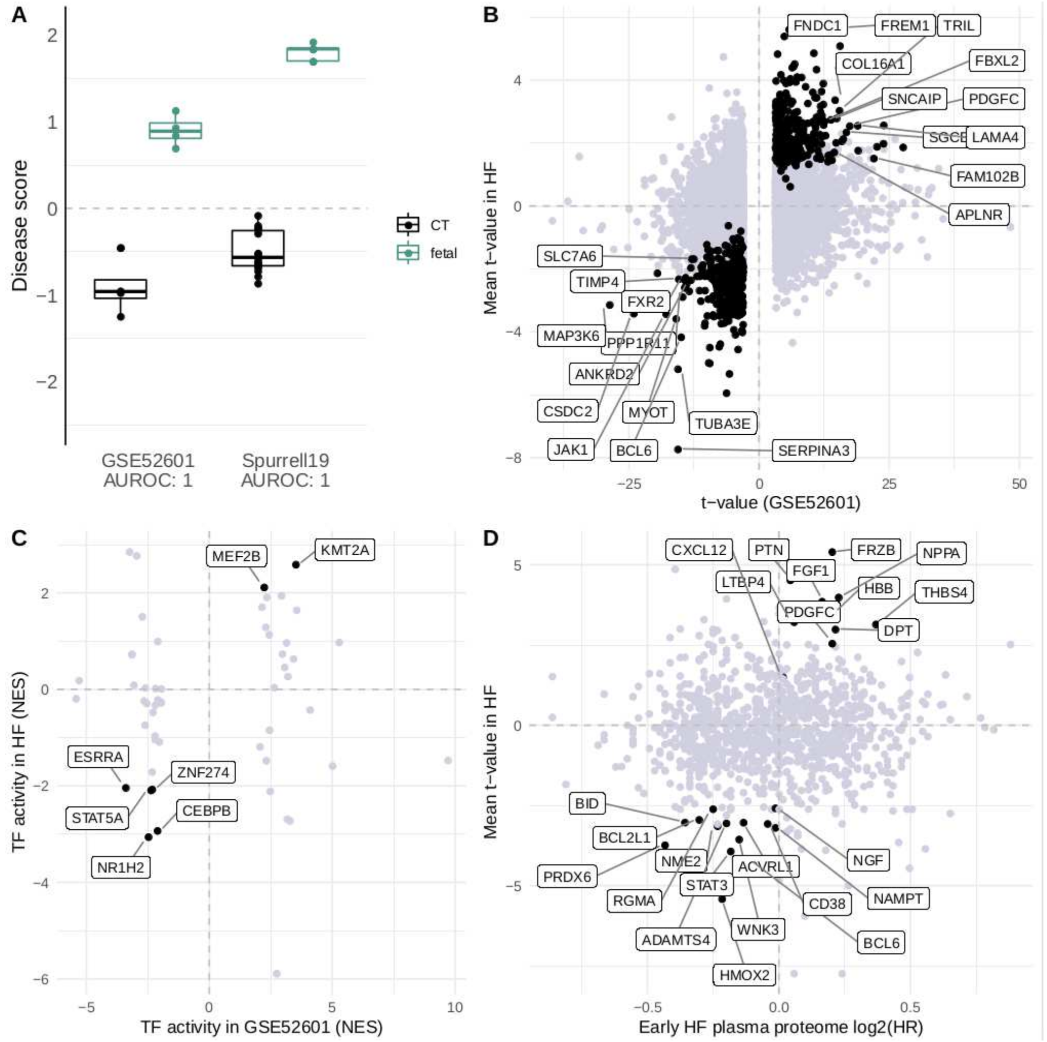


Figure 1.14 Heart Failure consensus signature (HF-CS) as a reference that complements independent studies.

A) Disease score calculation for fetal experiments Spurrell19 and GSE52601. CT, control (adult non failing heart samples); fetal, fetal heart samples. B) Significant genes in GSE52601 mapped to the HF-CS. Black dots indicate correlated genes in the enrichment leading edge. Labels indicate genes with a rank < 500 in HF-CS and adjusted p-value < 10e-4.3. C) Significant TFs in GSE52601 mapped to TFs derived from the HF-CS. Black dots and labels indicate significant and correlated TFs in GSE52601 and HF-CS. D) Plasma proteome of early HF patients mapped to the HF-CS. All plasma proteins are displayed. Black dots and labels indicate correlated proteins with a rank < 500 in the HF-CS. This data was analyzed by me. *Reprinted from* [102].

1.5 Sharing the HF-CS with the community

The last part of this project was to make our work findable, accessible, interpretable and reusable. For this, I imagined two user types, a bioinformatician and a clinician or biologist. For the former we provided all processed data from this study in an accessible data repository (https://zenodo.org/record/3797044#.XsQPMy2B2u5), and all scripts in a public code repository (https://github.com/saezlab/HF_meta-analysis). However, to familiarize with the code and data structures might be a hurdle for some scientists that might prevent the utilization of our work. For this reason we employed with the help of Christian Holland a shiny app that can be accessed via a user interface (https://saezlab.shinyapps.io/reheat/). I conceptualized this web page by defining main use cases for clinicians:

1. **Gene query**: Users can select genes of interest to explore their role in the consensus signature of heart failure (HF-CS).
    a. **Expression** of queried gene(s), as reported by single studies. Expression is displayed as a t-value resulting from the individual differential expression analysis
    b. **Ranking** of queried gene(s) in HF-CS. The lower the rank the more consistent the deregulation of that gene was observed in all studies
    c. **Mean t-value** of queried gene(s) compared to mean t-value of all genes in the HF-CS
    d. **Raw data** of queried genes (from individual studies and from meta-analysis) to download
2. **Enrichment of gene signatures:** Users can upload their own signatures to test quickly whether they are enriched in the HF-CS and see which genes are driving this enrichment.

3. **Explore the whole consensus signature of heart failure**: Results of individual studies are provided as well as the consensus statistics. Raw results can be queried and downloaded.
4. **Functional Characterization of the heart failure consensus signature (HF-CS):** Query signaling pathway activities, transcription factor activities, biological processes and miRNAs, that characterize the HF-CS.
5. **Study overview:**The heart failure consensus signature integrates results from 16 different studies. Relevant information are displayed:
    a. **Study overview** with reference to individual studies
    b. Comparison of **sample size** and availability of clinical information
    c. **Gender and age** distribution
    d. Comparison of **Gene coverage**

With these efforts I hoped to provide a useful resource for the cardiovascular research community. For instance, our work was so far successfully used as a reference for TRPM channel expression in HF [136], for ALAS1, SUCLG1/SUCLA2 expression [137] and as an approach to filter for HF-relevant genes from independent *in vitro* experiments [138].

1.6 Discussion and Conclusion

In this study, I presented a comprehensive meta-analysis of the HF transcriptome, analyzing and comparing 16 data sets, and a total of 916 samples, which constitutes the largest report of HF transcriptome to date. HF is a complex disorder on both the clinical and genetic levels. As such, the published work in myocardial transcriptomics represents a heterogeneous picture of transcriptional regulation in the heart with little agreement on key regulated genes. In the studies included in this meta-analysis, clinical heterogeneity is compounded by wide variability in analysis pipeline, study design, tissue protocol, and patient selection. Our work showed that despite these difficulties, combining the insights of these studies provides an opportunity not only to robustly evaluate their reproducibility, but also to gain a more complete picture of transcriptional regulation [102].

The presented study combines gene expression data from microarray and sequencing technologies. While the measurements of both technologies differ fundamentally, we demonstrated that similar biological profiles can be captured. We focused on comparing and combining differential expression results across studies, as opposed to integrating all samples in a single data set. This framework prioritized molecular differences between phenotypes that are similar in independent patient cohorts and allowed us to reuse and review a large patient cohort to create the HF-CS [102].

Our results suggest that the magnitude of changes in mean expression of marker genes depends highly on the study. We observed a 5% agreement of the top 500 differentially expressed genes between studies. This disagreement cannot be explained by differences in gene coverage or technologies, since the intersection of profiled genes in all studies is ≈70%. However, patterns of gene coexpression are stable and comparable among cohorts, regardless of their sample size, technology, and variability, allowing for their integration. Unexpectedly, studies with fewer than 10 patients were still able to effectively capture similar patterns of gene deregulation as studies with >200 patients. This highlights the importance of representative patient sampling, since it may compensate for sample size. Moreover, we observed that consistent coexpression patterns were shared among etiologies, suggesting that conserved disease mechanisms converge in end-stage HF [102].

Important limitations of our study relate to the data used. In this meta-analysis, we included only public data sets from published studies. Since most of the studies lack complete descriptions of the individuals included in their cohorts, it is unfeasible to estimate how much of the clinical and demographic diversity of patients with HF is covered in our curation. As the necessity of studying HF in clinically ramified subgroups is becoming evident [75] the impact of comorbidities, medication, and disease phenotype on the gene transcription profile needs to be considered. To test how the reported gene expression patterns associate with severity and progression, a deeper patient characterization is required. With this work, we aimed to encourage the community in the field to open the dialogue about secure data-sharing standards and more inclusive and transparent study designs [102]. Although clinical information was very sparse, we found little variation in gene expression to be associated with those variates that were reported. The agreement of gene expression trends as well as the expected similarity of HF patients regarding their late disease stage together indicate that a common end-stage disease landscape exists. We observed that the gene expression changes in late stage HF overshadow other possible patterns. Thus, besides the improvement of sample annotations, novel strategies for acquiring cardiac samples are needed to estimate more reliably possible etiology or subcohort specific characteristics of ventricular remodeling to ultimately enable a more personalized treatment of hf patients. Possible avenues are the improvement of low-risk sample acquisition [139] and standardization of sample sharing and annotation.

We built the user-friendly free platform ReHeaT (Reference of the Heart Failure Transcriptome; https://saezlab.shinyapps.io/reheat/) to facilitate further use of the HF-CS. We propose two ways in which the HF-CS can be exploited. First, the genes, TF, and pathways provide a rich resource for interpreting and understanding the transcriptional landscape of HF. Second, the HF-CS can be used as a trustworthy reference of HF to assist in hypothesis building or confirmation [102]. Below, we discuss in detail both approaches.

First, I aimed at the biological interpretation of the HF-CS and I will briefly discuss interesting findings. These functional insights, however, still require experimental validation to confirm their relevance.

MME encodes for Neprilysin, a transmembrane zinc-metalloendopeptidase that, if measured in blood, predicts an increased risk of recurrent cardiovascular admissions in ambulatory patients with heart failure[140] . More relevant, Neprilysin inhibitors (i.e drug combinations of sacubitril/valsartan) are a rather new member of the physicians arsenal of systolic heart failure treatment [141] . Neprilysin is involved in the degradation of A-type natriuretic protein (encoded by **NPPA**), a relevant serum marker for HF diagnosis. The meta ranking reports highly consistent MME (meta-rank: 88) and NPPA (meta-rank: 293) upregulation which encourages the search for novel therapeutic targets and biomarkers of heart failure in the meta-ranking.

NANOG and **SOX2** are key regulatory TFs that help to maintain pluripotency in stem cells. Although functional myocardium is regenerated throughout a human lifetime [142], cardiac progenitor cells probably are not the primary resource of cardiomyocyte regeneration [143,144]. The role of cardiac progenitor cells is still poorly understood, but the footprints of NANOG and SOX2 could indicate that cardiac progenitor cells are kept in a pluripotent state, forestalling cardiomyocyte differentiation. Cardiomyocytes themselves are proposed to be able to regenerate themselves, an ability that decreases with age [145]. We detected activity of **MEIS1** and **MEIS2**, which contribute to the curbing of cardiomyocyte differentiation: Deletion of MEIS1 in mice or inhibition of MEIS2 in rats resulted both in the re-activation of adult cardiomyocyte mitosis [128,129] . I therefore propose that these mechanisms could be active in human failing hearts as well, as the regenerative power of the myocardial syncytium is impaired.

ARNT (HIF1b) interacts with hypoxia response elements by heterodimerization with multiple TFs including HIF1a. The cardiac specific depletion of ARNT resulted in an increased fatty acid oxidation leading to improved cardiac function in mice [128,129]. The reported ARNT activity in the meta-ranking could indicate that this mechanism might be part of the pathological gene expression pattern in human heart failure and therefore constitute a promising target of intervention.

We found footprints of **MEF2A**, **MEF2C** activities. All MEF2 family members are expressed during cardiac development and have been described as part of the fetal reprogramming in adult heart failure [128,146]While it's complete depletion in mice lead to prenatal death, a later siRNA knock down approach in murine heart failure model attenuated cardiac hypertrophy [128,146,147], illustrating its physiological role during development and pathological role during adulthood. We confirm consistent MEF2 footprints in CHF.

The classifier built with the diseases score calculation was able to correctly discern fetal heart samples from healthy adult heart samples in two different studies, indicating that the genes that drive gene expression variation in adult CHF also vary in fetal hearts compared to adult hearts, part of this pattern are the genes regulated by MEF2 TFs.

ZBTB7 is a proto-oncogene that is involved in a large variety of fundamental cellular functions like proliferation, apoptosis, migration and metabolism [148]. It exerts its action by regulating DNA chromatin structure and recruiting many other TFs or initiating their transcription [149]). ZBTB7A has mainly been studied in the field of cancer, where it was shown to act in tissue specific context either as tumor suppressor or activator. In this study ZBTB7A yielded strong positive footprints. This TF has not been introduced to the pathology of chronic heart failure yet, but, considering its immense complexity of downstream effects, could be one of the master regulators that might be responsible for a large share of the gene expression changes observed in failing hearts.

The **JAK STAT** pathway is activated by growth factors and cytokines and is an imperative regulator of cardiac development and inflammation. The role of JAK-STAT in CHF is ambivalently discussed, with evidence that JAK STAT is involved in physiological as well as pathophysiological cardiac hypertrophy, ischemic pre and post conditioning and cardiac fibrosis as reviewed previously [150,151] [152]. We report JAK STAT to be a significantly activated pathway in the meta ranking of end stage human heart failure. JAK STAT could therefore be a mechanistic feature common in end stage human heart failure and could be part of the cardioprotective loop that is activated for compensational purpose but on the long run injures cardiac via induction of pathophysiological cardiac hypertrophy and fibrosis.

Inflammation plays a fundamental role in the development of heart failure. **TNFa** levels are elevated in heart failure patients in relation to decreasing functional status of the patient. Clinical Trials targeting TNFa however failed to improve HF outcome [153,154]. We report a decreased signature of TNFa activation in failing myocardium compared to healthy tissue, which is accompanied by decreased NfKB, RELA activity. This might indicate that TNFa does not work directly on myocardial tissue in the failing heart leading to the absence of TNFa signaling.

The extracellular **purine metabolism** regulates the balance between ATP and Adenosine. The equilibrium is tightly regulated and impacts among others local immune response, ischemic preconditioning, thrombosis and vascular calcification. Further adenosine is believed, to be cardioprotective against HF via i) the attenuation of catecholamine release, β-adrenoceptor–mediated myocardial hypercontraction, and myocardial Ca2+ overload; ii) the increases in coronary blood flow; and iii) the inhibition of platelet and leukocyte activation [155]. In the meta ranking we find evidence of changes in actors involved in purine metabolism. CD73 is known to be a crucial regulator of extracellular Adenosine by catalyzing ADP to Adenosine dephosphorylation. We found that CD73 encoding NT5E gene (rank: 125) is consistently upregulated. Adenosine signaling is mediated by G-protein coupled receptors, which are regulated by among others, RGS4 protein. RGS4 overexpression is known to be found in hypertrophic hearts [156] and its induction in a mouse model lead to cardiac decompensation following transverse aortic constriction [157]. In our meta-analysis, RGS4 (rank: 240) is consistently upregulated and could mediate adenosine signaling in HF. The role of purine metabolism during HF has hitherto been poorly addressed, but could lead to impactful results in future investigations.

The second use case of the HF-CS is a proposed signature evaluation. In this part I demonstrated the utility of the HF-CS by integration with studies analyzing the fetal transcriptome and the plasma proteome from patients with HF. The two biological use cases of the signature matching, might be to find the commonality or divergence to the HF-CS.

In the case of the fetal reprogramming, it might be tempting to distinguish between those patterns in the HF-CS that relate to fetal program reactivation and which are novel programs. Detailed pathophysiology of this process is incompletely understood. Our analysis provides a plethora of genes and TFs that might shape the fetal response in HF. We detected established TFs like MEF2, but also identified a collection of less explored TFs including SOX2, ZBTB7A, NANOG, and ONECUT1.

The plasma proteome of patients with HF is used to identify circulating biomarkers. However, tracing the origin of measured candidates to the heart is often difficult. Here, I used the HF-CS to detect commonalities by filtered circulating proteins. This approach identified the established marker NPPA [158]. Other identified markers include Wnt modulators SFRP1 and FRZB; the latter has been associated with HF outcome before[159]. We also identify CXCL12 to be of potential myocardial origin, which is associated with stroke [160] and acute HF [161] HAPLN1, MATN2, and COL8A1 constitute extracellular matrix components with, to date, an unknown role in HF. To suggest cardiac tissue specificity of candidates, we assessed protein expression in cardiac tissue. As a result of this, we propose CCDC80 as a promising HF biomarker candidate, which has been suggested to be secreted by cardiomyocytes in response to pressure overload before [162,163]. BID also displayed reasonable cardiac tissue specificity but has not been studied in the context of HF yet. Other genes with reported protein expression included MAP2K1, MRC2, JAK2, and LTBP4. These candidates could represent biomarkers of pathophysiological relevance and potential clinical utility [102].

We proposed that the utility of data integration with more independent studies is highly promising. Especially with transcriptomic technologies developing toward single-cell and spatial resolution, this resource could help to confirm cell type–specific elements in a large HF population. Additionally, etiology-specific responses could be derived by comparing differences of different cohorts with our proposed consensus signature. As more data are released, the resource described in this work will be updated to be a trustful reference of the transcriptome of HF [102].

In summary, I demonstrated the feasibility of combining gene expression data sets from different technologies, years, and centers in a biologically meaningful way. I highlighted the importance of data reviewing and contextualisation of individual studies with prior knowledge and data. Further, I identified knowledge gaps in earlier stage HF patient collective and HFpEF and provided suggestions for experimental follow up studies. As the number of cardiovascular high-throughput studies increases, the need for structured data integration is evident. I provide a reference for this purpose that is applicable to many other research topics within the cardiovascular field [102].

Chapter II - Fibrosis in murine HF models

2.1 Background

HFpEF comprises a complex and multifactorial interplay of the disease promoting risk factors, such as hypertension, obesity, metabolic syndrome, chronic inflammation, and aging. Suitable animal models were missing until a few years ago [21,164], when a two-hit mouse model combining a 60% high-fat diet with inhibition of the constitutive nitric oxide synthase by Nω-nitro-l-arginine methyl ester (L-NAME) recapitulated metabolic and hypertensive stress in HFpEF. Analysis of this model led to major mechanistic insights in the pathophysiology of hypertrophy and cardiac immunometabolic alterations in HFpEF[21,165,166] and potential drug targets. Since these studies focused predominantly on cardiomyocyte hypertrophy and metabolism[7], little knowledge was gathered about the distinct role of cardiac interstitial cells and their cross-talk in ventricular stiffening and fibrosis[7,165].

Single-cell RNA sequencing (scRNAseq) allows for the quantification of transcriptional changes of individual cells and description of cell phenotype heterogeneity. Consequently, scRNAseq opened the door for fundamental insights into cellular heterogeneity, developmental biology and molecular disease processes in the cardiovascular field [71,167,168]. Thus, its application to a HFpEF model could shed light on the cellular disease mechanisms but, to our knowledge, no such study exists.

Here I present a scRNAseq analysis of the ventricular interstitium in mice receiving L-NAME and high fat diet (further called HFpEF model) in early stages of diastolic dysfunction. I compare fibroblast phenotypes and disease signatures by integration with scRNAseq data from other HF models that are phenotypically closer to HFrEF and identify HFpEF specific patterns of fibroblast activation. I characterize HFpEF associated fibrotic signatures and compare them with human bulk references, providing new pathophysiologic hypotheses relevant for HFpEF fibrosis.

This chapter is part of a manuscript that is currently in revision in a peer review journal. I wrote this manuscript together with Laura Wienecke. I conceived, implemented, presented and discussed all the analysis performed in this chapter, if not stated otherwise. Laura Wienecke performed all experiments in this chapter, with assistance by Maura M. Zylla, Niklas Hartmann and Florian Sicklinger. Julio Saez-Rodriguez and Florian Leuschner supervised this project and provided minor edits to the manuscript.

2.2 Single cell RNAseq of the murine HFpEF model

2.2.1 Disease model description

To mimic and study HFpEF, we used the established two-hit mouse model that induces metabolic and hypertensive stress by 60% high-fat diet and L-NAME, respectively[21]. From 7 weeks of dietary intervention onwards, a diastolic dysfunction phenotype was observed echocardiographically under preservation of systolic left ventricular function (Figure 2.1A). Body and heart weight, normalized to tibia length, increased concordantly indicating obesity and cardiac hypertrophy (Figure 2.1A). To describe this early remodeling, we isolated cardiac interstitial cells after 7 weeks by MACS dead cell depletion and FACS sorting of live and metabolically active cells (Figure 2.1B). We performed scRNAseq with the 10x Chromium droplet based platform to analyze cellular transcriptomic changes within cardiac ventricular interstitial cells of two control and two HFpEF murine hearts.

2.2.2 Data processing and quality control

The resulting single-cell RNA-seq FASTQ files were processed using CellRanger provided by 10x genomics with the help of Volker Ast from the core facility of the Mannheim University. I then processed the scRNAseq count data in sample wise manner with the following filters: >300 Feature numbers, <25% mitochondrial genes, <1% ribosomal genes and >500 RNA counts. Doublet scores were calculated with the R-package *scDblFinder*[169] and only predicted singlets were kept. I further calculated a dissociation score by estimating expression of dissociation associated gene expression[170] with *Seurat*'s [171] *AddModuleScore* function and I removed cells above the 99% quantile. Data was log-normalized. Samples were clustered individually by selecting the 3,000 highest variable genes with the *FindVariableFeatures* function from the *Seurat* package. From the intersection of these lists, the top 3,000 genes were selected to calculate principal components (PCs). Top 30 PC embeddings were adjusted with *harmony* R-package, with samples as covariates.

To identify the main celltypes captured, I applied the nearest neighbor approach and graph based clustering (i.e. Louvain algorithm) implemented in *Seurat* to cluster cells and to stepwise test optimal cluster resolution (from 0.1 to 1.6 in 0.1 steps) by computing silhouette widths. This identified initially 14 distinct clusters. I removed four clusters that were inconclusive for different reasons, i.e. high expression of mitochondrial genes, expression of multiple cell type markers, and/or consistently low RNA and Feature counts. After removal the integration process was repeated and a final atlas was created and I retained expression profiles of 6,132 cells described by 15,046 genes (mean UMI coverage per cell: 2,838).

Next, I annotated celltype clusters by calculating and interpreting cluster markers. I calculated these with the *FindMarkers* function with default parameters (i.e. wilcoxon test) in *Seurat* and cell types were manually annotated based on known canonical markers. The ten clusters represented major cell types of the murine heart. I identified two fibroblast clusters (Col1a1+ and Wif1+), endothelial cells (EC) (Pecam1+), natural killer cells (Gzma+), macrophages (CD68+), T-effector cells (CD8+) and T-helper cells (CD4+), B-cells (CD19+), granulocytes (S100a9+), smooth muscle cells and pericytes (Acta2+) (Figure 2.1D).

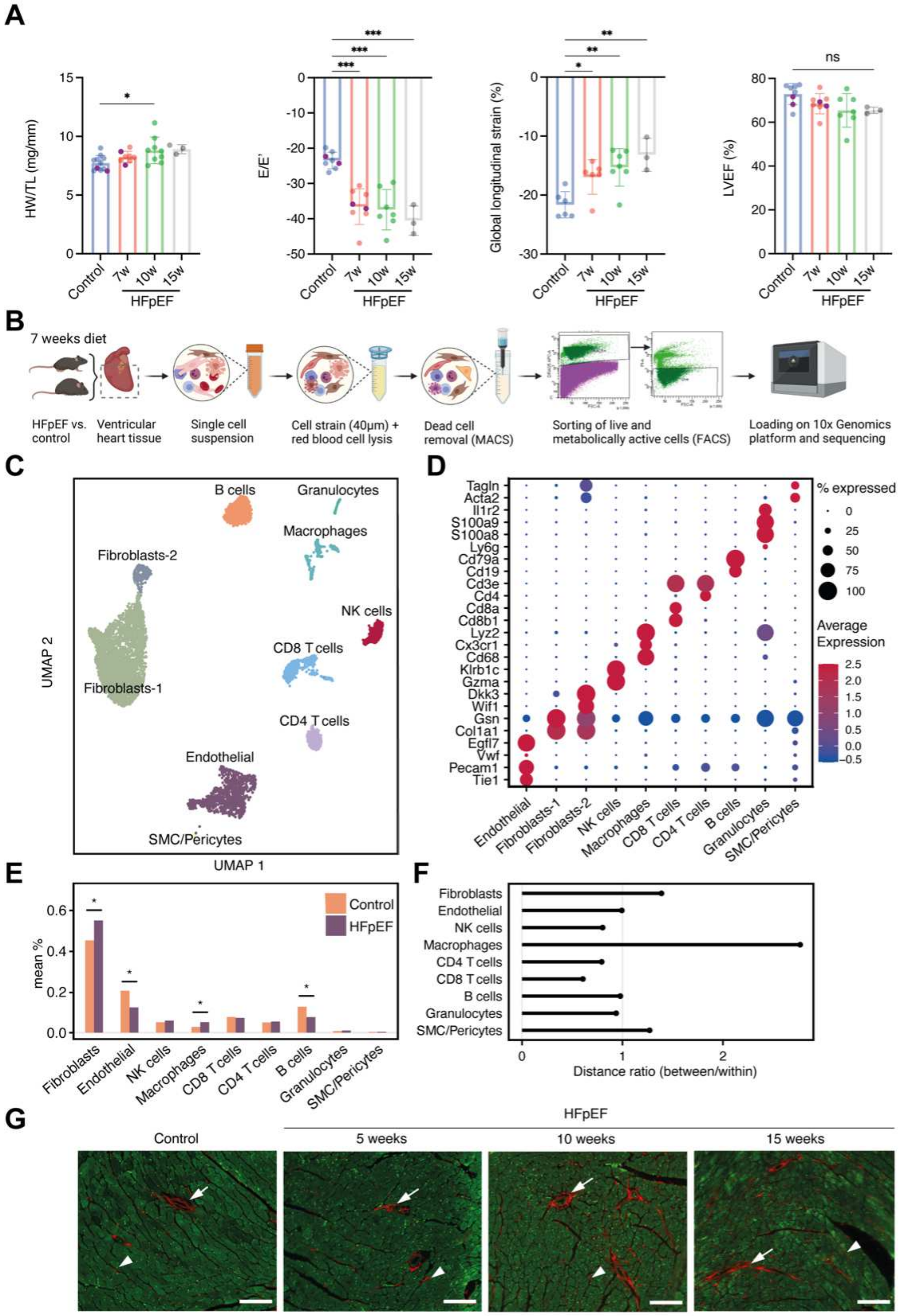


Figure 2.1. Study model and cell type assignment.

A) Murine HFpEF model characterization by ratio of heart weight to tibia length (HW/TL) and echocardiographic hallmarks (E/E', global-longitudinal strain and LVEF), purple data points represent the animals used for single-cell RNA sequencing (scRNAseq). Statistical analysis performed by one-way ANOVA, bar graphs indicate mean±SD, *p<0.05, **p<0.01, ***p<0.001. ns= not significant, LVEF= left ventricular ejection fraction, w= weeks. B) Schematic summary of experimental setup for scRNAseq experiments using mice after 7 weeks of HFpEF or control diet. C) UMAP embeddings of normalized scRNAseq data after processing and filtering. D) Marker gene expression for cell type assignment. E) Cell type composition of main cell types as mean percentage per group (n=2 per group), compared between HFpEF and control mice. *p<0.05, p-values were calculated via label permutation. F) Cosine distance ratios of pseudobulked cell type profiles. Median between group distance is divided by median within group distance. G) Representative Picrosirius-Red stainings of fibrotic fibers and perivascular fibrosis (red) from control and different stages of HFpEF heart sections. Imaging performed in 594 nm (Picrosirius-Red) and 488 nm (cardiomyocyte autofluorescence) channels. Scale bars in the right bottom corner indicate 100µm length. Panels A,B,G were generated by Laura Wienecke.

2.2.3 Cell type composition and molecular profiles suggest fibroblast and macrophage involvement in cardiac remodeling

Next, I addressed whether the remodeling that we observed on phenotype level was associated with compositional changes. Since our study design was limited to a two by two comparison, I could not perform statistical tests on sample level. I thus evaluated the significance of compositional changes via label permutation.

I tested if cell type composition changes between groups are meaningful by implementing a permutation approach to estimate a null distribution. For each individual cell, I considered the sample it came from and the cell type label it was assigned. From this table, I created 1000 permutations of the sample label. For each permutation run I calculated the cell proportions for each sample and calculated the mean proportion per cell per group (ct, hf), from which the difference in cell proportion was calculated as test statistic. By calculating the proportion per sample and not per group, I kept unequal cell numbers in samples. The resulting 1000 random cell proportion differences were an estimate for a null distribution. All distributions passed the Shapiro-Wilk test for normality ($p > 0.05$). I calculated the area under the normal curve from the mean and standard deviation of the null distribution to estimate the probability of observing the actual measured proportional difference. This composition analysis yielded a modest but significant increase of fibroblasts and macrophages and decrease of B-cells and endothelial cells (Figure 2.1E).

As cell type compositions are not independent and therefore only partially informative of the importance of a cell type for remodeling, I addressed which cell types displayed variation in their molecular profiles. For this I computed cosine distances between pseudobulk profiles per cell type to assess whether the variation of gene expression between experimental groups was higher than the variability expected within a single group[54]. First, highly variable features were calculated per cell type with *FindVarFeature* function from *Seurat* and the top 3000 features were selected for distance calculation. For each cell type and sample, pseudobulk profiles were TMM normalized and voom transformed with the *edgeR* and *voom* R-package and cosine similarities were calculated. I calculated median sample distances within groups and between groups to assess the distance ratio. Cell types with distance ratio below 1 show higher sample distances between groups than within groups and are candidates for differential gene expression analysis. I found that macrophages displayed the highest ratio of 'between to within group distance' followed by SMC and fibroblasts. ECs and B-cells did not display high disease associated variability, suggesting that their relative decrease in proportion is not associated with fundamental gene expression changes or high within group variation. I further applied *Augur*, a classifier-based cell type prioritization method[172], to identify distinguishable cell types. This yielded the highest performance for macrophages and endothelial cells, followed by fibroblasts (data not shown). L-NAME treatment directly targets ECs, expected to induce direct transcriptional changes, possibly also leading to EC depletion.

Taken togther, the positive compositional change and the molecular differences suggest that fibroblasts and macrophages are important contributors to the early HFpEF associated remodeling and phenotype. HFpEF is known to be accompanied by interstitial fibrosis[14] which we confirmed histologically in myocardial tissue at various time points in the HFpEF model (Figure 2.1G). While fibrosis at 7 weeks can be described as an early state, collagen deposition and ECM remodeling is present thus, the underlying fibroblast phenotype is of high interest to better understand HFpEF-related cardiac fibrosis

2.3 Atlas of fibroblast activation in murine heart failure

Cardiac fibroblasts accomplish a wide range of biological functions, crucial for tissue homeostasis and architecture[173]. In various types of heart failure, cardiac fibrosis represents a major axis of reparative and adverse remodeling. The activation of fibroblasts has often been described as a process involving TGFbeta and myofibroblasts transdifferentiation and knowledge about etiology specific activation patterns is missing. After establishing that fibroblasts are involved in early HFpEF remodeling on phenotype and molecular level, I contextualized and compared HFpEF fibroblast activation across four murine HF models, i.e. early myocardial infarction (MI), late MI, Angiotensin II (AngII) and HFpEF.

In this section, I describe how I integrated fibroblast data sets (section 2.3.1) to comprehensively define major common phenotypes in health and disease (section 2.3.2). Next, I derived and compared fibroblast gene expression signatures from each HF etiology (section 2.3.3). Finally, I addressed the division of labor between fibroblast states and described gene expression patterns regarding fibroblast activation involving fibroblast phenotypes (section 2.3.4).

2.3.1 Study integration of cardiac fibroblasts

I compared HFpEF fibroblast activation with other cardiac fibrotic disease etiologies by integrating our single-cell data with two other single-cell studies: firstly, a model for cardiac hypertrophy by hypertensive stress induced by two weeks of angiotensin II (AngII) administration[174] and secondly, an acute myocardial infarction model[168]. Both models were used to study heart failure and included a replacement or reactive or reparative fibrosis (Figure 2.3A).

My analysis strategy for the atlas integration was to reprocess each study with the same computational pipeline, identify fibroblasts via clustering and annotation of cluster markers and then integrate the annotated fibroblasts from all studies. More specifically, the raw FASTQ files for the two additional 10x Genomics scRNAseq datasets were processed with the same cell ranger pipeline as described above. Sample integration was performed via canonical correlation analysis as implemented in *Seurat*. Unsupervised clustering and cluster marker assessment was used to identify fibroblasts in each study by selecting Col1a1+, Pdgfra+ and Gsn+ cells (Figure 2.2A-C and D-E) which were subset to perform study integration.

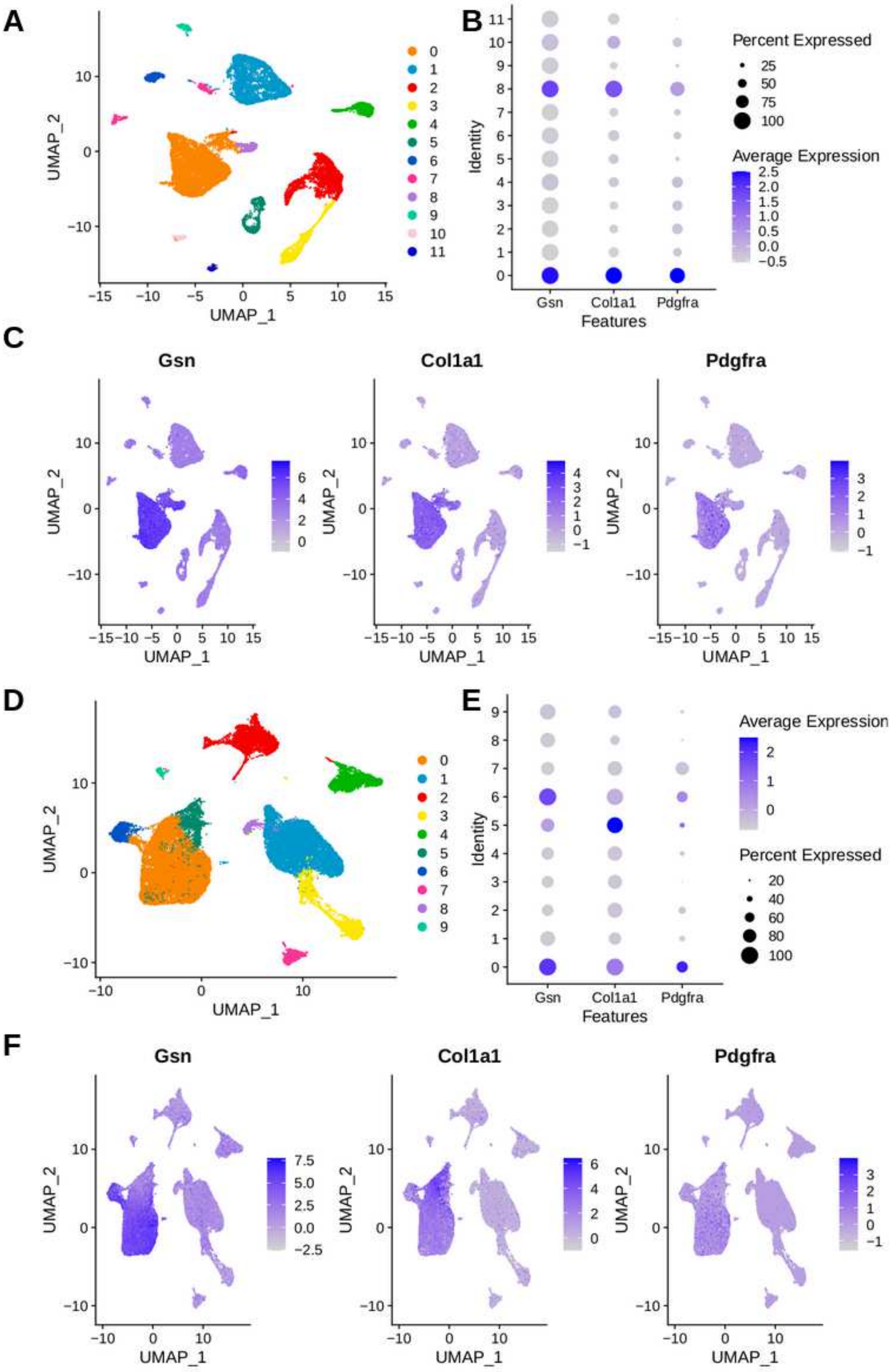


Figure 2.2 Fibroblast annotation in AngII and MI model.

AngII model (A-C) and MI model (D-E). A, D) UMAP embedding of full cell atlas after processing, filtering and clustering. B+C) Expression of fibroblast marker genes (Gsn, Col1a1, Pdgfra) in the AngII model per cluster. Cluster 0 and 8 were subset for downstream analysis. E+F) Expression of fibroblast marker genes (Gsn, Col1a1, Pdgfra) in the MI model per cluster. Cluster 0, 5 and 6 were subset for downstream analysis.

Then, I integrated fibroblast cell data from three datasets via calculating highly variable features in each data set, using 3000 overlapping features of all datasets. I used *Harmony* with study and sample ID as covariates for data set integration. Downstream analysis was performed as described above. To evaluate integration performance I ensured that each study contributed cells to each cluster. In addition, to quantify batch effects from different studies, samples and experimental groups, I calculated a batch mixing score based on average silhouette width[51]. A score of 1 represents a balanced integration while 0 represents strong batch effect conservation. The Integrated fibroblasts atlas yielded a batch mixing score of ~0.99 for study labels, ~1 for group labels and ~0.97 for sample labels.

This analysis resulted in an integrated atlas of 26,455 cardiac fibroblasts, capturing a wide spectrum of phenotype diversity across HF models.

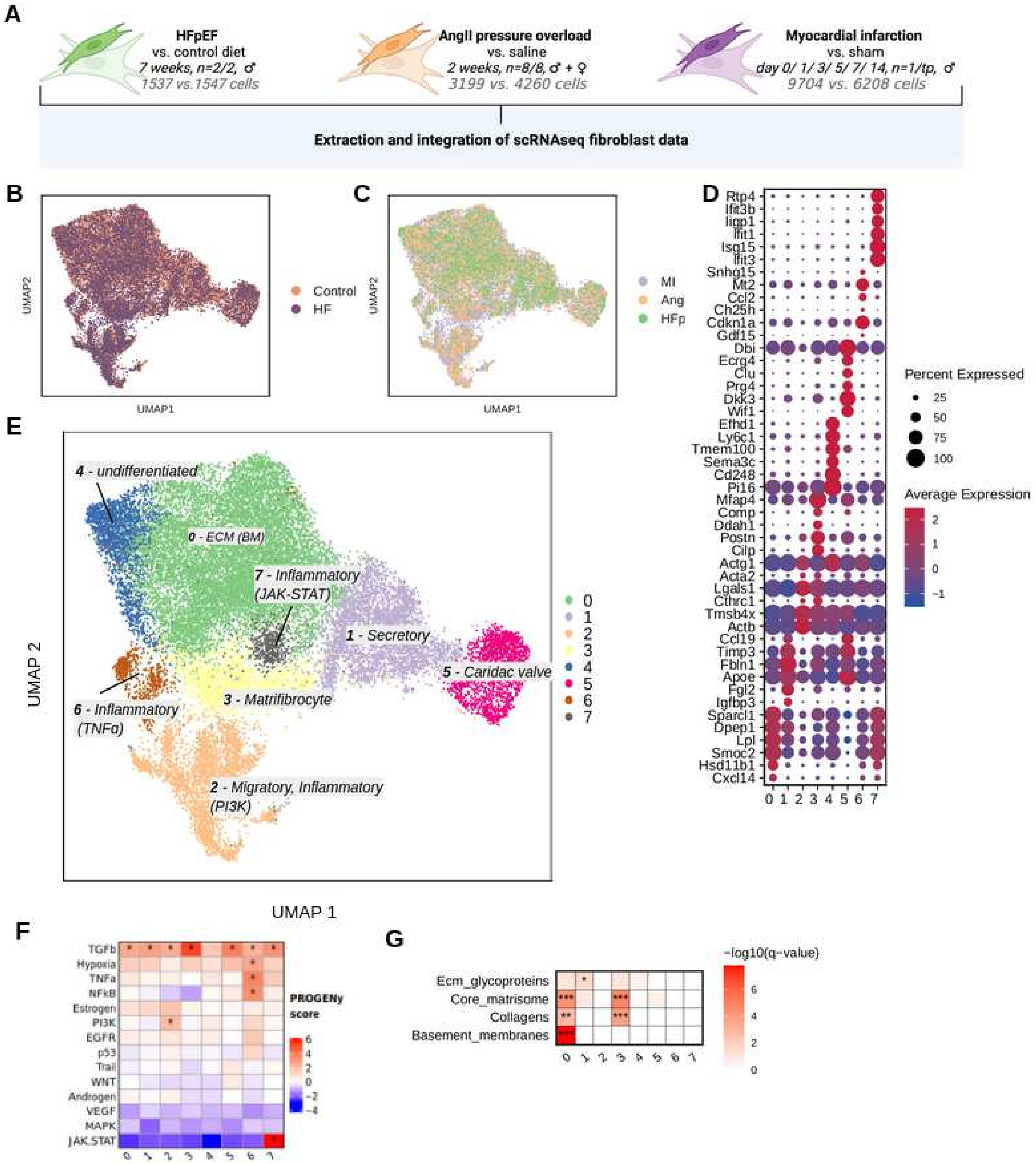


Figure 2.3 Integrated atlas of cardiac fibroblast phenotypes from different disease models.

A) Schematic of murine HFpEF and HFrEF (AngII and MI) fibroblast studies. B+C) UMAP embeddings of integrated fibroblasts, colored by disease (HF, Heart Failure) vs. control (B) and study (C). D) Overview of top cluster marker expression of integrated fibroblast states (IFS). E) UMAP embeddings of fibroblasts colored by cluster with annotations derived from functional interpretation of cluster markers. F) Estimated pathway activities with PROGENy based on cell state marker gene expression (x-axis). G) Gene set enrichment of extracellular matrix gene sets in cell state markers. Hypergeometric test with BH correction, *q <0.05, **q<0.01, ***q<0.001). AngII= angiotensin II, HF=

heart failure, MI= myocardial infarction.

The integrated analysis of the three data sets allowed now for the joint annotation of fibroblast phenotypes or often called cell states. Previous work has reported cardiac fibroblast phenotype diversity at the single-cell level in healthy and diseased hearts [168,174–178], but a consensus of main cell states is missing.

I identified eight integrated fibroblast cell states (IFS) by performing unsupervised clustering as described before while every study contributed to all cell states (Figure 2.4A). In this section I aimed to characterize these IFS by suggesting their functional niche based on their molecular profiles.

First, I calculated IFS marker genes (Figure 2.3D, Figure 2.4B) and compared them with a cross organ fibroblast atlas[177] (Figure 2.4C). This atlas can be regarded as a fibroblast reference to jointly define fibroblast phenotypes from different organs. Thus, by comparing this atlas, I could suggest which of the IFS could constitute cardiac specific fibroblasts. I found that IFS 0 (Col15a1+), 3 (Comp+) and 4 (Pi16+) displayed high marker overlap (hypergeometric test $p<0.01$) which suggest that these states might represent fibroblast phenotypes shared across organs. Conversely IFS 1, 2, 5, 6 and 7 displayed weaker and/or ambiguous associations and could represent cardiac specific fibroblast phenotypes.

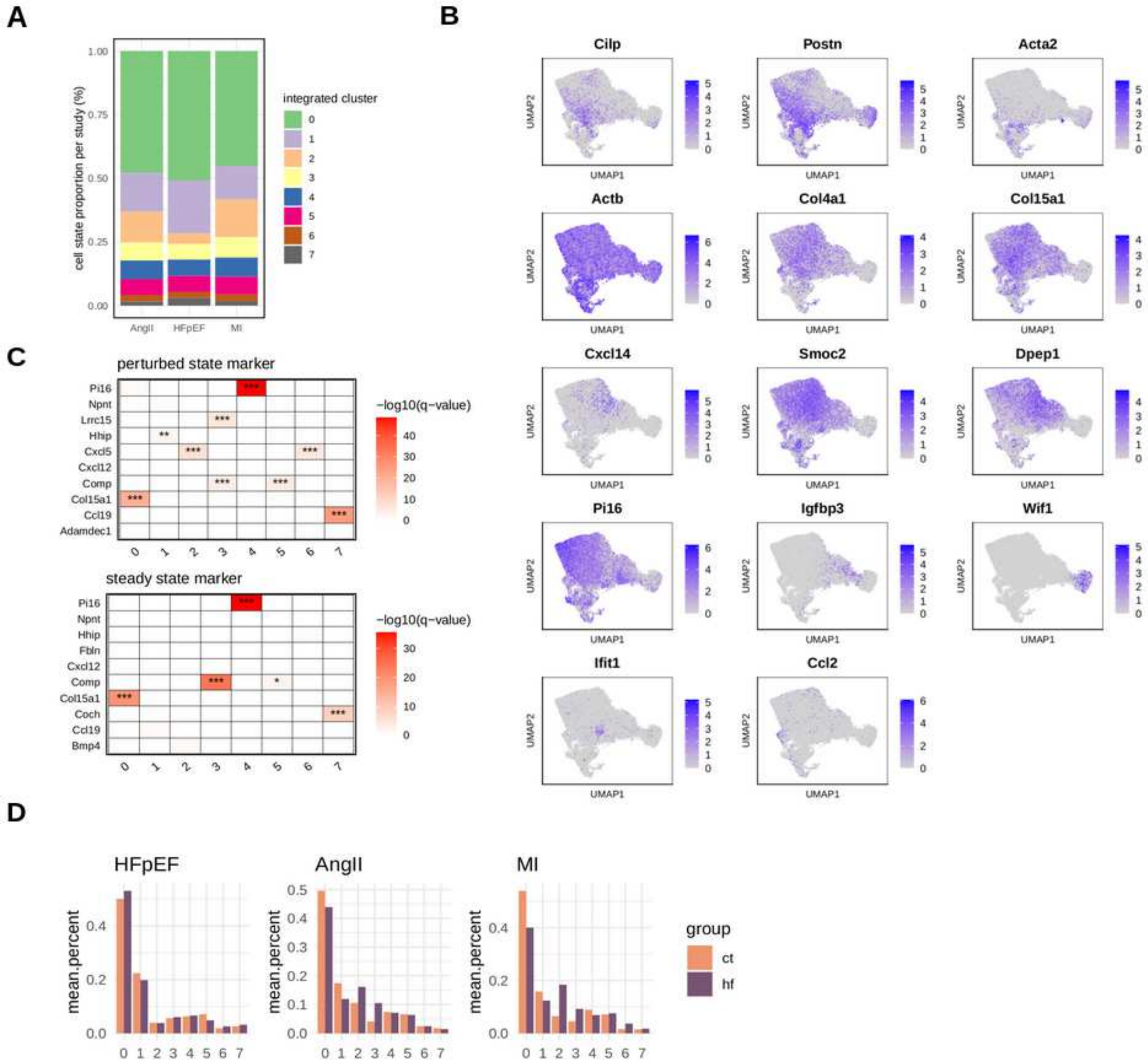


Figure 2.4 Functional characterization of Integrated fibroblast states (IFS).

A) Composition of cell states per study in percent. B) Cell state marker expression in the fibroblast atlas. C) Comparison of top 100 IFS marker with top 100 marker from cross organ fibroblast atlas. Hypergeometric test with Benjamini Hochberg correction. ***p<0.001, **p<0.01, *p<0.05. D) Mean sample composition of IFS by study and by group.

Next, to characterize IFS roles, I performed pathway activity (Figure 2.3F) and gene set enrichment analysis (Figure 2.3G). **IFS 0** fibroblasts were the most abundant cell type in every dataset (Figure 2.4D) and have been described as homeostatic fibroblasts characterized by Col15a1, Dpep1 expression. **IFS 4** fibroblasts are characterized by Pi16 expression and constituted adventitial stromal cells that might accomplish a reservoir function for downstream fibroblast differentiation [177,179]. The **IFS 3** can be termed matrifibrocytes and are characterized by Cilp, Thbs4, Comp and Postn expression [174,178]. Pathway analysis indicated that IFS 3 demonstrated highest TGFβ activity (Figure 2.3F), which highlights the pro-fibrotic potential of this cell state. ECM remodeling is a major operation of fibroblasts and was assessed by enrichment of ECM related gene sets[180] (Figure 2.3G), suggesting that **IFS 0** and **IFS 3** fibroblasts were the main ECM producers: both were characterized by expression of collagens (e.g. Col5a3, Col6a3) and core matrisome related genes (e.g. Ltbp2, Col8a1, Cilp), while IFS 0 uniquely expressed genes associated with the basement membrane (e.g. Col4a1, Lamb1, Hspg2, Col15a1).

I identified three IFS with inflammatory profiles: **IFS 2, IFS 6** and **IFS 7**. **IFS 2** appeared to be a heterogenous group of fibroblasts that were partly characterized by Acta2 and Actb expression which are myofibroblast characteristics, as well as pro inflammatory genes involved in antigen processing and representation (Psmd8, Psma6, Vamp8) and Chaperonin containing T-complex polypeptide (CCT) genes (CCT3, CCT7, CCT4, CCT8) that have been associated with proliferative and fibrotic tissue remodeling [181–183]. Furthermore IFS 2 exhibited highest PI3K pathway activity which has been shown to enable fibroblast migration[184,185]. **IFS 6** fibroblasts were characterized by pro-inflammatory NFκB and TNFα signaling (Figure 2.3F) and cytokine expression of Ccl2, Cxcl5 suggesting that IFS 6 participated in immune cell attraction. **IFS7** cells formed a small cluster that every study contributed to with a comparatively small number of cells and was characterized by JAK-STAT activity and interferon-γ related gene expressions (Ifit3, Isg15). JAK-STAT pathway has been linked to fibroblast activity in rheumatoid arthritis [186,187] and osteoporosis [188] but its function in cardiac fibroblast is unclear.

IFS 5 was characterized by Wif1 and Dkk3 expression. In the heart, Wif1+ cells were shown to localize to the cardiac valves and their adjacent hinge regions [189]. **IFS 1** was characterized by a secretory gene expression pattern including insulin like growth factor1 (Igf1) and fibrinogen-like protein 2 (Fgl2) which control cardiomyocyte growth [190,191], Insulin like growth factor binding proteins (Igfbp3, Igfbp4) which are age related factors[192] that can modulate Igf function, and a set of Glycoproteins like Fibulin-1 (Fbln1), extracellular matrix protein 1 (Ecm1), Matrix-gla protein (Mgp).

In summary, the phenotype atlas of murine cardiac fibroblasts could be described as inflammatory states IFS 2 (PI3K activity), IFS 6 (TNFa activity), IFS 7 (JAK-STAT activity); ECM producing states IFS 0 (basement membrane), IFS3 (collagens); stromal fibroblasts IFS4, structural valve fibroblast IFS5 (Wif1) and secretory fibroblasts IFS1 (Growth factors and Glycoproteins) (Figure. 2.3E).

2.3.2 Comparison of fibroblast signatures between murine HF models

To functionally compare fibroblast activation between study models, I performed differential gene expression analysis by comparing disease to control fibroblasts for each study independently to avoid cross batch comparisons. Since the MI study included multiple timepoints, I separated the samples by calculating a signature of early (days 3, 5 and 7) and later remodeling (days 14 and 28).

Similar to compositional analysis, the limited study design prevented a pseudobulked differential expression analysis due to difficulties in estimating dispersion at sample level. Thus I chose a cell-level differential expression analysis approach. To control for different absolute numbers of cells per sample, I subsampled the total number of cells to the lowest cell number in a sample. For these cells I calculated differentially expressed genes with the Wilcoxon test implemented in *Seurat's* FindMarker function. To ameliorate sampling effects I repeated this subsampling process 5 times, and reported the gene intersection of genes with Benjamini Hochberg corrected p-value <0.05 and absolute log2FC >0.1. This approach was applied to every study independently.

The signatures contained different intersections between studies (Figure 2.5A), containing a small core set of upregulated (Timp1, Col1a1, Loxl1 and Sparc) and downregulated genes, common to all disease models. When quantifying the intersections, I found little overlap in general, except for AngII and late MI signatures (Figure 2.5B). As discussed in chapter I of this thesis, the agreement on sets of DEGs can depend on technical factors and obfuscate transcriptional similarity. Hence, I compared the direction of regulation by correlating fold change regulation of the disease signatures between studies (Figure 2.5C). Interestingly, the HFpEF signature did not correlate with AngII while displaying weak agreements with early and late MI. Again, the strongest correlation was found between AngII and late MI fibroblasts.

In HFpEF fibroblasts I detected 74 upregulated genes that included fibrosis related genes such as Col1a1, Col1a2, Col4a1 and genes involved in metabolism such as Angptl4, Slit3, Pcolce, and Ace (Figure 2.5D). Col4a1 is an important component of the basement membrane and its accumulation over time in the HFpEF model was confirmed immunohistologically and indicated a Col4a1 pattern of interstitial sheathing of cardiac cells (Figure 2.5E). Angiopoietin-like 4 (Angptl4) is a lipoprotein lipase inhibitor that was not expressed in control fibroblasts, but was induced in HFpEF.

For further characterization of the signatures, I enriched annotated gene sets from the MSIG DB database. Fibrosis signatures across models contained major ECM related gene sets (Figure 2.5F). The HFpEF signature was uniquely characterized by heat shock factors, crosslinking of collagens, basement membrane and laminin components, and thrombospondin-type-1-repeat (TSR) glycosylations, but contained less components related to elastic fibers than AngII and MI (Figure 2.5G). Next, I used log fold change regulation of target genes to infer upstream transcription factor (TF) activities (Figure 2.5H). Hsf1, Ppara and Pparg are suggested to be relevant TFs in the HFpEF model. Notably, Pparα and Pparg positively regulate Angptl4 and other genes related to metabolic remodeling and adipogenesis, Interestingly all models displayed high Smad3 activity, possibly linking this TF to cardiac fibrosis. Hif1α displayed activity in HFpEF and early MI. When comparing pathway activities (Figure 2.5I), TGFβ showed strong activation in the AngII and late MI model while in early MI fibroblasts, proinflammatory TNFα and NFκB as well as hypoxia pathways are active. In HFpEF, none of the tested pathways was highly activated, however, displaying only modest activity of TGFβ and TNFα.

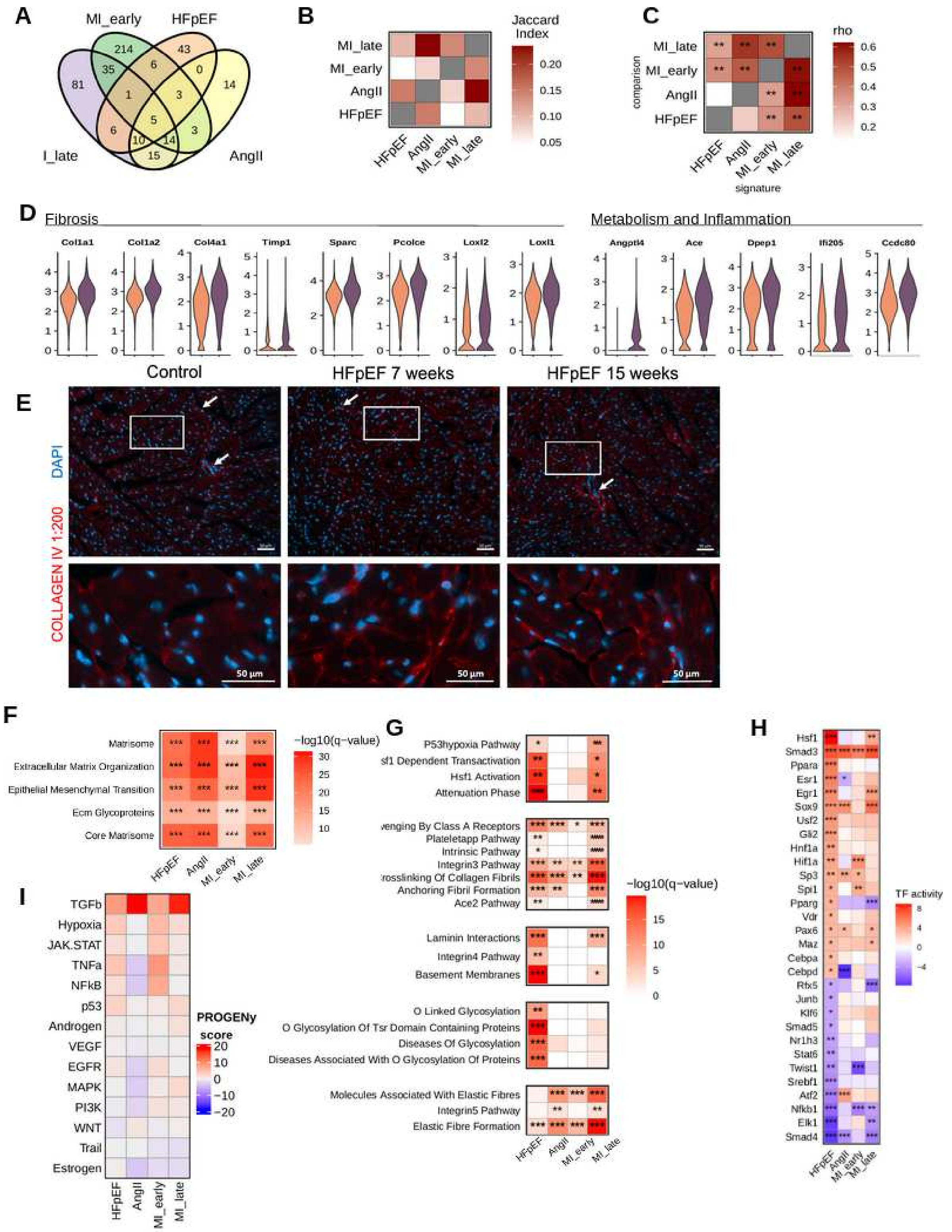


Figure 2.5 Comparison and interpretation of study specific fibroblast disease signatures.

A) Comparing intersection of upregulated (heart failure vs control) genes between studies (Venn diagram) and B) intersection quantification via jaccard index. C) Comparison of direction of regulation between studies. Pearson correlation was calculated between log2FC vectors of signature genes in pairwise comparisons. **p<0.01. D) upregulated genes E) Immunofluorescence images of collagen IV (red) and DAPI (blue) staining of left ventricular heart sections. Lower panels show magnifications of the areas marked by white boxes. White arrows indicate capillaries or larger blood vessels. Scale bars in the right bottom corner indicate 50µm length. F+G) Heatmap of geneset overrepresentation in study specific fibroblast disease signatures. F) Common genesets between signatures, G) selected gene sets to highlight HFpEF signature characteristics. Each heatmap represents a group of similar genesets. q-value = Benjamini Hochberg corrected p-value from hypergeometric test, *q <0.01, **q<0.001, ***q<0.0001 H) Estimated TF activities with DoRothEA based on effect size (avg log2 fold change) of target genes within each study. I) Estimated pathway activities with PROGENy based on effect size of footprint genes compared between studies. Panel E was generated by Laura Wienecke.

2.3.3 Connecting fibroblast signatures with phenotypes

In the previous sections I defined and characterized IFS across mice models, and interpreted study specific disease signatures. To link both aspects, I asked how fibroblasts from different IFS partake in fibroblast activation, and, importantly, if there are different activation patterns between studies.

I conceptualized different patterns of disease signature upregulation in regard to IFS (Figure 2.6A). First, I distinguished between a composition and a transcriptional shift. The former describes a rather stable expression within an IFS but is accompanied by an compositional increase of that IFS. Thus the quantification of this gene would suggest an upregulation. On the other hand, transcriptional shifts constitute an upregulation without compositional increase. Here, I proposed to distinguish between an upregulation emphasized within an IFS (state dependent) or within many or all IFS (state independent). These terms were used to describe gene expression patterns in fibroblasts, however, it was expected that these categories were not exclusive.

To investigate possible compositional shifts, I calculated IFS composition changes between control and diseased mice per study (Figure 2.6B). In HFpEF, composition changes are faint and only IFS 0 and 6 expanded slightly, while in early MI the highest compositional dynamics were observed with expansion of IFS 2, 3, 5 and 6. Late MI remodeling displayed more similar characteristics to the AngII model with increase of IFS 3 and 2. To demonstrate that these compositional shifts impacted the disease signatures I mapped the disease signatures to IFS state markers (Figure 2.6C). I found that IFS 0 shared markers with the HFpEF signature while IFS 3 with AngII, late and early MI signatures, IFS 2 and 6 with early MI signature only. This suggested that a compositional shift was apparent in all mouse models, however, different emphases of IFS between HF models was evident. Interestingly no other model shared the importance of IFS 0 with HFpEF, which could possibly be an important and unique feature of metabolic cardiac fibrosis.

To investigate possible transcriptional shifts, I addressed whether fibroblasts from IFS that did not share disease signature genes or increase in composition nevertheless contributed to the remodeling. I calculated gene set scores for the different disease signatures for all fibroblasts and quantified via AUROCs how well cells within the same IFS could be distinguished regarding their control and disease label (Figure 2.6D). In general, all AUROCs were higher than 0.5. This indicated that the signatures were increasingly expressed across IFS and thus a characteristic of a transcriptional shift was apparent in every study model. However, the highest median AUROCs were achieved in the MI (early and late) models while HFpEF and ANGII models displayed a less pronounced transcriptional shift. This might be explained by the acuteness and intensity of tissue stress after MI as opposed to the chronic stimuli of AngII administration and HFpEF diet. In addition, a different IFS responsiveness could be observed: While in HFpEF IFS 7, 0 and 4 display highest AUROCs, the highest transcriptional shift in the other models could be found consistently in IFS 3 and 7 (Figure 2.6D). Furthermore, IFS 5 fibroblasts were the least responsive IFS in all study models and thus probably less relevant for disease related remodeling and could fulfill rather homeostatic function.

I concluded that those IFS that displayed i) state marker overlap and ii) compositional increase and iii) displayed a high within-state-transcriptional shift, represent the crucial fibroblast phenotypes in each study. Those states were IFS 0 in HFpEF, IFS 2, 3 and 6 in early MI and IFS 3 in late MI and AngII. However, besides this prioritization, all IFS partook in the cardiac remodeling and displayed transcriptional shifts.

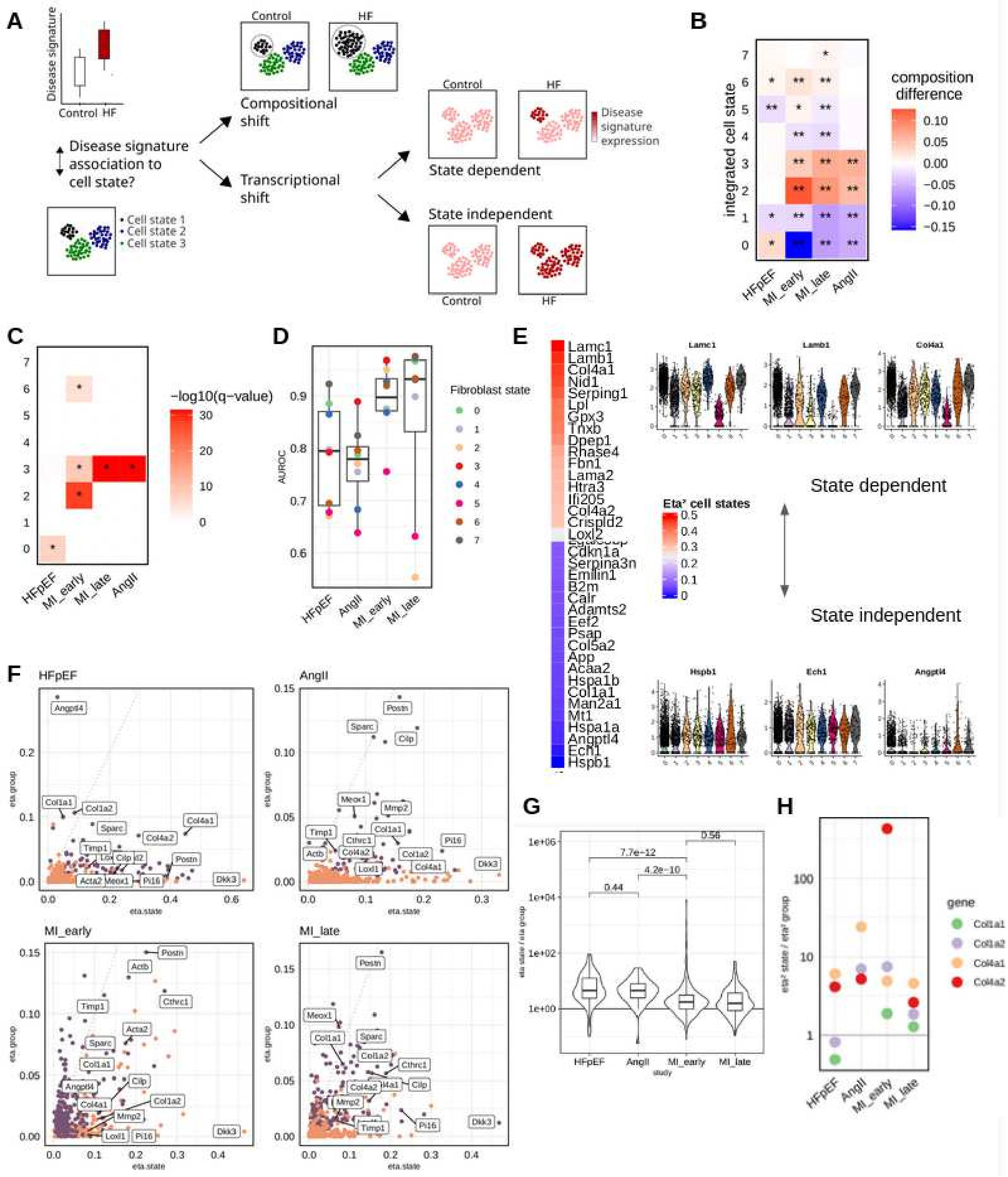


Figure 2.6 Transcriptional shifts in cardiac fibroblasts.

A) Schematic of different expression patterns in regard to cell states that could yield an upregulation of a disease signature. Compositional shifts by expanding cell number are distinguished from transcriptional shifts via uniform (state independent) or non-uniform (state dependent) upregulation of disease signatures. B) Composition change of IFS between control and heart failure group per study. P-values calculated via label permutation. *p < 0.05, **p < 0.01. C) Hypergeometric test of disease specific fibroblast signatures (x-axis) and top 100 IFS marker (y-axis), *p < 0.05 D) Gene set scores of study specific signatures (x-axis) were used to calculate the area under the receiver operator curve (AUROC, y-axis) between control and diseased cells within every fibroblast cell state (color). E)

HFpEF signature expression in regard to integrated fibroblast states. Explained variance (eta² values) of gene-wise ANOVAs (gene ~ cell state): Variance in single cell gene expression as explained by fibroblast state categories. Violin plots display normalized expression values of three genes with lowest (lower panel) and highest (upper panel) variance explained by cell state. F) Explained variance (eta² values) by cell state on x-axis and explained variance by disease class (gene ~ disease class) on y-axis. Violett dots are part of the disease signature. G) Quantification of differences in state dependent regulation of disease signatures across heart failure models. The ratio of the explained variance by cell state and disease class was calculated for each HF model and its disease signature. Wilcoxon test p-values are shown. H) The ratio of explained variance by state and disease class compared for collagens I and IV.

After establishing commonalities and differences regarding transcriptional and compositional shifts in the HF models, I next aimed to decompose disease signatures regarding their state dependency. This could help to elucidate for a given gene whether it is expressed in state dependent or state independent manner (see Figure 2.6A). Biologically, this might characterize gene programs that represent IFS related functions and those that are general cell responses. To quantify this dependency, I fit ANOVA models for each gene of the disease signatures, by modeling their expression value by IFS category and calculated the explained variance (eta^2 values) of those models (Figure 2.6E). In HFpEF, I found that genes related to the basement membrane (Lamc1, Lamb1, Col4a1, Nid1) were expressed rather state dependent, while metabolism associated (Angptl4, Ech1, Man2a, Acaa2) and fibrosis associated genes (Col1a1, Col1a2, Timp1, Mmp1) were rather state independently expressed.

To compare these state dependent expression patterns between studies, I now calculated eta^2 values from ANOVAs for state and group labels for all studies separately (Figure 2.6F). Known state markers like DKK3 (IFS 5) or Pi16 (IFS4) displayed high state dependency and low group dependency in all studies, serving as examples of genes that were state markers but without disease involvement. POSTN displayed high group and state related variability in all models except HFpEF, and thus represented a state marker with high disease association, further highlighting the important role of IFS3 in non-HFpEF fibrosis. I calculated the ratio of variance explained by state and group to compare state dependent expression between studies (Figure 2.6G). The two chronic models (HFpEF & AngII) again displayed a more state-dependent transcriptional shift compared to the MI (late & early) fibroblasts (Wilcoxon test, p < e10).

To compare fibroblast activation for single genes of interest I selected the basement membrane related Col4a1 and Col4a2 which displayed high state dependent expression in all models (Figure 2.6H). Interestingly, main collagens of the ECM (Col1a1, Col1a2) were expressed state-dependently in AngII and MI (early & late) models but were state independently expressed in HFpEF (Figure 2.6G). This could indicate that fibrosis due to collagen I deposition in HFpEF might not be related to state composition shift (such as differentiation to IFS3) but remained a global fibroblast task.

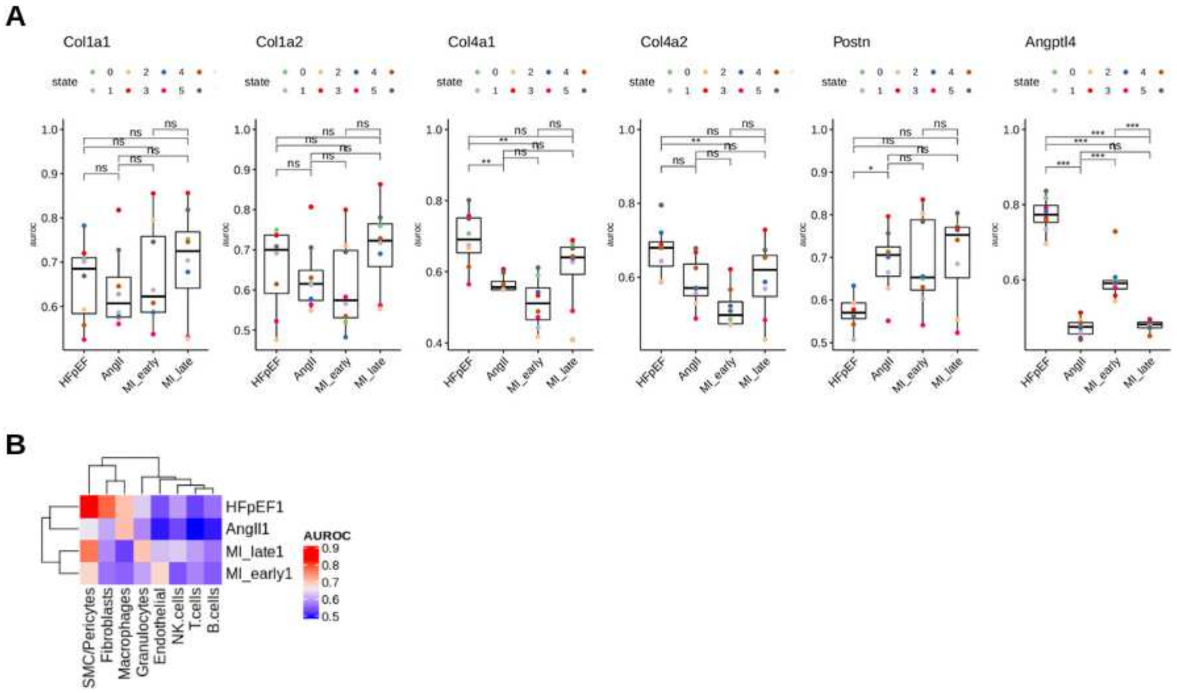


Figure 2.7 Single gene expression pattern across IFS.

A) Quantification of IFS separability with AUROC based on expression of Col1a1, Col1a2, Col4a1, Col4a2, Postn and Angtpl4 compared between models. B) Assessing possible background expression. Fibroblast signatures (y-axis) are used to test separability within the HFpEF data set of other cell types by calculating AUROCs.

Lastly, I assessed whether some of the discussed key genes that were state-dependently expressed were also part of a transcriptional shift, by calculating AUROCs for within state regulation of single genes (Figure 2.7A). I found that Col1a1 and Col1a2 were in all models part of a transcriptional shift, showing highest upregulation within IFS3 in non HFpEF models. Col4a1 and Col4a2, although state dependent expressed in all models, displayed a high transcriptional shift in most IFS in HFpEF. This further elucidated that genes that were expressed in a division of labor between fibroblasts (such as collagen IV in IFS 0 or collagen I in IFS 3) were also upregulated by other IFS in the respective disease context. In addition, Angptl4 displayed low state dependent variance and a high transcriptional shift in all IFS in HFpEF, possibly rendering it a key candidate for general metabolic fibroblast stress.

Differential gene expression analysis could be confounded by background gene expression that could be associated with increased cell dissociation in diseasesed tissue affecting contrast comparison. To ensure that the discussed disease signatures were not confounded by background expression, I tested whether other cell types in our single cell data could be separated on the basis of these genes (Figure 2.7B). Indeed, low AUROCs were found for all cell types except SMC/Pericytes which were very few cells and thus I interpreted this as a possible noise signal. Furthermore, the discussed low correlation and overlap of signatures between HFpEF and other HF signatures (section 2.3.2) was again demonstrated here by showing that other disease signatures failed to separate HFpEF fibroblasts from control.

In conclusion, I described fibroblast activation as a mixture of compositional and transcriptional shifts in all HF models. However, in the MI models, acute tissue remodeling was associated with a stronger transcriptional shift than in chronic remodeling induced via AngII or L-NAME/HFD. The compositional shift lead to a prioritization of IFS and their functional niche. The prioritized IFS upregulated their state markers in the disease setting (See POSTN in Figure 2.7). However, cells from non-prioritized IFS too upregulated those patterns, but to a lesser extent (see collagen IV in Figure 2.7). I found that some genes were state-dependently or independently expressed. I decomposed the HFpEF signature by extracting genes which were part of a general fibroblast response including metabolic genes and protein stress genes. Furthermore, the division of labor for crucial collagen I synthesis was shifted from a general fibroblast task in HFpEF to an IFS task in the other models which was possibly associated with the extent of observed fibrosis in tissue staining.

2.4 Corroborating fibroblast signatures in human data

2.4.1 Fibroblast signature detection in human HF

I have identified key disease related gene programs in the murine HFpEF fibroblasts, and associated IFS 0 as the most responsive fibroblast state in HFpEF. To investigate whether these patterns constituted a detectable disease signal in human heart failure, I curated myocardial bulk transcriptomic signatures acquired from HFrEF and HFpEF patients. For HFrEF, I relied on the HF-CS that I derived in chapter I. For HFpEF there were very few public data sets available which represents a major challenge in HFpEF research. I re-analyzed 5 patients that underwent coronary artery bypass graft surgery and met the echocardiographic and diagnostic criteria of HFpEF[193]. I selected top upregulated genes from both bulk resources and performed overrepresentation analysis with the fibroblast disease signatures (Figure 2.8A, left panel). The murine AngII and late MI signatures displayed a highly significant overlap with the human HFrEF bulk reference, while murine HFpEF signatures were moderately enriched in the human HFpEF bulk reference. Next, I addressed whether this overlap of disease signals between mouse and human could also be recovered for IFS markers (Figure 2.8A, right panel). I found the most significant overlap of IFS 3 with human HFrEF and of IFS 0 with human HFpEF. This suggested that the presented fibroblast signatures from AngII, MI and HFpEF models are partially conserved across species (mouse to human) as well as across data modalities (single-cell to bulk RNAseq).

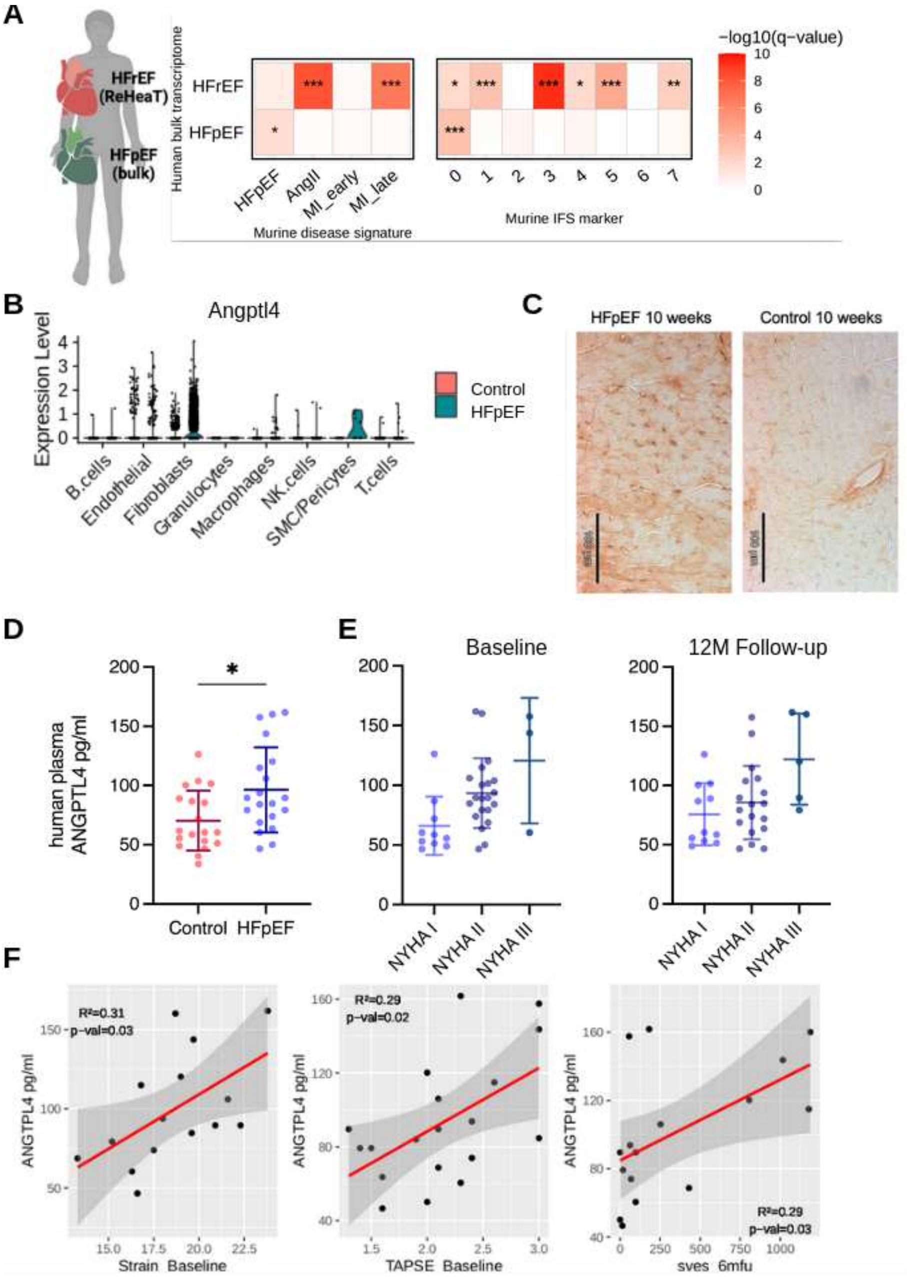


Figure 2.8 Corroborating findings in human data.

A) Corroboration of murine fibroblast signatures in human myocardial samples. Human HFpEF and HFrEF studies were curated and top differentially upregulated genes were selected (y-axis). Gene set intersection with fibroblast disease signatures from different study models (left-panel) or fibroblast state marker (right panel) (Hypergeometric test). AngII= angiotensin II model, HFpEF= heart failure with preserved ejection fraction, MI= myocardial infarction. q-value = Benjamini Hochberg corrected p-value, *q <0.05, **q<0.01, ***<0.001. B) Angptl4 normalized gene expression among different cardiac interstitial cell types derived from the scRNAseq data showing control (left columns) and HFpEF (right columns) samples separately per cell type. C) Immunohistochemistry DAB stainings of Angptl4 in control, HFpEF 10 week-diet hearts and of the fibrotic zone 28 days after myocardial infarction (MI). D) Circulating levels of ANGPTL4 in human plasma samples of HFpEF and age matched controls measured by sandwich ELISA. n=19/20, Mann-Whitney U test, *p<0.05. E)

ANGPTL4 plasma levels in relation to the NYHA functional class of all recruited patients. ANOVA, p-value <0.05, n= 10/21/3 in baseline and n= 11/18/5 in 12 months (12M) follow-up. F) Subanalysis in the HFpEF patient collective. Correlating clinical parameters to ANGPTL4 circulating levels in HFpEF patients (n Strain = 15, n SVES = 19, n TAPSE =16) with simple linear regression. Strain= global longitudinal strain, SVES= supraventricular extrasystoles, TAPSE=tricuspid annular plane systolic excursion. Plots in C,D, E display mean±SD. Panels C, D and E were generated by Laura Wienecke.

2.4.2 Angptl4 as a possible biomarker for HFpEF

Angptl4 is functionally linked to inflammation, metabolism and fibrosis. Angptl4 was induced with the highest fold change regulation in HFpEF fibroblasts and is regulated by PPARa and PPARy transcription factors, which were both predicted in HFpEF fibroblasts. Furthermore, I identified Angptl4 as a gene expressed as part of a state independent transcriptional shift in HFpEF fibroblasts. When compared to other cell types in our data set, only ECs displayed moderate Angptl4 expression (Figure 2.8B). To confirm Angptl4 accumulation in cardiac tissue, we confirmed its protein expression in the murine HFpEF model (Figure 2.8C) and compared it to late remodeling after MI, as a proxy for cardiac remodeling in HFrEF, where no significant Angptl4 levels were found (data not shown).

Thus, we hypothesized that Angptl4 might be a promising candidate to be involved in HFpEF pathophysiology and evaluated whether Angptl4 could serve as a biomarker detectable in human plasma. We analyzed circulating levels of Angptl4 in 20 plasma samples of HFpEF and 20 non-HFpEF (control) patients. All patients were diagnosed for symptomatic atrial fibrillation and screened for HFpEF by echocardiography, stress echocardiography, NT-proBNP, and HFA-PEFF-score[194]. Plasma samples were analyzed by ELISA, which revealed significantly higher circulating Angptl4 levels in HFpEF (Figure 2.8D). Angptl4 levels increased significantly in higher NYHA stages in all patients (Figure 2.8E). A subanalysis in HFpEF patients revealed that high ANGPTL4 levels correlated positively to global-longitudinal strain and TAPSE as markers of left and right ventricular function in HFpEF patients (Figure 2.8F). In addition, counts of supraventricular extrasystoles in holter ECGs at 6- and 12- (p=0.014, r=0.618) months follow-up were related to ANGPTL4 at baseline (prior to atrial fibrillation ablation) in HFpEF patients exclusively (Figure 2.8F).

2.5 Discussion and Conclusion

In this study, I provided a first comprehensive characterization of the interstitial cardiac remodeling in a two-hit HFpEF mouse model on single-cell level. I confirmed the murine resemblance of critical HFpEF phenotype features and systematically assessed transcriptional disease response on a cell type level. Deterioration of cardiac diastolic function was accompanied by increased interstitial fibrosis as assessed by histology. This phenotype was associated with a pro-fibrotic gene program in fibroblasts. By integrating single cell atlases of two murine HF models, I identified common and unique characteristics of fibroblast activation in HFpEF, AngII stimulation and MI. I further provided a functional interpretation of disease model specific signatures, as well as their corroboration in human transcriptome data and suggested that Angptl4 could serve as a potential biomarker for HFpEF in humans.

It can be hypothesized that among fibroblasts, different cell states correlate with tissue functions and/or spatial niches[195,196]. However, a consensus and nomenclature of cell states has not been accomplished yet, in part due to shortcomings of the concept of cell states describing a continuity and distinguishing between a more transient functional nature of a state or a cell differentiation[195]. By integrating multiple studies, I provided a catalog of main cardiac fibroblast cell phenotypes in cardiovascular disease. I linked conserved cell states to disease models and found that MI gives rise to a set of fibroblast cell states including myofibroblasts (IFS 2) and matrifibrocytes (IFS 3) together with other proinflammatory states (IFS 7 and 6). In contrast, AngII treatment mainly exerts fibrosis via matrifibrocytes. HFpEF fibrosis might differ fundamentally from respective HFrEF remodeling processes, as I found little disease signals involving matri- and myofibroblasts. Instead, I identified homeostatic IFS 0 fibroblasts to be the main cell state linked to early HFpEF fibrosis. These cells were characterized as important contributors to ECM production, especially of the basement membrane. The basement membrane represents a highly active ECM that underlies ECs, SMCs in the heart but also provides a scaffold that connects the interstitial ECM with cardiomyocytes[197]. Functionally, it plays an important role in angiogenesis, mechanotransduction and cell differentiation[198]. The role of the basement membrane in HFpEF has not been sufficiently explored yet, but its modulation of laminins has been suggested to cause gene expression changes in cardiomyocytes related to increased stiffening[199].

The not occurring myo- and matrifibrocyte activation in HFpEF has been suggested before by the evidence of ECM production in the absence of TGFβ signaling and presence of metabolic stimulation[14]. Our here presented data further support this hypothesis by demonstrating that no relevant FAP protein expression and FAPi PET-CT tracer uptake was observed in HFpEF hearts, whereas FAP was previously described as a marker of fibroblast activation in acute MI and AngII/PE [200,201]. In addition, I confirmed the relevance and specificity of the described fibrotic signatures including IFS 0 state markers in myocardial transcriptomes of human HFrEF and HFpEF. These results suggested the capability of mouse models to partially mimic human disease.

Besides compositional shifts, I found a strong expression shift in most fibroblast states in HFpEF, suggesting that the defined disease signatures are in part a general cellular response that is executed by many fibroblasts across cell states. I reported that these expression shifts as previously described [54] were also found in other heart failure models, suggesting that composition shifts alone do not explain fibroblast activation in cardiac fibrosis. In HFpEF, the gene program associated with an expression shift was characterized by common fibrotic response and metabolic genes including ANGPTL4 while the basement membrane related genes were more state specific gene programs and subsequently associated with the composition shift.

ANGPTL4 is linking inflammatory, metabolic and fibrotic mechanisms mainly by acting as a secreted protein, but it also controls intracellular lipoprotein metabolism and energy homeostasis by inhibiting lipoprotein lipase in divergent tissues[202]. The expression is induced by fasting and hypoxia under the control of several transcription factors including PPARs, glucocorticoid receptors and HIF1α. ANGPTL4 levels were previously reported to be associated with the risk of coronary artery disease, atherosclerosis and type 2 diabetes[203,204]. I observed an upregulation of ANGPTL4 in cardiac fibroblasts during HFpEF, next to activity of transcription factors Ppara and Pparag. In a patient collective with atrial fibrillation, ANGPTL4 correlated with functional capacity, HFpEF disease development and burden of supraventricular extrasystoles, but correlated positively with global-longitudinal strain. ANGPTL4 might exhibit beneficial or detrimental functions as reported in HFrEF and atrial fibrillation models[203,205]. Thus, further research is necessary to illustrate its mechanistic role for HFpEF.

The main limitations of our study related to the sample size of the single-cell experiment. Subtle disease changes, such as gene programs occurring in more rare cell types or cell states were probably not detectable. At the same time, our statistical approach for differential expression analysis might result in a higher rate of false positives than more robust approaches that rely on higher sample size. Our study design focused on early changes of the HFpEF remodeling. As a longer dietary regimen leads to further disease progression, we cannot provide insights into potential dynamics of the reported cellular disease signatures. Thus, a potential role of matrifibrocytes during later stages of this HFpEF model was not addressed. However, validation in human bulk RNA-Seq demonstrated that matrifibrocyte markers were upregulated in human HFrEF, but not in HFpEF patients. Additional validation of these findings in large human HFpEF studies could not be accomplished, due to the small number of publicly available data sets of gene and protein expression in human HFpEF.

Common fibrotic pathways are active across pathologies, organs and species and include hallmark signaling mediated by TGFβ, integrins, cytokines and vasoactive substances[206], resulting in increased ECM production and reparative tissue replacement. Pharmacomodulation of these major fibrotic axes has been mainly unsuccessful in the past, partially because of their fundamental impact on global tissue homeostasis. As more details about differences in fibrotic signaling and fibroblast phenotypic heterogeneity are accumulating[177,207], better targeted antifibrotic therapies might come within reach[19]. In cardiac fibrosis, inhibition of RAAS is a crucial treatment option in HFrEF but not HFpEF[208,209]. Our study might provide important insights into the mechanism of this lacking therapy response in HFpEF patients. In parallel, a mutual activation occurs in the cross-talk with fibroblasts. We observed in this context especially an expansion of the Col15a1+ pool that might contribute to a pattern of interstitial fibrosis along basement membrane scaffolds. Future research is necessary to further delineate these fibrotic mechanisms and identify potential targets for HFpEF therapy.

Taken together, this is the first description of adverse interstitial remodeling in HFpEF on a single-cell level. My work provides new insights into distinct and common features of cardiac fibrosis in heart failure and might serve as a valuable source for the scientific community to identify disease specific treatment strategies for HFpEF in the future.

Chapter III - Comorbidity space of HF patients

3.1 Background

HF patients suffer from a wide range of comorbidities, which are considered important for HF development and progression [90]. In the pathogenesis of HFpEF, comorbidities have been suggested as causal factors [7,13] and could possibly be linked to genetic etiology. Treatment of comorbidity has also been shown to have beneficial effects of cardiac physiology [91], emphasizing the potential to address HF subtypes through their comorbidities. However, the modeling of multiple diseases is currently not attainable in animal models to circumvent assessment of patients.

Systems medicine attempts to model disease in a holistic manner. One facet of this, network medicine, is used to analyze complex systems such as patients, organs or cells via network representation [92,93]. Comorbidity networks represent diseases as nodes, connected via edges based on co-occurrence in patients. These networks can be used to define disease modules or explore topological changes between patient cohorts [94–97]. Previous work has shown that disease comorbidity is also often linked to shared disease genes that locate close together in gene-based networks like protein-protein interaction networks [97,101]. This observation is often the basis of network-based gene prediction, where novel disease genes are predicted based on network proximity to known disease genes.

Cardiovascular diseases are particularly suited for system medicine approaches due to the typical multi-organ involvement [210] and multifactorial etiology [211]. To date, such approaches to study HFpEF have been limited, though the comorbidity driven pathophysiology of HFpEF makes it a promising subject. In addition, despite the technological advances in multi-omics, functional genomics knowledge of HFpEF remains limited, possibly due to difficulties of biopsy acquisition in HFpEF patients [12] and heterogeneity of HFpEF patients [212].

In this chapter, I applied a network medicine approach to describe comorbidity patterns in HFpEF and investigate a genetic background associated with these patterns. I first demonstrated that comorbidity profiles vary between HFpEF and HFrEF patients and derived distinct comorbidity profiles for each cohort. Then, I built a comorbidity network that contained disease clusters relevant for HF patients. The construction of a multilayer heterogeneous network by integration of prior knowledge resources allowed me to translate the comorbidity profiles into a gene signature for HFpEF. I then corroborated this signature in the cardiac transcriptome of a murine HFpEF model. This network medicine approach allowed me to identify distinct comorbidity profiles and novel genetic patterns in HFpEF.

I conceived, conceptualized and implemented all analysis and figures in this chapter. Hauke Hund provided access to the data warehouse and Rebecca Levinson supervised this project. Alberto Valdeolivas provided feedback and helpful discussion on graph analysis. This chapter is part of a manuscript that is currently in revision in a peer reviewed journal [213]. I wrote the manuscript while Julio Saez-Rodriguez, Rebecca T. Levinson and Alberto Valdeolivas provided edits.

3.2 Defining the study population and their comorbidities

3.2.1 The study population

I derived the study population from a research data warehouse containing data from patients that visited the Department of Cardiology, Angiology, and Pneumology at Heidelberg University Hospital, Heidelberg, Germany [214]. Heidelberg University Hospital acts as a tertiary care center for the surrounding region, specializing in the treatment of cardiomyopathy. From this data warehouse I identified patients with HF, visiting between 01.01.2008 and 01.01.2021. The study protocol was approved by the local ethics committee. I defined HF as two or more HF-relevant International classification of disease, version 10 (ICD-10) diagnosis codes (I50*, I11.0, I13.0, I13.2, I42.0, I42.5, I42.8, I42.9, I25.5) or at least one HF-relevant diagnosis and at least one of the following criteria: i) elevated N-terminal pro b-type natriuretic peptide (NTproBNP) (>120 ng/ml), ii) recorded New York Heart Association functional class, iii) echocardiography based E/e’ >15 (ratio of early diastolic mitral inflow velocity to early diastolic mitral annulus velocity), iv) echocardiography or MRI-based left ventricular ejection fraction (LVEF) <50%, v) documented loop diuretic. Patients with HF before age 40, those with a diagnosis of inheritable cardiomyopathy (I42.1-I42.4, I42.6, I42.7), and heart transplant patients (Z94.3) were excluded from the HF cohort. Within the HF cohort, I identified HF subtypes, based on echocardiographic or MRI-based LVEF. Patients with LVEF >= 50% were labeled HFpEF, LVEF 40-50% HFmrEF (Heart failure with mid range ejection fraction) and =<40% HFrEF (Figure 3.1).

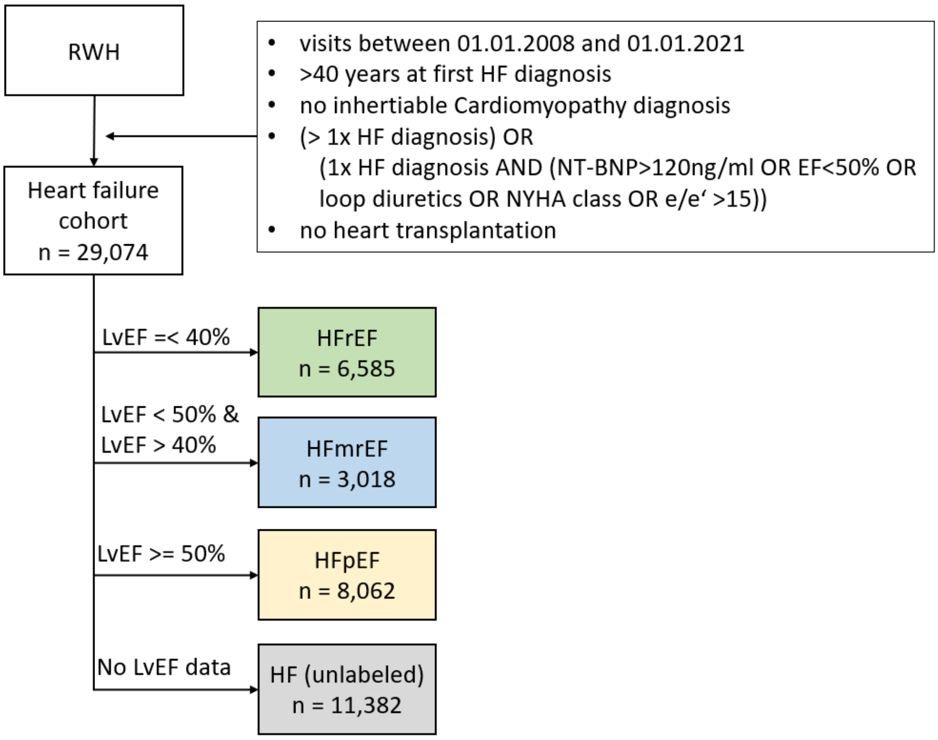

RWH
visits between 01.01.2008 and 01.01.2021
>40 years at first HF diagnosis
no inhertiable Cardiomyopathy diagnosis
(> 1x HF diagnosis) OR
(1x HF diagnosis AND (NT-BNP>120ng/ml OR EF<50% OR loop diuretics OR NYHA class OR e/e‘ >15))
no heart transplantation
Heart failure cohort n = 29,074
LvEF =< 40%
HFrEF n = 6,585
LvEF < 50% & LvEF > 40%
HFmrEF n = 3,018
LvEF >= 50%
HFpEF n = 8,062
No LvEF data
HF (unlabeled) n = 11,382

Figure 3.1 Patient Cohort description.

Phenotyping algorithm to define HF cohorts. HF patients were selected with hospital visits over a time span of 13 years at the University Hospital Heidelberg. I defined a general HF cohort by selecting patients with either two or more HF relevant ICD-10 codes or one HF relevant ICD-10 code and one additional HF relevant clinical characteristic, yielding 29,047 HF patients. Based on LvEF I subclassified HF patients to HFrEF, HFmrEF or HFpEF. RWH, Research Data Warehouse; HF, Heart failure; LvEF , left ventricular ejection fraction; e/e’ is the ratio between early mitral inflow velocity and mitral annular early diastolic velocity on echocardiography. *Reprinted from* [213]*.*

After applying the described phenotyping algorithm, my study population consisted of 29,047 patients with HF (Figure 3.1). This cohort consisted of three sub cohorts, HFpEF (8,062 patients), HFrEF (6,585 patients) and HF with mid-range ejection fraction (HFmrEF) (3,018 patients) based on LVEF. 11,382 patients in the HF cohort lacked a subtype label (HF- unlabeled).

Next, I compared available clinical features of these cohorts. HFpEF patients were more often female compared to HFrEF patients (35 vs 25%, p<0.01) (Table 3.1). However, we did not observe a significant difference in body mass index (Median [IQR] = 26.8 [24.2, 30.0] vs 26.5 [24.1, 30.1] for HFpEF vs HFrEF, p=0.9) or age (Median [IQR] = 70[61, 88] for HFpEF vs 70 [60,70] for HFrEF, p=0.5). When phenotypic data were available, cholesterol, LDL, HDL and blood pressure values were higher in HFpEF patients compared to HFrEF, while NT-proBNP values were higher in HFrEF patients. Comorbidity burden measured by Elixhauser index was slightly lower in HFpEF than HFrEF patients, as previously reported [215]. HFpEF patients were intubated (8.5% vs 15%, p<0.001) or received an implantable cardioverter-defibrillator (16% vs 26%, p<0.001) less frequently than HFrEF patients, suggesting that the HFrEF cohort was at a later stage of HF.

			HF Subtypes				
Variable	N	Overall	HFpEF	HFmrEF	HFrEF	p-value*	p-value+
		(N= 17,665)	(N=8,062)	(N=3,018)	(N=6,585)		
Sex	17,617					<0.001	<0.001
Female		5,247 (30%)	2,822 (35%)	790 (26%)	1,635 (25%)		
Male		12,370 (70%)	5,228 (65%)	2,218 (74%)	4,924 (75%)		
Age (years)	17,665	70 (60,78)	70 (61, 78)	70 (59,77)	70 (60, 78)	0.093	0.5
BMI	9,132	26.8 (24.2, 30.0)	26.8 (24.2, 30.0)	26.9 (24.2, 29,9)	26.5 (24.1, 30.1)	0.08	0.9
Systolic BP(mmHg)	5,146	148 (134, 160)	150 (139, 164)	148 (135, 160)	140 (127, 154)	<0.001	<0.001
Diastolic BP (mmHg)	5,146	84 (76, 92)	85 (78, 93)	84 (76, 93)	82 (73, 90)	<0.001	<0.001
LDL (mg/dL)	12,270	87 (69, 110)	88 (69, 110)	91 (72, 113)	84 (67, 106)	<0.001	<0.001
HDL (mg/dL)	12,368	44 (36, 54)	46 (38, 56)	43 (36, 53)	40 (34, 50)	<0.001	<0.001

Triglycerides (mg/dL)	13,859	112 (85, 153)	112 (85, 151)	112 (85, 156)	113 (85, 156)	0.11	0.006
Cholesterol (mg/dL)	13,577	160 (135, 188)	163 (138, 190)	164 (140, 192)	153 (129, 183)	<0.001	<0.001
N PheCodes	17,665	13 (8, 21)	13 (9,21)	12 (7, 19)	14 (9, 22)	<0.001	0.088
Elixhauser Index	17,665	5.39 (2.72)	5.36 (2.68)	5.09 (2.70)	5.56 (2.76)	<0.001	<0.001
Intubated	17,665	1,766 (10.0%)	552 (6.8)	257 (8.5%)	957 (15%)	<0.001	<0.001
ICD Implantation	17,665	3,213 (18%)	1,007 (12%)	468 (16%)	1,738 (26%)	<0.001	<0.001
PCI	17,665	9,116 (52%)	4,267 (53%)	1,554 (51%)	3,295 (50%)	0.002	<0.001
log(NT-BNP)	6,169	2.99 (2.45, 3.53)	3.07 (2.53, 3.55)	3.07 (2.53, 3.55)	3.45 (2.96, 3.88)	<0.001	<0.001
All continuous values displayed as Median (IQR) except for Elixhauser Index which is Mean (SD)							
All dichotomous values displayed as N (%)							
* kruskal-wallis p-value across all subtypes							
+ Wilcoxon-rank sum or chi-squared p-value for HFpEF vs HFrEF							

Table 3.1. Clinical characteristics of HFrEF, HFmrEF and HFpEF cohorts.

Descriptive statistics of HFrEF, HFmrEF and HFpEF cohorts. F, female; m, male; BMI, body mass index; BP, blood pressure; LDL, low density lipoprotein; HDL, high density lipoprotein; ICD, implantable cardioverter defibrillator; PCI, percutaneous coronary intervention; NT-BNP, N-terminal pro b-type natriuretic peptide. All numerical values are median (IQR), Elixhauser index is mean (SD). *Reprinted from* [213].

3.2.2 The comorbidity space

After defining the patient cohort, I had to define the feature space of comorbidities. In the data warehouse, comorbidities were recorded in ICD-10 format. To keep the feature space spares, I summarized similar diseases by mapping ICD-10 codes to Phenome-wide association scan codes (PheWas codes). I accessed "https://phewascatalog.org/" in October 2021 and downloaded "PheCodeMap 1.2". From a total of 7,817 unique ICD-10 codes in our dataset, 3,030 could be mapped directly to PheCodes. To further improve this coverage I performed an additional stepwise mapping: the 4,787 ICD-10 codes that could not be mapped were shortened to 4 characters (e.g. N18.89 was shortened to N18.8) and mapped again to PheCodes. Unmapped ICD-10 codes after this step were shortened to 3 characters (e.g. N18.8 shortened to N18) and mapped again to PheCodes, resulting in a total of 6,676 mapped ICD-10 codes. Most frequent ICD-10 codes that were not mapped to PheCodes included codes from the Z-chapter that were considered outside the area of interest of this study and were discarded after this step. Total coverage of the mapping was 85% (mapped ICD-10 codes/unmapped ICD-10 codes) resulting in 1,481 unique PheCodes (Figure 3.2A).

Next, I selected comorbidities based on prevalence in my cohort. I calculated patient frequencies in the HF cohort for these PheCodes and only analyzed PheCodes with at least 50 patients, defining the final feature space of 569 PheCodes (Figure 3.2B).

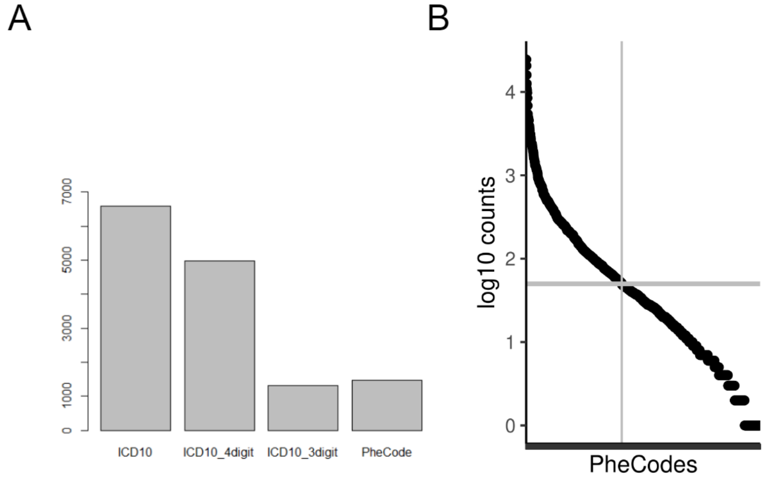


Figure 3.2 ICD10 code mapping.

A) Number of features recorded in the general HF cohort. PheCodes and 3-letter ICD10-codes reduce the feature space to a comparable feature size. B) Overview of all recorded PheCodes and their frequency (log10 transformed). PheCodes with a prevalence of at least 50 patients (horizontal gray line) resulted in 569 Phecodes (vertical gray line). These PheCodes were used in all downstream analysis. *Reprinted from* [213].

3.3 Assessing distinct comorbidity profiles between HF-cohorts

After defining the patient cohort and the comorbidity feature space, I posed the question, whether HF subcohort labels were associated with an unsupervised estimate of variance in the comorbidity space. For this, I applied multiple correspondence analysis (MCA), which provided a ranking of cohort comparisons based on associated variance in comorbidity profiles (section 3.3.1). In the next section I then derived the most distinctive comorbidities for the pronounced HFpEF/HFrEF contrast (3.3.2). Then, I investigated whether the HFpEF/HFrEF comorbidity profiles might depend on observed confounders (section 3.3.3), and finally provided a clinical interpretation (section 3.3.4). The concept of this analysis is visualized in (Figure 3.3A).

3.3.1 High variation in comorbidity profiles is associated with HFpEF/ HFrEF subtype

I expected differences in the composition of comorbidity profiles between HF subtype cohorts. To quantify this variance, I applied MCA and estimated the variance associated with sub-cohort labels and clinical features (Figure 3.3A). Disease profiles of HFpEF, HFrEF and HFmrEF cohorts were captured as binary variables (0 - patient has no record, 1 - patient has a record of disease) of 569 (PheCodes). In this feature space (569 comorbidities x 17,665 HF patients) I performed MCA (R-package *FactoMiner* [216]). Each MCA dimension was then tested for association with clinical covariates with linear regression models for binary categorical (e.g. MCA-dimension 1 ~ Sex) and continuous covariates (e.g. MCA-dimension 1 ~ Age). For each covariate I summed the variance associated to all significantly associated dimensions (p-value <0.05) as an estimate for the total associated variance.

Device implantation was the feature most strongly associated with variance in comorbidity profiles (Figure 3.3B). When comparing HF subtypes, HFpEF and HFrEF cohort labels were associated with a high degree of explained variation (39.5%). HFmrEF patients seemed to be in an intermediate state, as they displayed lower variance when compared to HFpEF (25.2%) and HFrEF (18.6%). Sex and age were each associated with high variance (37.9% and 44.4%, respectively) as expected. In summary, this analysis approach identified a pronounced contrast between comorbidities in HFpEF and HFrEF patients.

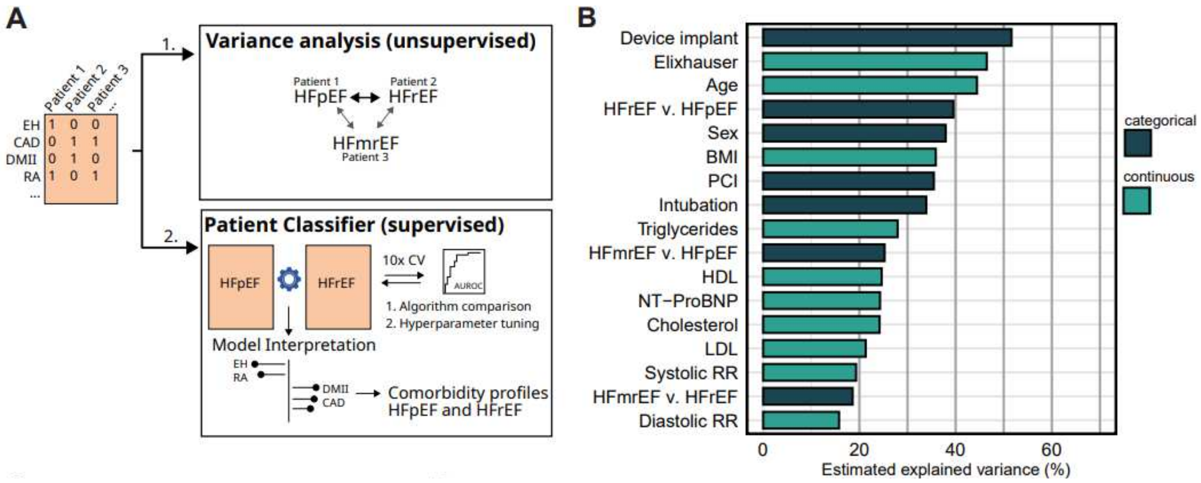


Figure 3.3 Comparison of comorbidity profiles in heart failure subtypes.

A) Scheme of analysis. EH, essential hypertension; CAD, coronary artery disease; DMII, Diabetes mellitus type II; RA, rheumatoid arthritis. B) Multiple correspondence analysis of comorbidity profiles of HFpEF and HFrEF cohort. MCA dimensions were tested for association with clinical covariates and summed up to estimate total explained variance. *Reprinted from* [213].

3.3.2 Deriving HFpEF and HFrEF comorbidity profiles

Next, to explain and interpret the variance between HFpEF and HFrEF, I derived distinct comorbidity profiles for both cohorts. For this purpose, I fit random forest and elastic net classifier models with the 569 comorbidities as features to distinguish between HFpEF and HFrEF (Figure 3.3A). Models were trained in R with R-packages *tidymodels* using model engines from *glmnet* [217] and *ranger* [218].

For hyperparameter tuning, I performed 10 fold cross validation of 90-10% training-test splits and selected hyperparameter values yielding highest mean AUROC. Hyperparameters include the ratio of L1/L2 regularization and penalty for elastic net and number of variables (mtry) and number of trees in random forest (Figure 3.4A, B). The highest achieved mean AUROCs were 0.778 for the random forest and 0.777 for the elastic net model, indicating that the random forest's ability to model more complex interactions between comorbidities did not improve classifier performance substantially. I compared the most important features and found that they were shared in both models (Figure 3.4C).

Next, because elastic net parameter estimates can provide both magnitude and direction, I selected the elastic net model to assign HFpEF and HFrEF a distinctive set of comorbidities. The elastic net model contained 196 non-zero comorbidity coefficients. To find the most discriminant comorbidities for HFpEF and HFrEF, I next performed a forward selection with an L1-regularized logistic regression model of the 196 non-zero features from the elastic net model. Finally, I found that the model with 100 comorbidities yielded a cross validated AUROC of 0.780 (Figure. 3.4D). 71 and 29 comorbidities from this model were assigned to HFpEF or HFrEF, respectively, which I will refer to as their comorbidity profiles.

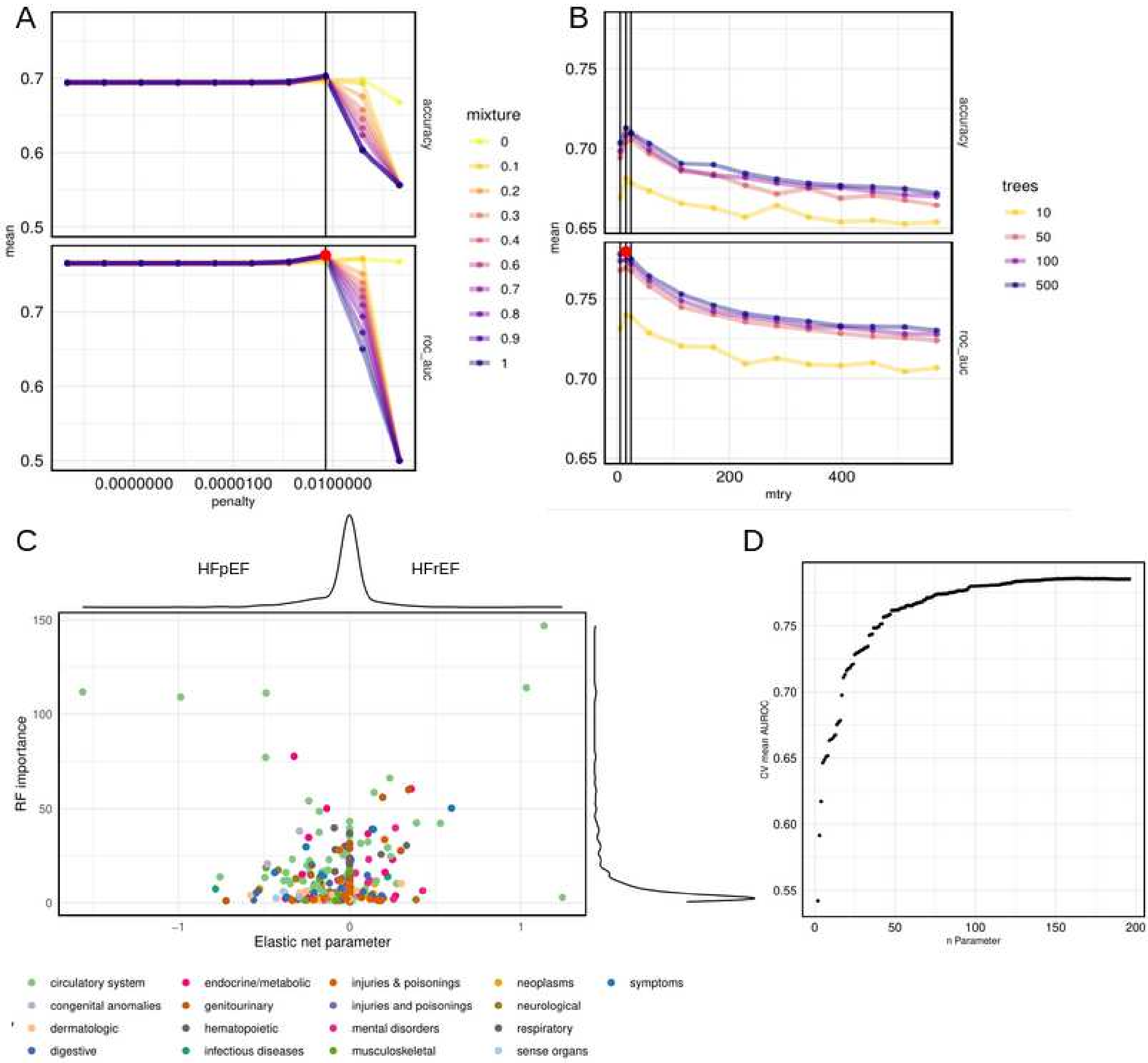


Figure 3.4 Patient classifier training.

A) Hyperparameter tuning for the elastic net model. Mixture of L1 and L2 penalty (color) and penalty value (x-axis) are compared by accuracy and AUROC. Each hyperparameter combination was assessed via 10-fold CV-splits. B) Hyperparameter tuning for random forest model. Number of trees (color) and mtry (x-axis) are compared by accuracy and AUROC. Each hyperparameter combination was assessed via 10-fold CV-splits. C) Comparison of random forest feature importance (y-axis) with Elastic net coefficient estimates (x-axis). D) Forward selection training of L1- regularized logistic regression by stepwise including parameters to the model (x-axis) and estimating a 10-fold CV AUROC (y-axis). *Reprinted from* [213].

3.3.3 Comorbidity profile assignment compared for effects of age, sex, time to HF diagnosis and time of recording

The derived comorbidity profiles might have been influenced by factors such as age, sex, time of visit or time relative to HF diagnosis. I therefore investigated whether these factors influence the assignment of these 100 comorbidities to HF stubtype by fitting a series of logistic regression models in different data subsets.

Age and Sex

I selected the comorbidity profiles of HFpEF and HFrEF (i.e. top 100 comorbidities from the patient classifier) and tested each in an independent logistic regression model while including age and sex as covariates for association with HF subtype label (HFpEF/HFrEF ~ comorbidity + sex + age). I fit these models on the full cohort data and found that the coefficients assigned to the comorbidities to be consistent with the patient classifier assignments (Figure 3.5A, column “full data”).

Time to HF diagnosis

I investigated whether the comorbidity profiles of HFpEF or HFrEF were different in regard to the time point of the patient’s first HF diagnosis. I calculated the time to HF diagnosis in months and found that most comorbidities were recorded within a year of the first HF diagnosis, which is most likely related to the nature of routine clinical care data from a tertiary care provider than with the true time point of comorbidity occurrence. Nevertheless, HFpEF patients received their comorbidities later than HFrEF patients (Wilcoxon test, p-value <0.05) (Figure 3.5B). I next subset the data for each patient to his comorbidities recorded at least six months before (pre HF) or after (post HF). In these subsets I again fitted the logistic regression model (HFpEF/HFrEF ~ comorbidity + sex + age) and found a conserved assignment of comorbidities. This indicated that comorbidity assignment to HFpEF or HFrEF was rather independent of the time point of HF diagnosis (Figure 3.5A).

Date of comorbidity assignment

Recording of comorbidities is subject to clinical practice that may change over time. I thus compared the dates of comorbidity assignments and found that HFpEF patients recorded more recent diagnosis compared to HFrEF patients (Figure 3.5C). To investigate if this difference in time of recording impacted our comorbidity profile assignment I stratified our observation window into three time blocks. I again fitted the logistic regression models (HFpEF/HFrEF ~ comorbidity + sex + age) for each time block separately and observed that the assignment of most comorbidities was mainly independent of the observation window (Figure 3.5A).

In summary, I found that the derived comorbidity profiles of HFpEF and HFrEF yielded mostly consistent patterns independent of the discussed factors.

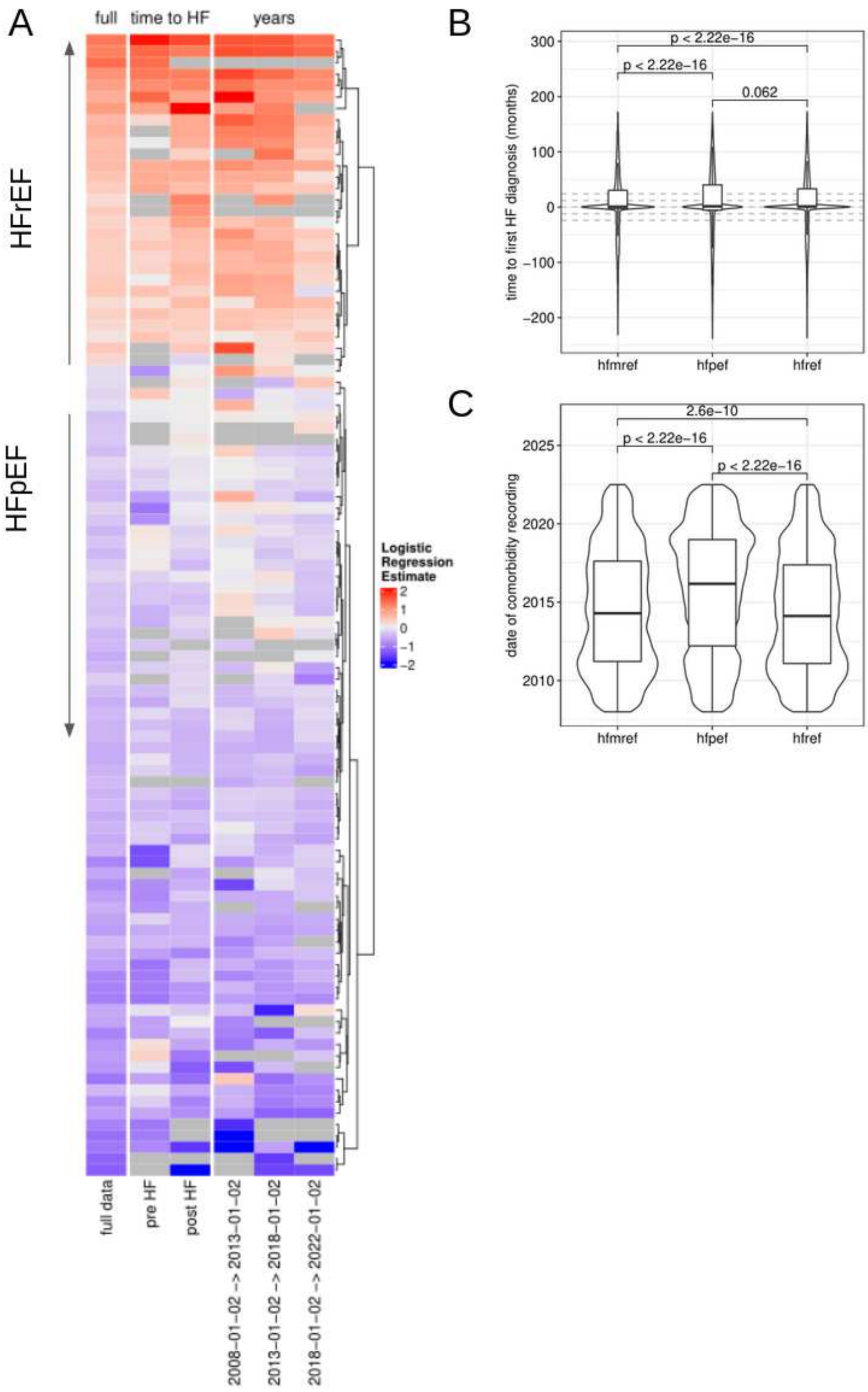


Figure 3.5 Time to HF and time of comorbidity profile assignment.

A) Comparing the comorbidity assignment to HF subtypes (y-axis) in different data subsets (x-axis). Logistic Regression models were fit without regularization for each comorbidity separately to predict HF subtype labels (HFpEF/HFrEF ~ comorbidity + age + sex). First column contains the comorbidity estimates from the full data set. Second block displays comorbidity estimates from the data subset to comorbidities with earliest diagnosis at least six months before (pre HF) or six months after the first HF diagnosis (post HF). The third block displays comorbidity estimates from data subset to three different observation windows. B) Distribution of the time in months between earliest comorbidity diagnosis and first HF diagnosis. C) Distribution of the date of comorbidity recordings between HF sub cohorts. B+C display p-values from unpaired, two sided wilcoxon test. *Reprinted from* [213].

3.3.4 Interpreting HFpEF and HFrEF comorbidity profiles

After deriving HFpEF and HFrEF comorbidity profiles and assessing influence of known confounders, I sought to interpret the profiles from a clinical perspective. The HFpEF profile (15 disease categories) was more diverse than the HFrEF profile (10 disease categories) and included comorbidities from the digestive disease, hematopoietic and neoplastic disease categories (Figure 3.6A). Cardiovascular disease was the most important category in both profiles, accounting for 48.2% of the sum of parameter estimates in HFrEF and 38.3% in HFpEF. In HFpEF, important comorbidities included hypertensive and pulmonary heart disease, essential hypertension, inflammatory cardiac conditions (pericarditis, myocarditis), sleep apnea, osteopenia, neoplasms (multiple myeloma, breast cancer, metastasis in digestive systems) and rheumatoid disorders. The HFrEF comorbidity profile was characterized among others by myocardial infarction, ischemic heart disease, tobacco abuse, mitral valve disease, coma and cardiogenic shock, neurological disorders (vascular dementia, cerebral edema), chronic kidney disease and diabetes type II (Figure 3.6B).

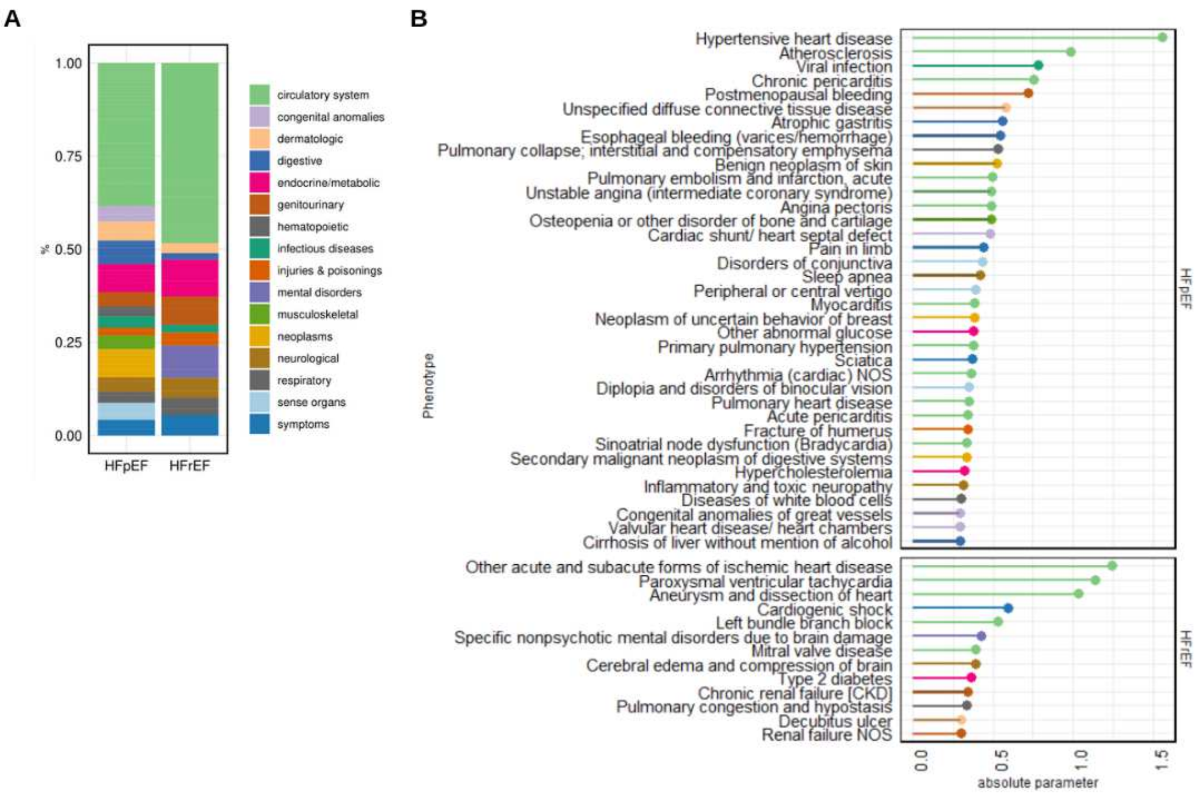


Figure 3.6 Patient classifier interpretation.

A) Proportions of the sum of parameter estimates of top 100 comorbidities of the patient classifier model, colored by disease categories. B) Top 50 comorbidities of the patient classifier. The parameters are the absolute fitted values of the coefficients in the elastic net model for each comorbidity of the patient classifier separated by association to HFpEF (top) or HFrEF features (bottom). Colors indicate disease category using the same color legend as in panel B. *Reprinted from* [213].

In conclusion, the observed variation in comorbidity profiles between HFpEF and HFrEF was analyzed by interpreting patient classifiers. The derived features captured known subtype associations such as typical etiologies of HF including hypertensive heart disease (with HFpEF) and ischemic heart disease (with HFrEF) but also more novel and understudied comorbidity associations such as breast cancer or rheumatoid arthritis with HFpEF.

3.4 The HFnet

The comorbidity profiles in the previous section describe comorbidities that were distinctive in our cohort for HFpEF or HFrEF patients. I next addressed whether the correlative association between these comorbidities can provide insights into disease occurrence for these cohorts. For this reason I built a comorbidity network as previously described [96–98,219,220]. I posed four questions which are schematically shown in Figure 3.7:

1. Does the relationship between comorbidities depend on the subcohort?
2. To what extent are disease relationships specific to my cohort?
3. Which are the central diseases/disease groups in my cohort?
4. Can I summarize correlative clusters of diseases and associate those to higher order illness concepts? And do these clusters capture cohort differences?

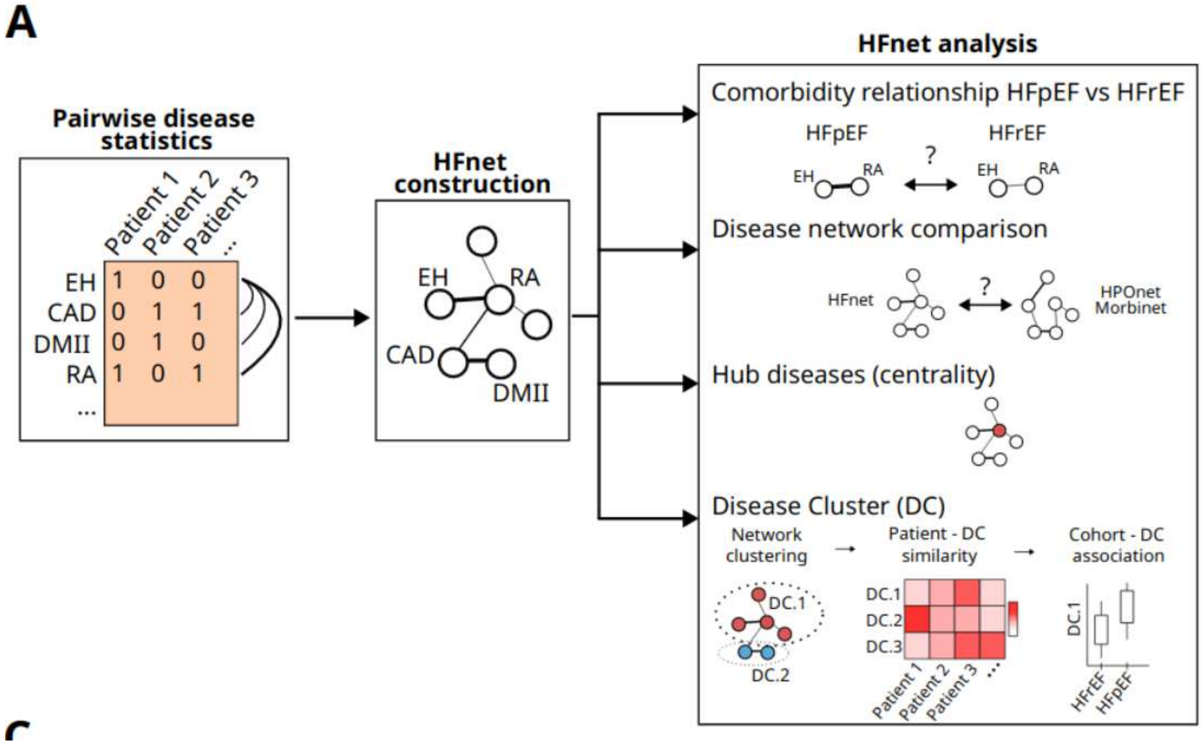
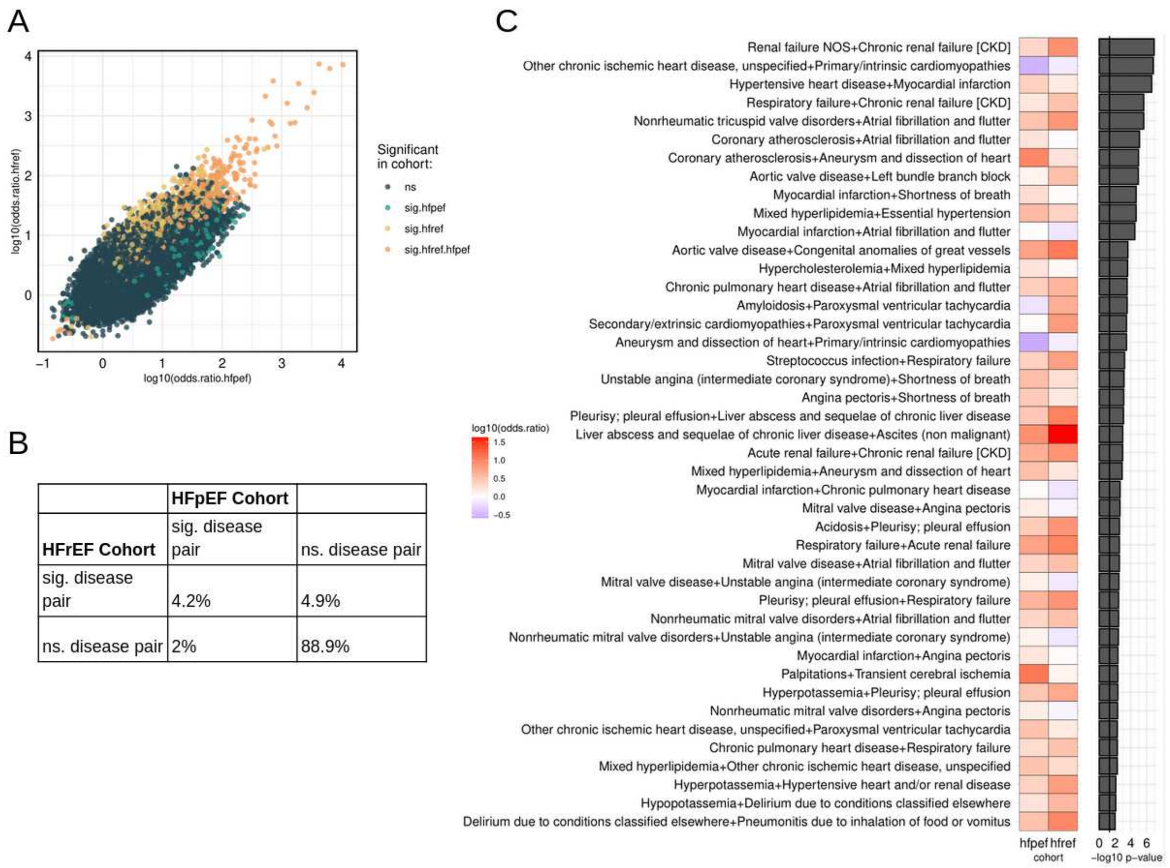


Figure 3.7 Scheme of comorbidity network analysis.

A) Scheme of comorbidity network analysis. EH, essential hypertension; CAD, coronary artery disease; DMII, Diabetes mellitus type II; RA, rheumatoid arthritis. *Reprinted from* [213].

3.4.1 Comorbidity relationships compared between HFpEF and HFrEF

While certain comorbidities were distinctive for HFpEF or HFrEF, it was unclear whether the disease relationships that built the HFnet also depend on the HF subtype. When comparing odds ratios for each disease pair from both cohorts, I found a high concordance (Figure 3.8A, B). Only 33 disease pairs had significantly different odds ratios between HFpEF and HFrEF (Breslow-Dayes test for Homogeneity of odds ratio with BH correction $p<0.01$) (Figure 3.8C), suggesting that in the vast majority of cases, the co-occurence of two diseases did not depend on whether it was assessed in HFpEF or HFrEF patients.



	HFpEF Cohort	
HFrEF Cohort	sig. disease pair	ns. disease pair
sig. disease pair	4.2%	4.9%
ns. disease pair	2%	88.9%

Figure 3.8 Comparison of comorbidities between HFpEF and HFrEF cohort.

A) Pairwise disease odds ratios and Fisher's exact test p-values (with BH correction) were calculated for HFpEF and HFrEF patient cohorts separately. Pearson correlation of odds ratios between both cohorts was ~1 with p<0.01. B) Comparison of all tested disease pairs at fisher test p<0.0001. C) Breslow dayes test for homogeneity of odds ratios was applied and significant disease pairs (p< 0.01) are shown with log odds ratios for each cohort (heatmap) and log transformed corrected p-value (barplot). *Reprinted from* [213].

3.4.2 HFnet construction and comparison

After finding that disease relationships between HFpEF and HFrEF were highly concordant, I decided to use the full HF-cohort to learn correlative disease structures.

In disease comorbidity networks, nodes represent diseases while edges represent statistical association of co-occurrence, resulting in the graphical depiction of comorbidities as diseases that are statistically dependent. In detail, I selected edges using Fisher's exact test for estimating statistical dependence and its Benjamini-Hochberg (BH) corrected p-value ($p < 0.0001$) to discard non-significant disease pairs and keep a more sparse network structure. To determine strength of association I calculated ϕ-correlation, which can be interpreted as a Pearson correlation for binary variables. I selected all edges with positive correlation (Figure 3.9A). To account for bias in ϕ-correlations towards high frequent diseases, I scaled the values by dividing by mean correlation values for every disease and assigned these values as edge weights [221] (Figure 3.9B).

The resulting significant disease - disease relationships were assembled to form an undirected and weighted heart failure comorbidity network (HFnet) consisting of 569 nodes and 19,347 edges (Figure 3.9C), with edge weights defined by a statistical dependency of co-occurrence for each disease pair.

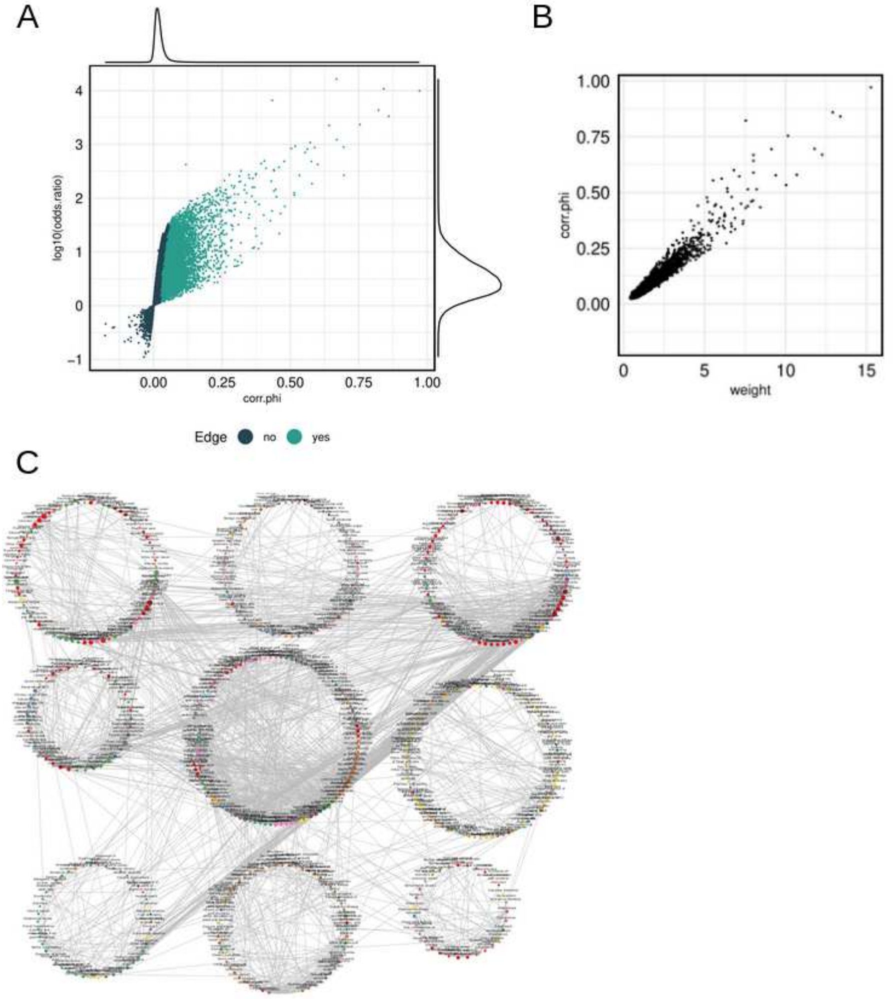
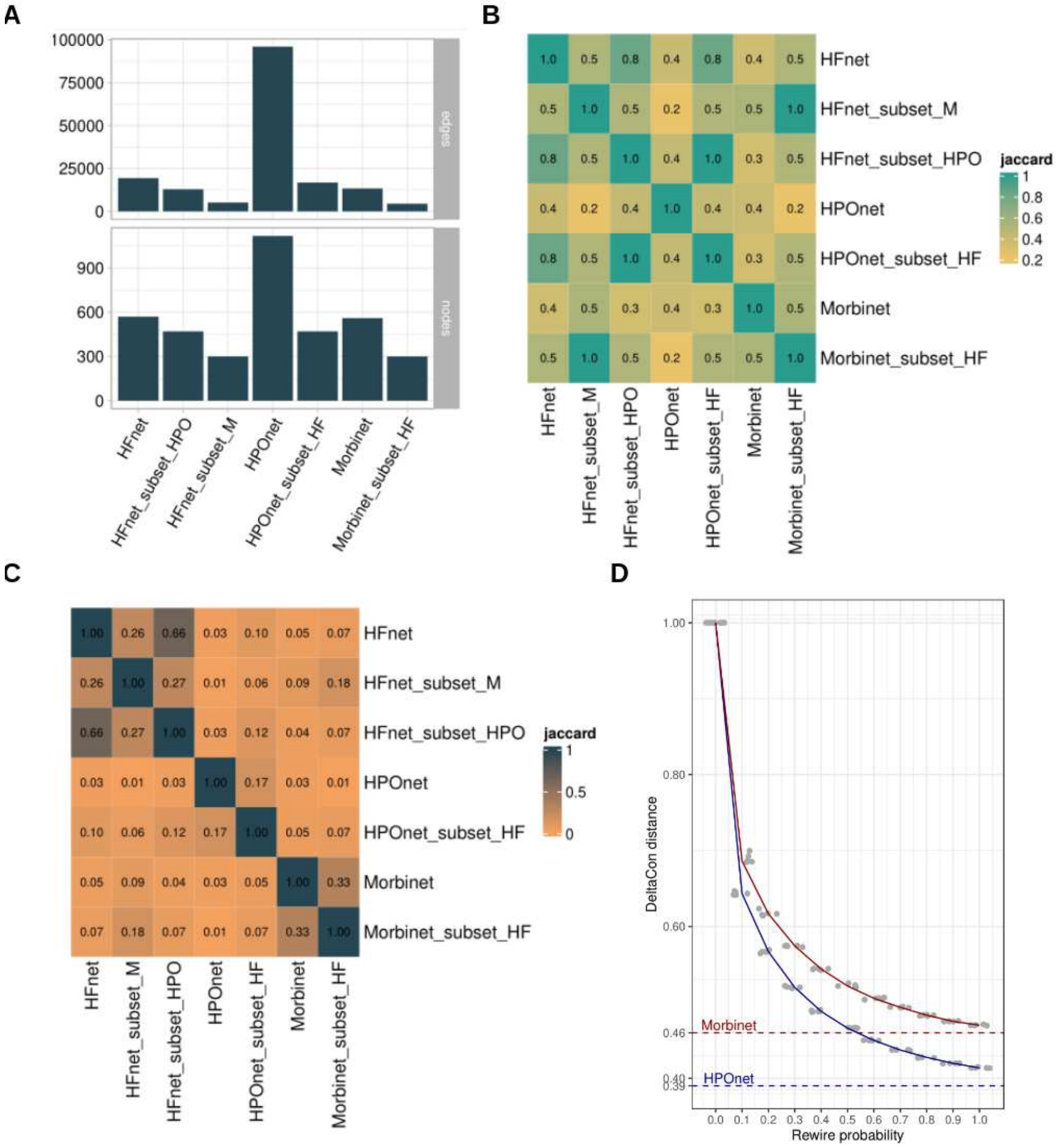


Figure 3.9 HFnet overview.

A) Phi correlation and log odds ratio of pairwise disease comparison in the general HF cohort. B) Adjustment of phi-correlation. Correlation coefficients were scaled by disease. As low prevalence diseases are expected to result in lower phi-correlation coefficients and scaling by mean coefficient can address this effect. The calculated weights were used as edge weights in the HFnet. C) HFnet plotted and clustered by disease cluster (DC). DCs are arranged in rows (e.g first row : DC1, DC2, DC3). *Reprinted from* [213].

My first question was, whether the HFnet constituted a unique wiring of diseases or predominantly captured generalizable disease relationships. To investigate this I analyzed and compared two additional disease networks: a human phenotype ontology network (HPOnet) where two diseases are connected if they are phenotypically similar and Morbinet [96], another comorbidity network from a large patient cohort but without a cohort defining disease.
I downloaded network data from Morbinet (https://shiny.odap-ico.org/morbinet, accessed October 2021), using a threshold of OR>1 and fisher p value of p<0.01 and mapped the ICD-10 disease ontology to PheCodes. Then, I built a phenotypic disease network, where two diseases are connected, if they share a similar phenotype based on the human phenotype ontology (HPO). To construct this network, I downloaded the HPO from https://hpo.jax.org/app/ (accessed October 2021) and mapped disease ontologies to Phecodes. Disease similarity was calculated with Lin's methods implemented in the *OntologySimilartiy* R-package [222]. The full distance matrix resulting from the ontology similarity was used to create a fully connected network with edge weight representing ontology similarity. From this network I extracted the backbone [223] with the implementation in the *corpustools* R-package. This backbone extraction is based on an assumed null distribution of local edge weights where based on an alpha level (here 0.05) edges can be extracted that are unlikely to fall into that distribution. After subsetting each network to the same nodes as the HFnet, the number of edges was comparable (Figure 3.10A).

Now I could compare the three disease networks. Jaccard index based edge similarity of HFnet and Morbinet was very modest with 0.18 and of HFnet and HPOnet was 0.12 (Figure 3.10B,C). I then calculated network similarities with the DeltaCon algorithm to capture conserved node affinities between networks [224]. HFnet and Morbinet displayed again a higher similarity (0.46) than HFnet and HPOnet (0.39) (Figure 3.10D). This suggested that comorbidity correlation was not completely redundant with phenotype similarity. The differences between Morbinet and HFnet indicated that many disease relationships in the HFnet could be specific for HF patients.



	HFnet	HFnet_subset_M	HFnet_subset_HPO	HPOnet	HPOnet_subset_HF	Morbinet	Morbinet_subset_HF
HFnet	1.0	0.5	0.8	0.4	0.8	0.4	0.5
HFnet_subset_M	0.5	1.0	0.5	0.2	0.5	0.5	1.0
HFnet_subset_HPO	0.8	0.5	1.0	0.4	1.0	0.3	0.5
HPOnet	0.4	0.2	0.4	1.0	0.4	0.4	0.2
HPOnet_subset_HF	0.8	0.5	1.0	0.4	1.0	0.3	0.5
Morbinet	0.4	0.5	0.3	0.4	0.3	1.0	0.5
Morbinet_subset_HF	0.5	1.0	0.5	0.2	0.5	0.5	1.0

	HFnet	HFnet_subset_M	HFnet_subset_HPO	HPOnet	HPOnet_subset_HF	Morbinet	Morbinet_subset_HF
HFnet	1.00	0.26	0.66	0.03	0.10	0.05	0.07
HFnet_subset_M	0.26	1.00	0.27	0.01	0.06	0.09	0.18
HFnet_subset_HPO	0.66	0.27	1.00	0.03	0.12	0.04	0.07
HPOnet	0.03	0.01	0.03	1.00	0.17	0.03	0.01
HPOnet_subset_HF	0.10	0.06	0.12	0.17	1.00	0.05	0.07
Morbinet	0.05	0.09	0.04	0.03	0.05	1.00	0.33
Morbinet_subset_HF	0.07	0.18	0.07	0.01	0.07	0.33	1.00

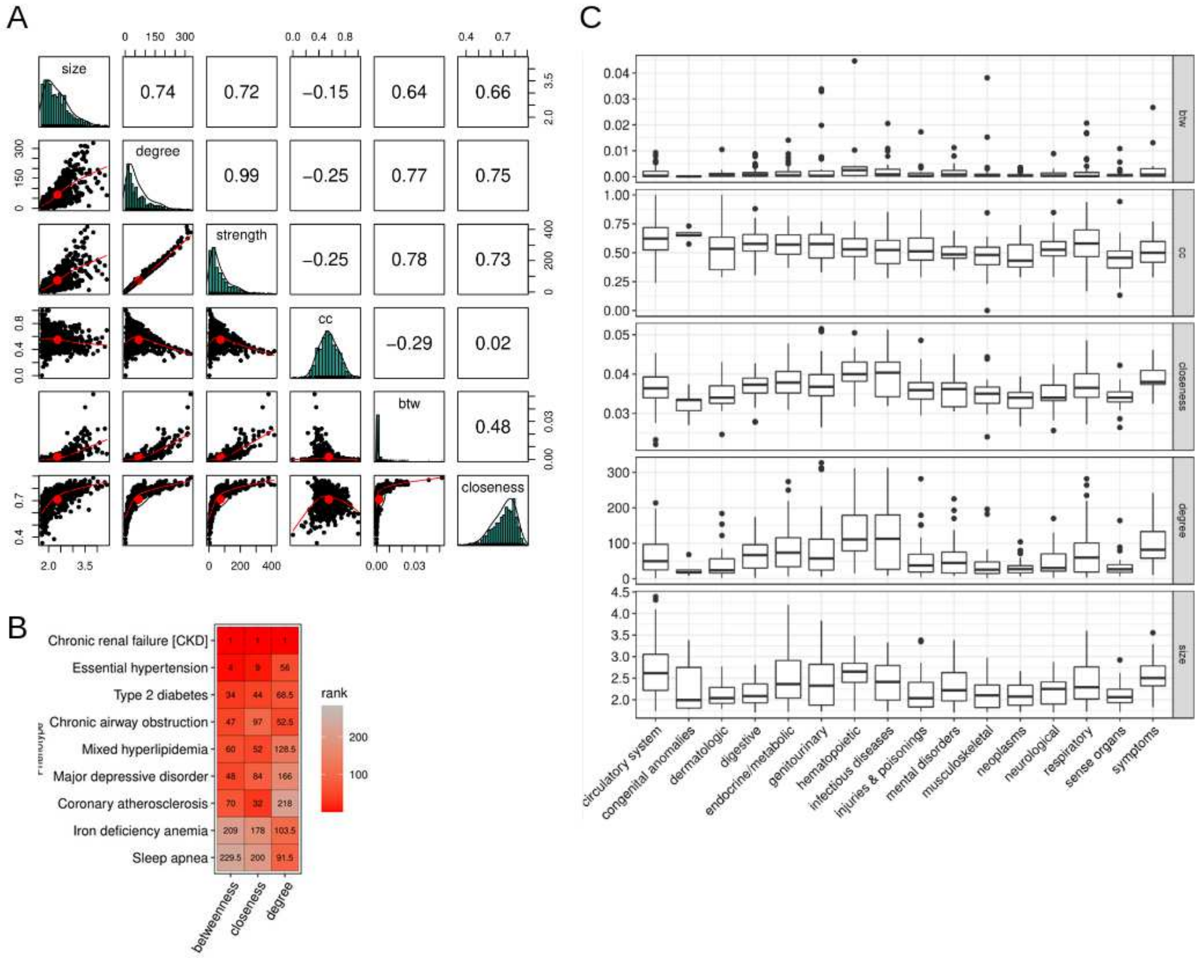
Figure 3.10 Comparison of disease networks.

A) Number of edges (top panel) and number of nodes (bottom panel) compared for different disease networks. Each disease network was also subset to the same nodes of the HFnet for comparability. B) Jaccard comparison of nodes. C) Jaccard comparison of edges. D) DeltaCon similarity (y-axis) compared to rewiring probabilities of the HFnet (x-axis). We used subsets of the HFnet with Morbinet (red) and with HPOnet (blue) and rewired each subset five times with a given probability (x-axis) and computed DeltaCon similarity with the original HFnet. Dashed lines indicate HFnet and Morbinet and HFnet and HPOnet similarities. *Reprinted from* [213].

3.4.3 Central diseases in the HFnet

Next, I analyzed the centrality of diseases. Diseases which were most frequently reported could be considered the network hubs, as indicated by their high node degree and their closeness and betweenness centrality scores (Figure 3.11A). My network captured and centralized on well-known HF comorbidities (41,42), like chronic kidney disease, which by multiple metrics was the main HFnet hub (Figure 3.11B). I tested whether disease categories were associated with node centrality measures, and found that closeness and degree centrality were both significantly associated with the disease category (Figure 3.11C) (Kruskal Wallis $p<0.01$). Infectious and hematopoietic diseases had the highest median centrality scores (betweenness, closeness and degree), indicating that patients with diseases from these categories were typically suffering from many comorbidities. Diseases affecting the circulatory system had the highest prevalence as was expected with a HF centered cohort (Figure 3.11C).

I calculated network node characteristics, such as betweenness, closeness and degree centrality, and transitivity with the *igraph* R-package. To calculate metrics based on graph distance I replaced weights for edge i (W_i) with a new edge score S_i : $Si = max(W) - Wi.$



Phenotype	betweenness	closeness	degree
Chronic renal failure [CKD]	1	1	1
Essential hypertension	4	9	56
Type 2 diabetes	34	44	68.5
Chronic airway obstruction	47	97	52.5
Mixed hyperlipidemia	60	52	128.5
Major depressive disorder	48	84	166
Coronary atherosclerosis	70	32	218
Iron deficiency anemia	209	178	103.5
Sleep apnea	229.5	200	91.5

Figure 3.11 Comparison of centralities.

A) Comparison of local (node-wise) graph theory based metrics. Size, log10(prevalence) per disease; degree, number of edges per disease; strength, sum of edge weights per disease, cc, cluster coefficient, number of connected vs unconnected first order neighbors per disease; btw, betweenness centrality (fraction of shortest paths with the node vs without the node); closeness, closeness centrality (inverse of the sum of distances to all the other vertices in the graph). Upper half displays Pearson's correlation between metrics. B) Important comorbidities of HF compared via centrality rankings in the HFnet. C) Graph metrics compared by disease category in the HFnet. Kruskal Wallis test p < 0.01 for all metrics except betweenness centrality (btw). Size, log10(prevalence); degree, number of edges; cc, cluster coefficient, number of connected vs unconnected first order neighbors; btw, betweenness centrality (fraction of shortest paths with the node vs. without the node); closeness, closeness centrality (inverse of the sum of distances to all the other vertices in the graph). *Reprinted from* [213].

3.4.4 Disease Cluster analysis

Network communities represent densely connected subgraphs and can be helpful to summarize network topology. To identify disease clusters (DCs) within the network I applied different clustering algorithms. I assumed that if different clustering algorithms detect similar structures, these structures could be more reliable. I compared different cluster algorithms based on the shared information between the assigned cluster labels (normalized mutual information and adjusted Rand Index) as well as modularity scores and module size and number. The Leiden algorithm [225] achieved the highest adjusted Rank index and normalized mutual information scores when compared to other tested algorithms (Figure 3.12A-E). I then tested different resolution parameters of the Leiden algorithm and selected a parameter of 1.1 that yielded multiple clusters and maintained a high network modularity score and high normalized mutual information values (Figure 3.12F).

At this resolution the Leiden algorithm identified nine disease clusters (DCs).

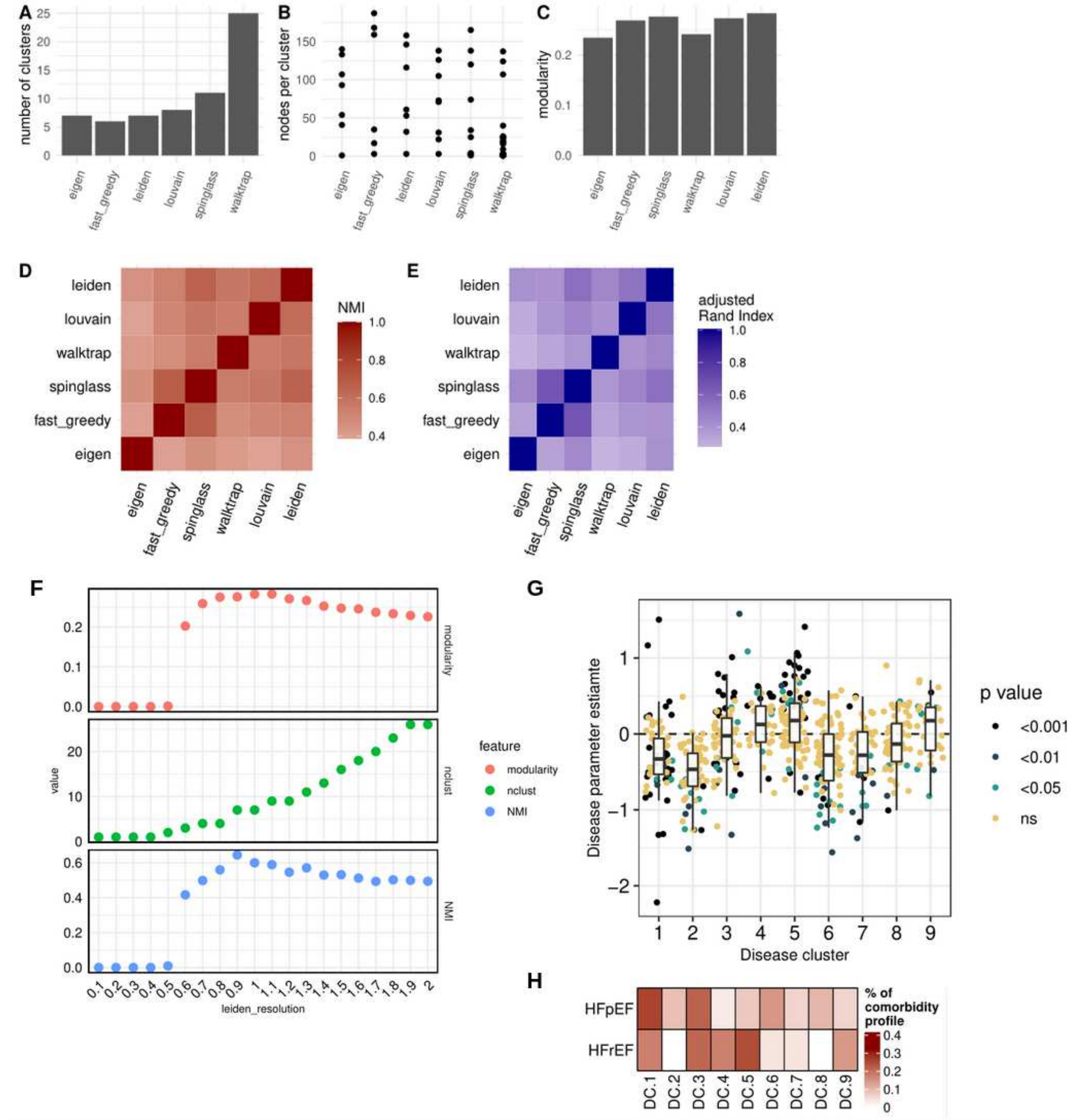


Figure 3.12 Comparison of clustering algorithms in the HFnet.

A) Comparison of number of clusters by algorithms. B) Number of nodes per cluster. C) Network modularity achieved by algorithm. D&E) Comparison of similarity of node assignment between clustering algorithms with normalized mutual information (D) and adjusted rand index (E). F) Comparison of modularity , cluster number and mean normalized mutual information (with other algorithms) by resolution parameter in leiden algorithm. G) Comparisons of single disease parameters from logistic regression models for HFpEF/HFrEF contrast that were controlled for sex. Disease parameter estimates are on y-axis and significance of the parameter is color coded. H) Composition of the comorbidity profiles from the patient classifier (rows) in disease clusters (DCs). *Reprinted from* [213].

The nine DCs were partially grouped by disease categories (Figure 3.13B) and I labeled DCs by manually reviewing disease composition (Table 3.2). For instance, DC1 and DC3 contained the majority of cardiovascular diseases. While DC1 contained cardiovascular diseases with vascular etiology (EH, CAD, MI) and included metabolic and endocrine diseases, DC3 contained valve disorders and arrhythmias (Figure 3.13C).

DC	Label	Important nodes
DC1	Cardiac / endocrine / respiratory diseases	EH, MI, COPD, Hyperlipidemia, Hypothyroidism
DC2	Sensory / ophthalmologic / skin disease	Cataract, Macular degeneration, Melanomas of skin
DC3	Cardiovascular disease with heart focus	Valve disease, Congenital anomalies, Arrhythmias
DC4	Vascular / renal / diabetic diseases	DM II, CKD, Atherosclerosis
DC5	Critical Illness / complications	Infectious disease, Organ failures
DC6	Rheumatoid / osteological / psychiatric diseases	Osteoporosis, Osteopenia, RA, Depression
DC7	Gastroenterological diseases	Gastritis, Diverticulitis, Cirrhosis
DC8	Neoplastic / hematopoietic diseases	Breast cancer, Aplastic anemia, Lymphomas
DC9	Neurological / vascular neurological diseases	Stroke, Dementias, Epilepsy

Table 3.2. Overview of disease clusters.

Manual labeling of disease clusters (DC) by characterizing most central and prevalent diseases in each cluster. EH, Essential Hypertension; MI, Myocardial Infarction; COPD, Chronic Obstructive Pulmonary Disease; DM II, Diabetes Mellitus Type II; CKD, Chronic Kidney Disease; RA, Rheumatoid Arthritis. *Reprinted from* [213].

After defining the DCs, I hypothesized that DCs represent facets of the subcohort specific HF comorbidity spectrum, and I therefore tested whether DCs capture demographic or HF subtype related characteristics. I quantified the similarity of an individual patient's comorbidity profile with each DC by calculating Jaccard indices and tested for differences between patient cohorts (Figure 3.13B).

In age stratified analyses I found that all DCs, except DC 7, were more similar to 60-80 year old (n= 16,54) compared to 40-59 year old patient's (n =5,973) comorbidity profiles. This could indicate a general increase of comorbidity burden with age or that with age come increasingly consistent comorbidity profiles between individuals. The 80+ cohort (n= 6,527) was less similar to DC1 and significantly more similar to DC3, DC5 and DC9 profiles compared to 40-60 year old patients. When comparing female and male patients I found that DC6 and DC2 yielded the highest similarity differences, respectively. Comparing HFpEF with HFrEF patients, I found that DC1, DC2, DC6 and DC8 were more similar to HFpEF patients, while DC3, DC4 and DC5 were suggested to be similar to HFrEF patients. As DC1 and DC6 also captured sex-related comorbidity differences, I investigated further, whether DC6 diseases are more prevalent in HFpEF independent of sex. For this I fit logistic regression models for each disease predicting HFpEF/HFrEF while controlling for sex (Figure 3.12G). Again, DC1, DC2 and DC6 contained more diseases prevalent in HFpEF while DC3, DC4, and DC5 diseases were more prevalent in HFrEF. In addition, this analysis also suggested that many diseases in DC7 and DC8 too had higher prevalence in HFpEF.

I further compared the comorbidity profiles from the patient classifier from section 3.3, by mapping them to DCs which yielded a qualitatively similar DC to HF subtype association (Figure 3.12H). No DC was positively associated with HFmrEF. Instead, HFmrEF patients were less similar to DC1 and DC6 than HFpEF patients and less similar to DC3, DC4 and DC5 than HFrEF patients.

In general, I found that aggregating comorbidity profiles (569 dimensions) to DC similarity (9 dimensions), allowed me to capture differences among patient cohorts in regard to sex, age and HF subtype in meaningful disease groups.

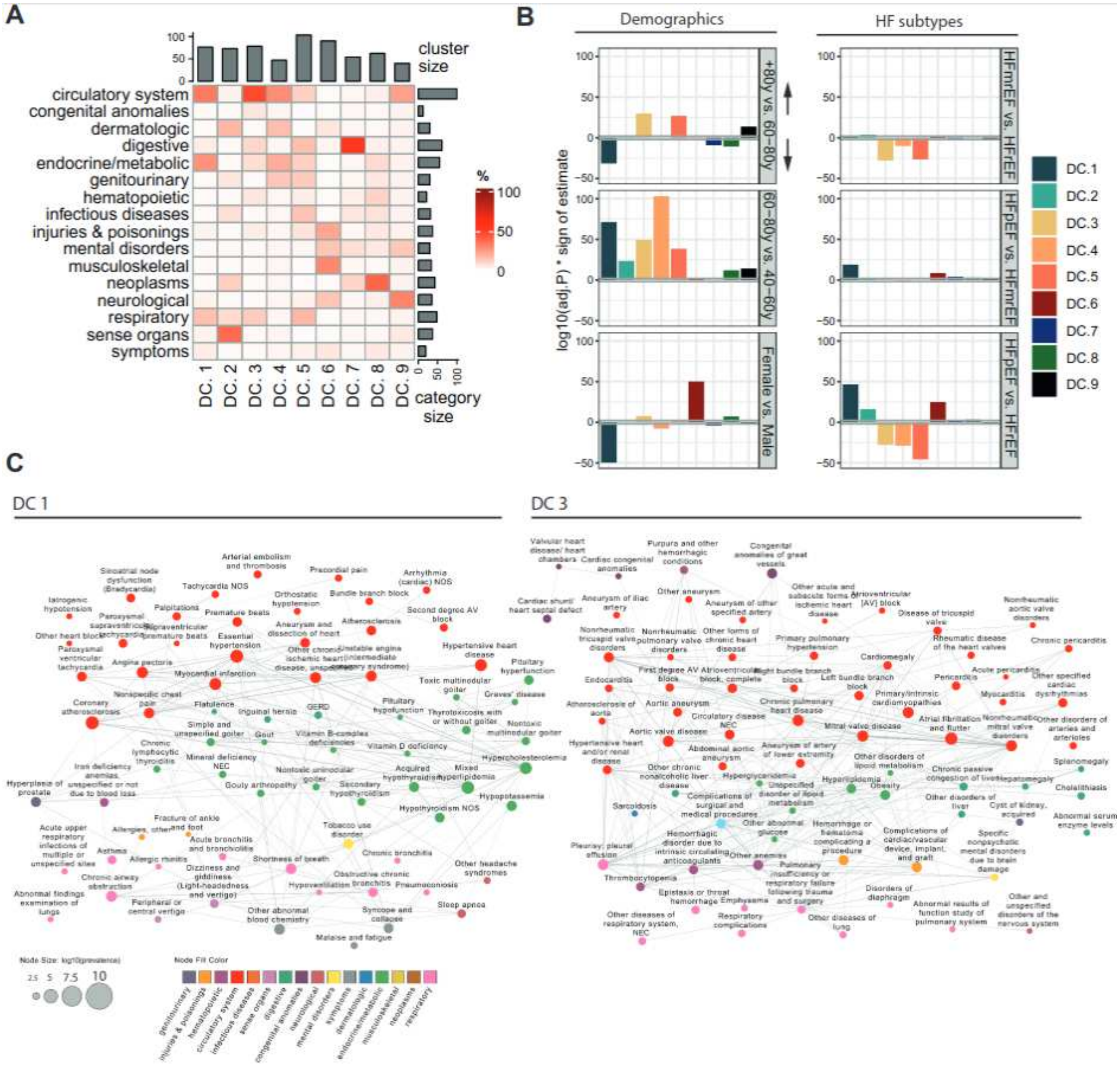


Figure 3.13 The heart failure comorbidity network (HFnet).

A) Disease category composition of disease clusters (DCs) in the HFnet. Number of nodes per cluster in top barplot and number of disease per category in side barplot. B) Comparison of patient cohorts based on DC similarity. Jaccard indices were calculated between each patient and each DC, then unpaired two sided wilcoxon rank test was applied to compare different patient cohorts. The log transformed p-value was multiplied by the sign of the test estimate for visualization purposes such that positive values indicate higher cluster similarity with the first cohort of the contrast label. Patient cohorts were selected by age stratification, sex and HF subtype. C) Subgraphs of the HFnet visualized (left DC1, right DC3). Node size relates to prevalence, edge width to scaled phi-correlation, node color to disease category. Only edges with highest weights were plotted for visibility. *Reprinted from* [213].

3.5 The HFhetnet and gene prediction for HFpEF

Biomedical research has yielded significant knowledge of disease gene associations, which can be harnessed to extrapolate novel disease gene relationships. HFpEF is a comorbidity driven syndrome and I hypothesized that the identified HFpEF comorbidity profile could be translated to a genetic profile consisting of possibly recurrent genetic associations to these comorbidities. In this part of my study, I first integrated multiple biomedical databases to construct a cardiac specific multi-layer disease and gene network (section 3.5.1). I then estimated the success of this network to recover known disease-gene associations (section 3.5.2), used the HFpEF comorbidity profile to identify the most commonly associated genes (section 3.5.3) and, finally, corroborated gene predictions in independent experimental data (section 3.5.4)

3.5.1 Building the Heart Failure Heterogeneous Network (HFhetnet)

Gene-gene association

To consider multiple layers of gene organization, I constructed a multi-layer gene network from different sources.

Omnipath [226] is a meta resource of a multitude of biological knowledge databases, and I curated a network by connecting two genes if a resource provides a co-membership for a signaling pathway. I used the number of resources that reported a relationship as an estimate for the confidence in the relationship, which I introduced as edge weights in the Omnipath layer.

The protein-protein interaction (PPI) network was constructed based on the union of publicly available data from experimental and literature curated data [227].

Gene Ontology (GO) gene networks have been constructed before, and I used the GO networks constructed by [228].

Each gene network was reduced to remove loops and multiple edges. To filter for genes relevant in cardiac tissue, gene networks were subset to genes expressed in the human heart on RNA or protein level. For protein expression I used proteomic data [229], where I selected all peptides that were detected in the human heart and used the leading gene associated with each peptide. For Gene expression I selected genes that were detected in heart tissue in the Genotype-Tissue Expression (GTEx) Project v8 with a *transcript per million* value > 1. To ensure that genes that become activated in diseased hearts were also captured, I also included the genes that were captured in the meta-analysis of my study in Chapter I [102].

Disease-gene association

Next, I used DisGeNET, a resource containing disease-gene associations, to connect the HFnet with the gene network. I downloaded DisGeNet v7.0 [230] and mapped the ICD-10 codes in DisGeNet to PheCodes. To ensure that the most frequent diseases in our cohort were mapped, I selected the most frequent 3 digit ICD-10 codes that were not mapped to the DisGeNet gene set and performed manual annotation via Unified Medical Language System (UMLS) IDs for 23 disease entities. E.g. PheCode *427.2 (*Atrial Fibrillation) was manually mapped to the UMLS ID *C000423.* I only included disease-gene associations with a DisGeNet confidence score >0.29. This cut-off was chosen, such that either one curated source or multiple experimental sources were necessary for disease-gene associations.

I connected 400 diseases of the HFnet with a total of 4,044 genes via 20,170 edges. As the HPOnet constructed earlier had a small intersection with the HFnet and appeared to capture a different type of disease relationship, it was included as an additional disease layer in our network.

HFhetnet

The presented HFhetnet is an assembly of the data driven comorbidity relationships (HFnet) and six biomedical databases resulting in a total of 13,572 nodes and 181,529 edges. Its main structure is set up by two biological networks (disease layer and gene layer) that each consist of two or four network layers (respectively) (Figure 3.14A). The two disease networks were the smallest when comparing node numbers (Figure 3.14B). However, edge density was much higher resulting in centralisation of these networks compared to the gene layers. Within gene layers, the ontological layers displayed the highest transitivity, as well as tendency to connect to hub genes (degree assortativity). To assess research bias in the gene networks I calculated Pearson's correlation between the number of abstracts in PubMed mentioning a gene and the gene's network degree per layer, and found that only the pathway layer (Omnipath) displayed significant correlation (p-value <0.05). This is related to a biomedical research bias towards the investigation of a small number of genes [231]. Thus, the integration of experimental and ontological data can ameliorate the centrality of over-studied genes.

In summary, I constructed the HFhetnet by integrating various prior knowledge resources to incorporate genetic information. The different network layers of the HFhetnet capture unique node relationships and display particular network topologies.

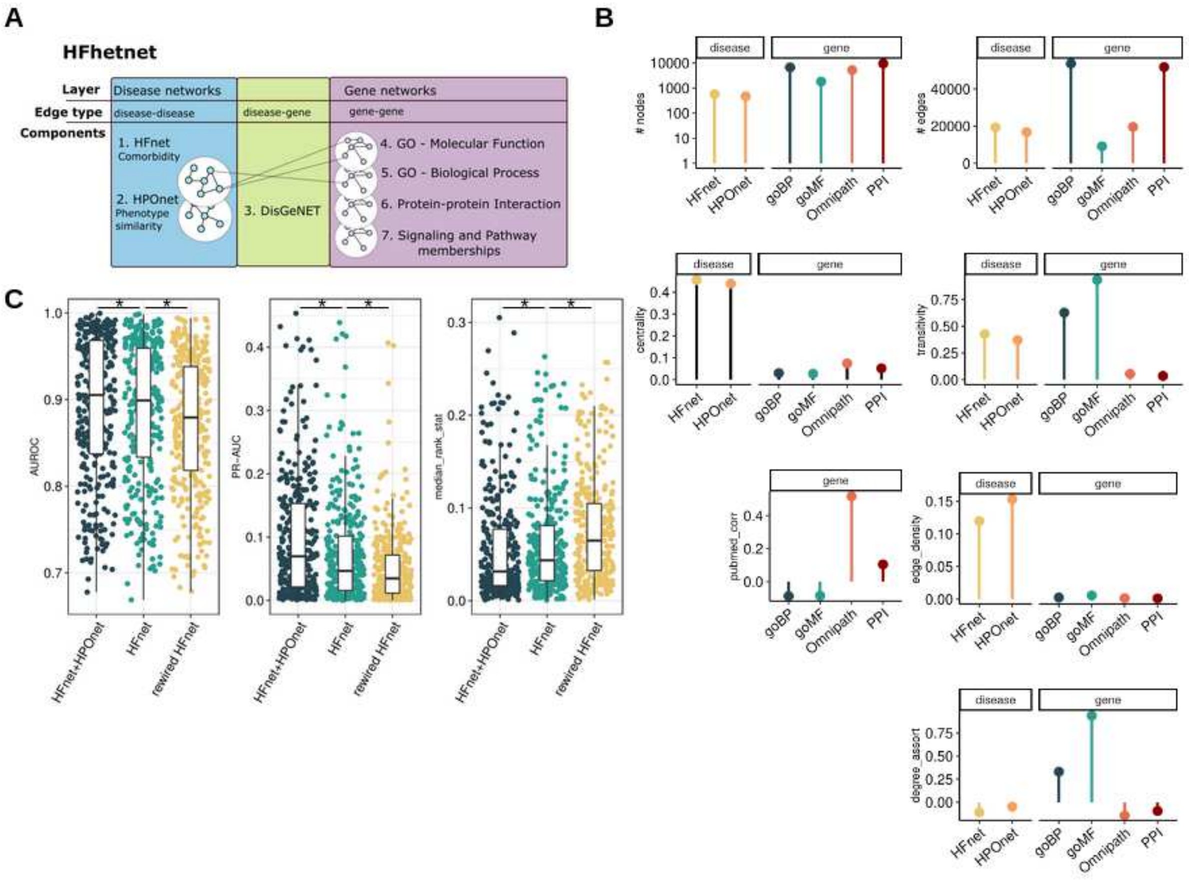


Figure 3.14 HFhetnet overview and .

A) Schematic overview of HFhetnet and its different layers built by including seven independent data sources. B) Characterization of network layers by size (number of nodes and edges), edge density (percentage of possible edges), degree centrality, global transitivity (average probability of the neighbors of a node being connected), degree assortativity (preference of nodes to connect with nodes of similar degree) and literature bias (i.e. Gene degree/pubmed score correlation). C) Leave one out cross validation results for all diseases with two or more DisGeNET links. I compared the performance of gene set recovery with different versions of the HFhetnet by modifying only the disease network. I compared HFnet + HPOnet (i.e. the original HFhetnet), only the HFnet (without HPOnet), and a rewired HFnet. Outliers are not plotted for visualization purposes. Paired, two-sided Wilcoxon test, * p<0.001. AUC-PR Area under the Precision/Recall Curve, AUROC, Area under the Receiver operator curve. GO, Gene Ontology; HPO, human phenotype ontology. *Reprinted from* [213].

3.5.2 Estimating the success of disease-gene prediction within the HFhetnet

To estimate the potential of the HFhetnet to predict disease - gene relationships we estimated the success of predicting known disease genes. The rationale behind this approach is the guilt-by-association principle that assumes that functionally related genes are also associated in the network context. Extending this notion to heterogeneous networks, this principle can be interpreted as a disease being associated with relevant disease genes through its position in the network. To quantify this property and predict genes from diseases within the HFhetnet, I applied a network propagation algorithm developed for multi-layer networks (Random walk with restart on Multiplex heterogeneous networks; [232]). This algorithm is an extension of the random walk algorithm, that tries to find a stationary distribution of probabilities that a node is visited when a random walk on the network is initiated in a set of seed nodes.

To quantify the property of gene prediction, I applied a the RWR-MH in a leave-one-out validation design to assess whether known disease genes can be recovered after removing the direct edges that connected them to a disease: For a given disease that was present in the HFhetnet and directly linked to two or more genes, I attempted to predict those genes after removing the direct links from the HFhetnet and running the RWR-MH with the disease as seed node. The position of the target genes in the resulting probability ranking was then assessed with different metrics to estimate success of disease gene recovery.

To evaluate gene prediction success, I used three different metrics: median rank ratio, AUROC and AUC-PR (Area under the precision recall curve).

For the median rank ratio, I calculated the median rank of the target gene set in the RW ranking and divided this rank by the total length of the ranking. This metric is close to 0 if the genes are located towards the top, and close to 1 if they are located close to the bottom of the ranking.

AUROCs and AUCPRs were calculated with the R-package *pROCroc*. Each target gene was considered as a true positive, others were true negatives and the assigned RW probabilities were used to calculate area under the ROC and PR curves. AUC-PRs tend to be very low, due to the high number of non-target genes in the top of the ranking that leads to a drop in precision. This in part is wanted for disease gene prediction, as these true negative genes could rather be unknown potentially relevant candidates. AUROCs can be inflated when small gene sets are recovered at the top of the ranking.

Finally, I performed this analysis by comparing the impact of three variations of the disease layer: i) HFnet + HPOnet (original HFhetnet), ii) only HFnet, and iii) a rewired HFnet. Gene prediction worked best in the original HFhetnet (median AUROC 0.91, median AUC-PR 0.07, and median rank ratio 0.03) (Figure 3.14C). This performance dropped for every metric when removing the HPO layer or when rewiring the HFnet (paired two-sided Wilcoxon's rank sum test p<e-10). The rewired HFnet still performed better than random, which might be explained by i) high edge density in the HFnet and ii) the large size of the unaltered gene-gene and disease-gene network in comparison to the smaller HFnet.

Prediction success correlated weakly but significantly with gene set size. In addition, neither disease prevalence nor DisGeNET confidence scores were significantly correlated with prediction success, suggesting that frequent diseases could not be predicted better than less frequent diseases. Prediction performance depended on disease category (Kruskal-Wallis test p-value <0.01 for all metrics). Respiratory, neurological, genitourinary and cardiovascular diseases performed best.

In summary, I found that within the HFhetnet, the disease genes remain close via the disease's connection through its comorbidities or phenotypically similar neighbors. Thus I concluded that HFnet and its extension, HFhetnet, capture meaningful disease-disease, disease-gene and gene-gene relationships, which can be exploited for predicting a disease's genetic profile through its comorbidities.

3.5.3 Predicting genes associated with comorbidity profiles of HFpEF and HFrEF

In the first part of this study I found that HFpEF and HFrEF patients were distinguishable based on their comorbidity profiles (section 3.3). I then demonstrated that diseases within the HFhetnet are located in network proximity to their respective disease-genes (section 3.4). Leveraging both insights, I hypothesized that genes located close to the HFpEF and HFrEF comorbidity profiles could yield novel candidates for the respective HF subtype.

Thus, I applied the RWR-MH algorithm with the HFpEF and HFrEF comorbidity profiles as seed nodes. This yielded two vectors of RW probabilities for each comorbidity profile. The top 500 genes yielded non-zero probability values for each profile (Figure 3.15A, B).

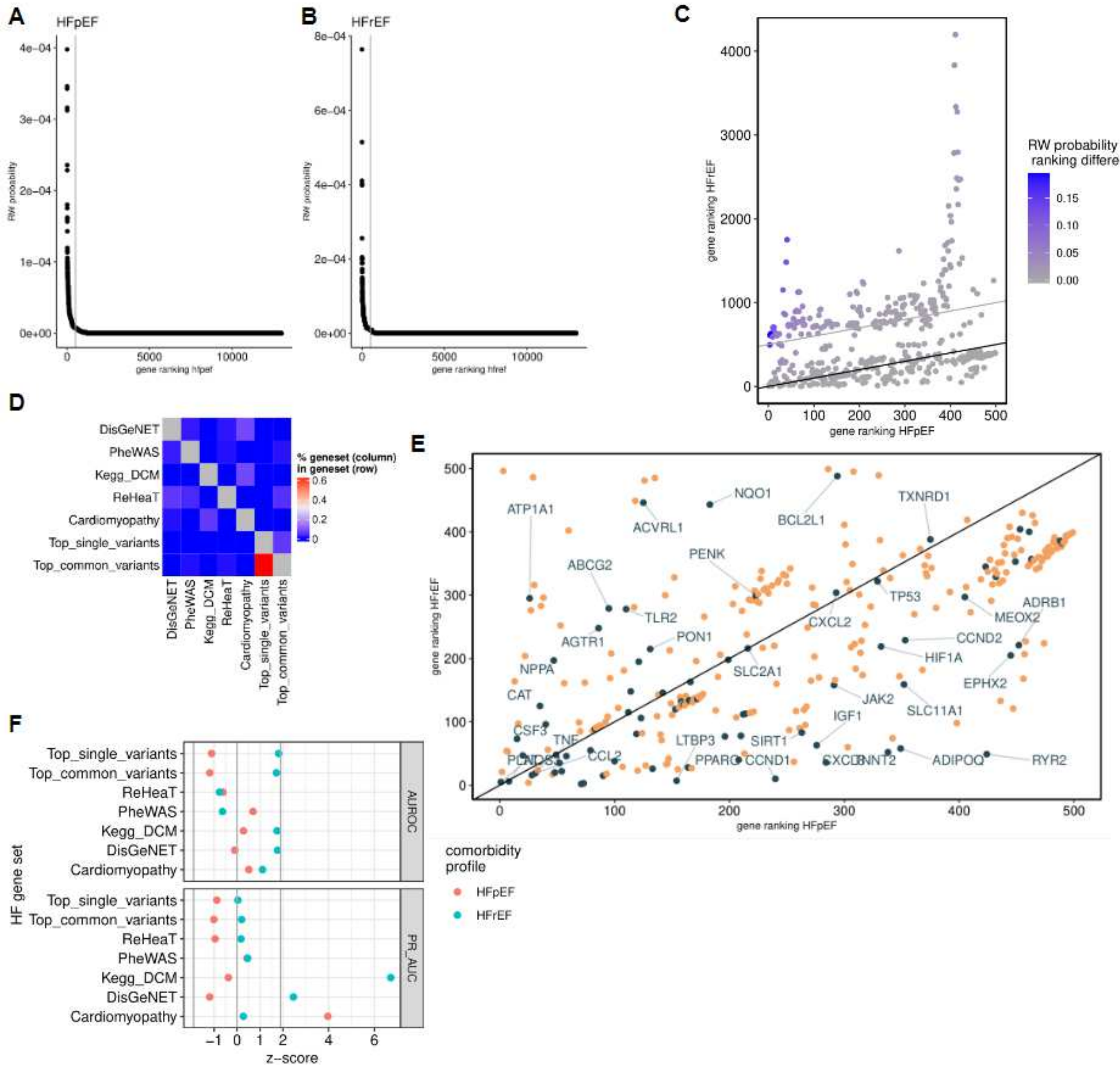


Figure 3.15 HF subtype gene prediction.

A+B) Random walk (RW) with restart in a multiplex heterogeneous network was applied with comorbidity profiles for HFpEF and HFrEF as seed nodes. RW probability distribution of all genes in the HFhetnet is shown for A) HFpEF and B), HFrEF comorbidity profile. C) Gene prioritization for HFpEF. Top 500 HFpEF genes are ranked (x-axis) and compared to their ranking in the HFrEF vector (y-axis). Color is the RW probability for HFpEF multiplied by the ranking difference. This calculation yields a new gene ranking that prioritizes HFpEF specific genes. D) Comparison of gene rankings within the top 500 genes of HFpEF (x-axis) and HFrEF (y-axis). Genes that are known to be associated with heart failure are colored and labeled. E) Comparison of intersections of HF gene sets demonstrating a low redundancy. F) HF geneset recovery (assessed with area under the receiver operator (AUROC) and area under the precision-recall curve (PR_AUC)) with 1000 random comorbidity profiles from the HFnet were used to generate null distributions. The geneset recovery values from the real HFpEF and HFrEF comorbidity profiles were then z-transformed. *Reprinted from*[213].

To assess whether the resulting gene rankings recapitulate known HF genes, I curated a set of HF related genes from various prior knowledge sources and independent datasets, which had only little intersection (Figure 3.15D). I selected prior knowledge sources including 1) DisGeNet genes associated to Heart Failure with confidence score >0.29; 2) Literature curated [233], 3) Kegg disease database, dilated cardiomyopathy related pathways and data driven resources including 4) Cardiovascular Disease Knowledge portal, top common variants for Heart Failure; 5) Cardiovascular Disease Knowledge portal, top single variants for Heart Failure (https://cvd.hugeamp.org/, accessed August 2022); 6) ReHeaT top 500 conserved genes from end stage heart failure meta-analysis [102]. 7) PheWAS gene sets associated with Heart Failure (p<0.05, Odds ratio>1) (https://phewascatalog.org/ , accessed August 2022).

I then assessed whether these genesets could be recovered within the HFpEF and HFrEF gene rankings (Figure 3.16A). Gene sets that were retrieved from experimental data (Gene expression, PheWAS, GWAS) performed worse in the predictions. Next, I compared these prediction results with random comorbidity profiles and found that the HFrEF profile associated with Kegg's dilated cardiomyopathy (DCM) (z-score AUROC 1.77; z-score PR-AUC 6.7) and DisGeNETs HF genes (z-score AUROC 1.76; z-score PR-AUC 2.46) (Figure 3.15F). This indicated that the HFrEF comorbidity profile which was more cardiac centered was within the HFhetnet closer to prior knowledge of HF genes. In general, well known genes relevant in HF were recovered for both, HFpEF and HFrEF comorbidity profiles including NPPA, NPPB, TNFa, NOS2, NOS3, CCL2, IL1B, LMNA, TTN (Figure 3.15E).

The randomization of comorbidity profiles suggest that HF gene sets tend to be rather close to the disease layer in general. Thus, instead of using the HFpEF probability ranking to suggest novel candidates I calculated a prioritization score for HFpEF and HFrEF, which punishes highly ranked genes from both rankings. Thus, genes that are close to the disease layer in general might be not highly specific for the selected comorbidity profile and receive a lower prioritization score. More specifically, I calculated $Gi = Pi^{RW} * |\Delta Ri|$. G is gene prioritization score, P is RW based probability, ΔR is rank difference between HFpEF and HFrEF rankings for gene i (Figure 3.15C).

I found that MMP1, MHY7 and DAPK1 received the highest scores (Figure 3.16B). Other candidates included genes functionally involved in fibrosis (e.g. LOX), metabolism (XDH), transcriptional regulation (ATF6), coagulation (THBD), oxidative stress (NOS1). I reviewed potentially interesting candidates for their biological function and possible association to HF (Table 3.3).

Rank	Gene symbol	Gene name	Functional group	Gene Function	Role in HF (exemplary or putative)	References
17	PCSK5	Proprotein convertase subtilisin/kexin type 5	Cell differentiation	Mediates post translational endoproteolytic processing	Cleaves GDF1, heart development	234,235
9	NKX2-5	NK2 Homeobox 5	Cell differentiation	TF expressed in precursor cardiac cells, involved in cardiac development	Heart development, activated in HF involved in hypertrophy	236,237
31	GATA5	GATA binding protein 5	Cell differentiation	TF, involved in embryonic development of the heart	Possible role in cardiac hypertrophy, linked to dilated CM	238
42	GATA3	GATA binding protein 3	Cell differentiation	TF, involved in embryonic development of the heart and immune cell differentiation	Role in T-cell recruitment	239,240
67	JAG1	Jagged canonical Notch ligand 1	Cell differentiation	Interacts with four receptors in the mammalian Notch signaling pathway	Possible protective role against PAH and diabetic CM, involved in regenerative capacity of the heart	241–243
62	NOTCH1	Notch receptor 1	Cell differentiation	Regulates interactions between physically adjacent cells through binding of Notch family receptors to their cognate ligands	Possible protective role against PAH and diabetic CM, involved in regenerative capacity of the heart	241–243
13	CITED2	Cbp/p300 interacting transactivator with Glu/Asp rich carboxy-terminal domain 2	Cell differentiation	TF that controls pluripotency among others	Cardiomyocyte pluripotency	244
23	KLKB1	Kallikrein B1	Coagulation	Factor XII activation	Interacts with LDL, possible HF biomarker	245–247
27	TBXA2R	Thromboxane A2 receptor	Coagulation	G protein-coupled thromboxane A2 receptor that induces platelet aggregation and regulate	Involved in endothelial homeostasis, angiogenesis	248–250

				hemostasis		
34	THBD	Thrombomodulin	Coagulation	Endothelial-specific type I membrane receptor that binds thrombin.	Angiogenesis and anticoagulant , possible HF biomarker	251,252
11	DAB2IP	DAB2 interacting protein	Endothelial dysfunction	Ras GTPase-activating protein (GAP)	Associated with endothelial dysfunction in atherosclerosis	253,254
77	ESR1	Estrogen receptor1	Endothelial dysfunction	Estrogen receptor and ligand-activated TF	Regulates NOS activity among many other effects	255
7	ATF6	Activating transcription factor 6	ER-stress	Unfolded protein response	Activates XBP1 which modulates ER-stress	256–258
3	MMP1	Matrix metallopeptidase 1	Fibrosis	Proteases involved in degrading of ECM	involved in cardiac fibrosis	259
25	FGFR2	Fibroblast growth factor receptor 2	Fibrosis	Receptor tyrosine kinases that promote mitogenic signal mediators that induce cell proliferation and survival	Involved in cardiac fibrosis	260
33	LOX	Lysyl oxidase	Fibrosis	Cross-linking of collagens	Possible role in myocardial stiffness	261
2	MYH7	Myosin heavy chain 7	Hypertrophy	Part of the thick filament in cardiac muscle, involved in contraction	Hypertrophy marker, also reported in HFpEF	262,263
4	SUZ12	SUZ12 polycomb repressive complex 2 subunit	Hypertrophy	Core PRC2 (polycomb repressive complex 2) protein	Mediates long non coding RNA Ahit induced cardiac hypertrophy	264
65	FSTL1	Follistatin like 1	Hypertrophy	Extracellular glycoprotein	Protective against hypertrophy	265,266
79	RGS5	Regulator of G protein signaling 5	Hypertrophy	RGS proteins are involved in the regulation of heterotrimeric G proteins by acting as GTPase activators	Protects from cardiac hypertrophy and fibrosis	267,268
89	CD14	Cluster of differentiation 14	Inflammatio n	Membrane glycoprotein primarily expressed by myeloid cells that plays a key role in inflammation	Possible HFpEF plasma biomarker	269
47	SHARPI N	SHANK associated RH domain interactor	Inflammatio n	Displacing talin from the integrin cytoplasmic domain	Involved in integrin Inactivation and NF-κB Signaling	270,271
90	CD209	Cluster of	Inflammatio	Expressed by macrophages and	Involved in rheumatoid	272

		differentiation 209	n	dendritic cells	arthritis	
36	MVK	Mevalonate kinase	Metabolism	Cholesterol metabolism	Fibrotic effects, Rho family small GTPase activity modulation	[273,274]
1	DAPK1	Death associated protein kinase 1	Oxidative stress	Calcium/calmodulin-dependent serine/threonine kinase	Protection from oxidative stress in MI	[272]
18	XDH	Xanthine dehydrogenase	Oxidative stress	Oxidative metabolism of purines	Linked to nitroso-redox balance, also possible plasma marker	[275]
69	NOS1	Nitric oxide synthase 1	Oxidative stress	Synthesizing nitric oxide from L-arginine	Inhibition possibly linked to protection of diastolic dysfunction	[276,277]
14,19	GSTZ1, GSTT1	Glutathion-S-transferases	Oxidative stress	Catalyzing the conjugation of the reduced form of glutathione to xenobiotic substrates for the purpose of detoxification	Related to oxidative stress in cardiac tissue	[278,279]

Table 3.3 Potential HFpEF candidates based on network proximity to comorbidity profiles.

Genes are sorted by functional groups. Rank indicates prediction rank for HFpEF. TF, transcription factor; CM, Cardiomyopathy; PAH, pulmonary arterial hypertension. *Reprinted from* [213].

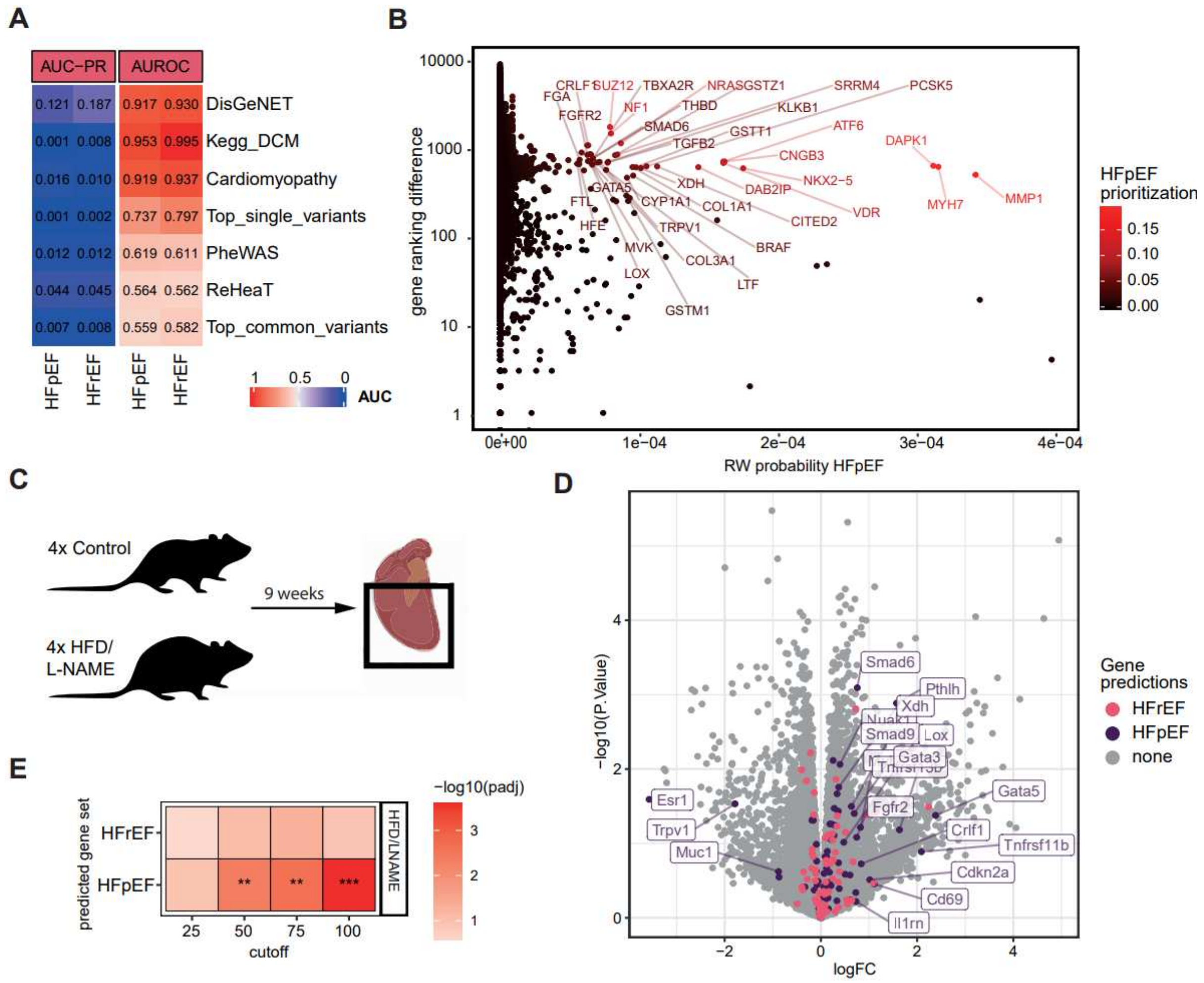


Figure 3.16. HFpEF gene prediction.

A) AUROC and AUC-PR for different HF related gene sets in random walk probability vectors based on HFpEF and HFrEF comorbidity profiles. Prior knowledge gene sets are DisGeNET, Kegg pathway for dilated cardiomyopathy (DCM), Cardiomyopathy (literature curated). Data based gene sets are PheWAS, ReHeaT and GWAS variants. B) Prioritizing genes for HFpEF that are close to HFpEF comorbidity profiles in the HFhetnet and also display high ranking differences when compared to gene predictions based on HFrEF comorbidity profiles. C) Scheme of experimental design for murine model of HFpEF by HFD/L-NAME diet. Cardiac ventricles were harvested after 9 weeks and bulk transcriptomics were collected. D) Volcanoplot displaying gene expression regulation in the murine HFpEF model compared to control. Labeled genes display HFpEF predicted genes from human comorbidity profiles. D) Predicted HF genes from comorbidity analysis were enriched in gene-level t-statistics of murine differentially expression analysis comparing disease with control. Gene set enrichment p.value. ***p <0.001. **p < 0.01. *Reprinted from* [213].

3.5.4 Corroboration of HFpEF gene candidates in independent experimental data

Finally, I asked whether I could recover candidate genes in a molecular disease signature relevant for HFpEF derived from orthogonal data. For this I collaborated with Prof. Johannes Backs and Dr. Mark Pepin. They generated RNAseq data from a murine model of HFpEF which successfully recapitulated important HFpEF phenotypes including preserved ejection fraction and diastolic dysfunction [21]. I received aligned RNAseq count data and performed downstream analysis by first characterizing HFpEF transcriptomic changes, and second comparing the comorbidity based gene prediction.

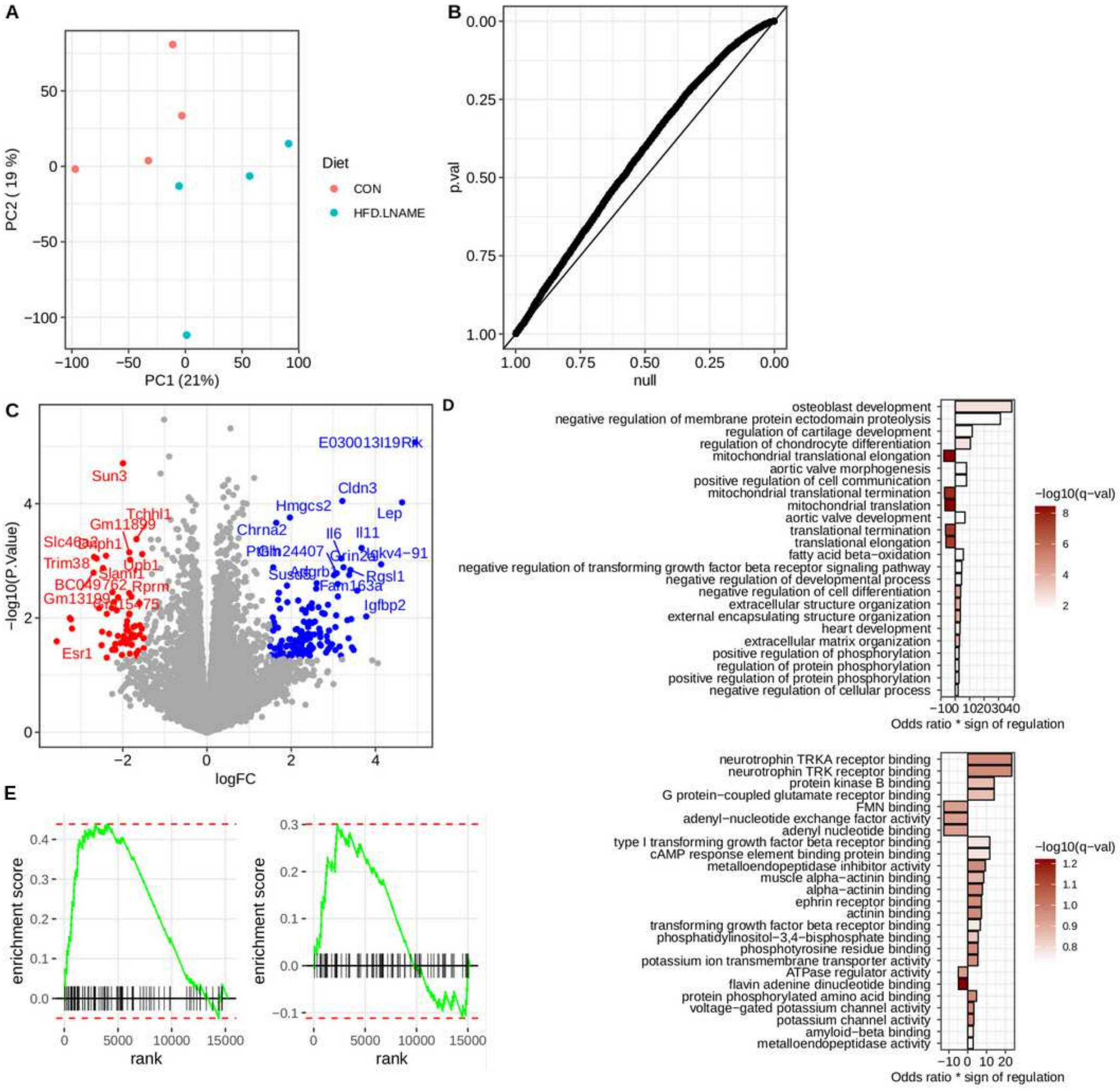


Figure 3.17 Myocardial gene expression in L-NAME/HFD.

A) Principal component analysis embedding transcriptomes of murine control (CON) and HFpEF samples (HFD.LNAME). B) Q-Q plot displaying deviation of gene-level p-values from a null model. C) Volcanoplot displaying up and down regulated genes. D) Overrepresentation analysis of GO terms, upper panel GO biological process, lower panel GO molecular function. Enrichment was performed by selecting up and down regulated genes and enriching them separately by calculating Odds Ratio and hypergeometric tests within ontology gene sets. For visualization purposes I multiplied the Odds Ratio by sign of regulation. E) Visualization of the running sum calculated in gene set enrichment analysis for the comorbidity predicted gene sets. Left panel are top 100 HFpEF predicted genes, right panel are top 100 HFrEF predicted genes. Gene ranks were ordered by t-statistic values. *Reprinted from* [213].

I found that variation in the transcriptome associated with the HFpEF model on PC1 and PC2 (Figure 3.17A) and after performing differential expression analysis (Figure 3.17B), I found that gene regulation associated with processes involving fibrosis and inflammation (Figure 3.17C, D).
I then ranked differentially expressed genes and performed enrichment analysis of the HFpEF and HFrEF gene predictions using different cut-offs (Figure 3.16D, Figure 3.17E). I found that the top 50 to 100 HFpEF predicted genes displayed significant enrichment in overexpressed genes in the murine HFpEF model, while the HFrEF predicted genes were not enriched. Fibrosis related genes like LOX, SMAD9, and PTHL and hypertrophy related genes like GATA5, GATA3 and MYH7 could be recovered, among others (Figure 3.16D). This suggested that the genes derived from human HFpEF comorbidities associate with relevant gene expression patterns in a HFpEF related disease signature.

3.6 Discussion and Conclusion

In this study I conducted a systems level analysis of comorbidities in a large retrospective cohort of HF patients. I derived clinically relevant insights by comparing comorbidity profiles between HFpEF and HFrEF patients, and biological insights by defining genes associated with HFpEF and HFrEF comorbidity profiles.

Patient clustering has been previously shown to yield novel subgroups of HFpEF defined by multivariate similarity [9–11]. In contrast, the clustering of features (i.e. comorbidities) can inform about patterns of co-occurring disease groups. Our study demonstrated that this approach can be useful to interpret comorbidity profiles. The aggregation of co-occurrence patterns of diseases can help to organize illness into different levels of clinical concepts like organs (DC7- Gastrointestinal tract), illness severity (DC5 - intensive care) or disease categories (DC8 - cancer). This aggregation via network clustering may also reduce multiple testing burdens and provide insights into the relevance of low prevalence diseases where comparisons for a single disease may be problematic.

In the patient classifier, HFpEF was characterized by a larger number of comorbidities with lesser emphasis on cardiac disorders. This supports the hypothesis of HFpEF as a comorbidity driven systemic syndrome [2,280]. I found that hypertensive heart disease was the most discriminant feature for HFpEF, which has been viewed as a major etiology for diastolic HF involving cardiac hypertrophy and myocardial stiffness [281,282]. In contrast, ischemic etiologies including myocardial infarction characterized HFrEF consistent with other studies [283].

I identified novel disease associations with HFpEF such as neoplastic diseases including breast cancer. HF related hospitalizations in breast cancer survivors have been recently associated more with HFpEF than with HFrEF [284], though the reasons for this remain incompletely elucidated [285]. The association to other cancerous diseases remains largely unexplored and should be addressed in future studies. Another interesting aspect of HFpEF patient comorbidities were the high similarity to DC6, which contained rheumatic, osteologic and mental diseases. Systemic inflammatory diseases could be a driving factor for HFpEF and rheumatic disease could constitute a pathophysiologic linkage [280,286–288]. Bone mineralization also has been reported to be lowered in HFpEF patients [289] and is a symptomatic link to postmenopausal endocrinology [290]. While mental health has been studied in the context of HF extensively, differences between HFpEF and HFrEF are largely unexplored. The joint clustering of these disease complexes and their similarity to female patients provides a potential link between female sex and HFpEF. Future work should further explore these relationships.

HFpEF and HFrEF clearly displayed distinguishable comorbidity profiles. By contrast, HFmrEF, introduced as a unique form of HF in 2016 [208] appeared to be a combination of attributes from HFrEF and HFpEF. Thus, from the comorbidity perspective it may be a transitional state instead of a unique syndrome as suggested before [291].

I predicted an associated genetic profile from data driven HFpEF comorbidity profiles. This genetic profile indicates that HFpEF comorbidities are associated with recurrent patterns of genes involved in fibrosis, inflammation, cell differentiation, metabolism and oxidative stress. As an example, the Glutathion-S-transferases, NOS1 and Xanthine dehydrogenase (XDH), were identified by our network. XDH catalyzes the rate limiting step in purine metabolism producing uric acid [292] and previous literature supports both the role of serum uric acids in heart failure [275] and plasma XDH activity as relevant for adverse clinical outcomes in HFpEF [293]. Nitric oxide synthase (NOS) has been proposed to contribute to endothelial dysfunction in HFpEF [276,277] and NOS1 inhibition was recently associated with recovery of diastolic dysfunction in a murine model resembling HFpEF[294]. Glutathione-S-transferases (GSTM1, GSTT1, GSTZ1) are antioxidant enzymes and polymorphisms of these genes have been reported as potentially relevant to HF and diastolic dysfunction [278,279]. This group of genes could constitute crucial gene candidates involved in comorbidity based HFpEF pathophysiology.

In general, HFpEF is likely to be a disease in which multiple genes and pathways contribute to the spectrum of phenotypes. Therefore, instead of using the disease-gene prediction to identify and validate individual genes, we have corroborated the overall effect of a spectrum of identified genes in murine gene expression data. In real-world populations, it is likely that the genetic heterogeneity of the syndrome will be influenced by the specific comorbidities that are well represented in each population. In previous disease-gene prediction studies, gene prediction was performed either by selecting multiple seed genes or single seed diseases [295,296]. I propose that our approach for gene inference based on data driven comorbidity profiles might be suitable for systemic syndromes where multimorbidity plays an important role like HF, and especially HFpEF. In addition, several data resources were generated in this study: i) HFpEF gene predictions, ii) HFhetnet and iii) murine HFpEF transcriptome to help facilitate future efforts to understand HFpEF related pathophysiology and benefit the research community.

This study is subject to several limitations. I analyzed routine clinical care data which is limited to the information captured i) in our hospital system and ii) at the hospital visits of a patient. Thus, possible non-observed confounders like socioeconomic status or health related behavior could not be taken into account. Further, I performed a cross-sectional analysis of comorbidity profiles and future studies are necessary to delineate different disease trajectories by considering the time of events. Another data limitation relates to the ICD-10 coding system which does not contain specific codes for HF subtypes. I determined subtypes using LVEF, which can be error prone [297] and might not fully provide a sufficient criterion for the HFpEF diagnosis [298]. Patients with more serious conditions will tend to visit a tertiary health care provider more often and thus could be overrepresented. This seemed to affect the contrast between HFrEF and HFpEF, as HFrEF patients had higher intubation prevalence and DC5 similarity. Furthermore, these limitations may have contributed to differences between this study population and other reports of HFpEF population characteristics. However, given the known heterogeneity of HFpEF and HFrEF [9–11], I believe these differences are plausible and a more granular approach to study HFpEF subtypes could be necessary to address inconsistent patient characteristics [299].

Many open questions remain regarding HFpEF pathophysiology and genetics [12]. Interdisciplinary and translational approaches are needed to account for the cross-organ disease involvement that is suggested to be critical in HFpEF. The increasing abundance of routine clinical care data and novel approaches like network medicine can provide novel insights and guidance for future experimental approaches.

Concluding Remarks

HF is a puzzling syndrome caused by a complex interplay of a multitude of pathways subjected to internal and external factors specific to individual patients. To fully appreciate this complexity, high throughput methods to generate unbiased molecular profiles together with deeply phenotyped clinical data is necessary. Thus, the analysis of high-dimensional and complex data has become an integral part of cardiovascular research. The need for highly interdisciplinary research between experts of cardiovascular biology, clinical research and bioinformatics is evident. This thesis attempted to connect these research fields by addressing clinically relevant questions regarding HF subtype characteristics in diverse biological and clinical data sets with a toolbox of statistical and machine learning methods.

In chapter I, I provided the first comprehensive review of public data sets of the bulk transcriptomic era in HF research. While previous meta-analysis often compared the intersections of reported up or downregulated genes, I showed that this strategy is a fallacy when used to judge the agreement of gene expression between studies. The consistency can be rather assessed by cross study disease score calculations or methods that compare transcriptome wide gene expression like gene set enrichment analysis. I showed that the consistency between HF studies is remarkably high, and further can be distilled to a gene consensus ranking. This ranking provides a gene expression reference that might have great translational impact for future experimental investigations that rely on identification of robust and generalizable targets for biomarkers or therapeutic approaches. A limitation of this generalization regards the represented limited heterogeneity of HF patients. Indeed, young, male patients with severe HFrEF were typically sampled, identifying our current knowledge gap in phenotypically diverging patient cohorts, including patients suffering from HFpEF.

In chapter II, a murine model was analyzed to address this knowledge gap in HFpEF. Here, I analyzed the single cell transcriptome of the interstitial cardiac cells of an early HFpEF mouse and identified fibroblast and macrophage to be key cell types. While fibroblast activation is a hallmark of HF in general, I compared different HF models to derive common and distinct patterns of fibroblast activation. I demonstrated that the typical collagen and extracellular matrix deposition is a feature of all models alike. However, additional gene regulation patterns characterized each model. In HFpEF, the metabolic and protein stress was a unique feature, together with basement membrane expression. I found that AngII and late post-MI fibroblast patterns displayed high similarity, supporting my findings in chapter I regarding similarity of DCM and ICM patients. I connected gene expression patterns to fibroblast phenotypes (i.e. fibroblast states) and I showed that the typical culprit of cardiac fibrosis, the matrifibrocyte, does only play a subsidiary role in the onset of fibrosis in HFpEF. Instead collagen deposition is a task performed by all fibroblast states. Furthermore, I derived that the activation of fibroblasts in the acute model of MI is driven by state-independent gene expression while the more chronic AngII and HFpEF models upregulated genes in dependency of fibroblast states i.e. in a manner of division of labor. Here, a possible difference between AngII and late post-MI fibroblast was demonstrated. The description of these fibroblast activation patterns, might be an important step to a more tissue centric understanding of fibroblast phenotypes contributing to different aspects of remodeling and thus enable more targeted and better adjusted antifibrotic therapy. For instance, Angtpl4 was derived as a marker of HFpEF fibroblast activation and could be recovered in human HFpEF plasma. However, the studied mouse model of HFpEF simplifies the complex disease etiology by using two pathological stimuli (i.e. HFD diet and L-NAME) and thus human studies that account for the systems level disorder that characterizes HFpEF are needed.

In chapter III, I addressed the multi-organ state of HFpEF patients by investigating comorbidity patterns. By analyzing the recorded disease codes of HF patients, I demonstrated that HFpEF and HFrEF can be distinguished by their comorbidity profiles alone. To characterize the distinctive comorbidities, I interpreted a patient classifier that yielded comorbidity associations of HFpEF including rheumatoid diseases, bone diseases and neoplasms. Furthermore, by constructing and analyzing a joint network of HF patient derived comorbidities, I showed that distinct disease groups tend to co-occur and form disease clusters. These clusters characterized the HFpEF and HFrEF cohort by capturing and summarizing distinct illness concepts. Thus, I found that HFrEF patients were more severely ill and suffered from more heart centered comorbidities while HFpEF patients suffered from more multi-organ diseases. Additionally, I showed that these comorbidity patterns can be used to infer associated genetic profiles that yielded a comorbidity based gene signature that aligned with transcriptomic changes in the mouse model.

In summary, I comparatively analyzed HFpEF and HFrEF on multiple scales by using diverse data sources gathered from mice and humans. In this work I followed an analysis strategy of contextualizing findings that I will briefly revisit. When comparing healthy with diseased samples, a strong disease signal can often be observed e.g. in transcriptome data, even when randomizing genes. This strength of signal is related to a diseased organ state affecting i) most of the cells in the tissue and ii) gene expression with almost genome-wide impact. In the case of HFpEF and HFrEF, the contrasting analysis enabled me to decipher this strong universal disease signal to detect more subtype specific patterns. To this end, I routinely incorporated various data sets that capture different aspects of HF for this task. This approach relied on available data, and thus might not be applicable in less intensely researched topics. However, even in well funded HF research, the data availability is highly limited, which relates to the lack of community standards for data sharing and annotation. While patient privacy concerns have to be considered, often motivation of individual data exploitation prevents open sharing. Thus, my thesis not only demonstrated the usefulness of integrating diverse prior knowledge resources and independent data sets, but it also highlighted the importance for the research community to facilitate data access, possibly by providing stricter publication rules enforcing or at least encouraging data publication. Besides improving reproducibility, the analysis of millions of data points is often a task that cannot be performed exhaustively within a single project but should rather be used as a resource to address diverse research questions.

Outlook

In this section, I will provide a brief outlook on future research efforts at the interface between data science and HF research. I summarized and reviewed the state of this field in a peer-reviewed manuscript [22]. Extracts and structure of this paragraph are cited from the review here.

To date, many challenges for HF research remain, as reflected by high mortality and morbidity rates and limited treatment options. With the acknowledgement of the heterogeneity of the HF syndrome, a crucial step was taken towards a more complete comprehension. Nevertheless, HFpEF and HFrEF phenotypes only insufficiently describe the diversity of pathways leading to HF onset and progression. Hence, evaluating more ramified patient HF phenotypes will provide valuable insights into more individualized pathomechanisms, following the notion of so-called personalized medicine. This might help us to consider more and more factors of a patient's disease course and bring us much closer to the goal of treating the right patient with the right treatment at the right time. Thus, the direction that HF research has taken towards (big) data collection and bioinformatic evaluation promises to advance our knowledge substantially. However, major issues of this approach include biases in the collection and analysis of that data. As we rush to collect ever larger sample sizes, we should pause to carefully consider whether we are merely enthralled by ever increasing data samples (so-called data chauvinism [300]) or whether the biological question is best answered by data of the type and quality available. For many omics technologies, the number of features considered requires large samples, or the noise introduced will result in inferior model fitting. In other cases, a large sample size can be less informative if the sampling is of lower quality, for instance if non-probabilistic sampling was applied [122]. Thus, many omic studies, especially those analyzing sparse myocardial tissue, suffer from small patient cohorts that can not compensate for the biological and clinical variability. A large-scale effort to acquire and comprehensively characterize relevant tissue samples with a variety of omics techniques would ameliorate this issue and potentially provide greater insight into the biology of HF. Such efforts have proven valuable in other areas, most notably in oncology (e.g. The Cancer Genome Atlas Program (TCGA)). In clinical data analysis we must balance the desire to find subsets of patients that share characteristics, with the goal of making sure that all patients benefit from the potential of precision medicine. Concerns about sampling bias, data missingness, and measurement error in big data, and especially big clinical data, are all relevant to research in HF [301–303]. These data quality concerns are also important because they will directly affect the output of machine learning analyses [22,304].

Lastly, despite the excitement about big data analysis and bioinformatics, the ultimate goal in medicine must always be to improve human health. Physicians should receive additional training allowing them to appropriately evaluate the potential of novel algorithmic tools in clinical care [305]. To successfully implement precision medicine approaches based on omics and big data technologies, clinicians will need to understand the strengths and weaknesses of methodologies and have confidence in their relevance to disease. The role of data science and bioinformatics in HF prevention and treatment necessitates a multi-disciplinary discussion where physicians are needed to take a leading role.

Glossary

ANOVA	Analysis of variance
AngII	Angiotensin II
BH	Benjamini and Hochberg
CM	Cardiomyocyte
DC	Disease cluster
ECG	echocardiogram
ECM	Extracellular matrix
ELISA	Enzyme-linked Immunosorbent Assay
GO	Gene Ontology
HF	Heart failure
HFD	High fat diet
HFmrEF	Heart failure with mid range reduced ejection fraction
HFnet	Heart failure comorbidity network
HFhetnet	Heart failure heterogeneous network
HFpEF	Heart failure with preserved ejection fraction
HFrEF	Heart failure with reduced ejection fraction
HF-CS	Heart failure consensus signature
IFS	Integrated fibroblast state
L-NAME	N(ω)-nitro-L-arginine methyl ester
LVEF	Left ventricular ejection fraction
MCA	Multiple Correspondence Analysis
MI	Myocardial Infarction
NYHA	New York Heart Failure Association (Classification)
PCA	Principal component analysis
PPI	Protein-protein interaction
RWR-MH	Random walk with restart on multilayer heterogeneous network
SVES	Supraventricular extrasystole
TF	Transcription Factor
TAPSE	Tricuspid annular plane systolic excursion

Bibliography

1. Mann, D. L. & Bristow, M. R. Mechanisms and models in heart failure: the biomechanical model and beyond. *Circulation* **111**, 2837–2849 (2005).
2. Pfeffer, M. A., Shah, A. M. & Borlaug, B. A. Heart failure with preserved ejection fraction in perspective. *Circ. Res.* **124**, 1598–1617 (2019).
3. Yusuf, S. *et al.* Effects of candesartan in patients with chronic heart failure and preserved left-ventricular ejection fraction: the CHARM-Preserved Trial. *Lancet* **362**, 777–781 (2003).
4. Dougherty, A. H., Naccarelli, G. V., Gray, E. L., Hicks, C. H. & Goldstein, R. A. Congestive heart failure with normal systolic function. *Am. J. Cardiol.* **54**, 778–782 (1984).
5. Dunlay, S. M., Roger, V. L. & Redfield, M. M. Epidemiology of heart failure with preserved ejection fraction. *Nat. Rev. Cardiol.* **14**, 591–602 (2017).
6. Holstiege, J., Akmatov, M. K., Störk, S., Steffen, A. & Bätzing, J. Higher prevalence of heart failure in rural regions: a population-based study covering 87% of German inhabitants. *Clin. Res. Cardiol.* **108**, 1102–1106 (2019).
7. Mishra, S. & Kass, D. A. Cellular and molecular pathobiology of heart failure with preserved ejection fraction. *Nat. Rev. Cardiol.* **18**, 400–423 (2021).
8. Samson, R. & Le Jemtel, T. H. Therapeutic stalemate in heart failure with preserved ejection fraction. *J. Am. Heart Assoc.* **10**, e021120 (2021).
9. Nagamine, T. *et al.* Multiscale classification of heart failure phenotypes by unsupervised clustering of unstructured electronic medical record data. *Sci. Rep.* **10**, 21340 (2020).
10. Gulea, C., Zakeri, R. & Quint, J. K. Model-based comorbidity clusters in patients with heart failure: association with clinical outcomes and healthcare utilization. *BMC Med.* **19**, 9 (2021).
11. Woolley, R. J. *et al.* Machine learning based on biomarker profiles identifies distinct subgroups of heart failure with preserved ejection fraction. *Eur. J. Heart Fail.* **23**, 983–991 (2021).
12. Shah, S. J. *et al.* Research priorities for heart failure with preserved ejection fraction: national heart, lung, and blood institute working group summary. *Circulation* **141**, 1001–1026 (2020).
13. Simmonds, S. J., Cuijpers, I., Heymans, S. & Jones, E. A. V. Cellular and Molecular Differences between HFpEF and HFrEF: A Step Ahead in an Improved Pathological Understanding. *Cells* **9**, (2020).
14. Sweeney, M., Corden, B. & Cook, S. A. Targeting cardiac fibrosis in heart failure

with preserved ejection fraction: mirage or miracle? *EMBO Mol. Med.* **12**, e10865 (2020).

15. Nakamura, M. & Sadoshima, J. Mechanisms of physiological and pathological cardiac hypertrophy. *Nat. Rev. Cardiol.* **15**, 387–407 (2018).
16. Hinderer, S. & Schenke-Layland, K. Cardiac fibrosis - A short review of causes and therapeutic strategies. *Adv. Drug Deliv. Rev.* **146**, 77–82 (2019).
17. Zile, M. R. *et al.* Myocardial stiffness in patients with heart failure and a preserved ejection fraction: contributions of collagen and titin. *Circulation* **131**, 1247–1259 (2015).
18. Kanagala, P. *et al.* Relationship between focal and diffuse fibrosis assessed by CMR and clinical outcomes in heart failure with preserved ejection fraction. *JACC Cardiovasc. Imaging* **12**, 2291–2301 (2019).
19. Henderson, N. C., Rieder, F. & Wynn, T. A. Fibrosis: from mechanisms to medicines. *Nature* **587**, 555–566 (2020).
20. Wynn, T. A. & Ramalingam, T. R. Mechanisms of fibrosis: therapeutic translation for fibrotic disease. *Nat. Med.* **18**, 1028–1040 (2012).
21. Schiattarella, G. G. *et al.* Nitrosative stress drives heart failure with preserved ejection fraction. *Nature* **568**, 351–356 (2019).
22. Lanzer, J. D., Leuschner, F., Kramann, R., Levinson, R. T. & Saez-Rodriguez, J. Big data approaches in heart failure research. *Curr. Heart Fail. Rep.* **17**, 213–224 (2020).
23. Lee, T. I. & Young, R. A. Transcriptional regulation and its misregulation in disease. *Cell* **152**, 1237–1251 (2013).
24. Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. *Nat. Rev. Genet.* **20**, 631–656 (2019).
25. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
26. Willyard, C. New human gene tally reignites debate. *Nature* **558**, 354–355 (2018).
27. Dugourd, A. & Saez-Rodriguez, J. Footprint-based functional analysis of multiomic data. *Current Opinion in Systems Biology* **15**, 82–90 (2019).
28. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* **102**, 15545–15550 (2005).
29. Väremo, L., Nielsen, J. & Nookaew, I. Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. *Nucleic Acids Res.* **41**, 4378–4391 (2013).
30. Badia-I-Mompel, P. *et al.* decoupleR: ensemble of computational methods to infer

biological activities from omics data. *Bioinformatics Advances* **2**, vbac016 (2022).

31. Yang, J. *et al.* Decreased SLIM1 expression and increased gelsolin expression in failing human hearts measured by high-density oligonucleotide arrays. *Circulation* **102**, 3046–3052 (2000).
32. Raghow, R. An "omics" perspective on cardiomyopathies and heart failure. *Trends Mol. Med.* **22**, 813–827 (2016).
33. Kim, G. H., Uriel, N. & Burkhoff, D. Reverse remodelling and myocardial recovery in heart failure. *Nat. Rev. Cardiol.* **15**, 83–96 (2018).
34. Peterzan, M. A., Lygate, C. A., Neubauer, S. & Rider, O. J. Metabolic remodeling in hypertrophied and failing myocardium: a review. *Am. J. Physiol. Heart Circ. Physiol.* **313**, H597–H616 (2017).
35. Liu, Y. *et al.* RNA-Seq identifies novel myocardial gene expression signatures of heart failure. *Genomics* **105**, 83–89 (2015).
36. Louzao-Martinez, L. *et al.* Characteristic adaptations of the extracellular matrix in dilated cardiomyopathy. *Int. J. Cardiol.* **220**, 634–646 (2016).
37. Mann, D. L., Topkara, V. K., Evans, S. & Barger, P. M. Innate immunity in the adult mammalian heart: for whom the cell tolls. *Trans. Am. Clin. Climatol. Assoc.* **121**, 34–50; discussion 50 (2010).
38. Alimadadi, A., Munroe, P. B., Joe, B. & Cheng, X. Meta-Analysis of Dilated Cardiomyopathy Using Cardiac RNA-Seq Transcriptomic Datasets. *Genes (Basel)* **11**, (2020).
39. Sharma, U. C., Pokharel, S., Evelo, C. T. A. & Maessen, J. G. A systematic review of large scale and heterogeneous gene array data in heart failure. *J. Mol. Cell. Cardiol.* **38**, 425–432 (2005).
40. Barth, A. S. *et al.* Reciprocal transcriptional regulation of metabolic and signaling pathways correlates with disease severity in heart failure. *Circ. Cardiovasc. Genet.* **4**, 475–483 (2011).
41. Toro-Domínguez, D. *et al.* A survey of gene expression meta-analysis: methods and applications. *Brief. Bioinformatics* **22**, 1694–1705 (2021).
42. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).
43. Browaeys, R., Saelens, W. & Saeys, Y. NicheNet: modeling intercellular communication by linking ligands to target genes. *Nat. Methods* **17**, 159–162 (2020).
44. Holland, C. H. *et al.* Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. *Genome Biol.* **21**, 36 (2020).
45. Argelaguet, R. *et al.* Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.* **14**, e8124 (2018).

46. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* **13**, 599–604 (2018).

47. Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C. & Teichmann, S. A. The technology and biology of single-cell RNA sequencing. *Mol. Cell* **58**, 610–620 (2015).

48. Selewa, A. *et al.* Systematic Comparison of High-throughput Single-Cell and Single-Nucleus Transcriptomes during Cardiomyocyte Differentiation. *Sci. Rep.* **10**, 1535 (2020).

49. Yekelchyk, M., Guenther, S., Preussner, J. & Braun, T. Mono- and multi-nucleated ventricular cardiomyocytes constitute a transcriptionally homogenous cell population. *Basic Res. Cardiol.* **114**, 36 (2019).

50. Argelaguet, R., Cuomo, A. S. E., Stegle, O. & Marioni, J. C. Computational principles and challenges in single-cell data integration. *Nat. Biotechnol.* **39**, 1202–1215 (2021).

51. Luecken, M. D. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).

52. Rukhlenko, O. S. *et al.* Control of cell state transitions. *Nature* **609**, 975–985 (2022).

53. Trapnell, C. Defining cell types and states with single-cell genomics. *Genome Res.* **25**, 1491–1498 (2015).

54. Petukhov, V. *et al.* Case-control analysis of single-cell RNA-seq studies. *bioRxiv* (2022).

55. Laehnemann, D. *et al.* 12 Grand Challenges in Single-Cell Data Science. (2019) doi:10.7287/peerj.preprints.27885v3.

56. Chaudhry, F. *et al.* Single-Cell RNA Sequencing of the Cardiovascular System: New Looks for Old Diseases. *Front. Cardiovasc. Med.* **6**, 173 (2019).

57. Ackers-Johnson, M., Tan, W. L. W. & Foo, R. S.-Y. Following hearts, one cell at a time: recent applications of single-cell RNA sequencing to the understanding of heart disease. *Nat. Commun.* **9**, 4434 (2018).

58. Molenaar, B. & van Rooij, E. Single-Cell Sequencing of the Mammalian Heart. *Circ. Res.* **123**, 1033–1035 (2018).

59. Meilhac, S. M. & Buckingham, M. E. The deployment of cell lineages that form the mammalian heart. *Nat. Rev. Cardiol.* **15**, 705–724 (2018).

60. Miranda, A. M. A. *et al.* Single-cell transcriptomics for the assessment of cardiac disease. *Nat. Rev. Cardiol.* (2022) doi:10.1038/s41569-022-00805-7.

61. Skelly, D. A. *et al.* Single-Cell Transcriptional Profiling Reveals Cellular Diversity and Intercommunication in the Mouse Heart. *Cell Rep.* **22**, 600–610 (2018).

62. Wolfien, M. *et al.* Single-Nucleus Sequencing of an Entire Mammalian Heart: Cell

Type Composition and Velocity. *Cells* **9**, (2020).

63. Tabula Muris Consortium *et al.* Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367–372 (2018).
64. Xiong, H. *et al.* Single-Cell Transcriptomics Reveals Chemotaxis-Mediated Intraorgan Crosstalk During Cardiogenesis. *Circ. Res.* **125**, 398–410 (2019).
65. Asp, M. *et al.* A Spatiotemporal Organ-Wide Gene Expression and Cell Atlas of the Developing Human Heart. *Cell* **179**, 1647-1660.e19 (2019).
66. Li, G. *et al.* Single cell expression analysis reveals anatomical and cell cycle-dependent transcriptional shifts during heart development. *Development* **146**, (2019).
67. Phansalkar, R. & Red-Horse, K. Techniques converge to map the developing human heart at single-cell level. *Nature* **577**, 629–630 (2020).
68. Cui, Y. *et al.* Single-Cell Transcriptome Analysis Maps the Developmental Track of the Human Heart. *Cell Rep.* **26**, 1934-1950.e5 (2019).
69. Litviňuková, M. *et al.* Cells of the adult human heart. *Nature* **588**, 466–472 (2020).
70. Chaffin, M. *et al.* Single-nucleus profiling of human dilated and hypertrophic cardiomyopathy. *Nature* **608**, 174–180 (2022).
71. Koenig, A. L. *et al.* Single-cell transcriptomics reveals cell-type-specific diversification in human heart failure. *Nat. Cardiovasc. Res.* **1**, 263–280 (2022).
72. Datta, S., Bernstam, E. V. & Roberts, K. A frame semantic overview of NLP-based information extraction for cancer-related EHR notes. *J. Biomed. Inform.* **100**, 103301 (2019).
73. Zeng, Z., Deng, Y., Li, X., Naumann, T. & Luo, Y. Natural Language Processing for EHR-Based Computational Phenotyping. *IEEE/ACM Trans Comput Biol Bioinform* **16**, 139–153 (2019).
74. Sheikhalishahi, S. *et al.* Natural language processing of clinical notes on chronic diseases: systematic review. *JMIR Med. Inform.* **7**, e12239 (2019).
75. Iorio, A., Pozzi, A. & Senni, M. Addressing the heterogeneity of heart failure in future randomized trials. *Curr. Heart Fail. Rep.* **14**, 197–202 (2017).
76. Altman, R. B. & Ashley, E. A. Using “big data” to dissect clinical heterogeneity. *Circulation* **131**, 232–233 (2015).
77. Desai, R. J., Wang, S. V., Vaduganathan, M., Evers, T. & Schneeweiss, S. Comparison of machine learning methods with traditional models for use of administrative claims with electronic medical records to predict heart failure outcomes. *JAMA Netw. Open* **3**, e1918962 (2020).
78. Adler, E. D. *et al.* Improving risk prediction in heart failure using machine learning. *Eur. J. Heart Fail.* **22**, 139–147 (2020).

79. Ahmad, T. *et al.* Machine learning methods improve prognostication, identify clinically distinct phenotypes, and detect heterogeneity in response to therapy in a large cohort of heart failure patients. *J. Am. Heart Assoc.* **7**, (2018).

80. Angraal, S. *et al.* Machine learning prediction of mortality and hospitalization in heart failure with preserved ejection fraction. *JACC Heart Fail.* **8**, 12–21 (2020).

81. Cikes, M. *et al.* Machine learning-based phenogrouping in heart failure to identify responders to cardiac resynchronization therapy. *Eur. J. Heart Fail.* **21**, 74–85 (2019).

82. Tabassian, M. *et al.* Diagnosis of heart failure with preserved ejection fraction: machine learning of spatiotemporal variations in left ventricular deformation. *J. Am. Soc. Echocardiogr.* **31**, 1272-1284.e9 (2018).

83. Acharya, U. R. *et al.* Deep convolutional neural network for the automated diagnosis of congestive heart failure using ECG signals. *Appl. Intell.* 1–12 (2018) doi:10.1007/s10489-018-1179-1.

84. Ambale-Venkatesh, B. *et al.* Cardiovascular Event Prediction by Machine Learning: The Multi-Ethnic Study of Atherosclerosis. *Circ. Res.* **121**, 1092–1101 (2017).

85. Nirschl, J. J. *et al.* A deep-learning classifier identifies patients with clinical heart failure using whole-slide images of H&E tissue. *PLoS ONE* **13**, e0192726 (2018).

86. Inan, O. T. *et al.* Novel wearable seismocardiography and machine learning algorithms can assess clinical status of heart failure patients. *Circ. Heart Fail.* **11**, e004313 (2018).

87. Stehlik, J. *et al.* Continuous Wearable Monitoring Analytics Predict Heart Failure Hospitalization: The LINK-HF Multicenter Study. *Circ. Heart Fail.* **13**, e006513 (2020).

88. Meghani, S. H. *et al.* The conceptualization and measurement of comorbidity: a review of the interprofessional discourse. *Nurs. Res. Pract.* **2013**, 192782 (2013).

89. Scully, J. L. What is a disease? *EMBO Rep.* **5**, 650–653 (2004).

90. Pandey, A. *et al.* Temporal trends in prevalence and prognostic implications of comorbidities among patients with acute decompensated heart failure: the ARIC study community surveillance. *Circulation* **142**, 230–243 (2020).

91. Palmiero, G. *et al.* Impact of SGLT2 inhibitors on heart failure: from pathophysiology to clinical effects. *Int. J. Mol. Sci.* **22**, (2021).

92. Maron, B. A. *et al.* A global network for network medicine. *NPJ Syst. Biol. Appl.* **6**, 29 (2020).

93. Hu, J. X., Thomas, C. E. & Brunak, S. Network biology concepts in complex disease comorbidities. *Nat. Rev. Genet.* **17**, 615–629 (2016).

94. Khan, A., Uddin, S. & Srinivasan, U. Comorbidity network for chronic disease: A

novel approach to understand type 2 diabetes progression. *Int. J. Med. Inform.* **115**, 1–9 (2018).

95. Guo, M. *et al.* Analysis of disease comorbidity patterns in a large-scale China population. *BMC Med. Genomics* **12**, 177 (2019).

96. Aguado, A., Moratalla-Navarro, F., López-Simarro, F. & Moreno, V. MorbiNet: multimorbidity networks in adult general population. Analysis of type 2 diabetes mellitus comorbidity. *Sci. Rep.* **10**, 2416 (2020).

97. Hidalgo, C. A., Blumm, N., Barabási, A.-L. & Christakis, N. A. A dynamic network approach for the study of human phenotypes. *PLoS Comput. Biol.* **5**, e1000353 (2009).

98. Cruz-Ávila, H. A., Vallejo, M., Martínez-García, M. & Hernández-Lemus, E. Comorbidity networks in cardiovascular diseases. *Front. Physiol.* **11**, 1009 (2020).

99. Brunson, J. C., Agresta, T. P. & Laubenbacher, R. C. Sensitivity of comorbidity network analysis. *JAMIA Open* (2019) doi:10.1093/jamiaopen/ooz067.

100. Sadegh, S. *et al.* Lacking mechanistic disease definitions and corresponding association data hamper progress in network medicine and beyond. *Nat. Commun.* **14**, 1662 (2023).

101. Goh, K.-I. *et al.* The human disease network. *Proc Natl Acad Sci USA* **104**, 8685–8690 (2007).

102. Ramirez Flores, R. O. *et al.* Consensus Transcriptional Landscape of Human End-Stage Heart Failure. *J. Am. Heart Assoc.* **10**, e019667 (2021).

103. Litvinukova, M. *et al.* Cells and gene expression programs in the adult human heart. *BioRxiv* (2020) doi:10.1101/2020.04.03.024075.

104. Hannenhalli, S. *et al.* Transcriptional genomics associates FOX transcription factors with human heart failure. *Circulation* **114**, 1269–1276 (2006).

105. van Heesch, S. *et al.* The translational landscape of the human heart. *Cell* **178**, 242-260.e29 (2019).

106. Sweet, M. E. *et al.* Transcriptome analysis of human heart failure reveals dysregulated cell adhesion in dilated cardiomyopathy and activated immune pathways in ischemic heart failure. *BMC Genomics* **19**, 812 (2018).

107. Kittleson, M. M. *et al.* Gene expression analysis of ischemic and nonischemic cardiomyopathy: shared and distinct genes in the development of heart failure. *Physiol. Genomics* **21**, 299–307 (2005).

108. Tarazón, E. *et al.* RNA sequencing analysis and atrial natriuretic peptide production in patients with dilated and ischemic cardiomyopathy. *PLoS ONE* **9**, e90157 (2014).

109. Read, D. F. *et al.* Single-cell analysis of chromatin and expression reveals age- and

sex-associated alterations in the human heart. *BioRxiv* (2022) doi:10.1101/2022.07.12.496461.

110. Kong, S. W. *et al.* Heart failure-associated changes in RNA splicing of sarcomere genes. *Circ. Cardiovasc. Genet.* **3**, 138–146 (2010).

111. Molina-Navarro, M. M. *et al.* Differential gene expression of cardiac ion channels in human dilated cardiomyopathy. *PLoS ONE* **8**, e79792 (2013).

112. Greco, S. *et al.* MicroRNA dysregulation in diabetic ischemic heart failure patients. *Diabetes* **61**, 1633–1641 (2012).

113. Yang, K.-C. *et al.* Deep RNA sequencing reveals dynamic regulation of myocardial noncoding RNAs in failing human heart and remodeling with mechanical circulatory support. *Circulation* **129**, 1009–1021 (2014).

114. Barth, A. S. *et al.* Identification of a common gene expression signature in dilated cardiomyopathy across independent microarray studies. *J. Am. Coll. Cardiol.* **48**, 1610–1617 (2006).

115. Pepin, M. E. *et al.* DNA methylation reprograms cardiac metabolic gene expression in end-stage human heart failure. *Am. J. Physiol. Heart Circ. Physiol.* **317**, H674–H684 (2019).

116. Kim, E. H. *et al.* Differential protein expression and basal lamina remodeling in human heart failure. *Proteomics Clin. Appl.* **10**, 585–596 (2016).

117. Schiano, C. *et al.* Heart failure: Pilot transcriptomic analysis of cardiac tissue by RNA-sequencing. *Cardiol. J.* **24**, 539–553 (2017).

118. Francis, R. & Lewis, C. Myocardial biopsy: techniques and indications. *Heart* **104**, 950–958 (2018).

119. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

120. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).

121. Korotkevich, G. *et al.* Fast gene set enrichment analysis. *BioRxiv* (2016) doi:10.1101/060012.

122. Meng, X.-L. Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *Ann. Appl. Stat.* **12**, 685–726 (2018).

123. Nakao, K., Minobe, W., Roden, R., Bristow, M. R. & Leinwand, L. A. Myosin heavy chain gene expression in human heart failure. *J. Clin. Invest.* **100**, 2362–2370 (1997).

124. Petretto, E. *et al.* Integrated genomic approaches implicate osteoglycin (Ogn) in the

regulation of left ventricular mass. *Nat. Genet.* **40**, 546–552 (2008).

125. Wittchen, F. *et al.* Genomic expression profiling of human inflammatory cardiomyopathy (DCMi) suggests novel therapeutic targets. *J. Mol. Med.* **85**, 257–271 (2007).

126. Laugier, L. *et al.* Whole-Genome Cardiac DNA Methylation Fingerprint and Gene Expression Analysis Provide New Insights in the Pathogenesis of Chronic Chagas Disease Cardiomyopathy. *Clin. Infect. Dis.* **65**, 1103–1111 (2017).

127. Akat, K. M. *et al.* Comparative RNA-sequencing analysis of myocardial and circulating small RNAs in human heart failure and their utility as biomarkers. *Proc Natl Acad Sci USA* **111**, 11151–11156 (2014).

128. Dirkx, E., da Costa Martins, P. A. & De Windt, L. J. Regulation of fetal gene expression in heart failure. *Biochim. Biophys. Acta* **1832**, 2414–2424 (2013).

129. Wu, R. *et al.* Cardiac-specific ablation of ARNT leads to lipotoxicity and cardiomyopathy. *J. Clin. Invest.* **124**, 4795–4806 (2014).

130. Mann, D. L. Innate immunity and the failing heart: the cytokine hypothesis revisited. *Circ. Res.* **116**, 1254–1268 (2015).

131. Guo, X. *et al.* Cardioprotective Role of Tumor Necrosis Factor Receptor-Associated Factor 2 by Suppressing Apoptosis and Necroptosis. *Circulation* **136**, 729–742 (2017).

132. Papathanasiou, S. *et al.* Tumor necrosis factor-α confers cardioprotection through ectopic expression of keratins K8 and K18. *Nat. Med.* **21**, 1076–1084 (2015).

133. Egerstedt, A. *et al.* Profiling of the plasma proteome across different stages of human heart failure. *Nat. Commun.* **10**, 5830 (2019).

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

135. van der Pol, A., Hoes, M. F., de Boer, R. A. & van der Meer, P. Cardiac foetal reprogramming: a tool to exploit novel treatment targets for the failing heart. *J. Intern. Med.* **288**, 491–506 (2020).

136. Vandewiele, F. *et al.* TRPM4 inhibition by meclofenamate suppresses Ca2+-dependent triggered arrhythmias. *Eur. Heart J.* **43**, 4195–4207 (2022).

137. Takada, S. *et al.* Succinyl-CoA-based energy metabolism dysfunction in chronic heart failure. *Proc Natl Acad Sci USA* **119**, e2203628119 (2022).

138. Santos, G. L. *et al.* Using different geometries to modulate the cardiac fibroblast phenotype and the biomechanical properties of engineered connective tissues. *Biomater. Adv.* **139**, 213041 (2022).

139. Liu, X. & Song, J. The application of autopsy and explanted heart samples in scientific research. *Cardiovasc. Pathol.* **59**, 107424 (2022).

140. Núñez, J. *et al.* Serum neprilysin and recurrent admissions in patients with heart failure. *J. Am. Heart Assoc.* **6**, (2017).

141. McMurray, J. J. V. *et al.* Angiotensin-neprilysin inhibition versus enalapril in heart failure. *N. Engl. J. Med.* **371**, 993–1004 (2014).

142. Lázár, E., Sadek, H. A. & Bergmann, O. Cardiomyocyte renewal in the human heart: insights from the fall-out. *Eur. Heart J.* **38**, 2333–2342 (2017).

143. Maliken, B. D. *et al.* Gata4-Dependent Differentiation of c-Kit+-Derived Endothelial Cells Underlies Artefactual Cardiomyocyte Regeneration in the Heart. *Circulation* **138**, 1012–1024 (2018).

144. Li, Y. *et al.* Genetic lineage tracing of nonmyocyte population by dual recombinases. *Circulation* **138**, 793–805 (2018).

145. Amini, H., Rezaie, J., Vosoughi, A., Rahbarghazi, R. & Nouri, M. Cardiac progenitor cells application in cardiovascular disease. *J. Cardiovasc. Thorac. Res.* **9**, 127–132 (2017).

146. Pereira, A. H. M. *et al.* MEF2C silencing attenuates load-induced left ventricular hypertrophy by modulating mTOR/S6K pathway in mice. *PLoS ONE* **4**, e8472 (2009).

147. Filomena, M. C. & Bang, M.-L. In the heart of the MEF2 transcription network: novel downstream effectors as potential targets for the treatment of cardiovascular disease. *Cardiovasc. Res.* **114**, 1425–1427 (2018).

148. Constantinou, C. *et al.* The multi-faceted functioning portrait of LRF/ZBTB7A. *Hum Genomics* **13**, 66 (2019).

149. Ramos Pittol, J. M., Oruba, A., Mittler, G., Saccani, S. & van Essen, D. Zbtb7a is a transducer for the control of promoter accessibility by NF-kappa B and multiple other transcription factors. *PLoS Biol.* **16**, e2004526 (2018).

150. Xuan, Y. T., Guo, Y., Han, H., Zhu, Y. & Bolli, R. An essential role of the JAK-STAT pathway in ischemic preconditioning. *Proc Natl Acad Sci USA* **98**, 9050–9055 (2001).

151. Wagner, M. A. & Siddiqui, M. A. Q. The JAK-STAT pathway in hypertrophic stress signaling and genomic stress response. *JAKSTAT* **1**, 131–141 (2012).

152. Boengler, K., Hilfiker-Kleiner, D., Drexler, H., Heusch, G. & Schulz, R. The myocardial JAK/STAT pathway: from protection to failure. *Pharmacol. Ther.* **120**, 172–185 (2008).

153. Chung, E. S. *et al.* Randomized, double-blind, placebo-controlled, pilot trial of infliximab, a chimeric monoclonal antibody to tumor necrosis factor-alpha, in patients with moderate-to-severe heart failure: results of the anti-TNF Therapy Against Congestive Heart Failure (ATTACH) trial. *Circulation* **107**, 3133–3140

(2003).

154. Mann, D. L. *et al.* Targeted anticytokine therapy in patients with chronic heart failure: results of the Randomized Etanercept Worldwide Evaluation (RENEWAL). *Circulation* **109**, 1594–1602 (2004).

155. Chen, Y. & Bache, R. J. Adenosine: a modulator of the cardiac response to stress. *Circ. Res.* **93**, 691–693 (2003).

156. Rogers, J. H. *et al.* RGS4 causes increased mortality and reduced cardiac hypertrophy in response to pressure overload. *J. Clin. Invest.* **104**, 567–576 (1999).

157. Quast, C., Alter, C., Ding, Z., Borg, N. & Schrader, J. Adenosine Formed by CD73 on T Cells Inhibits Cardiac Inflammation and Fibrosis and Preserves Contractile Function in Transverse Aortic Constriction-Induced Heart Failure. *Circ. Heart Fail.* **10**, (2017).

158. Houweling, A. C., van Borren, M. M., Moorman, A. F. M. & Christoffels, V. M. Expression and regulation of the atrial natriuretic factor encoding gene Nppa during development and disease. *Cardiovasc. Res.* **67**, 583–593 (2005).

159. Askevold, E. T. *et al.* Secreted Frizzled Related Protein 3 in Chronic Heart Failure: Analysis from the Controlled Rosuvastatin Multinational Trial in Heart Failure (CORONA). *PLoS ONE* **10**, e0133970 (2015).

160. Schutt, R. C., Burdick, M. D., Strieter, R. M., Mehrad, B. & Keeley, E. C. Plasma CXCL12 levels as a predictor of future stroke. *Stroke* **43**, 3382–3386 (2012).

161. Döring, Y., Pawig, L., Weber, C. & Noels, H. The CXCL12/CXCR4 chemokine ligand/receptor axis in cardiovascular disease. *Front. Physiol.* **5**, 212 (2014).

162. Iaccarino, D. *et al.* P316Expression and functional role of Ccdc80 in developing heart and in cardiomyopathies. *Cardiovasc. Res.* **103 Suppl 1**, S57-8 (2014).

163. Blanton, R. M., Cooper, C., Hergruetter, A., Aronovitz, M. & Calamaras, T. D. Abstract 154: CCDC80 functions as a protein kinase GI substrate and is secreted by cardiac myocytes. *Circ. Res.* **121**, (2017).

164. Withaar, C., Lam, C. S. P., Schiattarella, G. G., de Boer, R. A. & Meems, L. M. G. Heart failure with preserved ejection fraction in humans and mice: embracing clinical complexity in mouse models. *Eur. Heart J.* **42**, 4420–4430 (2021).

165. Schiattarella, G. G. *et al.* Immunometabolic Mechanisms of Heart Failure with Preserved Ejection Fraction. *Nat. Cardiovasc. Res.* **1**, 211–222 (2022).

166. Travers, J. G. *et al.* HDAC inhibition reverses preexisting diastolic dysfunction and blocks covert extracellular matrix remodeling. *Circulation* **143**, 1874–1890 (2021).

167. Abplanalp, W. T., Tucker, N. & Dimmeler, S. Single-cell technologies to decipher cardiovascular diseases. *Eur. Heart J.* **43**, 4536–4547 (2022).

168. Forte, E. *et al.* Dynamic Interstitial Cell Response during Myocardial Infarction

Predicts Resilience to Rupture in Genetically Diverse Mice. *Cell Rep.* **30**, 3149-3163.e6 (2020).

169. McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst.* **8**, 329-337.e4 (2019).

170. O'Flanagan, C. H. *et al.* Dissociation of solid tumor tissues with cold active protease for single-cell RNA-seq minimizes conserved collagenase-associated stress responses. *Genome Biol.* **20**, 210 (2019).

171. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *BioRxiv* (2020) doi:10.1101/2020.10.12.335331.

172. Skinnider, M. A. *et al.* Cell type prioritization in single-cell data. *Nat. Biotechnol.* **39**, 30–34 (2021).

173. Furtado, M. B., Costa, M. W. & Rosenthal, N. A. The cardiac fibroblast: Origin, identity and role in homeostasis and disease. *Differentiation.* **92**, 93–101 (2016).

174. McLellan, M. A. *et al.* High-Resolution Transcriptomic Profiling of the Heart During Chronic Stress Reveals Cellular Drivers of Cardiac Fibrosis and Hypertrophy. *Circulation* **142**, 1448–1463 (2020).

175. Hesse, J. *et al.* Single-cell transcriptomics defines heterogeneity of epicardial cells and fibroblasts within the infarcted murine heart. *eLife* **10**, (2021).

176. Ruiz-Villalba, A. *et al.* Single-Cell RNA Sequencing Analysis Reveals a Crucial Role for CTHRC1 (Collagen Triple Helix Repeat Containing 1) Cardiac Fibroblasts After Myocardial Infarction. *Circulation* **142**, 1831–1847 (2020).

177. Buechler, M. B. *et al.* Cross-tissue organization of the fibroblast lineage. *Nature* **593**, 575–579 (2021).

178. Farbehi, N. *et al.* Single-cell expression profiling reveals dynamic flux of cardiac stromal, vascular and immune cells in health and injury. *eLife* **8**, (2019).

179. Elmentaite, R., Domínguez Conde, C., Yang, L. & Teichmann, S. A. Single-cell atlases: shared and tissue-specific cell types across human organs. *Nat. Rev. Genet.* **23**, 395–410 (2022).

180. Naba, A. *et al.* The matrisome: in silico definition and in vivo characterization by proteomics of normal and tumor extracellular matrices. *Mol. Cell. Proteomics* **11**, M111.014647 (2012).

181. Yi, X. *et al.* Overexpression of chaperonin containing T-complex polypeptide subunit zeta 2 (CCT6b) suppresses the functions of active fibroblasts in a rat model of joint contracture. *J. Orthop. Surg. Res.* **14**, 125 (2019).

182. Araki, K. *et al.* Functional profiling of asymmetrically-organized human CCT/TRiC chaperonin. *Biochem. Biophys. Res. Commun.* **481**, 232–238 (2016).

183. Qiu, X. *et al.* Overexpression of CCT8 and its significance for tumor cell proliferation, migration and invasion in glioma. *Pathol. Res. Pract.* **211**, 717–725 (2015).

184. Onishi, K., Higuchi, M., Asakura, T., Masuyama, N. & Gotoh, Y. The PI3K-Akt pathway promotes microtubule stabilization in migrating fibroblasts. *Genes Cells* **12**, 535–546 (2007).

185. Welf, E. S., Ahmed, S., Johnson, H. E., Melvin, A. T. & Haugh, J. M. Migrating fibroblasts reorient directionality by a metastable, PI3K-dependent mechanism. *J. Cell Biol.* **197**, 105–114 (2012).

186. Emori, T. *et al.* Role of JAK-STAT signaling in the pathogenic behavior of fibroblast-like synoviocytes in rheumatoid arthritis: Effect of the novel JAK inhibitor peficitinib. *Eur. J. Pharmacol.* **882**, 173238 (2020).

187. Horiuchi, M. *et al.* Interferon-gamma induces AT(2) receptor expression in fibroblasts by Jak/STAT pathway and interferon regulatory factor-1. *Circ. Res.* **86**, 233–240 (2000).

188. Xu, L. *et al.* The participation of fibroblast growth factor 23 (FGF23) in the progression of osteoporosis via JAK/STAT pathway. *J. Cell. Biochem.* **119**, 3819–3828 (2018).

189. Muhl, L. *et al.* Single-cell analysis uncovers fibroblast heterogeneity and criteria for fibroblast and mural cell identification and discrimination. *Nat. Commun.* **11**, 3953 (2020).

190. Takeda, N. *et al.* Cardiac fibroblasts are essential for the adaptive response of the murine heart to pressure overload. *J. Clin. Invest.* **120**, 254–265 (2010).

191. Fan, C., Chen, H., Liu, K. & Wang, Z. Fibrinogen-like protein 2 contributes to normal murine cardiomyocyte maturation and heart development. *Exp. Physiol.* **106**, 1559–1571 (2021).

192. Francis, T. G., Jaka, O., Ellison‐Hughes, G. M., Lazarus, N. R. & Harridge, S. D. R. Human primary skeletal muscle‐derived myoblasts and fibroblasts reveal different senescent phenotypes. *JCSM Rapid Commun.* **5**, 226–238 (2022).

193. Das, S. *et al.* Transcriptomics of cardiac biopsies reveals differences in patients with or without diagnostic parameters for heart failure with preserved ejection fraction. *Sci. Rep.* **9**, 3179 (2019).

194. Zylla, M. M. *et al.* Catheter Ablation of Atrial Fibrillation in Patients with Heart Failure and Preserved Ejection Fraction. *Circ Heart Fail, in press* (2022).

195. Soliman, H. & Rossi, F. M. V. Cardiac fibroblast diversity in health and disease. *Matrix Biol.* **91–92**, 75–91 (2020).

196. Tallquist, M. D. Cardiac Fibroblast Diversity. *Annu. Rev. Physiol.* **82**, 63–78 (2020).

197. Boland, E., Quondamatteo, F. & Van Agtmael, T. The role of basement membranes in cardiac biology and disease. *Biosci. Rep.* **41**, (2021).

198. Del Monte-Nieto, G., Fischer, J. W., Gorski, D. J., Harvey, R. P. & Kovacic, J. C. Basic biology of extracellular matrix in the cardiovascular system, part 1/4: JACC focus seminar. *J. Am. Coll. Cardiol.* **75**, 2169–2188 (2020).

199. Hochman-Mendez, C., Curty, E. & Taylor, D. A. Change the laminin, change the cardiomyocyte: improve untreatable heart failure. *Int. J. Mol. Sci.* **21**, (2020).

200. Tillmanns, J. *et al.* Fibroblast activation protein alpha expression identifies activated fibroblasts after myocardial infarction. *J. Mol. Cell. Cardiol.* **87**, 194–203 (2015).

201. Aghajanian, H. *et al.* Targeting cardiac fibrosis with engineered T cells. *Nature* **573**, 430–433 (2019).

202. Aryal, B., Price, N. L., Suarez, Y. & Fernández-Hernando, C. ANGPTL4 in metabolic and cardiovascular disease. *Trends Mol. Med.* **25**, 723–734 (2019).

203. Zhu, X., Zhang, X., Cong, X., Zhu, L. & Ning, Z. ANGPTL4 Attenuates Ang II-Induced Atrial Fibrillation and Fibrosis in Mice via PPAR Pathway. *Cardiol. Res. Pract.* **2021**, 9935310 (2021).

204. Dewey, F. E. *et al.* Inactivating variants in ANGPTL4 and risk of coronary artery disease. *N. Engl. J. Med.* **374**, 1123–1133 (2016).

205. Yu, X. *et al.* Inhibition of cardiac lipoprotein utilization by transgenic overexpression of Angptl4 in the heart. *Proc Natl Acad Sci USA* **102**, 1767–1772 (2005).

206. Rockey, D. C., Bell, P. D. & Hill, J. A. Fibrosis--a common pathway to organ injury and failure. *N. Engl. J. Med.* **372**, 1138–1149 (2015).

207. Forte, E. *et al.* Adult mouse fibroblasts retain organ-specific transcriptomic identity. *eLife* **11**, (2022).

208. Ponikowski, P. *et al.* 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: The Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC)Developed with the special contribution of the Heart Failure Association (HFA) of the ESC. *Eur. Heart J.* **37**, 2129–2200 (2016).

209. Fraccarollo, D. *et al.* Additive improvement of left ventricular remodeling and neurohormonal activation by aldosterone receptor blockade with eplerenone and ACE inhibition in rats with myocardial infarction. *J. Am. Coll. Cardiol.* **42**, 1666–1673 (2003).

210. Lusis, A. J. & Weiss, J. N. Cardiovascular networks: systems-based approaches to cardiovascular disease. *Circulation* **121**, 157–170 (2010).

211. Bousquet, J. *et al.* Systems medicine and integrated care to combat chronic noncommunicable diseases. *Genome Med.* **3**, 43 (2011).

212. Joseph, J. *et al.* Genetic architecture of heart failure with preserved versus reduced ejection fraction. *Nat. Commun.* **13**, 7753 (2022).

213. Lanzer, J. D. *et al.* A Network medicine approach to study comorbidities in heart failure with preserved ejection fraction. *Res. Sq.* (2023) doi:10.21203/rs.3.rs-2429581/v1.

214. GMS | GMDS 2012: 57. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e. V. (GMDS) | Webservice zur sicheren Pseudonymisierung durch Datentreuhänder. https://www.egms.de/static/en/meetings/gmds2012/12gmds121.shtml.

215. Ergatoudes, C. *et al.* Non-cardiac comorbidities and mortality in patients with heart failure with reduced vs. preserved ejection fraction: a study using the Swedish Heart Failure Registry. *Clin. Res. Cardiol.* **108**, 1025–1033 (2019).

216. Lê, S., Josse, J. & Husson, F. FactoMineR : An *R* package for multivariate analysis. *J. Stat. Softw.* **25**, (2008).

217. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **33**, 1–22 (2010).

218. Wright, M. N. & Ziegler, A. ranger : A Fast Implementation of Random Forests for High Dimensional Data in*C++* and*R*. *J. Stat. Softw.* **77**, 1–17 (2017).

219. Klimek, P., Aichberger, S. & Thurner, S. Disentangling genetic and environmental risk factors for individual diseases from multiplex comorbidity networks. *Sci. Rep.* **6**, 39658 (2016).

220. Divo, M. J. *et al.* COPD comorbidities network. *Eur. Respir. J.* **46**, 640–650 (2015).

221. Chmiel, A., Klimek, P. & Thurner, S. Spreading of diseases through comorbidity networks across life and gender. *New J. Phys.* **16**, 115013 (2014).

222. Greene, D., Richardson, S. & Turro, E. ontologyX: a suite of R packages for working with ontological data. *Bioinformatics* **33**, 1104–1106 (2017).

223. Serrano, M. A., Boguñá, M. & Vespignani, A. Extracting the multiscale backbone of complex weighted networks. *Proc Natl Acad Sci USA* **106**, 6483–6488 (2009).

224. Koutra, D., Vogelstein, J. T. & Faloutsos, C. DELTACON: A Principled Massive-Graph Similarity Function. *arXiv* (2013) doi:10.48550/arxiv.1304.4657.

225. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).

226. Türei, D. *et al.* Integrated intra‐ and intercellular signaling knowledge for multicellular omics analysis. *Mol. Syst. Biol.* **17**, (2021).

227. Luck, K. *et al.* A reference map of the human binary protein interactome. *Nature*

580, 402–408 (2020).

228. Buphamalai, P., Kokotovic, T., Nagy, V. & Menche, J. Network analysis reveals rare disease signatures across multiple levels of biological organization. *Nat. Commun.* **12**, 6306 (2021).

229. Doll, S. *et al.* Region and cell-type resolved quantitative proteomic map of the human heart. *Nat. Commun.* **8**, 1469 (2017).

230. Piñero, J. *et al.* The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.* **48**, D845–D855 (2020).

231. Kustatscher, G. *et al.* Understudied proteins: opportunities and challenges for functional proteomics. *Nat. Methods* **19**, 774–779 (2022).

232. Valdeolivas, A. *et al.* Random walk with restart on multiplex and heterogeneous biological networks. *Bioinformatics* **35**, 497–505 (2019).

233. Czepluch, F. S., Wollnik, B. & Hasenfuß, G. Genetic determinants of heart failure: facts and numbers. *ESC Heart Fail.* **5**, 211–217 (2018).

234. Wesseling, M., de Poel, J. H. C. & de Jager, S. C. A. Growth differentiation factor 15 in adverse cardiac remodelling: from biomarker to causal player. *ESC Heart Fail.* **7**, 1488–1501 (2020).

235. Szumska, D. *et al.* Pcsk5 is required in the early cranio-cardiac mesoderm for heart development. *BMC Dev. Biol.* **17**, 6 (2017).

236. Warren, S. A. *et al.* Differential role of Nkx2-5 in activation of the atrial natriuretic factor gene in the developing versus failing heart. *Mol. Cell. Biol.* **31**, 4633–4645 (2011).

237. Arasaratnam, D. *et al.* The role of cardiac transcription factor NKX2-5 in regulating the human cardiac miRNAome. *Sci. Rep.* **9**, 15928 (2019).

238. Tayal, U., Prasad, S. & Cook, S. A. Genetics and genomics of dilated cardiomyopathy and systolic heart failure. *Genome Med.* **9**, 20 (2017).

239. Liu, X., Shi, G.-P. & Guo, J. Innate Immune Cells in Pressure Overload-Induced Cardiac Hypertrophy and Remodeling. *Front. Cell Dev. Biol.* **9**, 659666 (2021).

240. Blanton, R. M., Carrillo-Salinas, F. J. & Alcaide, P. T-cell recruitment to the heart: friendly guests or unwelcome visitors? *Am. J. Physiol. Heart Circ. Physiol.* **317**, H124–H140 (2019).

241. Croquelois, A. *et al.* Control of the adaptive response of the heart to stress via the Notch1 receptor pathway. *J. Exp. Med.* **205**, 3173–3185 (2008).

242. Fernández-Ruiz, I. Selective JAG1-NOTCH3 targeting shows potential for treating PAH. *Nat. Rev. Cardiol.* **19**, 433 (2022).

243. Zhao, Q. *et al.* Endothelium-specific CYP2J2 overexpression improves cardiac dysfunction by promoting angiogenesis via Jagged1/Notch1 signaling. *J. Mol. Cell.*

Cardiol. **123**, 118–127 (2018).

244. Yuan, T. & Krishnan, J. Non-coding RNAs in Cardiac Regeneration. *Front. Physiol.* **12**, 650566 (2021).
245. Wang, J.-K. *et al.* Ablation of Plasma Prekallikrein Decreases Low-Density Lipoprotein Cholesterol by Stabilizing Low-Density Lipoprotein Receptor and Protects Against Atherosclerosis. *Circulation* **145**, 675–687 (2022).
246. Dixit, G., Blair, J. & Ozcan, C. Plasma proteomic analysis of association between atrial fibrillation, coronary microvascular disease and heart failure. *Am. J. Cardiovasc. Dis.* **12**, 81–91 (2022).
247. Lau, E. S. *et al.* Cardiovascular biomarkers of obesity and overlap with cardiometabolic dysfunction. *J. Am. Heart Assoc.* **10**, e020215 (2021).
248. Eckenstaler, R. *et al.* A Thromboxane A2 Receptor-Driven COX-2-Dependent Feedback Loop That Affects Endothelial Homeostasis and Angiogenesis. *Arterioscler. Thromb. Vasc. Biol.* **42**, 444–461 (2022).
249. Butenas, A. L. E. *et al.* Thromboxane A2 receptors contribute to the exaggerated exercise pressor reflex in male rats with heart failure. *Physiol. Rep.* **9**, e15052 (2021).
250. Hariri, E. *et al.* Nonplatelet thromboxane generation is associated with impaired cardiovascular performance and mortality in heart failure. *Am. J. Physiol. Heart Circ. Physiol.* **323**, H248–H255 (2022).
251. Alonso, F., Dong, Y. & Génot, E. Thrombomodulin, an unexpected new player in endothelial cell invasion during angiogenesis. *Arterioscler. Thromb. Vasc. Biol.* **41**, 1672–1674 (2021).
252. Eidizadeh, A. *et al.* Biomarker profiles in heart failure with preserved vs. reduced ejection fraction: results from the DIAST-CHF study. *ESC Heart Fail.* (2022) doi:10.1002/ehf2.14167.
253. Zhang, J., Zhou, H. J., Ji, W. & Min, W. AIP1-mediated stress signaling in atherosclerosis and arteriosclerosis. *Curr. Atheroscler. Rep.* **17**, 503 (2015).
254. Huang, Q. *et al.* AIP1 suppresses atherosclerosis by limiting hyperlipidemia-induced inflammation and vascular endothelial dysfunction. *Arterioscler. Thromb. Vasc. Biol.* **33**, 795–804 (2013).
255. Sickinghe, A. A., Korporaal, S. J. A., den Ruijter, H. M. & Kessler, E. L. Estrogen contributions to microvascular dysfunction evolving to heart failure with preserved ejection fraction. *Front Endocrinol (Lausanne)* **10**, 442 (2019).
256. Yoshida, H., Matsui, T., Yamamoto, A., Okada, T. & Mori, K. XBP1 mRNA is induced by ATF6 and spliced by IRE1 in response to ER stress to produce a highly active transcription factor. *Cell* **107**, 881–891 (2001).
257. Fu, F. & Doroudgar, S. IRE1/XBP1 and endoplasmic reticulum signaling — from

basic to translational research for cardiovascular disease. *Curr. Opin. Physiol.* **28**, 100552 (2022).

258. Schiattarella, G. G. *et al.* Xbp1s-FoxO1 axis governs lipid accumulation and contractile performance in heart failure with preserved ejection fraction. *Nat. Commun.* **12**, 1684 (2021).
259. Li, L., Zhao, Q. & Kong, W. Extracellular matrix remodeling and cardiac fibrosis. *Matrix Biol.* **68–69**, 490–506 (2018).
260. Faul, C. Cardiac actions of fibroblast growth factor 23. *Bone* **100**, 69–79 (2017).
261. Rodríguez, C. & Martínez-González, J. The role of lysyl oxidase enzymes in cardiac function and remodeling. *Cells* **8**, (2019).
262. Agrawal, V. *et al.* Natriuretic peptide receptor C contributes to disproportionate right ventricular hypertrophy in a rodent model of obesity-induced heart failure with preserved ejection fraction with pulmonary hypertension. *Pulm. Circ.* **9**, 2045894019878599 (2019).
263. Gupta, M. P. Factors controlling cardiac myosin-isoform shift during hypertrophy and heart failure. *J. Mol. Cell. Cardiol.* **43**, 388–403 (2007).
264. Yu, J. *et al.* Long Noncoding RNA Ahit Protects Against Cardiac Hypertrophy Through SUZ12 (Suppressor of Zeste 12 Protein Homolog)-Mediated Downregulation of MEF2A (Myocyte Enhancer Factor 2A). *Circ. Heart Fail.* **13**, e006525 (2020).
265. Tanaka, K. *et al.* Follistatin like 1 Regulates Hypertrophy in Heart Failure with Preserved Ejection Fraction. *JACC Basic Transl. Sci.* **1**, 207–221 (2016).
266. Seki, M. *et al.* Acute and Chronic Increases of Circulating FSTL1 Normalize Energy Substrate Metabolism in Pacing-Induced Heart Failure. *Circ. Heart Fail.* **11**, e004486 (2018).
267. Qin, M. *et al.* Absence of Rgs5 prolongs cardiac repolarization and predisposes to ventricular tachyarrhythmia in mice. *J. Mol. Cell. Cardiol.* **53**, 880–890 (2012).
268. Li, H. *et al.* Regulator of G protein signaling 5 protects against cardiac hypertrophy and fibrosis during biomechanical stress of pressure overload. *Proc Natl Acad Sci USA* **107**, 13818–13823 (2010).
269. Al-Kindi, S. G. *et al.* Soluble CD14 and risk of heart failure and its subtypes in older adults. *J. Card. Fail.* **26**, 410–419 (2020).
270. De Franceschi, N. *et al.* Mutually Exclusive Roles of SHARPIN in Integrin Inactivation and NF-κB Signaling. *PLoS ONE* **10**, e0143423 (2015).
271. Lim, S. *et al.* Sharpin, a novel postsynaptic density protein that directly interacts with the shank family of proteins. *Mol. Cell. Neurosci.* **17**, 385–397 (2001).
272. Corker, A., Neff, L. S., Broughton, P., Bradshaw, A. D. & DeLeon-Pennell, K. Y.

Organized chaos: deciphering immune cell heterogeneity's role in inflammation in the heart. *Biomolecules* **12**, (2021).

273. Essandoh, K., Auchus, R. J. & Brody, M. J. Cardiac decompensation and promiscuous prenylation of small GTPases in cardiomyocytes in response to local mevalonate pathway disruption. *J. Pathol.* **256**, 249–252 (2022).

274. Xu, H. *et al.* Inhibition of the mevalonate pathway improves myocardial fibrosis. *Exp. Ther. Med.* **21**, 224 (2021).

275. Nishizawa, H., Maeda, N. & Shimomura, I. Impact of hyperuricemia on chronic kidney disease and atherosclerotic cardiovascular disease. *Hypertens. Res.* **45**, 635–640 (2022).

276. Mátyás, C. *et al.* Prevention of the development of heart failure with preserved ejection fraction by the phosphodiesterase-5A inhibitor vardenafil in rats with type 2 diabetes. *Eur. J. Heart Fail.* **19**, 326–336 (2017).

277. Cornuault, L., Rouault, P., Duplàa, C., Couffinhal, T. & Renault, M.-A. Endothelial dysfunction in heart failure with preserved ejection fraction: what are the experimental proofs? *Front. Physiol.* **13**, 906272 (2022).

278. Simeunovic, D. *et al.* Glutathione transferase P1 polymorphism might be a risk determinant in heart failure. *Dis. Markers* **2019**, 6984845 (2019).

279. Singh, M. M., Kumar, R., Tewari, S. & Agarwal, S. Association of GSTT1/GSTM1 and ApoE variants with left ventricular diastolic dysfunction in thalassaemia major patients. *Hematology* **24**, 20–25 (2019).

280. Franssen, C., Chen, S., Hamdani, N. & Paulus, W. J. From comorbidities to heart failure with preserved ejection fraction: a story of oxidative stress. *Heart* **102**, 320–330 (2016).

281. Tam, M. C., Lee, R., Cascino, T. M., Konerman, M. C. & Hummel, S. L. Current Perspectives on Systemic Hypertension in Heart Failure with Preserved Ejection Fraction. *Curr. Hypertens. Rep.* **19**, 12 (2017).

282. Hicklin, H. E., Gilbert, O. N., Ye, F., Brooks, J. E. & Upadhya, B. Hypertension as a Road to Treatment of Heart Failure with Preserved Ejection Fraction. *Curr. Hypertens. Rep.* **22**, 82 (2020).

283. Vedin, O. *et al.* Significance of ischemic heart disease in patients with heart failure and preserved, midrange, and reduced ejection fraction: A nationwide cohort study. *Circ. Heart Fail.* **10**, (2017).

284. Reding, K. W. *et al.* Lifestyle and cardiovascular risk factors associated with heart failure subtypes in postmenopausal breast cancer survivors. *JACC CardioOncol.* **4**, 53–65 (2022).

285. Saiki, H. *et al.* Risk of heart failure with preserved ejection fraction in older women

after contemporary radiotherapy for breast cancer. *Circulation* **135**, 1388–1396 (2017).

286. Huang, S. *et al.* The Association Between Inflammation, Incident Heart Failure, and Heart Failure Subtypes in Patients with Rheumatoid Arthritis. *Arthritis Care Res (Hoboken)* (2021) doi:10.1002/acr.24804.
287. Packer, M. Link Between Synovial and Myocardial Inflammation: Conceptual Framework to Explain the Pathogenesis of Heart Failure with Preserved Ejection Fraction in Patients with Systemic Rheumatic Diseases. *Card. Fail. Rev.* **6**, (2020).
288. Gevaert, A. B., Boen, J. R. A., Segers, V. F. & Van Craenenbroeck, E. M. Heart failure with preserved ejection fraction: A review of cardiac and noncardiac pathophysiology. *Front. Physiol.* **10**, 638 (2019).
289. Gao, H. *et al.* Sex- and race-specific associations of bone mineral density with incident heart failure and its subtypes in older adults. *J. Am. Geriatr. Soc.* (2022) doi:10.1111/jgs.18121.
290. Sabbatini, A. R. & Kararigas, G. Menopause-Related Estrogen Decrease and the Pathogenesis of HFpEF: JACC Review Topic of the Week. *J. Am. Coll. Cardiol.* **75**, 1074–1082 (2020).
291. Savarese, G., Stolfo, D., Sinagra, G. & Lund, L. H. Heart failure with mid-range or mildly reduced ejection fraction. *Nat. Rev. Cardiol.* **19**, 100–116 (2022).
292. Bortolotti, M., Polito, L., Battelli, M. G. & Bolognesi, A. Xanthine oxidoreductase: One enzyme for multiple physiological tasks. *Redox Biol.* **41**, 101882 (2021).
293. Watanabe, K. *et al.* Impact of plasma xanthine oxidoreductase activity in patients with heart failure with preserved ejection fraction. *ESC Heart Fail.* **7**, 1735–1743 (2020).
294. Yoon, S. *et al.* S-Nitrosylation of Histone Deacetylase 2 by Neuronal Nitric Oxide Synthase as a Mechanism of Diastolic Dysfunction. *Circulation* **143**, 1912–1925 (2021).
295. van Dam, S., Võsa, U., van der Graaf, A., Franke, L. & de Magalhães, J. P. Gene co-expression analysis for functional classification and gene-disease predictions. *Brief. Bioinformatics* **19**, 575–592 (2018).
296. Ata, S. K. *et al.* Recent advances in network-based methods for disease gene prediction. *Brief. Bioinformatics* **22**, (2021).
297. Pellikka, P. A. *et al.* Variability in Ejection Fraction Measured By Echocardiography, Gated Single-Photon Emission Computed Tomography, and Cardiac Magnetic Resonance in Patients With Coronary Artery Disease and Left Ventricular Dysfunction. *JAMA Netw. Open* **1**, e181456 (2018).
298. Pieske, B. *et al.* How to diagnose heart failure with preserved ejection fraction: the

HFA-PEFF diagnostic algorithm: a consensus recommendation from the Heart Failure Association (HFA) of the European Society of Cardiology (ESC). *Eur. Heart J.* **40**, 3297–3317 (2019).

299. Triposkiadis, F. *et al.* The continuous heart failure spectrum: moving beyond an ejection fraction classification. *Eur. Heart J.* **40**, 2155–2163 (2019).

300. Lee, E. W. J. & Viswanath, K. Big data in context: addressing the twin perils of data absenteeism and chauvinism in the context of health disparities research. *J. Med. Internet Res.* **22**, e16377 (2020).

301. Kaplan, R. M., Chambers, D. A. & Glasgow, R. E. Big data and large sample size: a cautionary note on the potential for bias. *Clin. Transl. Sci.* **7**, 342–346 (2014).

302. Wells, B. J., Chagin, K. M., Nowacki, A. S. & Kattan, M. W. Strategies for handling missing data in electronic health record derived data. *EGEMS (Wash. DC)* **1**, 1035 (2013).

303. Beaulieu-Jones, B. K. *et al.* Characterizing and managing missing structured data in electronic health records: data analysis. *JMIR Med. Inform.* **6**, e11 (2018).

304. Gianfrancesco, M. A., Tamang, S., Yazdany, J. & Schmajuk, G. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern. Med.* **178**, 1544–1547 (2018).

305. Moskowitz, A., McSparron, J., Stone, D. J. & Celi, L. A. Preparing a new generation of clinicians for the era of big data. *Harv. Med. Stud. Rev.* **2**, 24–27 (2015).