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
B.Sc, Ricardo Omar Ramirez Flores
born in: Mexico City, Mexico
Oral examination: 27.03.2023

Multi-scale molecular descriptions
of
human heart failure
using
single cell, spatial, and bulk transcriptomics

Referees: Prof. Dr. Rebecca Wade
Prof. Dr. Julio Saez-Rodriguez

*"To mom and dad: this is our **sacrifice***
*turned into a pair of **wings**"*

"To the ones who never thought
of distance in miles or years"

Abstract

Molecular descriptions of human disease have relied on transcriptomics, the genome-wide measurement of gene expression. In the last years the emergence of capture-based technologies have enabled the transcriptomic profiling of single cells both from dissociated and intact tissues, providing a spatial and cell type specific context that complements the catalog of gene expression changes reported from bulk technologies. In the context of cardiovascular disease, these technologies open the opportunity to study the inter and intra-cellular mechanisms that regulate myocardial remodeling. In this thesis I present comprehensive descriptions of the transcriptional changes in acute and chronic human heart failure using bulk, single cell, and spatial technologies. First, I describe the creation of the Reference of the Heart Failure Transcriptome, a resource built from the meta-analysis of 16 independent studies of human heart failure transcriptomics. Then, I report the first spatial and single cell atlas of human myocardial infarction, and propose a computational strategy to identify compositional, organizational, and molecular tissue differences across distinct time points and physiological zones of damaged myocardium. Finally, I outline a methodology for the multicellular analysis of single cell data that allows for a better understanding of tissue responses and cell type coordination events in cardiovascular disease and that links the knowledge of independent studies at multiple scales. Overall my work demonstrates the importance of the generation of reliable molecular references of disease across scales.

Zusammenfassung

Die Beschreibung menschlicher Krankheiten auf molekularer Ebene beruht auf der genomweiten Messung der Genexpression - dem Transkriptom. In den letzten Jahren hat das Aufkommen von *capture-based* Technologien die Erstellung von Transkriptom-Profilen einzelner Zellen sowohl aus dissoziiertem als auch aus intaktem Gewebe ermöglicht, wodurch eine räumliche und zelltypspezifische Einordnung geschaffen wurde, welche die erfassten Veränderungen der Genexpression von Bulk-Technologien zusätzlich ergänzt. Im Zusammenhang mit Herz-Kreislauf-Erkrankungen eröffnen diese Technologien die Möglichkeit, die inter- und intrazellulären Mechanismen zu untersuchen, die den Umbau des Herzmuskels steuern. In dieser Arbeit präsentiere ich umfassende Beschreibungen von Transkriptionsveränderungen bei akuter und chronischer menschlicher Herzinsuffizienz unter Verwendung von Bulk-, Einzelzell- und räumlichen Technologien. Zunächst beschreibe ich die Erstellung einer Referenz des Transkriptoms bei Herzinsuffizienz, die aus einer Meta-Analyse von 16 unabhängigen Transkriptom-Studien bei menschlicher Herzinsuffizienz resultiert. Anschließend beschreibe ich den ersten räumlichen und einzelligen Atlas des menschlichen Myokardinfarkts und schlage eine Berechnungsstrategie vor, um Unterschiede in der Zusammensetzung, der Organisation und den molekularen Geweben über verschiedene Zeitpunkte und physiologische Zonen des geschädigten Myokards hinweg zu identifizieren. Schließlich skizziere ich eine Methodik für die multizelluläre Analyse von Einzelzelldaten, die ein besseres Verständnis von Gewebereaktionen und zellulären Koordinationsereignissen bei Herz-Kreislauf-Erkrankungen ermöglicht und Erkenntnisse aus unabhängigen Studien auf verschiedenen Ebenen miteinander verknüpft. Insgesamt demonstriert meine Arbeit die Wichtigkeit, zuverlässige molekulare Anhaltspunkte für Krankheiten auf mehreren Ebenen zu erhalten.

Acknowledgements

The work presented in this thesis is the consequence of years of discussions and collaborative efforts with researchers across laboratories and institutions at different stages of my education. I want to thank my mentor, Julio Saez-Rodriguez, who has always supported me, not only by trusting my research interests since I was a bachelor student, but also by teaching me the importance of kindness to foster creativity. Julio's group throughout the last years became a source of motivation. I want to thank all my colleagues who contributed to my work or allowed me to contribute to theirs, particularly to Jan D. Lanzer, Jovan Tanevski, Attila Gabor, Christian Holland, Aurelien Dugourd, Daniel Dimitrov, Sophia Müller-Dott, and Pau Badia i Mompel.

I want to thank Informatics for Life funded by the Klaus Tschira foundation that financially supported this work. Moreover, I want to acknowledge the laboratory resources of Rafael Kramann and Christoph Kuppe that provided the experimental data from Chapter 3.

Finally, I want to acknowledge the undergraduate program on genomics sciences from the National Autonomous University of Mexico in my home country that taught me the relentlessness necessary in academia.

Table of contents

List of figures and tables

Chapter 1

Introduction

1.1 Introduction to the molecular biology of heart failure

The prevalence of heart failure in the world is around 38 million people, representing 1-2% of the adult population in developed countries [1]. Although there is a constant debate in the clinical community of what defines heart failure [2]-[3], in general terms, heart failure represents a complex clinical syndrome caused by structural and/or functional cardiac dysfunctions. These dysfunctions lead to low cardiac output, reduced blood flow and death. Known causes of heart failure are related to ischemic heart disease (e.g. coronary artery disease and myocardial infarction), hypertensive disease and/or valvular dysfunctions. Moreover, genetic, clinical, and environmental interactions contribute to the development of heart failure. Heart failure is an increased public health problem and while preventive actions have the biggest impact in reducing the affected population, poor prognosis is observed in affected individuals [1].

During the initial phases of heart failure, for example after myocardial infarction, widespread cardiac remodeling events are observed because of inflammatory and compensatory responses. Compensatory responses coordinate tissue level processes that maintain intact the heart's function despite structural dysfunctions caused by ischemic lesions, changes in blood pressure or congenital structural dysfunctions. These responses are tightly regulated, however they can become maladaptive and drive cardiac remodeling that impairs the heart's function. At the cellular level these compensatory events are mainly reflected in changes in the calcium kinetics that control cardiac muscle contraction, cytoskeleton deregulations that affect the sarcomere in cardiomyocytes, neurohormonal and adrenergic responses, reactivation of fetal programs, and fibrosis [4]–[6]. Changes in energy substrates are observed in early and late stages of heart failure from a slight increase of fatty acid oxidation to glucose use [7], demonstrating that different layers of biological information are linked during cardiac remodeling.

Some of the previously mentioned pathophysiological mechanisms are the usual target of current therapies in the clinic, such as Angiotensin-converting enzyme (ACE) inhibitors and β-adrenergic blocking agents, that antagonize the maladaptive effects of neurohormones and increased adrenergic drive [8]. Among other therapies, calcineurin and histone deacetylase inhibitors target hypertrophic growth [4] and stem-cell therapies aim to deal with the effects of apoptosis (eg. after myocardial infarction) [9]. Important circulating biomarkers for diagnosis and prognosis such as the N-terminal pro-B-type natriuretic peptide are also consequences of these mechanisms [2]. However, treatment response varies between individuals and a high number of heart failure patients need assist devices followed by heart transplantation [5].

The varied clinical profiles of the increased aging population, together with complex comorbidity and genetic interactions explain partially why some of the treatments are not universally effective [10]. An underlooked importance of heart failure patients with preserved ejection fraction (in other words, with no changes in blood pumping from the heart's left ventricle) [11], as well as a clear underrepresentation of studies of patients of underdeveloped countries or from black or hispanic communities [12], contribute to the missing pieces of the pathophysiology of heart failure. As I will discuss later in this chapter, the increased generation of knowledge requires community efforts to integrate and distribute knowledge to be able to prioritize, in the context of global health, studies that complete the profiling of heart failure's molecular biology.

To fully characterize the processes that regulate compensatory responses within the heart is necessary to understand the intra- and intercellular mechanisms that regulate cardiac remodeling. While cardiomyocytes comprise the largest volume of the heart, and thus the focus of heart failure research, they represent only one third of the cellular constituents of the organ. Other cell types such as fibroblasts, endothelial cells, pericytes, smooth muscle cells, and immune cells have major roles in heart's function and are involved in cardiac remodeling. Together, these cell-types maintain the tissue's function with homo- and heterotypic interactions via mechanical, chemical, and electrical processes both in short and long distances. Direct interactions through cell surface receptors or gap junctions combine with autocrine or paracrine signaling mechanisms where growth factors, hormones, cytokines and other ligands influence gene and protein expression ultimately coordinating cell's behaviors in the tissue (Figure CH1-1) [13]–[17].

In healthy hearts, cardiomyocytes are organized in sheets that attach to the extracellular matrix, a complex network of proteins including collagens, proteoglycans, glycoproteins, periostin, fibronectin, fibrillin, and hyaluronan [16]. The regulated contraction of cardiomyocytes is necessary for the blood pump and this is achieved by specialized cell-to-cell interactions called intercalated disks. Cardiomyocyte cross talk is achieved mainly through gap junctions, channels that allow the crossing of small molecules, including ions and small peptides. Cardiac fibroblasts are the cell-types responsible for maintaining the integrity of the extracellular matrix by regulating the production and secretion of the matrix components. This process ensures the integrity of the heart's structure by properly distributing the mechanical force in the myocardium. Fibroblasts are located throughout the matrix and have direct interactions with cardiomyocytes and cells from the vasculature network, that include endothelial cells, pericytes and smooth muscle cells [16]. Endothelial cells are located in the inner layer of blood vessels and represent roughly 60% of the non-myocyte population, outnumbering cardiomyocytes by a three to one ratio [15]. Cardiomyocytes are surrounded by a dense capillary network that ensures that their energetic demands are fulfilled, by facilitating the distribution of metabolic substrates and oxygen. Vascular smooth muscle cells form the wall of blood vessels and they regulate blood pressure and flow with contraction and relaxation mechanisms. Pericytes are located between endothelial and smooth muscle cells in larger vessels and regulate the homeostasis and permeability of the vasculature [15], [18]. Cross-talk between cells in the vasculature is necessary for angiogenesis and the deposition of the basal membrane, an extracellular matrix support of endothelial cells. Finally, immune cells including myeloid and lymphoid cells regulate inflammatory and reparative responses by communicating and interacting with vascular and muscle cells, and activating fibroblasts [19]. Despite the apparent specialized and compartmentalized functions of individual cell types in the heart, a multicellular organization is required for the organ’s function and homeostasis. For the same reason, compensatory and maladaptive responses should be understood in that specific context.

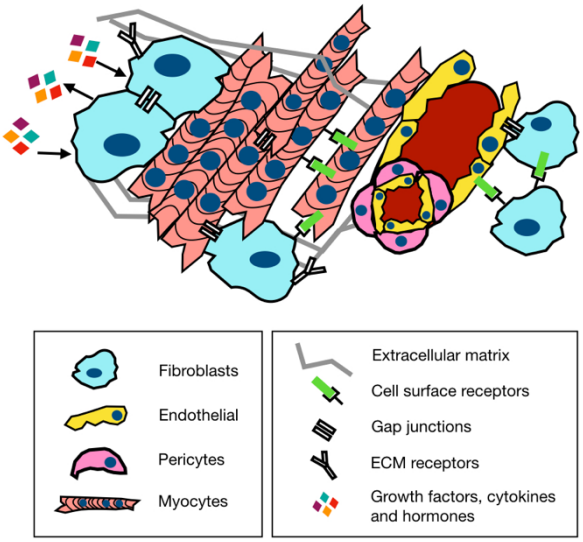

Figure CH1-1. Multicellular organization of the heart's tissue.
Homo- and heterotypic intercellular interactions regulate heart's function through cell surface receptors or gap junctions combined with autocrine or paracrine signaling mechanisms. Figure is adapted from [16]

During cardiac remodeling, cardiomyocyte autocrine signaling regulates hypertrophy, and paracrine signaling via cytokines and growth factors aids immune response and the reorganization of the vascular network [17]. Moreover, upon myocardial damage, fibroblasts transition to an activated state called myofibroblasts that increase the production of extracellular matrix and regulate the repair of cardiac tissue after the cardiomyocytes' death [20]-[21]. Nevertheless, impaired regulation of these multicellular processes lead to maladaptive responses that lead to heart failure. For example, uncontrolled responses of fibroblasts lead to fibrosis, impairing the electrical coupling of cardiomyocytes and generating mechanical stiffness that affect cardiac output [20], [21]. Pro-inflammatory and pro-fibrotic responses mediated by immune cells have an important role in maladaptive responses too [19], [22], [23]. Additionally, defective angiogenesis or deregulated vasodilation ultimately affects the cardiac metabolic homeostasis, since not enough resources can be provided to cardiomyocytes [24].

The main objective of this brief overview is to insist that a multicellular perspective is required to understand the molecular processes of heart failure. With this perspective, the pathophysiology of heart failure is described in terms of generic tissue responses and cell type specific responses, both coordinated by complex communication events between the same or different cell types. For the same reason, multiscale molecular profiling of bulk tissues and

single cells together with their spatial coordinates, becomes essential to understand the multicellular molecular mechanisms that ensure the function of organs. In the next section I will present some of the technological developments that allow for a multicellular understanding of tissues.

1.2 Multiscale profiling of disease

For the last 20 years, genome-scale data has been generated for different human diseases. Full transcriptomes, epigenetic marks, metabolic fluxes, signaling signatures, protein repertoires, chromosomal arrangements, among other molecular readouts of tissues, provide a catalog of individual processes that occur simultaneously in a time or spatial snapshot of a tissue [25]–[27]. Historically, technological developments were focused in generating -omics measurements in whole tissues (bulk profiling), however recent advances have been able to profile individual cells in dissociated or intact tissues [28]–[30]. Biomedical research has leveraged these technologies in combination with case-control experimental designs to describe differential processes from distinct biological layers. To link these apparently disjointed lists of processes, the field of systems biology has proposed distinct solutions that integrate them using mechanistic assumptions with prior knowledge [31]–[33], statistical multitable methods [34] or a combination of them [35].

Naturally, with the emergence of multiscale profiles, the integrative challenges of -omics data are not only limited to the connection of average signals of a single measurement across biological layers (eg. epigenomics, transcriptomics, and proteomics), but also focused at decomposing those signals as a complex interplay of single cells in space, where variable cell intrinsic processes are coordinated to maintain tissue's identity and function. Throughout this thesis I will provide an intuitive notion of how I approached this challenge in the context of cell-type specific gene expression deregulation during heart failure. A description of multiscale transcriptomics follows.

1.2.1 Bulk transcriptomics

Ribonucleic acid (RNA) transcription is one of the most profiled molecular readouts to understand genomic function. In case-control experimental designs, changes in the transcription of genes are usually understood as a consequence of the activation of signaling pathways and the effects of transcription factors. However, the transcriptional state of a tissue

does not completely correspond to its functional state given the low correlation between transcript and protein abundance [36] and other epigenetic [37]–[39], post-transcriptional [40], and post-translational mechanisms [41]. Nevertheless, when the transcriptional pathological state of a tissue is contrasted with its healthy counterpart, a potential set of differential cellular processes associated with a disease of interest still can be listed.

The initial steps of the experimental protocol for bulk transcriptomics requires the extraction of RNA from a tissue specimen, the enrichment of messenger RNA and synthesis of complementary DNA (cDNA) [42]. In DNA chips or microarrays [43], the transcriptome is then quantified as the level of hybridization that the cDNA of a sample has with fixed probes in the chip. The abundance of a transcript is quantified as light intensity with a dynamic range constrained by hybridization saturation levels. Even though microarrays required a reference transcriptome in order to be able to probe gene expression, their high throughput and low cost made them essential in the description of gene expression deregulation during disease [44]. With reduced costs and increased accessibility of sequencing technologies, microarrays were replaced with direct sequencing of short or long reads of the enriched library of cDNA from the sample (RNA-seq) [45]. This change in technologies allowed for the possibilities of *de novo* transcriptome assembly and isoform quantification. Technical details of microarrays and RNA-seq are out of the scope of this work and I refer to comprehensive descriptions by other authors [42], [44].

Typical computational workflows that follow transcriptome quantification in biomedicine are centered in the identification of differentially expressed genes between two or more distinct clinical groups. Sample normalization and batch effect correction usually precede the modeling of transcript abundance with variations of generalized linear models or non-parametric population comparison methods [46]–[48]. Gene level statistics from these models are used then to rank and identify potential biomarkers.

This analysis strategy is used routinely in almost every transcriptomics study, however a conceptual flaw in this approach is to treat each gene independently. Even in the most simplistic scenario of reducing gene expression regulation to the interplay of transcription factors, these factors cooperatively regulate the expression of more than a single gene. Thus, a better approach is to quantify regulatory events.

Enrichment analysis are statistical methods that group gene level statistics based on prior knowledge to estimate the likelihood of observing a coordinated gene expression regulatory event. Their main objective is to transform a disjointed list of differentially expressed genes into a collection of potential cellular processes that are being activated or repressed in the pathological state of the tissue [49]. Most of these methods are based on hypergeometric tests, rank statistics or empirical estimations of likelihoods based on permutations [50]–[52]. Regardless of the statistical method used, their hypothesis tests can be divided in two main classes: competitive or self-contained. Competitive hypothesis testing is used to compare if the expression coordination of a single gene set is greater or lower than the one observed from random collections of gene sets (usually of the same size). Self-contained hypotheses test if the observed coordination of a single gene set is associated with the variability of samples. To compute empirical p-values for self-contained hypothesis testing, a null distribution of the scores is created by repeatedly estimating the enrichment of gene sets from random population comparisons generated by shuffling sample group labels (eg. healthy and disease). Competitive hypotheses are the most used in enrichment analysis because they depend solely on a single vector of gene-level statistics compared to self-contained methods which are limited by the number of observations [51]. Since enrichment methods rely on prior knowledge coming from distinct sources with varied levels of quality, hypothesis weighting can be used to prioritize high confidence gene sets without losing the statistical power [53].

Enrichment methods can also be applied directly to gene expression data with the main objective to reduce highly dimensional expression data into a collection of interpretable latent variables. This dimensionality reduction increases the power of population comparisons, while keeping biological significance. Compared to other data-driven dimensionality reduction methods like principal component analysis or factor analysis, the estimated latent variables estimated from enrichment methods are not necessarily independent.

Usual gene sets used in enrichment analyses come from databases like KEGG or Reactome that curate gene sets associated with signaling pathways or cellular processes. However, these collections mainly group genes that are functionally relevant whenever they are translated to proteins as members of a pathway. As mentioned in the beginning of this section, the agreement between the functional state and the transcriptional state of a tissue is not true, and although these gene sets are useful to interpret transcriptomics data, this is mainly because genes belonging to a given cellular process are regulated collectively [54]. Under the assumption that

the quantified gene expression in a sample is a consequence of an usually unobserved collection of cellular high-order processes, gene set enrichment analysis can be used to trace these processes back. Gene sets used in these analyses are called functional footprints, because they capture direct transcriptomic responses upon "activities" of cellular processes [55]. For example, the functional footprint of a signaling pathway *X* is the collection of genes that consistently change upon perturbation of the pathway, rather than the members of the pathway itself [56]. Another example is the interpretation of transcription factor activities, where the transcriptomic levels of genes belonging to the regulon of a given transcription factor *Y* provide greater evidence of its functional relevance, compared to the expression of the transcription factor alone [57]. This change in perspective allows us to interpret transcriptomics data functionally and build the bases of more complex mechanistic models. Several efforts to collect functional footprints of pathway and cytokine activities [56], [58], transcription factors [59] and perturbations [60] are currently available.

Bulk transcriptomics has been essential in the description of the general heart tissue processes and their changes during disease, particularly because bulk technologies are of low cost and allow for the profiling of large patient cohorts possible. Nevertheless the transcriptomic profiles obtained from these studies are convoluted signals reflecting cell type composition and general and cell type specific molecular responses to the environment. Naturally, technological advancements evolved to profile single cells from tissues as I will describe in the next section.

1.2.2 Single cell transcriptomics

For the last 10 years, the emergence of plate-based and droplet-based methods to sequence the transcriptome of single cells has allowed the community to massively survey the human cell types in different organs [61]. Based on the idea that the transcriptional profile captures the lineage history of cells after development together with their functional variability, single cell transcriptomic sequencing complements histological and pathological tissue descriptions and expands the work of massive endeavors like the Human Protein Atlas [62]. Compared to bulk technologies, however, single cell profiling costs are considerably higher which limits the amount of independent samples in a single study. Additionally, transcriptome coverage is limited in most of the current technologies [63]. All of this together creates a trade-off that is characteristic of this type of data: increased number of measurements (cells), with increased

sparsity (low number of genes), in a nested structure of variability (patient, location, lineage, and function).

The experimental protocol of plate-based and droplet-based methods share the same objective of isolating cells and creating individual libraries for sequencing. However, what differentiates droplet methods is that they rely on microfluidics technologies that allow to isolate thousands of cells in individual barcoded droplets [63]. In my work I focused on data generated by the 10x Genomics Chromium droplet-based platform. For technical details I refer to the work by [64]. A limitation of droplet-based isolation is the size of the cells that can be captured and the amount of starting material needed. Thus as an alternative, single nuclei can be isolated and sequenced instead of single cells, with similar detection levels but with an overestimation of cell compositions of multinucleated cells and loss of sub localized transcripts [65], [66].

The computational challenges of single cell transcriptomics are different from the ones described for bulk transcriptomics because of the nested structure of the data. In the following sections I will describe the most current workflow for analyzing a single sample, multiple samples and case-control complex experimental designs. These descriptions follow read alignment to a reference genome or transcriptome and gene expression quantification. This task is more complicated than in bulk experiments because it involves cell calling, amplification error correction and read to gene and read to cell assignment [67].

Single sample analysis

Count matrices obtained after gene quantification still carry technical noise mainly from three sources: ambient noise, non-viable cells and doublets. The diluted suspension of cells required by droplet-based technologies adds cell-free mRNA to every droplet coming from ruptured or degraded cells, adding an off-target gene expression in cell types. This ambient noise can be corrected with computational methods that rely on the transcriptomic profiles of empty droplets at the expense of modifying the count matrix before any other downstream computational analysis [68], [69]. Additionally, non-viable cells are filtered out based on empirical thresholds of gene coverage, counts, mitochondrial gene expression (since they serve as markers of cell-death), and dissociation stress [70], [71]. Although the dilution of cell suspension aims at lowering the rate in which more than a single cell or nuclei is captured in a single droplet, a proportion of quantified droplets contain doublets, with abnormally high number of reads. Similarly as ambient correction, computational methodologies exist to identify potential

doublets and filter out those measurements [72]-[73]. These initial quality control metrics are tissue and context dependent and usually are confounded by biological signals. For the same reason, multiple iterations of these steps are performed to the data based on the "quality" of downstream analysis. Sequencing depth correction and count transformations to handle data heteroskedasticity are later performed to ensure proper variance stabilization and gene expression comparison [74].

Identification of highly variable genes followed by dimensionality reduction with principal component analysis precedes community detection and (graph-based) clustering analysis to identify groups of cells coming from similar lineages or functions [70]. Marker genes are identified with differential expression analysis with parametric or non-parametric tests. Corroboration with literature and prior knowledge allows to classify clusters of cells as "cell types" or "cell states", a coined term to refer to cells coming from an identical lineage, albeit with different functions. Clustering analysis is at the core of the definition of cell types and cell states in single cell transcriptomics, for the same reason, computational methods that assess the "quality" of a given clustering granularity have emerged recently [75]. The definition of cell states is an open debate that resembles the one in the field of evolutionary microbiology regarding the definition of species. The loose definition of cell states based solely on gene expression suffers from the previously discussed gap between the transcriptional and functional state of cells and the amplified technical effects of single cell sequencing. In general, the quality control steps mentioned in the first part of this section are adjusted based on the "biological" relevance of the definition of cell types and cell states from the filtered and adjusted atlas.

Functional transcriptomics tools to infer activities of transcription factors and signaling pathways (described in the "Bulk transcriptomics section") can also be applied robustly to single cell transcriptomics data [76]. Inference of gene regulatory networks is also possible when prior knowledge of regulons is not available [77]. Additionally, some computational tools exist to explore the gene expression dynamics of groups of cells that follow developmental or functional trajectories [78]-[79]. Methods to infer cell communication events between different populations of cells are also available. These methods use the co-expression of ligand-receptor pairs between two cell types or states as a proxy of likelihood of communication [80]-[81].

In summary, the main objective of the computational analysis of single cell transcriptomics data of a single sample consists in calling viable cells, classifying them based on cellular

ontologies and characterizing their functional variability. It is important to notice that the variability explored at this level only reflects a single sample and it does not necessarily reflect the variability observed in a complete population. Biomedical studies leveraging single cell technologies are increasing the number of profiled samples simultaneously which bring additional computational challenges.

Multiple sample analysis

The nested structure of single cell studies with multiple samples coming from a single condition requires the use of computational methods that match the observed variability between and within cell types. The main assumption is that sample specific effects in transcriptomics profiles (either technical or biological) can be corrected because there is a core covariance structure of lineage and functional gene expression markers. In other words, when profiling single cell expression profiles of a sample cohort, the main objective is to identify the cell types and their functional diversity that are consistent across samples [82]-[83].

The objective of integration and batch-correction methods is to project single cell data from multiple sources in a unified space where it is possible to estimate distances or create community graphs for clustering analysis. Some methods identify collections of mutual nearest neighbors between samples, either using the whole expression profile or reduced features, to reconstruct the shared space. Other methods regress out the effects of batches in a reduced space (eg. principal components) while maximizing the diversity of them in a given location of the shared space [84]. Additionally, deep learning methods leverage known or suggested cell type labels to facilitate the alignment and transfer of information between data sets [85]. Regardless of the core methodology used in integration, these methods provide a corrected shared space for all the cells coming from all samples or a mutual nearest neighbor graph, and/or an adjusted expression matrix [83]. Ultimately, this correction allows to define globally the different cell types and states in a patient cohort.

An important aspect of this integration task is the selection of shared highly variable genes. In low quality samples it is likely that the highly variable genes are associated with technical rather than biological variation. When this variability is handled improperly, wrong assignments of cell states or cell types are expected. Although integration methods have matured in the last years, proper balance of technical and biological correction is still a challenge. Proper integration is the basis of the analysis of single cell data in case-control or

other more complex experimental designs, since integration allows to define the processes that can be compared between patient conditions.

Case-control or other complex experimental designs

So far I have described the general quality control analysis that is usually performed to single cell transcriptomics samples and the workflow to identify cell types and their molecular variability in one or multiple samples. The challenge in complex experimental designs where multiple disease time points or conditions are profiled is to extract from all the observed variability across multiple cell-types, which processes are associated with clinical features. Even though it is possible to contrast every feature of each defined cell type between conditions, I advocate for a tissue-centric perspective [86] in which single cell data can be used to estimate in tissues: 1) cell type compositional changes and 2) transcriptional shifts within cell types.

The pathological state of a tissue can be described by changes in the compositions of cell types usually observed in its healthy counterpart. For example, upon ischemic injury in any tissue, it is expected an increased number of immune and stromal cells. With single cell data it is possible to contrast relative compositions between patient groups. However, two technical aspects are important to consider in this task. The first aspect is that measured cell type compositions of each sample are affected by the microfluidics technologies and tissue sampling. Proper experimental design allows to reduce the influence of these technical effects in population comparisons. The second aspect is that cell compositions are not independent (given the multicellular nature of tissues and the mathematical properties of proportions) and data transformations are needed before applying any type of statistical analysis. In compositional data analysis, to deal with dependencies between features, relative compositions are usually transformed to log ratios. Univariate or multivariate statistical methods are later applied to compare populations [86]. Alternative methods with a Bayesian perspective have been reported that ensure applicability in underpowered studies [87].

Complementary to compositional changes, cross-condition general transcriptional shifts can be observed across all cell types (eg. upon hypoxia or inflammation) together with cell type specific molecular responses (eg. fibroblast activation upon ischemic injury in tissues). To infer disease molecular signatures observed within each cell type, differential expression analysis between patient conditions is commonly used. However compared to bulk transcriptomics,

statistical tests are not applied directly to single cell observations. As I have continuously presented in this section, in complex experimental designs there is a nested structure of the data, in which it is assumed that the clinical condition is the factor with the major influence, followed by the individual (patient), the tissue sample, cell types and cell states. Since there is a bias in cell capture in independent samples, by simply comparing the profiles of single cells, there is a risk of capturing trends of single samples that contribute a higher number of cells in the whole dataset. Thus, to properly compare the distributions of gene expression between different conditions, the distribution of each condition should capture the variability across all individuals belonging to that class, rather than single cells. Although mixed-effects models are suitable for this type of experimental design, these models do not computationally scale well, and studies have reported that simpler strategies are more effective. One of these strategies is called *pseudobulking* in which the gene counts of all cells belonging to a given cell type or cell state of an individual are combined into a single profile. This strategy creates as many individual profiles as cell types or states in the data set. Once this process is done for all individuals, classic bulk transcriptomics models for differential expression analysis can be used. I refer to the work of [88] and [89], for a more comprehensive description of cross-condition analyses.

Alternative statistical methods to calculate transcriptional shifts specific to cell types across conditions, test the difference between the observed divergences of samples within a condition, compared to the ones observed between conditions. These methods assume that cell types with the greatest between-condition versus within-condition divergence ratio are the ones with the greatest functional change in a pathological state. This strategy allows to prioritize the characterization of cell states in a reduced number of cell types [86].

The emergence of cell states in a pathological condition is also a common cross-condition analysis. The assumption is that an observed functional profile of a cell type is mostly observed in a pathological process. For example, activated fibroblasts that have an increased production of extracellular matrix are expected to be in higher abundance in “tissue contexts” where this function is required. Even though this functional cell state may be traceable in healthy tissues, its relevance increases upon ischemic injury or fibrosis. In this situation, cross-condition compositional and transcriptional comparisons may converge. When contrasting the pseudobulked gene expression of a cell type of healthy and pathological contexts, it may be the case that the estimated differentially expressed genes associate with the cell state with the

greatest change in abundance between conditions. Nevertheless, this relationship is tissue dependent and is part of the discussion of cell state definitions.

A limitation of the analyses described above is that they treat each cell type as an independent entity of the tissue, disregarding the expected multicellular organization of molecular changes and tissue remodeling. Multi-view and multi-task learning approaches have been proposed recently to estimate multicellular programs, a "higher-order functional unit" that is composed of the combination of multiple cell type expression changes [90], [91]. These methods learn a latent space that simultaneously captures the variability structure of multiple cell types. Since these methods can be fully unsupervised, they also provide ways to evaluate different patient classifications. Additionally, cell type loadings of each latent space can be used to profile multicellular functions associated with specific disease processes. Similar ideas are applied to communication score matrices that capture ligand-receptor coexpression patterns between different pairs of cell types across samples [92]. The objective of the latter method is to associate the variability of gene expression across samples to cell communication events.

Despite the increased molecular resolution that single cell transcriptomics provide to the understanding of disease processes, one of its limitations is that the spatial context of each single cell is lost after tissue dissociation. Profiling the location of individual cells together with their gene expression can help to relate the observed variability of single cells to tissue structures. Additionally, coordinated multicellular responses can be explored either in terms of cell communications or general cell type responses.

1.2.3 Spatial transcriptomics

Initial efforts to profile the spatial organization of gene expression can be traced back to the 1980's with applications of single molecule fluorescence *in situ* hybridization on mRNAs. However, during the last 6 years, a series of technologies that profile highly multiplexed (almost genome-wide) spatially resolved gene expression data have emerged. All of these technologies balance three aspects: 1) the resolution of the measurements (from tissue regions to intra-cellular locations), 2) the number of measured features (from dozens to thousands) and 3) the effective profiled area.

Lundeberg, et al. classify spatial transcriptomics technologies in five different classes: i) technologies based on microdissected gene expression, ii) *in situ* hybridization technologies, iii) *in situ* sequencing technologies, iv) *in situ* capturing technologies, and v) *in silico* reconstruction of spatial data [93]. A complete revision of these technologies is out of the scope of this introduction and I refer to the works of [30], [93], [94] for detailed comparisons of the different technologies.

In this thesis I will focus on the description and analysis of one of the most recent *in situ* capturing technologies that was commercialized by 10x Genomics: the Visium platform. The experimental principle of Visium is to perform cDNA synthesis *in situ* followed by library preparation and sequencing *ex situ*. First, tissues are placed on slides with printed barcoded locations specifying the spatial coordinates. Then, tissues are fixed, stained, and imaged in these glass slides. A tissue permeabilization process allows mRNA to diffuse to the barcoded "spots" where reverse transcription happens before library preparation and sequencing. Post processing requires barcoded reads to be aligned to a reference transcriptome and assigned to spots, similarly as in single cell sequencing. Additionally, barcoded reads have to be aligned to the tissue images [95]. Each barcoded spot is 55 μm in diameter, with a distance of 100 μm between the center of two spots. The current capture area is 6.5 x 6.5 mm, with a total of 4992 total spots per capture area. It is reported that each spot can profile approximately 10 cells, however this is tissue dependent. The computational analysis of highly multiplexed spatial transcriptomics data is at its infancy, however most of the available tools and frameworks are motivated by two main objectives: 1) describe the tissue organization at the cellular and functional level and 2) associate those tissue descriptions to clinical covariates.

Describing the tissue architecture with spatial transcriptomics

Quality control of an individual slide requires calling viable spots with empirical filters based on the number of recovered genes and total gene counts. These filters can be adjusted based on the count metrics of spots that are not covered by tissues or complementary single cell datasets. Spot swapping, a technical artifact where mRNAs from neighboring spots contaminate the observed signal of each location, can additionally be corrected [96]. Compared to single cell data, in spatial transcriptomics, the relationship of mitochondrial gene expression with low quality locations is not clear, since functional areas with high metabolic rate (eg. muscle) can express mitochondrial genes. Spot gene expression is transformed to stabilize the variance and allow for within sample gene expression comparisons, similarly as in single cell

transcriptomics. Downstream computational analyses are focused on deconvoluting cell type compositional and functional signals of each individual location or leverage spatial information to identify organization patterns.

In Visium, each spot represents a mini bulk sample of multiple cells that belong to the same or different lineage. An important task in Visium data analysis is to determine the cellular composition of each spot to better describe the structural features of the tissue. Cell type deconvolution methods infer cell compositions of each location using a reference single cell transcriptomics data set. The most popular deconvolution methods available rely on negative binomial models [97], [98], matrix factorization [99], variational inference [100] and deep learning [101]. Moreover, functional transcriptomics approaches can be used to deconvolute cellular processes, such as pathway or transcription factor activities as presented in the bulk transcriptomics section. In combination, these methods allow linking structural and functional features in a single location.

Methods that leverage the spatial coordinates of the data have as main objectives the identification of spatial patterns and the inference of spatial relationships in different contexts. Spatially variable features are mainly identified with Moran's I, Laplacian scores, and gaussian process regression [94], [102]. These methods associate the expression of features to locations in single slides. Alternatively, community detection algorithms recover the most recurrent neighbors of a given categorical feature in space, usually a cell type label [103]. Expansions of this idea for continuous features focus on the definition of multi view predictive models where each view corresponds to the "estimated influence of predictors" in a given spatial, functional or cellular context [104], [105], [106]. The most general framework to estimate spatial relationships combines the predictions of different contextualized views in a late fusion explainable step. This allows to simultaneously estimate the contribution of each view and each predictor in the explanation of the variability of a given marker in space. The estimated interactions in each view then can be used as descriptive features of the spatial organization of multiple markers, expressions or activities within a single slide. Methods that specifically model ligand-receptor interactions are also available [107]-[108], however these represent a subtask of identifying spatial interactions of markers. Overall, these computational methods generate comprehensive descriptions of individual tissue samples and represent a structural and functional roadmap of the processes happening within a region of interest.

Cross-condition comparative analysis

Similarly as in single cell transcriptomics, cross-condition analysis in spatial transcriptomics depends on the definition of anchoring features that allow for a direct comparison of multiple samples that in principle can not be aligned. A collection of tissue samples coming from distinct individuals and/or conditions contains by design different "tissue architectures" even if these samples are collected from the same region of an organ of interest. The elements of the tissue architecture can be divided into compositional features and organizational features.

Compositional features of a tissue capture the relationships in abundance of cell types, tissue structures or cell functions. For example, when comparing ischemic tissue samples with healthy samples of any type of tissue, ischemic specimens will contain a "visible" increase in immune cells, cell death and scars (relative to intact tissue), compared to healthy tissues. Cell type compositions can be estimated from deconvolution methods and their comparisons resemble the cross-condition compositional analysis in single cell transcriptomics. Spot integration and clustering analysis (repurposed from single cell data analysis) can be used to define tissue structures without the need of dissection. These clusters of spots or "niches" represent the basic shared building blocks of the analyzed tissues and are characterized by similar functions or cell communities. Compositions of these niches can be compared following the same principles of compositional data analysis afterwards. Finally, effective areas of cellular functions within a tissue can be estimated using functional transcriptomic tools with self-contained hypothesis testing, similarly as what can be defined with bulk transcriptomics, followed by differential composition analysis. The limitation of compositional features is that they disregard the spatial arrangement of the elements which may encode biological relevant information.

Organizational features capture the spatial dependencies that exist between different cell types and/or cell functions in the whole tissue or in regions of interest, under the assumption that specific arrangements of cell types and functions define the hallmarks of the tissue. These features are estimated for each slide using community detection methods or multi-view spatial models, as described previously. When contrasting organizational features, the main objective is to identify in which condition is more likely to observe a spatial dependency between two or more features in the tissue. These differential interactions may occur because of remodeling events in large areas of the tissues, being tightly linked to compositional features. Additionally,

in situations where the overall cell-type composition of the tissue did not change but the arrangement of the cells did, differential interactions can also be traced with these methods.

1.2.4 Summary

Technological advances in molecular biology currently allow for a multiscale profiling of the transcriptome of tissues in healthy and pathological contexts. Bulk technologies provide descriptions of general transcriptional processes that relate to the emergence of cell-type-specific functional programs and compositional changes in tissues. Single cell technologies allow to deconvolute those general signals but reducing the size of patient cohorts. Finally, spatial technologies contextualize cell type expression variability in tissue structures and represent the anchoring data for bulk and single cell technologies.

1.3 On the necessity of molecular references

The promise of the technologies presented in the last section is the generation of immense volumes of data that allows to characterize molecular processes in tissues and their deregulations during disease. This promise carries the hopes of better patient stratification and personalized medicine. An important element for the success of multiscale molecular profiling of disease is the careful design of patient cohorts since it helps to correctly calculate the uncertainty of a given observation (eg. differential expression of a gene X in a cross-condition comparison) [109]. In the context of heart failure, variability in the disease presentation and progression, and the response to therapies can be explained by the complex interactions of comorbidities, clinical profiles, sex, and ethnic origins.

Unexpectedly, however, clinical and molecular characterization of heart failure has been focused in male patient populations of European descent, even though women and diverse ethnic groups compose a representative proportion of the people affected by heart failure. As a consequence, estimates on the prevalence and mortality of heart failure in “developing nations" are limited or unreliable. Additionally, with limited data from “heterogeneous” populations, it is unfeasible to assess the generalizability of treatments even if they are broadly used in the clinic [12], [21], [110].

Although this uneven progress in clinical and molecular profiling in underrepresented communities can be understood as another consequence of the social structure and the

disparities in health and technological services across the world, it is in the hands of current researchers (in privileged institutions) to change this. As properly stated by Bentley*, et. Al.*, diversity and inclusion in clinical and -omics research is important for two reasons: 1) To avoid genomic medicine to only benefit a privileged few and 2) to contribute in the understanding of the different genetic, social, and clinical factors that significantly impact the understanding of biology of diseases [111].

Among other actions, the creation of disease molecular references based on meta-analysis or systematic reviews, in my opinion, is a valuable step towards building an incremental knowledge bank that allows the fair and responsible use of high throughput technologies in diverse patient cohorts. I define a disease molecular reference as any effort that compiles the knowledge of independent omics studies of a disease of interest in a data-driven manner and that is accessible to the community. The main objective of molecular references is to unify the analysis of independent research groups, report the consistencies and differences in findings and provide a common ground to evaluate the diversity, both clinical and ethnic, of the profiled patients. Ideally, the objective of this strategy is not to monopolize knowledge, but rather to democratize it in a way that the community can easily access the data and results of complementary "apparently" similar studies and evaluate the necessity of generating a new omics dataset.

I consider it necessary to prioritize the collective effort of knowledge integration, if we want to fulfill the promises of personalized medicine, even at the expense of the so demanded novelty in academia. My hopes are that highly-qualified laboratories with access to advanced and costly technologies can take responsible actions when defining which patient populations to profile based on the available collective knowledge. This is considerably important in diseases like heart failure, where access to patient tissue specimens is limited given the impracticality of getting biopsies from healthy persons or heart failure patients.

In this PhD thesis I sought to present the most comprehensive molecular reference of heart failure bulk transcriptomics up to date that I compiled together with Dr. Jan D. Lanzer. Moreover, I describe the first detailed spatial and single cell transcriptomics reference of the cardiac remodeling events following myocardial infarction, one of the leading causes of heart failure. Altogether my research represents an initial multiscale description of heart failure transcriptomics that is easily accessible and expandable.

In Chapter 2, I describe the creation of ReHeaT (Reference of the Heart failure Transcriptome), a compilation of the last 15 years of human heart failure bulk transcriptomics. I discuss a simple but effective meta-analysis methodology based on coordinated gene expression events that demonstrate that despite clinical and technical diversity in independent patient cohorts, a consensus transcriptional response across all cohorts can be recovered.

In Chapter 3, I present the first spatial and single cell multi-omics map of human myocardial infarction. I discuss the technical aspects of the creation of this multiscale reference and propose a framework to perform cross-condition comparisons of the molecular, compositional, and organizational features of tissues. Furthermore, I describe the relationships between cell type heterogeneity and tissue organization during myocardium remodeling.

In Chapter 4, I propose a tissue-centric computational framework to analyze single cell transcriptomics data using multi view factor analysis. In this framework I compute multicellular gene expression programs that can be used in the unsupervised analysis of single cell data of patient cohorts. I apply this framework to multiple public single cell atlases of acute and chronic heart failure, including the spatial and single cell atlas of human myocardial infarction described in Chapter 3. I show how this approach allows us to infer the multicellular coordination of transcriptional responses in disease contexts and facilitates the integration of multiple studies across scales, including spatial and bulk transcriptomics.

Finally, I summarize this thesis and discuss future perspectives in the section “Concluding remarks”.

Chapter 2

Consensus transcriptional landscape of human end-stage heart failure

In this Chapter I describe the generation of a consensus bulk transcriptomics disease signature of heart failure associated with dilated and ischemic cardiomyopathies. In this work we compiled the information of 16 public studies that were available up to 2020. First, I justify the creation of this resource in the context of biomarker disparities across studies. Then I discuss a methodology that measures between studies the conservation of regulatory events that encompass more than a single gene. Finally, I present a meta-analysis performed with all these independent datasets and showcase the usefulness of its results to generate insights on the reactivation of fetal programs during heart failure and the cardiac origin of plasma biomarkers.

This Chapter is the product of the collaborative work with Jan D. Lanzer in the group of Julio Saez-Rodriguez. An edited version of this chapter has been peer-reviewed and published in [112], after being rejected by the editors of both the European Heart Journal and Circulation with the justification of not reaching priority despite positive peer reviews. Jan D. Lanzer guided the data curation, clinical interpretation of the patient cohorts, and biological relevance of the findings. I conceived and implemented most of the data analysis discussed in this chapter. Christian H. Holland implemented the web app that gives access to the results. Whenever I refer to collaborative efforts with Jan D. Lanzer, I will use the “we” pronoun. Julio Saez-Rodriguez and Rebecca Levinson supervised the project. The article was jointly written by Jan D. Lanzer and me, with style input from supervisors.

2.1 Human heart failure transcriptomics

The adverse cardiac remodeling events associated with heart failure are tightly coordinated by inter and intra-cellular signaling mechanisms across the organ. As discussed in the first chapter of this thesis, the consequences of these signaling mechanisms can be observed in the transcriptional state of a tissue. Thus, for the last 22 years the cardiovascular community has used high-throughput transcriptomics technologies to measure the changes in gene expression of the heart at the end-stage of heart failure, particularly of the left ventricle, whose objective is to pump blood rich in oxygen to the body. The first high-throughput transcriptomic study on failing human myocardial tissue, published in 2000, compared two non-failing and two failing hearts with microarrays and reported the expression of ~7,000 genes [113]. In subsequent years, additional studies exploring the transcriptional landscape of the failing heart followed, with RNA sequencing studies emerging in the mid-2010s increasing coverage to >20,000 genes. Large scale transcriptomic studies have helped elucidate the complexity of gene regulation in heart failure, notably in processes influencing cardiac hypertrophy [114], reverse remodeling [115] and cardiac metabolism [7].

However, most transcriptomic studies lack functional interpretations, discussing only a few differentially expressed genes as potential biomarkers, and disregarding subtle changes in patterns of variation and coexpression of multiple genes. Low sample sizes, selection bias and non-probabilistic sampling add uncertainty to measured transcriptional changes in single studies which difficult the knowledge transfer to larger patient populations. Additionally, lack of standards in transcriptomics studies in the community to report the clinical characteristics of patients, tissue protocols and data analysis add additionally confounding factors. Thus, an integrated analysis of multiple studies can allow us to quantitatively assess the robustness of the gene expression changes reported by individual studies and to prioritize molecular hallmarks less influenced by confounding factors. Several reports have attempted to compare heart failure gene expression studies [116]–[119] but, to my knowledge, a comprehensive analysis with employment of functional tools including transcription factor and pathway activity analysis has been lacking. These previous studies are limited to the meta-analysis of a small collection of studies of a single technology and do not compare the clinical characteristics of the distinct patient cohorts. Additionally, these studies do not provide easy access to their joint analysis or processed data sets.

In the following sections I will describe the creation of the most updated and comprehensive transcriptional signature of human end-stage heart failure based on bulk transcriptomics. First, I focus on the selection of studies and the initial comparisons of clinical and transcriptional data across patient cohorts. Later I demonstrate that disease footprints are more powerful than single biomarkers in identifying conserved regulatory events across heterogeneous patient cohorts. Then I describe the procedure used to generate a prioritized list of deregulation events with a meta-analysis and its characterization with functional transcriptomics tools. Finally, I showcase the usage of this consensus signature to validate the known reactivation of fetal transcriptional programs during heart failure and to explore the potential cardiac origin of plasma biomarkers measured with proteomics.

2.2 Study selection and global comparisons

We curated a collection of microarray and RNA-seq studies profiling human heart failure available in NCBI's Gene Expression Omnibus (GEO) database, the European Nucleotide Archive (ENA) and ArrayExpress. We used the following search terms: "heart failure", "ischemic cardiomyopathy", "dilated cardiomyopathy", "cardiac failure" and "heart disease".

Studies were included in our analysis if:

(i) Case samples came from biopsies of the heart's left ventricle of heart failure patients with either ischemic cardiomyopathy (ICM) or dilated cardiomyopathy (DCM) that were acquired during heart transplantation, left ventricular assist device implantation or surgical ventricular restoration.

(ii) Control samples were obtained from patients with non-failing hearts.

(iii) At least five samples were profiled (controls + heart failure patients).

(iv) Microarray data came from single channel chips and could be processed through automatic annotation pipelines.

(v) A publication or preprint with a detailed methodology was available.

The 16 studies we selected are presented in Table CH2-1. We found one additional study (vanHeesch19) from literature review. Studies from the database query results that did not match inclusion criteria due to differences in heart failure etiology, biopsy location or profiling

platform were used for further exploration of the disease score classifier (GSE10161, GSE4172, GSE76701, GSE84796, GSE9800, GSE52601) (Table CH2-2).

2.2.1 Variability in clinical reports and patient cohorts

The selected 16 studies consist of 263 control, 372 DCM and 281 ICM patient specimens and were published between 2005 and 2019. Studies profiled from 5 to 313 samples (Figure CH2-1B). Most of the studies lacked complete descriptions of the clinical and demographic characteristics of the patients included in their publications (Figure CH2-1A). This demonstrates the lack of transparency of the study design that impedes reproducibility of their findings and limits comparability. The minimum reports in 10 of 16 studies included age and sex information (Figure CH2-2A). Five studies included New York Heart Association (NYHA) classifications for heart failure patients.

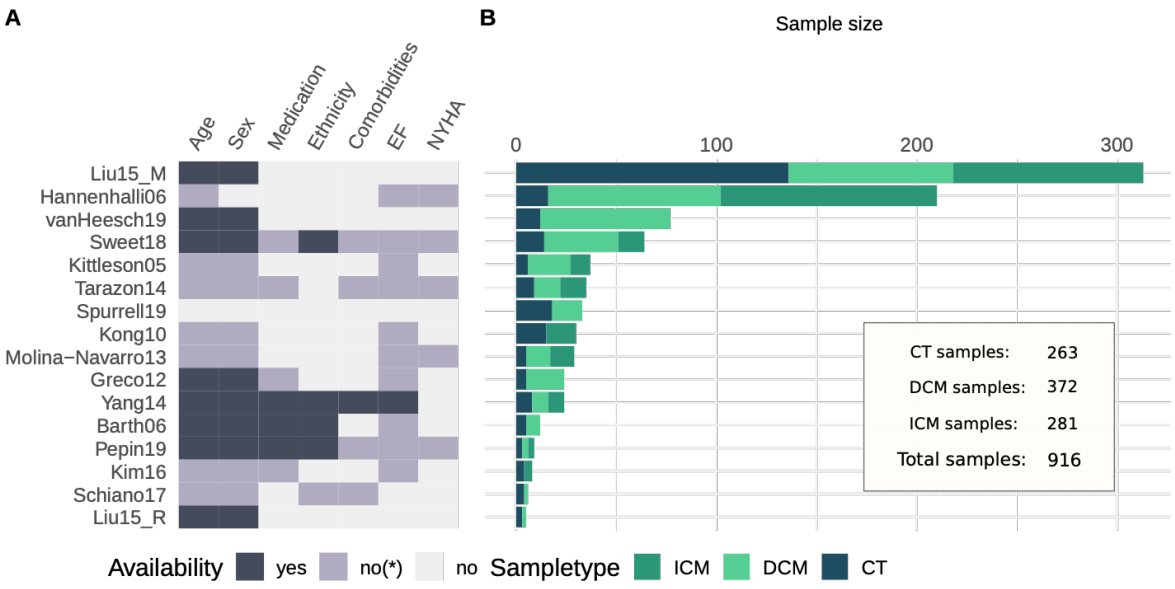

Figure CH2-1. Infographic of study information.
A, Sample information availability per study. yes, information per sample; no* = incomplete information or only summary statistics (i.e. information not applicable for modeling); no, no information available; EF, (left ventricular) ejection fraction; NYHA, New York Heart Association-Classification. **B**, Sample size comparison of studies. CT, Control; DCM, dilated cardiomyopathy; ICM, ischemic cardiomyopathy. Reprinted from [112].

Study ID	GEO ID	Samples (CT + HF)	Technology	Year	Country	Citation
Liu15_M	GSE57345	313	Microarray	2015	USA	[120]
Hannenhalli06	GSE5406	210	Microarray	2006	USA	[121]
vanHeesch19	not in GEO	77	RNAseq	2019	Germany	[122]
Sweet18	GSE116250	64	RNAseq	2018	USA	[123]
Kittleson05	GSE1869	37	Microarray	2005	USA	[124]
Tarazon14	GSE55296	35	RNAseq	2014	Spain	[125]
Spurrell19	GSE126573	33	RNAseq	2019	USA	[126]
Kong10	GSE16499	30	Microarray	2010	USA	[127]
Molina-Navarro13	GSE42955	29	Microarray	2013	Spain	[128]
Greco12	GSE26887	24	Microarray	2012	Italy	[129]
Yang14	GSE46224	24	RNAseq	2014	USA	[130]
Barth06	GSE3585	12	Microarray	2006	Germany	[131]
Pepin19	GSE123976	9	RNAseq	2019	USA	[132]
Kim16	GSE76701	8	Microarray	2016	USA	[133]
Schiano17	GSE71613	6	RNAseq	2017	Italy	[134]
Liu15_R	GSE57344	5	RNAseq	2015	USA	[120]

Table CH2-1. Collection of compiled studies.
16 data sets fulfilled the inclusion criteria. Sample size is displayed after processing. GEO, gene expression omnibus; CT, control; HF, heart failure. Reprinted from [112].

GEO ID	n (total)	HF etiology	Reason for exclusion	Technology	Citation	Year	Country
GSE10161	27	Aortic stenosis	HF etiology	Microarray	[135]	2008	Netherlands
GSE4172	12	Inflammatory DCM due to PVB19 infection	HF etiology & samples from right ventricle	Microarray	[136]	2006	Germany
GSE84796	17	Chagas disease	HF etiology	Microarray	[137]	2016	France
GSE9800	30	Eosinophilic myocarditis, alcoholic cardiomyopathy, hypertrophic cardiomyopathy, sarcoidosis, peripartal cardiomyopathy, ICM, DCM	HF etiology	Microarray	-	2007	Japan
GSE52601	20	ICM, DCM (additional 4 fetal samples)	Technical	oligonucleotide beads	[138]	2013	USA
GSE3586	28	DCM	Technical	Microarray	[131]	2013	Germany
GSE76701	8	ICM	Technical	Microarray	[133]	2016	USA

Table CH2-2. Collection of additional studies.
7 data sets were excluded from the main analysis. Sample size is displayed after processing. GEO, gene expression omnibus; CT, control; HF, heart failure. Reprinted from [112].

Despite these reporting problems, we assumed that given the circumstances of biopsy acquisition all heart failure biopsies came from patients who suffered a (close to) decompensated failing heart with reduced ejection fraction, justifying their inter-study comparability. As control samples, all studies included biopsies from donor hearts deemed unsuitable for transplant due to size disparities, blood (ABO) mismatch or other factors. The age of heart failure patients is noticeably younger than what would be expected, since heart failure prevalence increases with age (Figure CH2-2A). This might suggest that especially rapidly progressing heart failure cases which clinically justified early left ventricular assist device treatment or heart transplantation were recruited. In studies with NYHA class, the patients were either in class III or IV.

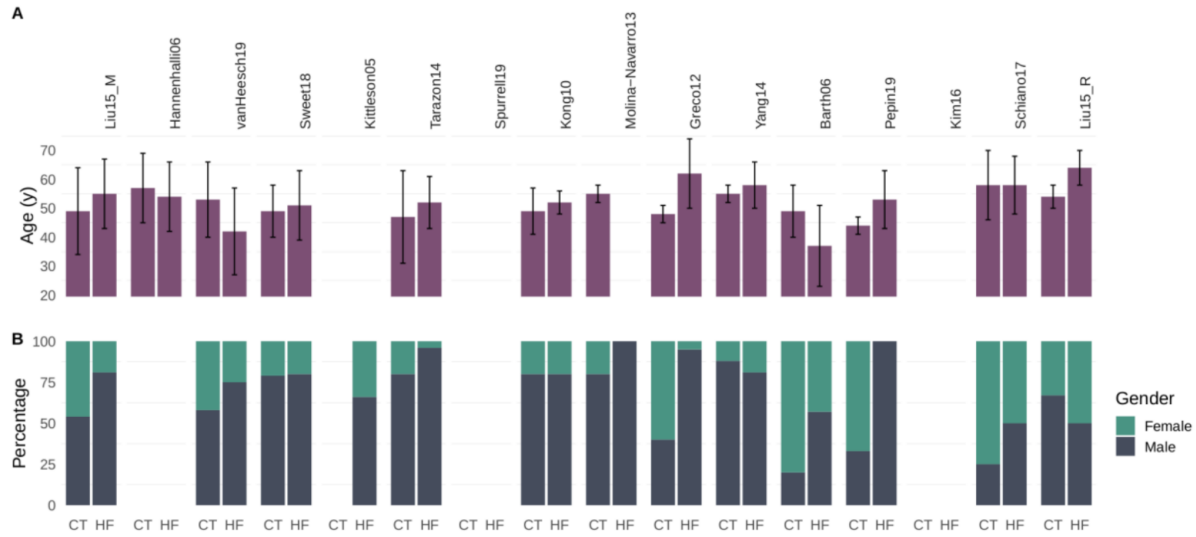

Figure CH2-2. Age and gender distribution per study.
A, Age distribution in years of control (CT) and heart failure samples (HF) per study. Displayed is mean and standard deviation. **B,** Gender of patients in % per study. Reprinted from [112].

Without clinical and demographic information it is impossible to estimate the uncertainty in the estimations of differential gene expression that are not solely confounded by other technical factors, but also confounded by biased sampling that decreases the effective sample size. For the same reason, a direct comparison of the molecular signatures of heart failure is timely and became the focus of my research.

2.2.2 Technical variability between studies

As mentioned in Chapter 1 there are fundamental differences between RNA sequencing and microarray data, particularly in the scale of the measured gene expression and gene coverage [139]. Additionally, data processing confounding factors can be added via gene expression transformations and models for differential expression analysis, thus overestimating the differences from the compiled data sets. To avoid increasing the already numerous missing and confounding factors in the comparison of the insights of the collected studies, I proposed an homogeneous processing framework for microarrays and RNA-Seq data.

Processing transcriptomic data sets

For microarray studies, I read CEL files with R's *oligo* package and normalized them using Robust Multi-array Average (RMA) [140]. Probes were annotated to their corresponding HUGO Gene Nomenclature Committee (HGNC) gene symbols using platform specific annotations. For duplicated measurements I summarized them with the mean. The measurements in microarrays are continuous, since they measure light intensity (See Chapter 1, section 1.2.1).

For RNA-Seq studies, we aligned reads using BioJupies [141]. BioJupies works with the ARCHS4 pipeline utilizing Kallisto to map reads onto the human GRCh38 cDNA reference. All studies were processed by Illumina platforms except for Tarazon14, which utilized AB 5500xl Genetic Analyzer. Here the nucleotide sequence is coded in color space that could not be handled by the BioJupies pipeline and the alignment of Tarazon14 was therefore performed with R's Rsubread package [142], TMM normalization factors were calculated with R's edgeR package [47]. All RNAseq datasets were transformed using voom from R's limma package to obtain continuous measurements comparable to microarrays [46].

Hannenhalli06 only provided processed data, but followed identical normalization methods. In the case of Kittleson05, processed data was used since raw available data was incomplete. Identical normalization procedures were followed. Read alignment was not performed for vanHeesch19 since they only provided raw transcript counts, but identical normalization procedures were followed. One sample from Liu15_R was excluded due to technical reasons.

For each experiment, sample quality was assessed by visually comparing the distribution of gene expression values. Multidimensional scaling was performed to visualize the separation of heart failure and control samples. No samples were excluded based on these metrics and no additional quality control was performed. Gene coverage after processing was comparable for all studies (mean jaccard index of ~0.67). A total of 14,041 genes were reported by at least 10 studies (Figure CH2-3).

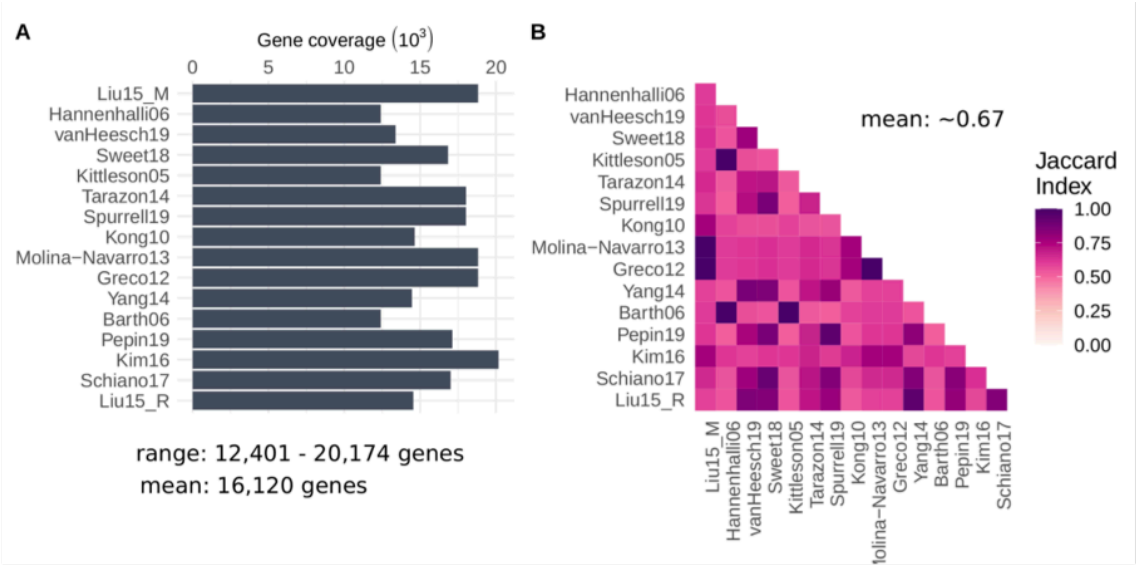

Figure CH2-3. Overview of gene coverage of studies included in meta-analysis.
A, Absolute gene coverage per study after processing. **B**, Pairwise comparison of covered genes measured with Jaccard Index. Reprinted from [112].

Gene expression global differences across studies

As I expected to observe study specific effects in gene expression despite consistent data handling for all datasets, first I quantified the global variation of gene expression on the union of all pre-processed datasets associated with study labels and heart failure etiology. I used principal component analysis (PCA) on different transformations of the unified expression matrices with the genes that were shared among all the studies to reduce the highly dimensional gene feature space and to be able to quantify the amount of explained variance associated with covariates of interest. Each principal component was tested for association with the study labels or heart failure using Analyses of Variance (ANOVAs) (p-value < 0.05).

Three types of data unification (with shared genes) were performed here:

i) All samples across all studies were collected in a single matrix after processing.

ii) z-transformed gene expression of heart failure patients across all studies were collected in a single matrix: For each gene in each heart failure patient, I subtracted from its gene expression the mean value of that gene in healthy patients within the patient cohort and then divided the result by the standard deviation of the healthy patients. I did this for each study individually before merging.

iii) Standardized gene expression of all samples in a single matrix after processing: I first standardized (mean = 0, sd = 1) all genes independently for each study including all samples and then merged them into a single matrix.

In a PCA of all unified gene expression values after processing (i), 85% of the variance of the samples was explained by the first two components representing study of origin and applied technology. In z-transformed heart failure samples (ii), 74% of the variance captured by the principal components associated with the study labels (ANOVA p-value <0.05). The difference of samples by study was better visualized when a t-SNE was performed to this data (Figure CH2-4). However, in gene standardized data (iii), the proportion of variance explained by study labels decreased to 0, suggesting that study or technical specific effects are associated with gene expression scales (Figure CH2-5).

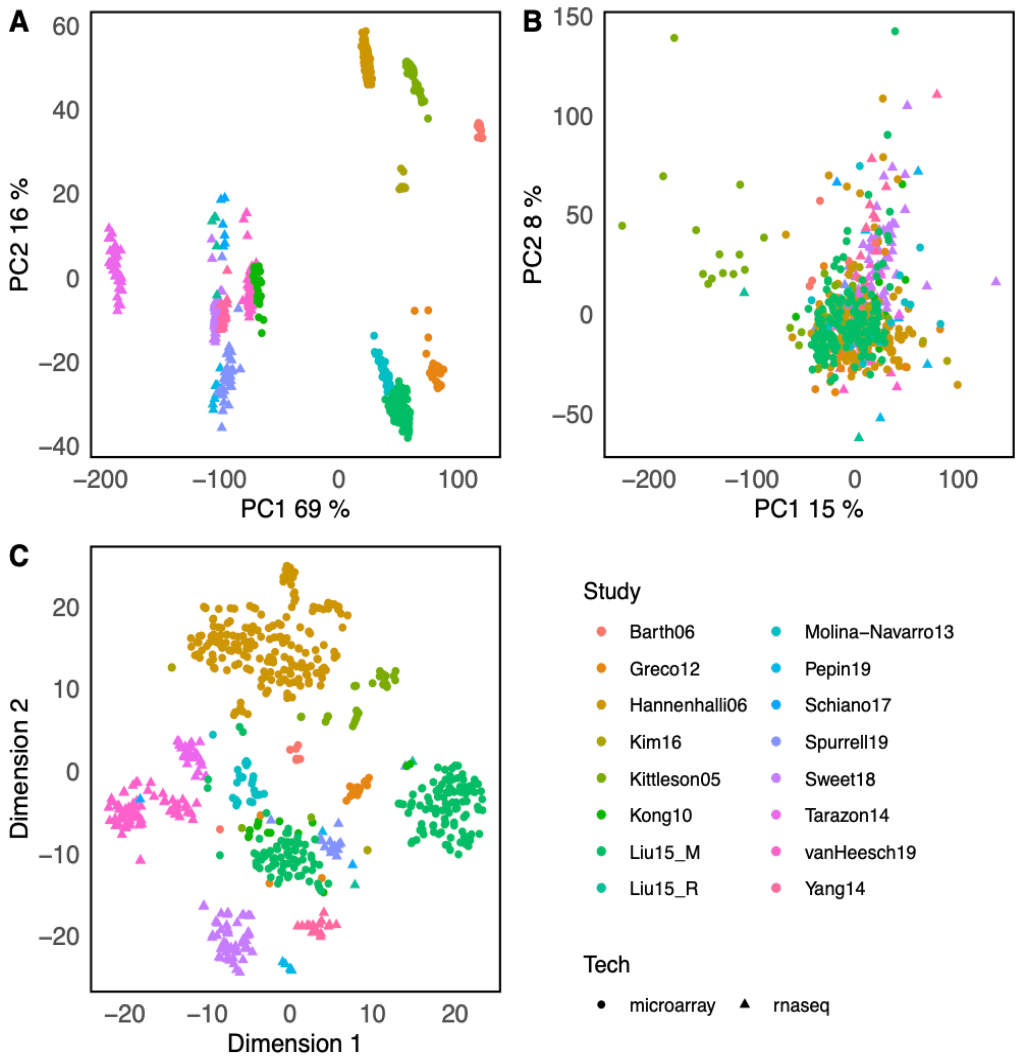

Figure CH2-4. Differences in samples included in the study.
A, First two components from a Principal Component Analysis (PCA) done to all samples, **B,** First two components from a PCA done to all z-transformed heart failure samples. **C,** t-distributed stochastic neighbor embedding of all z-transformed heart failure samples Reprinted from [112].

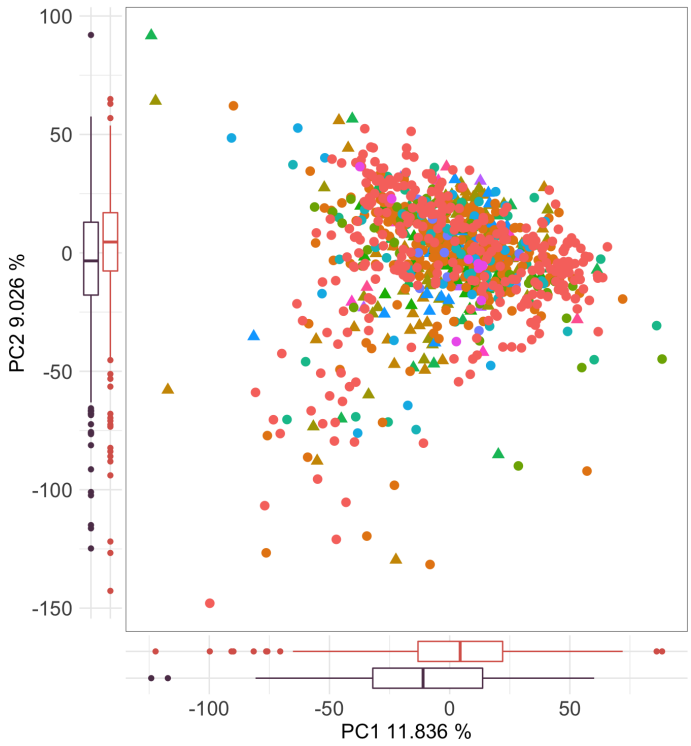

Figure CH2-5. Principal Component Analysis of all samples analyzed after gene standardization.
The scatter plot shows the first two principal components and the percentage of variance explained by them. Boxplots show the distribution of failing (red) and non-failing (purple) hearts. Reprinted from [112].

2.2.3 Gene expression variability associated with clinical covariates

To quantify how much of the variability of the samples within a study can be explained by the clinical covariates used in their differential expression analysis, I fitted linear models to a reduced data representation. For each study, first, I standardized its gene expression and performed dimensionality reduction using PCA. Then I tested each principal component for association with each covariate using linear models. If a covariate was associated with a principal component (p-value < 0.05), then I assigned the proportion of explained variance to it. I also applied the same methodology for heart failure patients only. Underestimations of proportion of explained variance are expected in small studies, since the number of evaluated principal components equals the number of samples. However this is a fair approximation for most of the studies.

Analysis of individual studies revealed that most of the variability of the patients can not be assigned to reported covariates (Figure CH2-6). Unmeasured variability may come from clinical, demographic, or genetic differences between patients, but also from differences in

tissue biopsies mostly associated with location and cell-composition. In studies with reported age and sex differences I observed different contributions of these covariates to the variability of patients, which highlighted the diversity in experimental designs. In the case of heart failure patients (Figure CH2-6B), compared to age (mean = 0.09, std.deviation = 0.08), or difference in sample acquisition (mean = 0.34, std.deviation = 0.04), etiology had a lower mean proportion of explained variance (mean = 0.0698, std.deviation = 0.0624). The variability in gene expression in heart failure patients may be explained by other clinically relevant features, but given the lack of patient information this could not be tested.

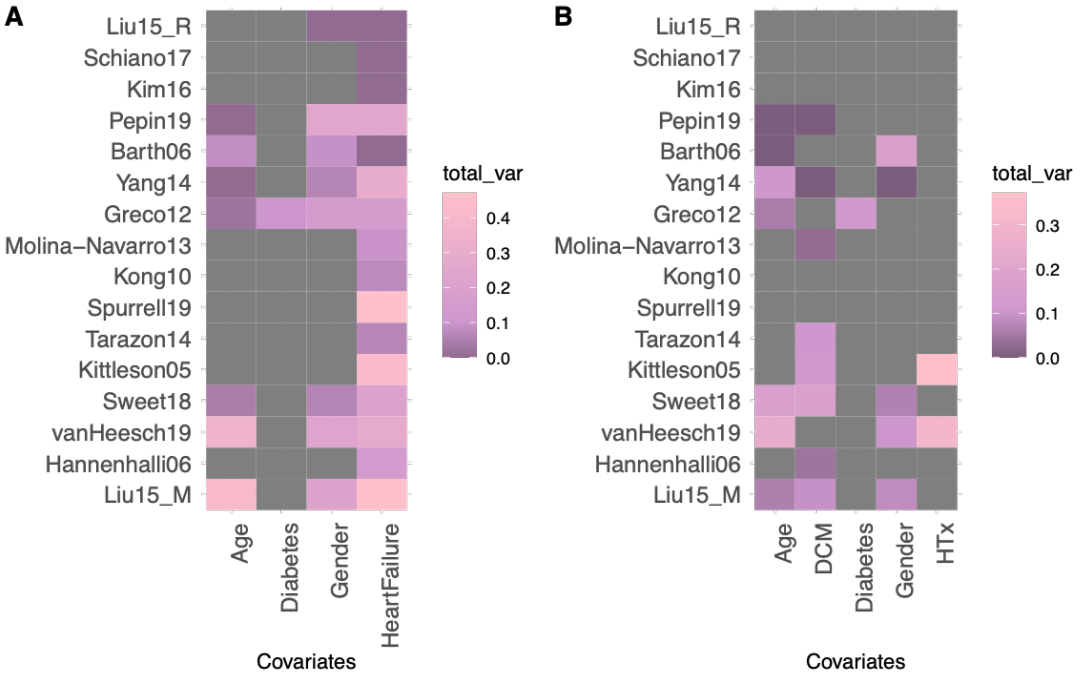

Figure CH2-6. Contribution of the covariates to the variability of individual studies.
Estimated proportion of explained variance associated with the different covariates used in the differential expression analysis in **A,** all patients and **B,** only heart failure patients. Gray tiles represent missing reported data. HTx, heart transplantation. Reprinted from [112].

2.2.4 Comparison of differentially expressed genes in heart failure across samples

To identify differentially expressed genes within each study, gene expression of the samples of control individuals and heart failure patients were compared using linear models with *limma* [46]. Gender, age, comorbidities, etiology, occasion of sample acquisition and technical batches were used as covariates for experiments that provided this information. Samples with incomplete clinical information from vanHeesch19 were excluded from the analysis to be able to account for the clinical information of the remaining samples in the differential expression analysis. I excluded the age information in the differential expression analysis of the samples

from Kim16. Here, excluding samples with unknown age information would have reduced the sample size drastically.

I evaluated the consistency in gene differential expression across the studies by comparing their transcriptional signatures using multiple metrics:

To explore how technical and sample variability affected gene level statistics of the differential expression analysis, I compared the distributions of p-values, t-values, and log fold changes associated with the coefficient capturing the difference between control and heart failure patients in the linear model. A strong difference in the distributions of t-values and p-values of the genes compared was visible in the largest study in our analysis (Liu15_M) (Figure CH2-7). This difference in distributions persisted after adjustment for all available clinical covariates, though it was consistent with expectations based on study sample size. These results together establish expected bias among datasets, likely dependent on technical differences rather than biology.

Then, I calculated the pairwise overlap of the top 500 differentially expressed genes of each study with Jaccard indexes. The mean Jaccard index of pairwise comparisons was 0.05, representing an almost null concordance at the gene level (Figure CH2-8A). This demonstrated the instability of gene expression biomarkers when put in context of different patient cohorts.

The lack of generalization of gene expression markers could be a consequence of the differences in patient sampling and technical procedures of individual studies, however, these markers may also capture to some extent a disease transcriptional footprint that could be effectively mapped in independent datasets. For this reason, I evaluated if the list of the differentially expressed genes of one study could be used as a predictor of heart failure in each other study, using sample classifications based on a disease score.

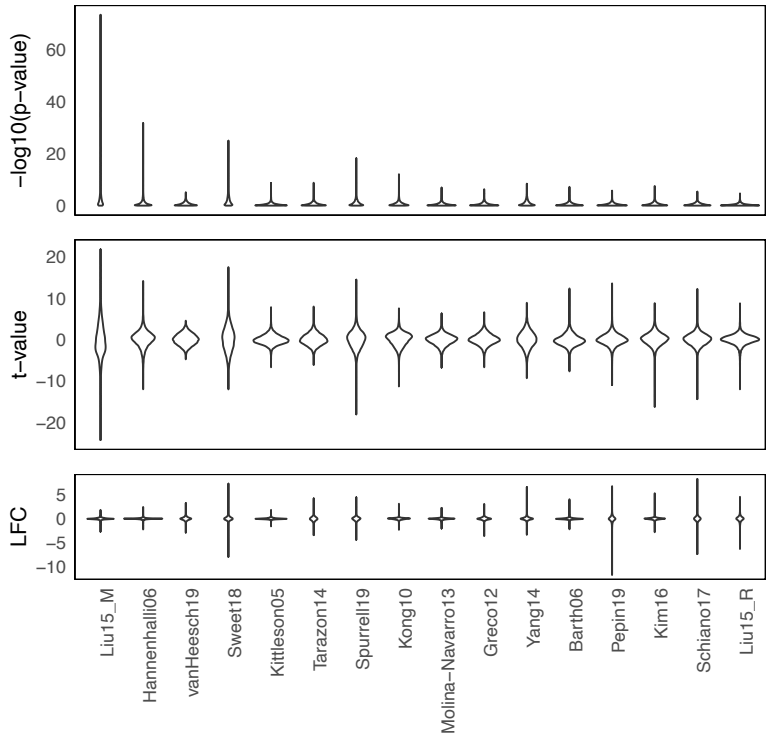

Figure CH2-7. Distributions of -log10(p-values), t-values and log2(fold-changes) [LFC] from the differential expression analysis of all genes measured in each study.
Reprinted from [112].

The disease score is an expression footprint based approach, inspired by "probability of expression" [143] and PROGENy [56], that compares the observed expression patterns in the samples of one experiment (B) with the expected disease patterns observed in an independent sample from another experiment (A). First, for an experiment *A*, *k* differentially expressed genes between the healthy and disease condition are defined using linear models. The t-values of these *k* genes are used as the expected disease pattern to be used for transfer learning. Then, for each sample *i* in experiment B, I calculate its disease score by making a linear combination of the t-values from these k genes with their expression values in sample i, for genes present in both the reference signature and the expression values (Figure CH2-9). Then, Standardized disease scores were used to classify heart failure patients in individual studies using as reference the disease patterns of all of the other studies. My assumption is that if two studies derive similar heart failure transcriptional signatures, then the disease score should effectively differentiate heart failure and healthy patients. In total, 16 disease classifiers were built corresponding to the t-values of the top 500 differentially expressed genes of each study included in the analysis. The area of the receiver operating characteristic curve (AUROC), where heart failure was used as a response variable, was used to test the accuracy of

classification of heart failure patients and used as a measurement of conservation of gene regulation patterns and similarity between studies.

When the top 500 differentially expressed genes of each study were used to build the disease score, the median AUROC of the classifications was 0.94, showing that despite technical differences, each study contained meaningful and complementary information (Figure CH2-8B). Studies that profiled only patients with ischemic forms of heart failure (eg. Kong10) effectively classified studies that profiled only dilated cardiomyopathy patients (eg. Spurrell19) (AUROC = 1) and vice versa (AUROC = 0.95). I observed no association between each study's mean AUROC and their technology (Wilcoxon test p-value = 0.72, Figure CH2-10A-left), sample size or estimated proportion of variance captured by heart failure (Pearson correlation = 0.17, 0.18, respectively; p-value > 0.4, Figure CH2-10B). The effectiveness of transcriptional footprints to transfer knowledge of independent cohorts can be explained by the fact that this approach focuses on differential trends of the direction of transcriptional regulation rather than differential levels of expression of a set of gene markers.

To confirm that the coordination of molecular responses is conserved among studies, I tested if the direction of deregulation of the top differentially expressed genes of each study were consistent with their direction in the rest of the studies. I separated the top differentially expressed genes of each study into up and downregulated genes, and enriched them into the sorted gene-level statistics of each of the other studies using Gene Set Enrichment Analysis (GSEA) [144]. Gene-level statistics of each study were sorted by their t-value.

The top 500 differentially up-regulated and down-regulated genes had a median enrichment score of 0.55 (Figure CH2-8C, upper panel) and -0.56 (Figure CH2-8C, lower panel), respectively. I observed a correlation between the AUROCs of the disease score classifications and the enrichment scores of differentially expressed genes (Pearson correlations 0.48 and -0.59; p-value<10e-15, for up and downregulated genes, respectively), supporting the idea that even though the size effects of heart failure relevant genes are dependent on the study (Figure CH2-8A), their direction of regulation is generally consistent (Figure CH2-8C), allowing their direct comparison.

I observed similar results when I selected different numbers of top genes (50, 100, 200, 500, and 1000) for both the disease score classifier and the enrichment analysis (Figure CH2-11)

These results suggest that the proper way to combine the evidence of the curated studies is by looking at the consistency of deregulation of collections genes and not at the dimension of the change in expression of a single one. I provided evidence to support the combination of evidence from multiple studies with various clinical and technical profiles. Finally, I propose a simple but effective framework to compare molecular signatures of transcriptomics profiles that can be used in different contexts.

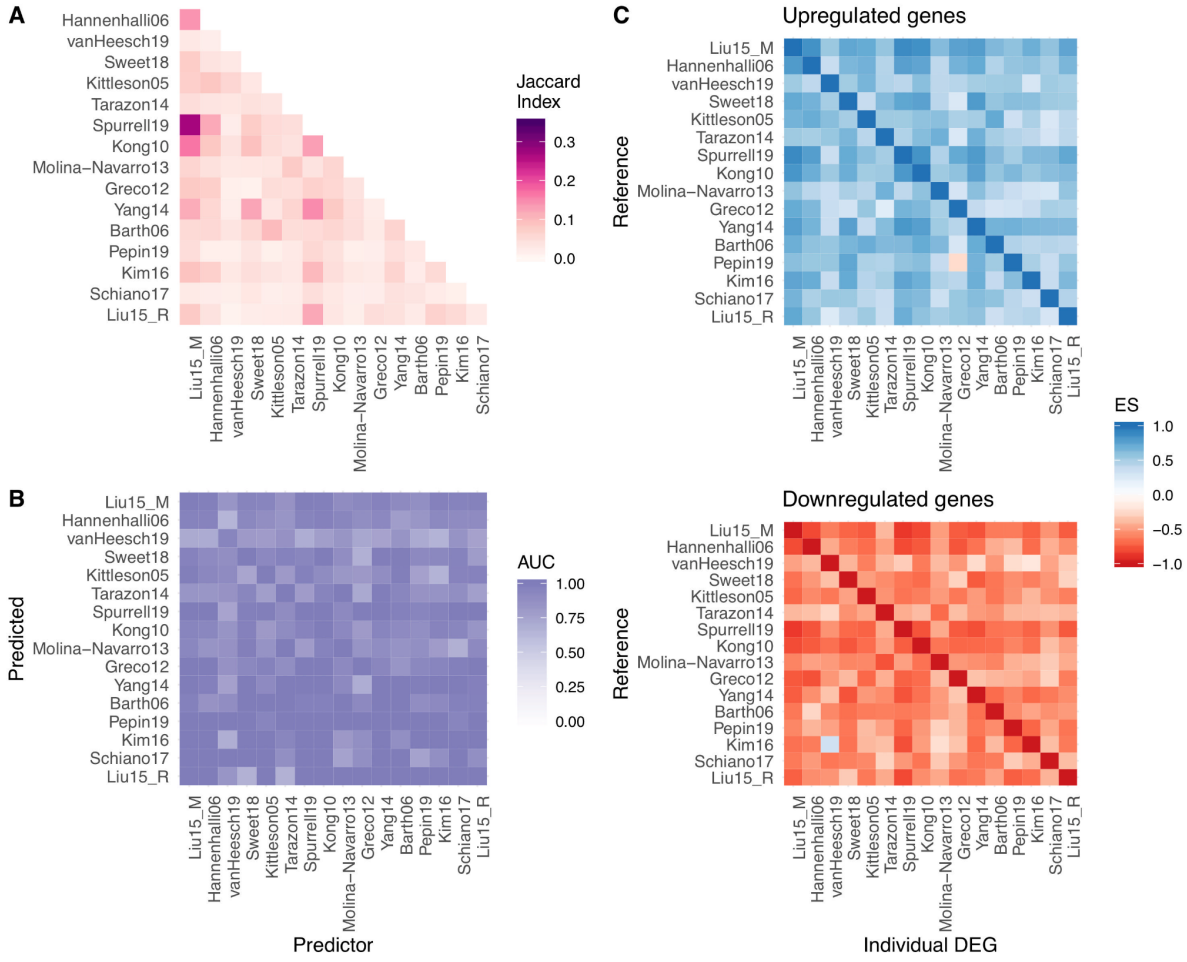

Figure CH2-8. Consistency of the transcriptional signal of end-stage HF among studies.
A, Pairwise comparison of the top 500 differentially expressed genes of each study using the Jaccard index. **B**, area under the receiver operating characteristic curve (AUC) of pairwise predictions using a disease 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. Reprinted from [112].

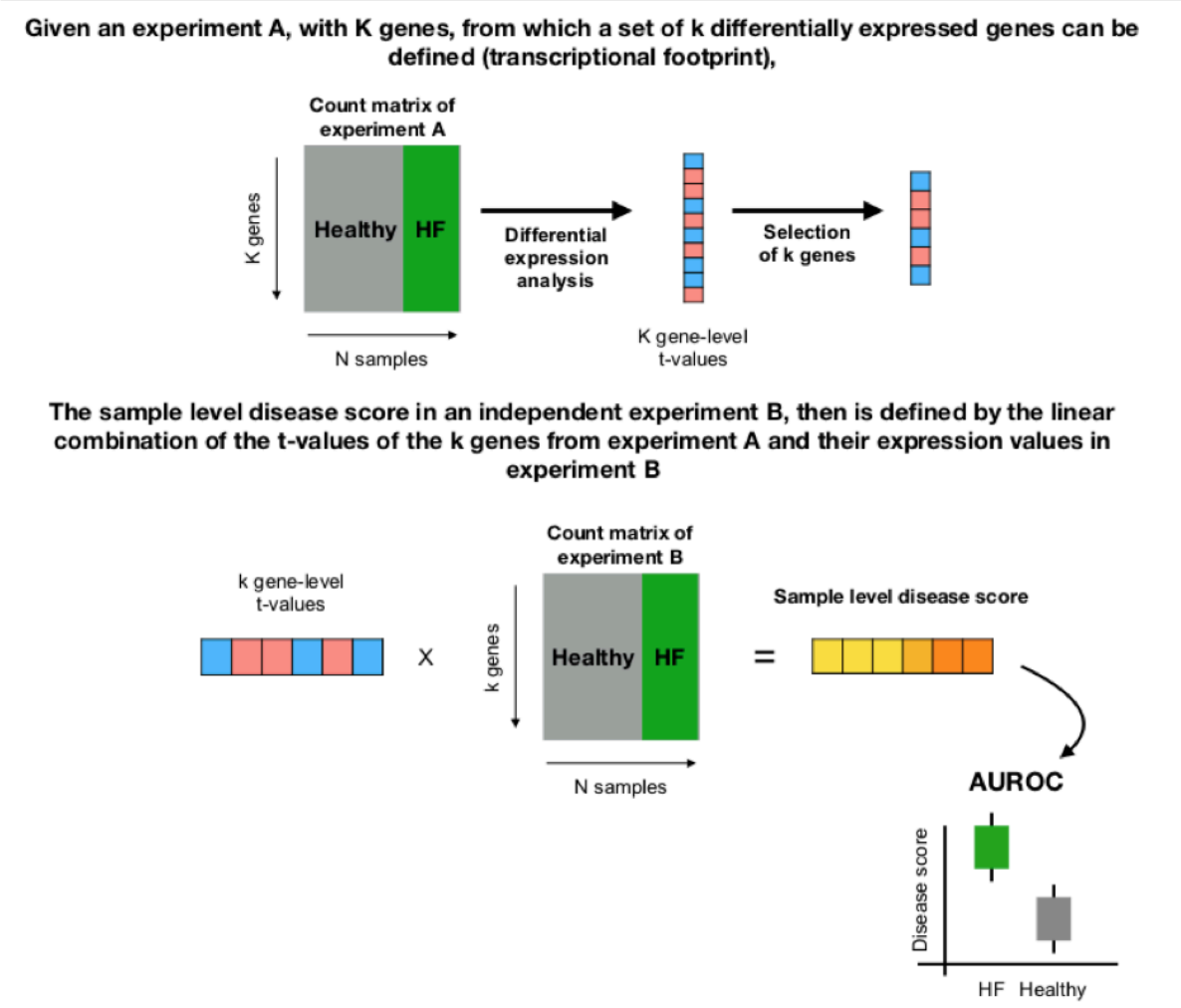

Figure CH2-9. Schematic representation of the disease score.
AUROC, area under the receiver operating characteristic. HF, heart failure. Reprinted from [112].

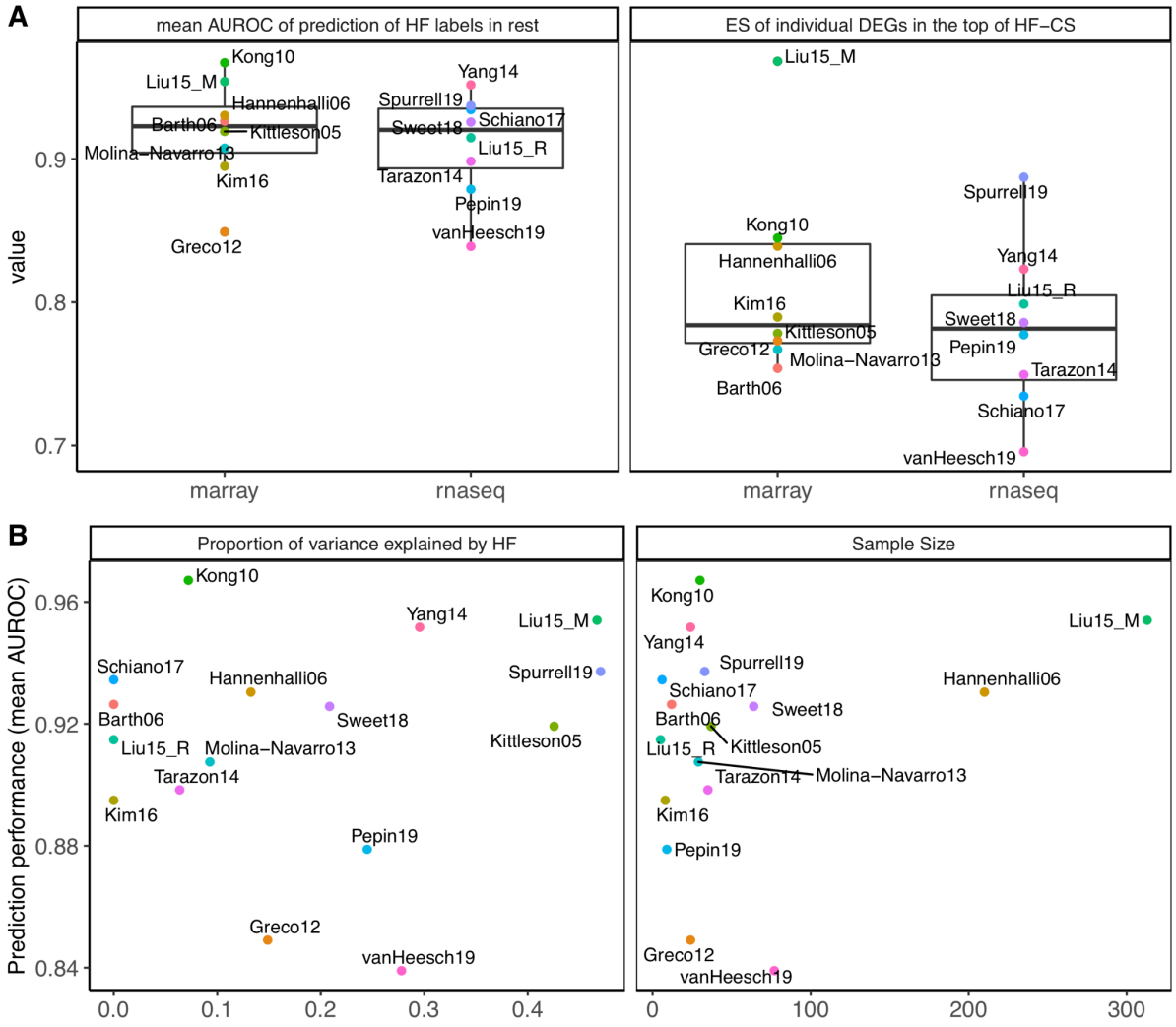

Figure CH2-10. Comparison of the studies included in the meta-analysis.
A, Distributions of study mean predictor performances using the disease score and enrichment of the differentially expressed genes of each study in the heart failure consensus signature (HF-CS) grouped by technology. In the left panel each dot represents the mean area under the receiver operating characteristic curve (AUROC) of the disease score classifier trained in a study and tested in the rest. In the right panel each dot represents the enrichment scores of the top 500 differentially expressed genes of the study in the HF-CS. **B,** Relationship between the predictive performance of each study and its proportion of explained variance associated with heart failure and sample size. Reprinted from [112].

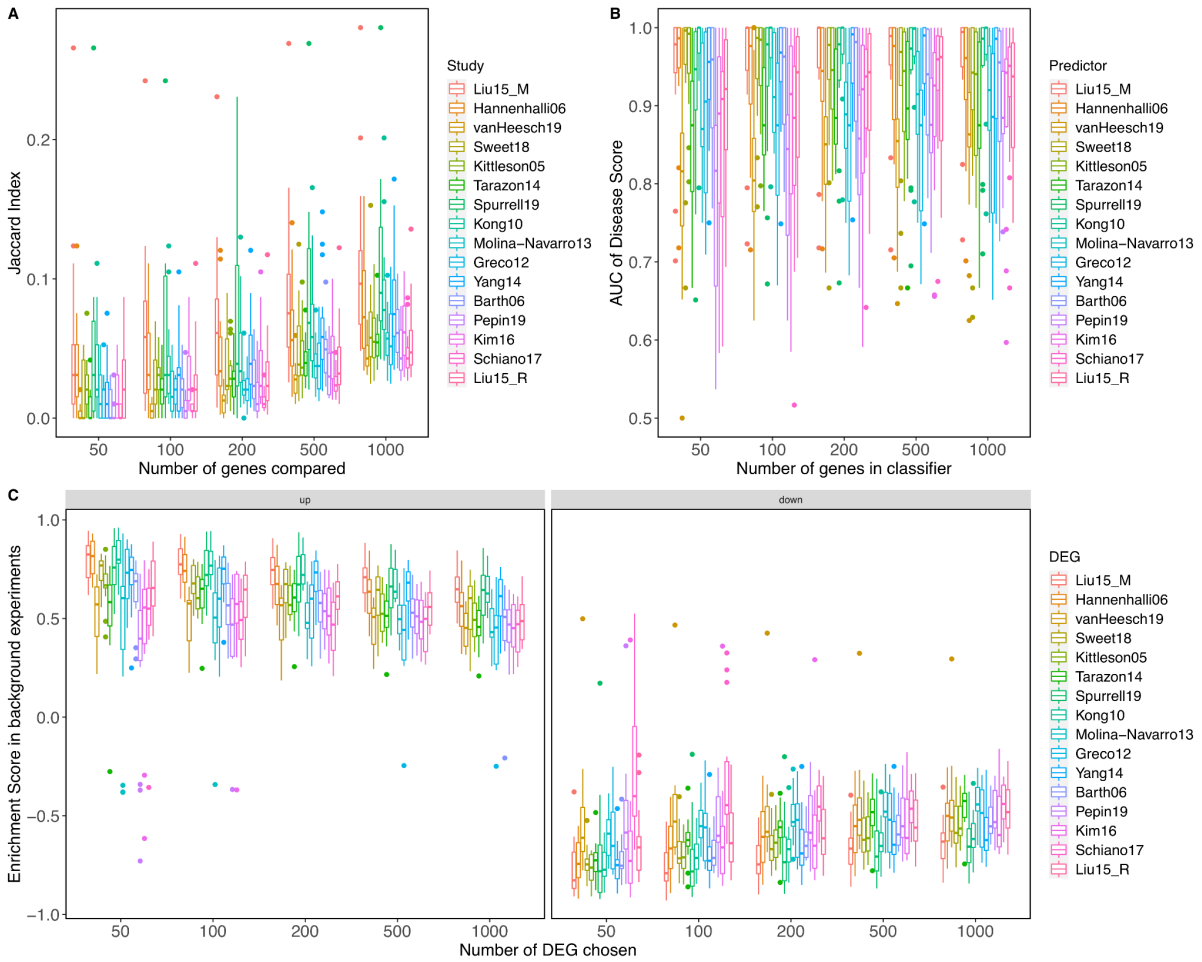

Figure CH2-11. Test of robustness of the replicability measures used to compare the studies included in the meta-analysis.
Each dot represents a pairwise comparison using:
A, Jaccard Index
B, Disease Score
C, Enrichment Score
Reprinted from [112].

2.3 Consensus signature of heart failure

I then performed an integrated meta-analysis of the transcriptional signatures of all the studies using the gene-level statistics of 14,041 genes. I combined the Benjamini-Hochberg (BH) corrected p-values of the differential expression analysis for all genes that were measured in at least 10 datasets using a Fisher's combined probability test. The degrees of freedom for the significance test of each gene were defined by the number of datasets that included it. I assumed that non-probabilistic sampling procedures happened in each study (which affects the effective sample size [109]), so no additional study weighting based on the number of profiled samples was used. This flexible hypothesis setting detects genes that have non-zero effect size in one or more studies and allows the comparison of heterogeneous datasets [145]. I ranked all genes by their meta-analysis BH corrected p-values to create a consensus signature (Figure CH2-12) that captures a gradient of consistently differentially regulated genes in end-stage heart failure across multiple studies regardless of their direction.

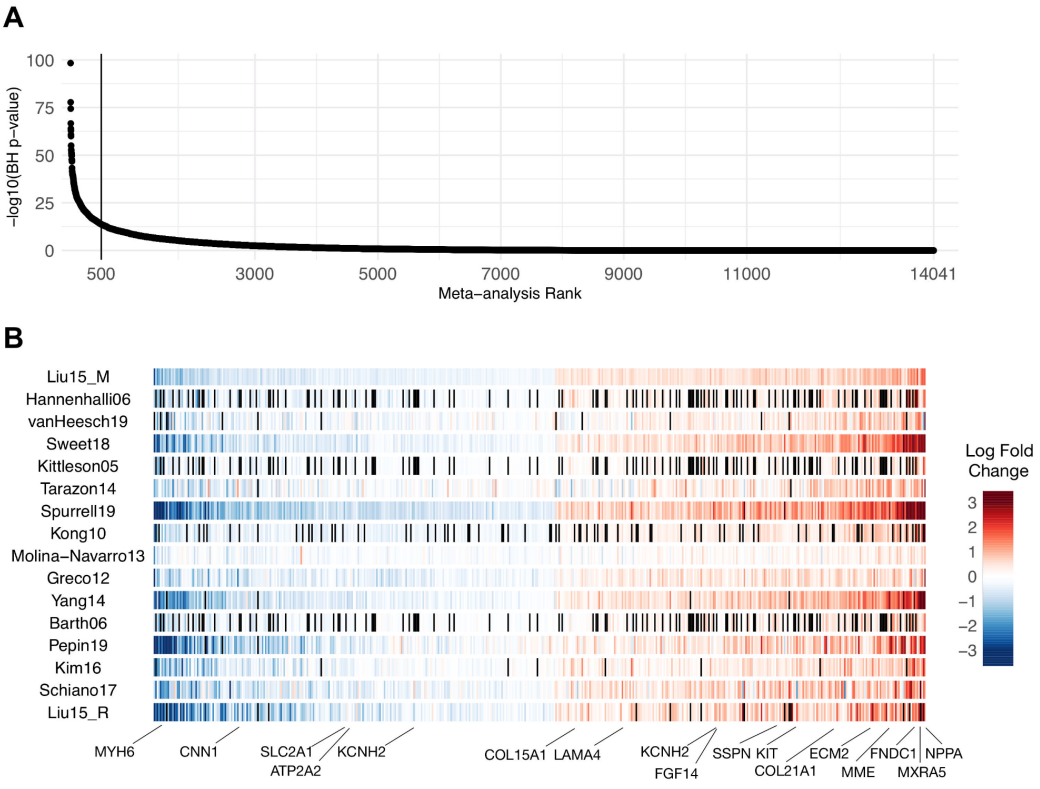

Figure CH2-12. Meta analysis of heart failure gene expression signatures.
A, Sorted −log10 (meta analysis BH p-values) of the 14 041 genes included in the Fisher combined test, representing the heart failure consensus signature (HF-CS). **B**, Top 500 genes sorted by their mean log fold change across all studies; black lines represent genes that were not measured in specific studies. A selection of heart failure marker genes are highlighted. BH indicates Benjamini–Hochberg. Reprinted from [112].

Among the top 500 genes in the heart failure consensus signature (Figure CH2-12B) we observed known heart failure markers such as MYH6, MME, CNN1, NPPA, KCNH2 and ATP2A2; extracellular associated proteins such as COL21A1, COL15A1, ECM2 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.

To test the relevance of the ranking, I evaluated the importance of the top genes of the meta-analysis ranking in the description of heart failure patients by repeating the classifications made with the disease score described before. Samples of each study were classified using a disease score using the first *n* or *total-n* genes in the meta-ranking and study-specific t-values. AUROCs were averaged for each predicted study and n ranged from 50 to the total number of genes in the meta-ranking I observed a constant decrease in the mean AUROCs of classifiers that excluded genes at the top of the consensus signature or included genes at the bottom (Figure CH2-13), confirming that a gradient of meaningful information is present in this ranking.

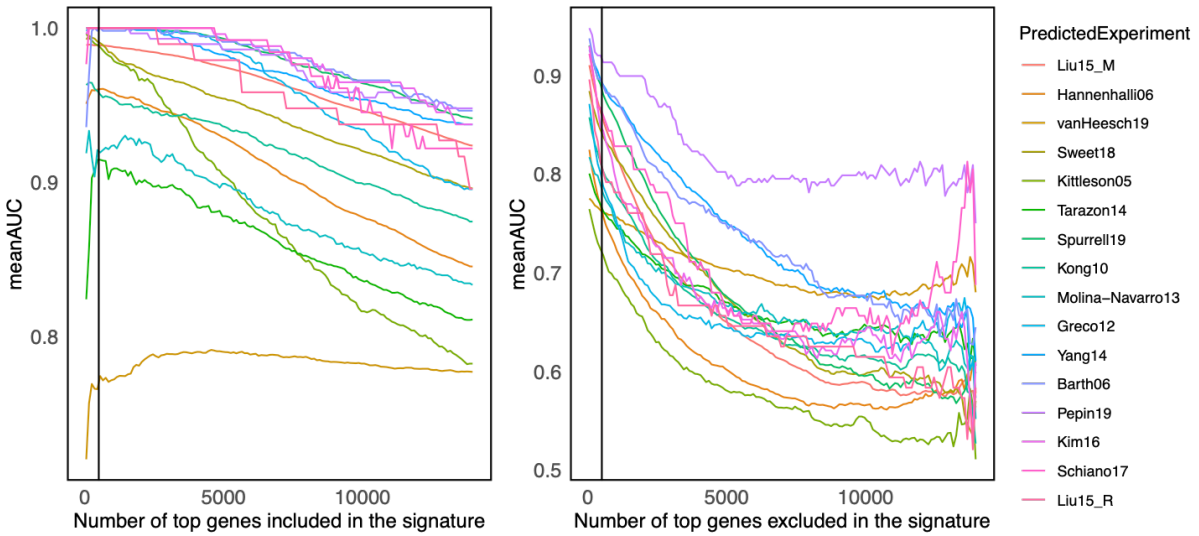

Figure CH2-13. Relevance of the genes of the consensus signature.
Mean area under the receiver operating characteristic curve (meanAUC) of predictions using the disease score with n (left panel) or total-n (right panel) genes of the consensus signature from the meta-analysis and gene-level statistics of all studies except the one being predicted to avoid overfitting. The line shows where I defined the cut-off for the rest of the tests (500). A general decrease of the meanAUC is observed as top genes of the meta-analysis are excluded from the calculation of the disease score. Reprinted from [112].

The contribution of each study to the meta-analysis was estimated with the enrichment score of its top 500 differentially expressed genes in the meta-ranking as calculated by Gene Set

Enrichment Analysis [144]. I found no correlation between the sample size of a study and the enrichment of its differentially expressed genes in the top of the heart failure consensus signature (Spearman correlation 0.24, p-value = 0.37), suggesting that proper experimental design and representative sampling could compensate for study size. Similarly, I found no association between the enrichment of differentially expressed genes of individual studies in the top of the heart failure consensus signature and the technology used (Wilcoxon test p-value = 0.4418, Figure CH2-10A-right), reflecting consistency for all studies.

To test the effect of each study in the final ranking of the heart failure consensus signature, I performed a leave-one-out (LOU) procedure. I repeated the meta-analysis 16 times, each time ignoring the values of one study at a time. Then I compared the similarity of the top 1000 genes of each LOU experiment and the original top 1000 genes of the heart failure consensus signature using a Jaccard Index. The LOU procedure demonstrated robustness of the signature (mean Jaccard index of the top 1000 genes = 0.91), although larger discrepancies were observed when the top 4 largest studies where ignored, as expected (mean Jaccard index of the top 1000 genes = 0.76).

To evaluate the added value of the meta-analysis, I tested if the selection of the top 500 genes from the heart failure consensus signature defined a better transcriptional signature of heart failure than signatures obtained from individual experiments of the same size. I tested if the AUROCs obtained from classifying heart failure samples using the top 500 genes from the heart failure consensus signature were greater than the ones coming from classifications made by the top 500 genes from individual studies using a Wilcoxon paired test. To show that the top genes of the consensus signature shared a more consistent direction of differential regulation than signatures coming from individual studies, I separated the 500 top genes from the consensus signature into up and downregulated independently for each dataset, and enriched them into the sorted gene-level statistics of each of the other studies using Gene Set Enrichment Analysis as in the previous section. I compared the enrichment scores of these pairwise comparisons to the ones obtained using the top 500 differentially expressed genes of individual experiments using a Wilcoxon paired test. An improvement in the AUROCs of classifiers based on the disease score was obtained (Wilcoxon paired test, p-value < 1x10e-16), and the top genes of the heart failure consensus signature were consistently more enriched in individual lists of differentially expressed genes than gene signatures from individual experiments (Wilcoxon paired test, p-value < 1x10e-16).

From a gene perspective, I measured how much of the variance of gene expression was associated with heart failure and other reported technical covariates for genes in the top of the ranking. First, I merged studies after processing and gene standardization. Independent two-way ANOVAs were fitted to each gene using disease status as a first factor, and for samples with available information, sample's study, transcriptional profiling technology, gender, age or occasion of sample acquisition, as a second factor. Each extra covariate was fitted in an independent model. The proportion of variability in gene expression explained by heart failure, controlled for other clinical and technical covariates, was greater for the top genes than for genes in a lower ranking in the heart failure consensus signature (Figure CH2-14). Gene standardization cancels the effect that the study of origin and technology have on gene expression (Figure CH2-5) and can be confirmed by the low eta-squared values in all genes (Figure CH2-14 upper panels).

Additionally, to evaluate the bias of the heart failure consensus signature towards dilated cardiomyopathy, I performed independent two-way analysis of variance (ANOVAs) to quantify the amount of explained variance in gene expression that could be accounted for by differences in heart failure etiology (Figure CH2-15). First, I selected 8 studies in our curation that profiled sufficient ICM and DCM patients (at least 3 patients of each etiology). Then, for each selected study I fitted to each gene an ANOVA with heart failure and etiology as covariates. Eta-squared values of each covariate were used as a proxy of the proportion explained variance. In all studies reporting expression of ICM and DCM patients, I observed a greater amount of explained variance associated with heart failure than to the etiology in the top 500 genes of the heart failure consensus signature.

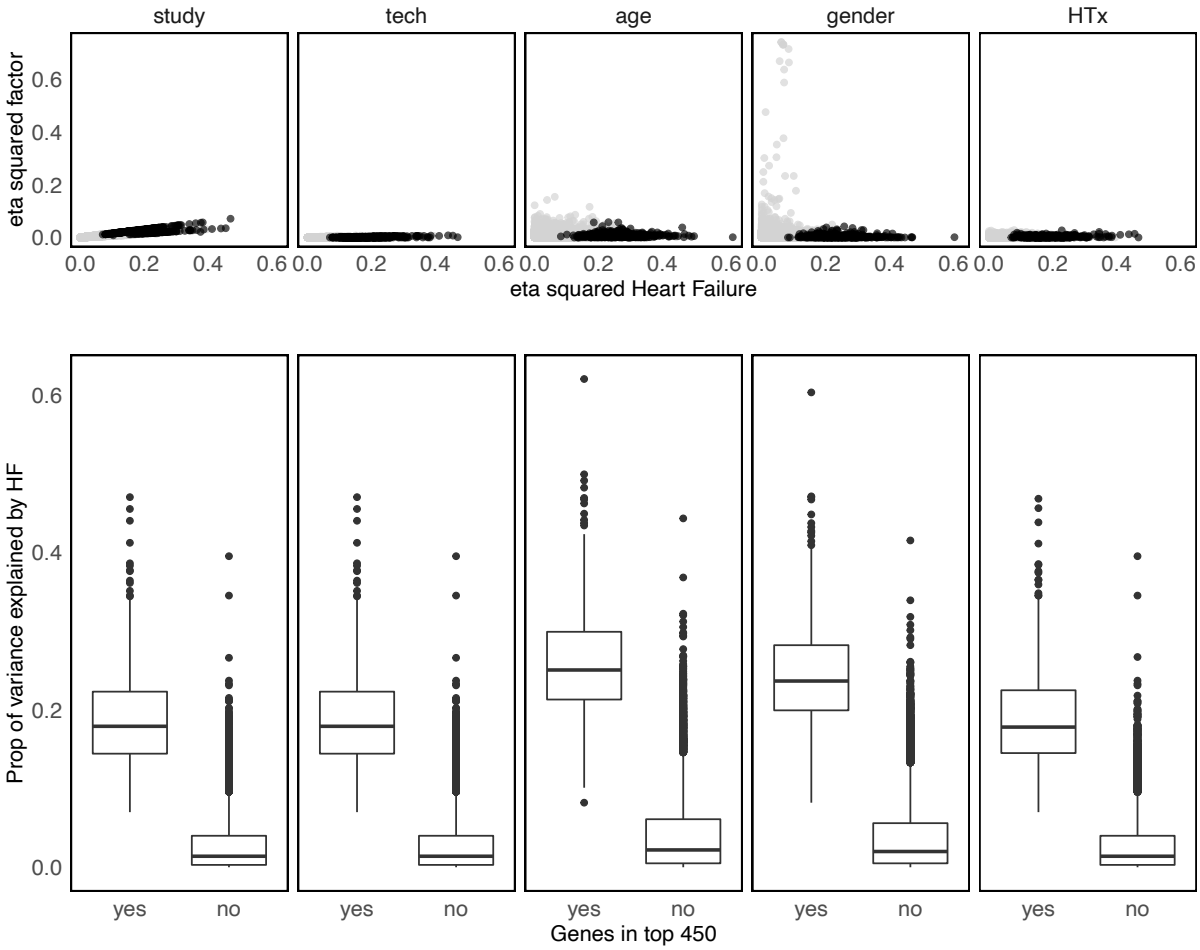

Figure CH2-14. Proportion of gene expression variance explained by heart failure (HF) and additional clinical and confounding factors.
Each vertical panel shows the results of an independent 2-way analysis of variance with HF and another clinical or technical covariate, from an integrated gene standardized data set that only included samples with available information. Upper panels show the proportion of explained variance from each factor as shown by their eta-squared values. Lower panels show the difference in the proportion of variance explained by HF between the top 500 genes of our consensus signatures and the rest. Reprinted from [112].

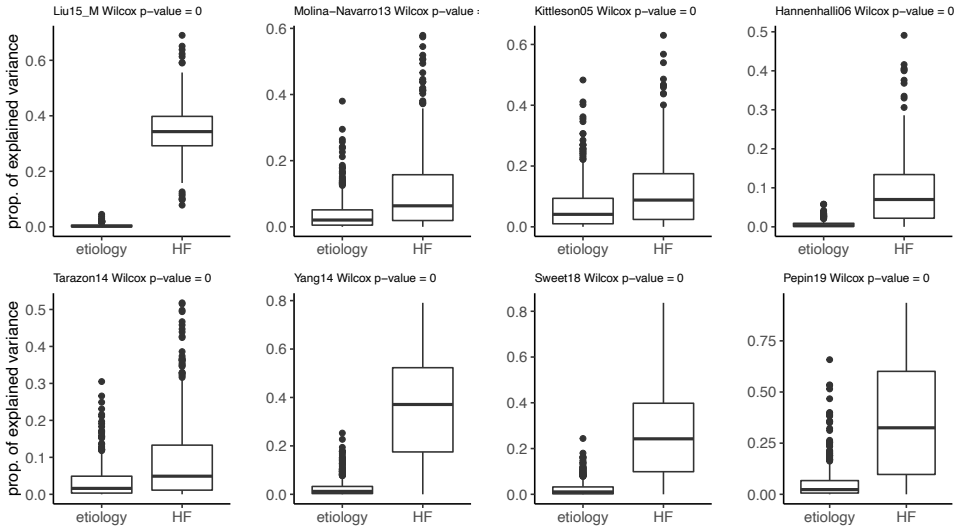

Figure CH2-15. Proportion of gene expression variance explained by heart failure (HF) and etiology (DCM [dilated cardiomyopathy] or ICM [ischemic cardiomyopathy]).
Each panel shows the results of independent 2-way ANOVAs fitted to the top 500 genes from the heart failure consensus signature with HF and DCM as covariates. Each dot represents a different gene and the y-axis is the eta-squared value of each covariate in the ANOVA model. Reprinted from [112].

2.3.1 Functional characterization of the heart failure consensus signature

Once determined that the consensus signature provides robust ranking, I functionally characterized it to identify biological processes that were consistently associated with end-stage heart failure among studies.

The *-log10(meta-analysis p-value)* of each gene was weighted by its mean direction of change in all studies to create a directed heart failure consensus signature. Gene Ontology (GO) terms, canonical and hallmark pathways from MSigDB (data downloaded in December 2019) [146] were tested for enrichment in the directed heart failure consensus signature with GSEA [144] using *fgsea* [147]. Gene sets with less than 15 or more than 300 genes were excluded from the GSEA analysis. Transcription factor and miRNA activities were estimated with *viper [148]* for human regulons obtained from DoRothEA [149] and the miRNA collection of targets from MSigDB [146], respectively. A, B, C and D regulons from DoRothEA with less than 20 genes were excluded from the viper analysis. The activity of signaling pathways was calculated with the top 100 footprint genes from PROGENy [56], [149]. Empirical p-values for PROGENy scores were calculated from pathways' null distributions calculated after permuting 1000 times the labels of the directed-meta-ranking. BH corrected p-values were calculated for each test. In

the case of MSigDB's gene sets, multiple test correction was performed to each analyzed collection (collection BH p-value) and to the union of all collections (global BH p-value).

I tested a total of 5998 gene sets of which 77 yielded an enrichment in the heart failure consensus signature (global BH, P<0.25). When each collection of gene sets was analyzed separately, 148 gene sets yielded an enrichment (collection BH, P<0.25). Top 50 results are displayed in Figure CH2-16A. Positively enriched gene sets predominantly relate to processes around the matrisome, while negatively enriched sets are associated with diverse processes many of which involve inflammation. From 343 tested transcription factors, 65 were differentially active or inactive (Figure CH2-16B, BH p-value<0.25). Among active TFs in heart failure are NANOG, SOX2, POU5F as well as MEF2C, MEF2A and MEIS2. Decreased TF activity was observed in MYNN, MTA2, ZEB2, and FOXP1. The estimated activities of TNFα, NFKB and Androgen receptor signaling were consistently reduced in the consensus signature (BH p-value<0.25, Figure CH2-16D). JAK-STAT was the only pathway from which I observed a consistent high activity (BH p-value<0.25, Figure CH2-16D). Out of 211 tested miRNA sets, 15 miRNA were enriched in the consensus signature (BH p-value <0.25) (Figure CH2-16C). Active miRNAs include *mir-137*, *mir-513*, *mir-105*, and *mir-3805P* while inactive miRNAs include *mir125* and *mir296*. Taken together, these results help us to interpret the consensus signature by characterizing the biological function of the higher ranked genes as well as their upstream regulation by pathways, transcription factors and miRNAs.

2.3.2 Leveraging the consensus signature to create insights

Generalization of the consensus transcriptional signature to other etiologies or technologies

I have shown throughout the last sections that the consensus heart failure molecular reference, created by identically processing and compiling different studies profiling diverse patient cohorts of ICM and DCM patients, captures conserved gene co-expression patterns associated to functional processes regulated by transcription factors and signaling activities.

To investigate if the genes of the consensus signature are participating in the pathological profile of heart failure due to other etiologies that are not ICM and DCM, and other technical variations, I leveraged the public heart failure transcriptome studies that did not match

inclusion criteria for the meta-analysis (Table CH2-2). I calculated the mean disease score of each sample of these excluded studies using the top 500 genes of the meta-ranking and the gene level statistics of the studies included in the meta-analysis. AUROCs were used to evaluate the ability of the disease score to differentiate between healthy and heart failure patients in each data set. I found that classification of heart failure was highly accurate in these studies as well (Figure CH2-17) suggesting that the top 500 genes still have discriminative power in those selected technical and clinical variations. Moreover, this hints that at the end-stage of heart failure the remodeled status of the heart converges regardless of the initial causes. Nevertheless, early disease molecular processes may be different across etiologies.

2.4 The heart failure consensus signature as a reference

The purpose of compiling the information of multiple patient cohorts is not solely to report the level of consistency in the knowledge generated, but also to serve as a molecular reference that can be easily used by the community to complement other studies. Thus, we propose a framework to use the generated heart failure consensus signature as a resource to build and confirm hypotheses associated with cardiovascular diseases. First, dysregulated features are identified in an independent -omics study. Next, a test for enrichment of these features is performed in the heart failure consensus signature using GSEA. Finally, highly consistent features between the study and consensus signature can be filtered by dysregulation direction and significance levels (Figure CH2-18A).

To showcase this strategy, we analyzed the plasma proteome of early and manifest heart failure patients from Egerstedt, *et al* [150] to trace their potential myocardial origin. We observed a clear enrichment of manifest heart failure proteins (GSEA p-value = 0.0001) and a modest enrichment of early heart failure proteins (GSEA p-value = 0.13) in the top of the heart failure consensus signature (Figure CH2-18B). 64 plasma proteins from manifest heart failure were part of the enrichment leading edge and agreed with the direction of transcriptional regulation (Figure CH2-18C). Candidate markers include the established heart failure marker NPPA and novel potential markers including CCDC80, BID, MAP2K1, MRC2, JAK2 and LTBP4.

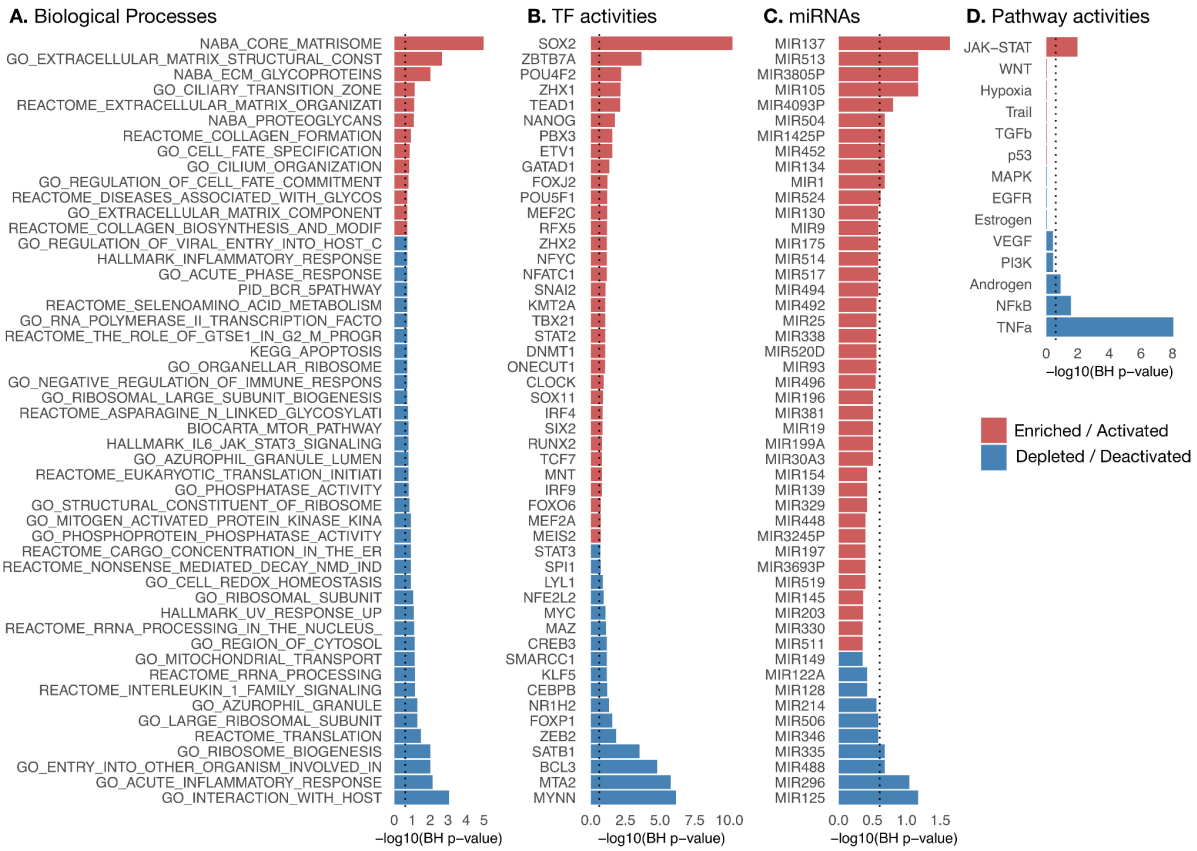

Figure CH2-16. Functional characterization of the HF-CS.
−log10 (BH P-values) coloured by direction of enrichment (A and C) or by direction of activation (B and D) of the top 50 (A) most enriched canonical and hallmark gene sets, (B) transcription factor activities, (C) miRNAs' targets, and (D) all signaling pathway activities. Dashed line indicates BH P=0.25. BH indicates Benjamini-Hochberg; HF-CS, heart failure consensus signature; and miRNA, micro RNA. Reprinted from [112].

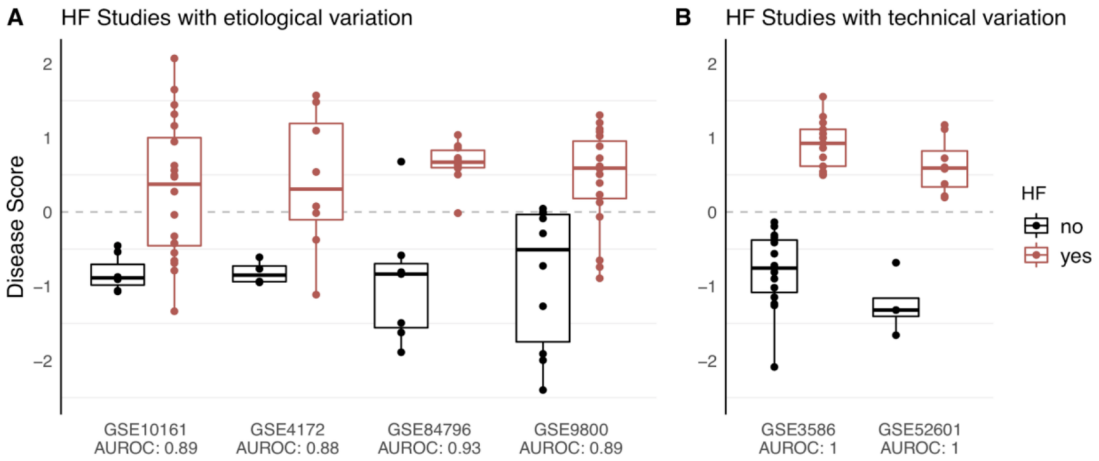

Figure CH2-17. Disease score calculation based on the top 500 genes from the consensus signature for diverse heart failure (HF) studies.
A, HF with diverse etiologies: aortic stenosis (GSE10161); PVB19 infection (GSE4172); chagas disease (GSE84796); eosinophilic myocarditis, alcoholic cardiomyopathy, hypertrophic cardiomyopathy, sarcoidosis, peripartum cardiomyopathy, ischemic cardiomyopathy (ICM) , dilatative cardiomyopathy (DCM) (GSE84796). **B,** HF studies with ICM and DCM samples but processed with different bioinformatic pipelines (GSE3586, GSE52601). Reprinted from [112].

Molecular remodeling processes in heart failure have been linked to the reactivation of fetal transcriptional responses [4]. For this reason, additionally, we tested if the consensus signature could be used to dissect the reactivation of fetal gene programs in heart failure by analyzing two public fetal cardiac transcriptomes (GSE52601, Spurrell19) and their estimated transcription factor activities (Figure CH2-18). Differential expression analysis and estimation of transcription factor activities of these two studies were performed as described before. Genes with a BH corrected p-value < 0.05 were tested for enrichment in the consensus signature and transcription factor activities with a p-value < 0.05 were compared to the ones estimated from the heart failure consensus signature. Fetal transcriptional signatures of both studies were enriched in the top rankings of the heart failure consensus signature (GSEA p-value < 0.01) (Figure CH2-18B). 221 of the top 500 genes from the consensus signature correlated with fetal genes reported by Spurrell19 (Figure CH2-18D) while 32 transcription factors correlated between the fetus heart and the heart failure consensus signature (Figure CH2-18E). Similar results were observed for GSE5260.

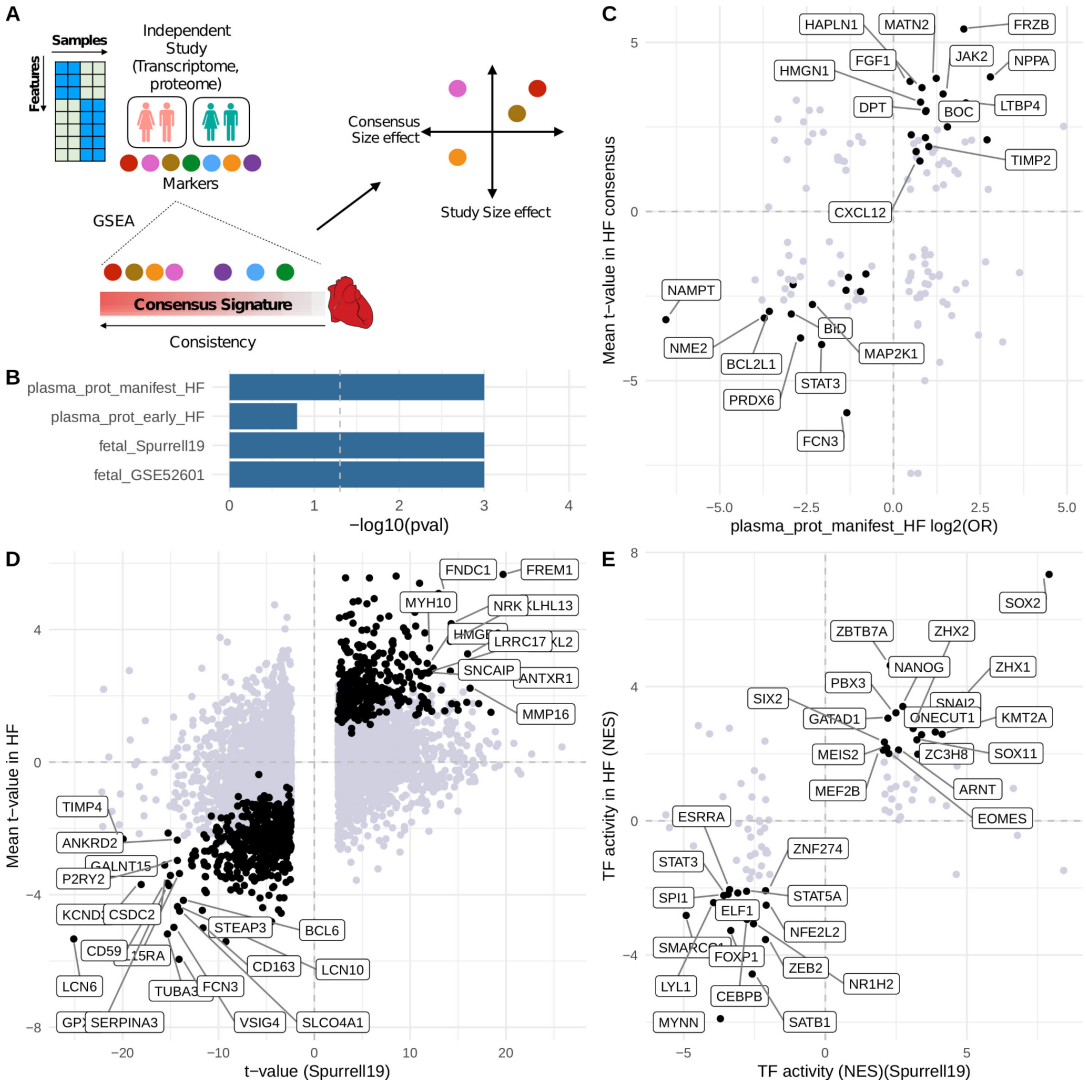

Figure CH2-18. Heart failure consensus signature (HF-CS) as a reference that complements independent studies.
A, Schematic of a suggested framework. Marker features from independent studies are enriched in the heart failure consensus signature (HF-CS) with gene set enrichment analysis (GSEA). Features that belong to the leading edge are further filtered, for example, by correlation or ranking in the HF-CS. **B**, Enrichment results of marker features from 4 individual studies. **C**, Plasma proteome of patients with HF mapped to the HF-CS. **D**, Fetal cardiac transcriptome (Spurrell19) mapped to HF-CS on gene level and (**E**) transcription factor (TF) level. Black dots in (**C** and **D**) indicate correlated features in the enrichment leading edge; labeled features in (**C** and **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. Reprinted from [112].

2.4.1 Creation of ReHeaT: The reference of the heart failure transcriptome

An important aspect of this work is that we made all the analysis and processed data public through GitHub (https://github.com/saezlab/HF_meta-analysis) and Zenodo (https://zenodo.org/record/3797044#.YwdrWexBy-a), to allow the community to re-use or expand our analysis and compilation easily. Moreover, we created an R shiny app called

ReHeaT (The reference of the heart failure transcriptome) hosted in https://saezlab.shinyapps.io/reheat/. This web service provides, for the first time, the immediate access of most of the results of the last 15 years of human heart failure bulk transcriptomics. There are five main functionalities in ReHeat:

1. **Gene query:** Assess the consistency of the deregulation in heart failure of a gene of interest in all 16 studies and explore its ranking in the consensus signature.
2. **Gene set enrichment analysis**: Assess whether a gene set of interest is enriched in the consensus signature.
3. Explore and download the **heart failure consensus signature**.
4. Explore the **functional interpretation (transcription factor and signaling pathway activities, enriched functional gene sets, and miRNA regulation)** of the consensus signature.
5. **Study overview**. Get an overview of the included studies from different perspectives.

With this effort we aimed to provide a molecular reference of human end-stage heart failure that is modular, accessible, transparent, and expandable.

2.5 Discussion and future perspectives

In this study I presented a comprehensive meta-analysis of heart failure, comparing 16 datasets and a total of 916 samples. Heart failure is known to be a complex disorder both on the clinical and genetic levels. As such, the published work in myocardial transcriptomics represents a heterogeneous picture of transcriptional regulation in the heart. In the studies included in this meta-analysis, clinical heterogeneity is compounded by wide variability in analysis pipelines, study design, tissue protocol, and patient selection. My work shows that despite these difficulties, combining these studies provides not only an opportunity to robustly evaluate their reproducibility, but to gain a more complete picture of transcriptional regulation.

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

cohort to create the heart failure consensus signature. However, a simplification of the transcriptome was necessary. I could not regard transcript isoforms nor non-coding transcripts in this analysis, since I focused on ~14.000 protein coding genes that were measured to similar extent by both technologies.

My results suggest that the magnitude of changes in mean expression of marker genes depends highly on the study. I observed a 5% of 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 less than 10 patients were still able to effectively capture similar patterns of gene deregulation as studies with more than 200 patients. This highlights the importance of representative patient sampling, since it may compensate for sample size. Moreover, I observed that consistent coexpression patterns were shared among etiologies, suggesting that conserved disease mechanisms converge in end-stage heart failure.

This study has several strengths and limitations. One strength is the added robustness to the gene dysregulation associations found in end-stage heart failure, that comes from integrating equally the evidence of a diverse collection of patient cohorts and technologies. In this meta-analysis I balanced the bias of the experimental design and the non-probabilistic sampling of each individual dataset, and increased the sample size. I reduced technical variance by standardizing the bioinformatics processing and analysis of each dataset, although reducing the number of genes that can be confidently reported. Another strength is that the estimation of transcription factor and signaling activities provides a functional catalog of new hypotheses that facilitates the interpretation of individual and disjointed gene associations with heart failure. In addition, these features are likely to be more robust than individual genes, as they integrate the information of multiple genes.

Important limitations of our study relate to the data used. In this meta-analysis, we only included public datasets that may under-represent the population of heart failure patients. A second limitation is the lack of reported demographic and clinical data in most of the collected datasets that didn’t allow us to estimate the impact of known comorbidities, drugs and genetic backgrounds in the transcriptional responses analyzed. Another important limitation of the

study, and of all studies using bulk RNA measurements, is that they do not allow the capture of cell-type specific contributions to the disease processes. Despite these limitations, I believe that the size of our meta-analysis and the novel application of transcription factor and signaling activity estimation tools provide important new information to the study of heart failure using transcriptomics. Furthermore, my methodology is general and, as more data becomes available, the meta-analysis can be expanded.

This study also highlights the importance of data sharing and reproducibility. In recent years studies that address curation, sharing, replication and comparability of experimental data sets have gained importance. Despite this movement, and the presence of data sharing platforms like ArrayExpress [151], the Gene Expression Omnibus [152], the European Nucleotide Archive [153], recount2 [154] , and BioJupies [141] since 2002, many heart failure gene expression studies conducted by the research community were not publicly available. Furthermore, many of those available studies lack sufficiently detailed clinical information of studied patients to robustly characterize disease phenotype and relate it to molecular perturbations. In the case of heart failure, the necessity of studying clinically ramified subgroups of the syndrome is becoming evident [10].

The investigation of differential gene expression between ICM and DCM samples did not yield any significant results, in contrast to reports of single studies, suggesting that myocardial remodeling in heart failure concludes in a final, molecular indistinguishable phenotype. As such, I provide a depiction of the common molecular perturbations in the failing myocardium and propose that our results reflect fundamental gene expression changes in heart failure of various etiologies and can serve as a resource for biomarker detection and hypothesis building or confirmation. For this purpose, we made our results available at https://saezlab.shinyapps.io/hgex_app. In this portal it is possible to query genes in the meta-ranking, obtain gene-level statistics coming from the 16 studies, enriched gene sets in the meta-ranking and calculate the disease score for new samples given an expression profile. With this resource we want to facilitate the collaborative crowdsourcing of the functional characterization of the transcriptional landscape of heart failure.

In summary, I demonstrated the feasibility of combining gene expression data sets from different technologies, years and centers in a biologically meaningful way. To my knowledge, this report represents the largest comprehensive meta-analysis of human heart failure gene

expression studies to date. As single cell and spatial profiles become widely available the obvious continuation of this work is to identify multi cellular gene regulation processes that explain the tissue remodeling effects associated with heart failure. For the same reason, this reference becomes essential to evaluate the translation of these findings in data with higher resolution.

Chapter 3

Dissecting the multicellular processes of human myocardial infarction with single cell and spatial transcriptomics

In this Chapter I describe the generation of the first multi-omics and multi-scale analysis of single cells together with spatial transcriptomics of human myocardial infarction. In this work I analyzed 31 specimens of 23 patients encompassing distinct physiological areas and time points after myocardial infarction and control myocardium. For a subset of 28 specimens, I present an integrated analysis of single cell and spatial transcriptomics. First, I discuss in greater detail the challenges of case-control multiscale analysis initially presented in Chapter 1. Then I describe methodological frameworks to properly integrate single cell and spatial transcriptomics data. Finally, I present the application of these frameworks to the heart specimens and discuss gained biological insights from the multi-scale analysis.

This chapter highlights the conceptual contributions that shaped the research paper peer-reviewed and published in [155]. The research paper from which this chapter is based was co-led by me, Christoph Kuppe, and Zhijian Li. Data generation and experimental work was organized by Christoph Kuppe, Rafael Kramann, and Hendrik Milting. Zhijian Li, supervised by Ivan Costa, performed the computational analysis focused on the integration of chromatin accessibility data with gene expression data that is not covered in this chapter. I conceived and implemented most of the data analysis presented in this text. Whenever I use the pronoun we, I refer to analyses that were done in collaboration with Zhijian Li or biological interpretations guided by Christoph Kuppe. Julio Saez-Rodriguez, Rafael Kramann, and Ivan Costa supervised this project.

3.1 Multi-scale analysis for the understanding of cardiac remodeling

3.1.1 Bridging the gap between tissue organization and functions with spatial transcriptomics

Spatially resolved genome-wide transcriptomes of intact tissue allow the study of gene regulation at unprecedented resolution. Emerging spatial barcoding-based technologies, such as 10x Visium, overcome the limitations of cell capture of single nucleus or single cell RNA-seq methods and increase the levels of multiplexing from immunofluorescence-based methods. In 10x Visium slides, the size of the captured area (55 μm) enables it to profile thousands of genes of up to 10 cells, creating collections of microenvironments that can be analyzed separately. Thus, in combination with single-cell profiling technologies, 10x Visium data is suitable to map single cell molecular processes in space and to study tissue function and organization at different levels of resolution.

The availability of paired single cell and spatial transcriptomics data allows to generate multi-scale molecular descriptions of the tissue architecture. On the one hand, it is possible to describe the organization of cell types in space, on the other, it is possible to assess to what extent the organization of cells is associated with their molecular variability. The inferred relationships of these two aspects of the tissue, organization and molecular variability, summarize partly the architecture of tissues.

Multi-scale molecular descriptions of tissues in disease contexts aim to contrast structural and organizational aspects of the tissue and associate them with clinical descriptions. Similarly as with digital pathology, the main objective is to identify tissue centric features that allow for better patient stratification or disease development estimations. Nevertheless, in comparison to histology based pathology only a limited set of patients can be currently profiled molecularly at multiple scales thus these technologies are used mainly to increase the resolution of the description of multicellular disease processes.

The generation of multi-scale descriptions also requires the proposal of new computational frameworks that: 1) Increase the resolution of spatial capture-based technologies, both at the compositional and functional level, 2) identify spatial patterns, 3) relate spatial variability with

cellular organization, and 4) define features that describe the spatial components of tissues to 5) perform cross-condition comparisons.

3.1.2 Multi-omic spatial map of human myocardial infarction

Cardiovascular diseases are the leading cause of mortality worldwide and myocardial infarction is the largest contributor to the number of deaths. Primary percutaneous coronary interventions and adjunctive therapies have increased patient survivability after myocardial infarction, nevertheless it is still the most common cause of heart failure. The understanding of the precise intra- and intercellular signaling mechanisms regulating cardiac remodeling processes associated with myocardial infarction can lead to better strategies to develop novel therapeutics. The role of the distinct cellular processes and tissue structures involved in the inflammatory and reparative responses following a heart attack are likely to be heterogeneous across patient groups, thus the necessity of a high resolution spatial map of distinct time points and regions of myocardial infarction patients and control myocardium [156].

To complement the work of several research groups describing the molecular diversity of the cells that conform the healthy heart [157] and distinct etiologies of heart failure [158]–[160], we combined the profiling of single nucleus gene expression (snRNA-seq), chromatin accessibility (snATAC-seq) and spatial transcriptomics (10x Visium) to study the events of cardiac tissue reorganization following myocardial infarction. In total we profiled 23 patients, encompassing 4 distinct physiological zones of myocardial infarction together with control myocardium from non-transplanted donor hearts. Myocardial infarction (MI) samples were taken from specimens of patients with acute myocardial infarction with necrotic tissue areas (ischemic zone, IZ), from the border zone (BZ), and the non-affected left ventricular myocardium (remote zone, RZ). We additionally profiled specimens of patients with ischemic heart failure that capture fibrotic zones and a later stage after myocardial infarction. This represented to our knowledge the largest multi-modal and multi-scale profiling of human myocardial infarction tissue, focused in distinct disease progression stages (Figure CH3-1). For details on the experimental protocols, tissue handling and processing, I refer to [155].

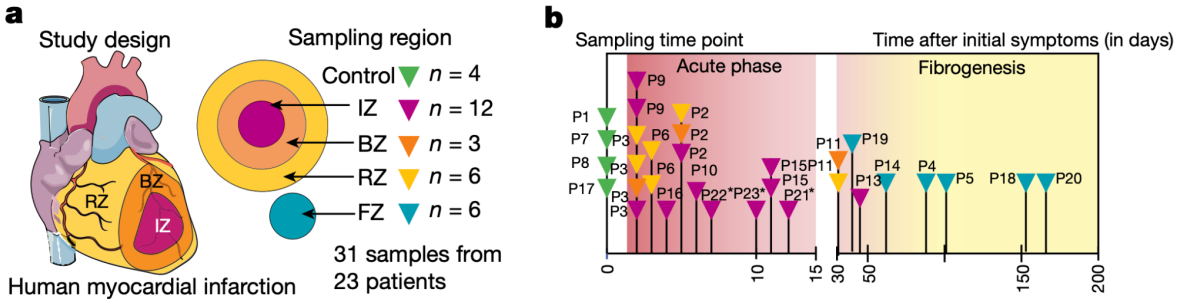

Figure CH3-1. Spatial multi-omic profiling of human myocardial infarction.
a, Schematic representation of the multi-omic human myocardial infarction atlas study design. The infarcted human heart is characterized by remote zone (RZ), border zone (BZ), ischemic zone (IZ) and fibrotic zone (FZ). The total number of samples (n) of each zone is shown. **b,** Sampling time point scheme showing time of analysis after initial onset of myocardial infarction symptoms of all samples divided into acute (day 1-15) and chronic (>day 30). * indicates snRNA-seq samples used for validation only (P21-23). Colors indicate sampling regions. Reprinted from [155]

3.2 Computational strategy to integrate multi-scale data

I devised a multi-omics and multi-scale integrative analysis spanning all scales (Figure CH3-2-3) motivated by three main objectives:

A) **Featurize the molecular and spatial components of tissues.**

As mentioned in Chapter 1, the first step of multi-scale modeling requires enumerating the distinct molecular, structural, and organizational features of the tissue. Structural features relate to compositions of cell types or cell neighborhoods (here proposed as basic building blocks of tissues). Organizational features refer to the way cell-types or neighborhoods are distributed in space (quantified as spatial variance) or the way they relate to each other in space (quantified as spatial dependencies). Molecular features refer to population level distributions of gene expression or functional activities estimated from single nucleus, single cell data or spatial locations. A second requirement for multi-scale descriptions is that the enumerated features are representative of a collection of samples. In other words, multi-scale features should be shared across samples and scales in a given cohort, and ideally in any collection of samples.

In this work I propose to define features that describe the molecular and spatial components of tissues as follows:

1. Major cell types: Usually defined with snRNA-seq data, these features refer to the identification of the major lineages of cell types that conform a profiled tissue. Each major cell type is a population of single nuclei or cells from which gene expression distributions can be estimated. Cell ontologies are usually predefined from prior knowledge (see Chapter 1).

2. Functional cell states: These features represent populations of single nuclei or cells of a given cell lineage (as defined above) with distinct functional profiles. Compared to cell types, a clear ontology of cell states can not be defined and usually these features are context and cohort specific (see Chapter 1).

3. Cell type or cell state compositions: Compositions can be directly quantified from single cell technologies or estimated from spatial transcriptomics data using deconvolution methods. Each sample is assigned a compositional vector with as many cell types or cell states as defined previously.

4. Spatially variable genes and functional processes: Gene level statistics obtained from the hypothesis test of spatially organized expression [161]. Functional enrichment can be assessed using overrepresentation analysis.

5. Cell type niches: Given that 10x Visium profiles spatially resolved mini-bulk samples, it is reasonable to assume that each spot represents a sample of a cellular neighborhood. A cellular neighborhood within a spot can be defined from the compositions of cell types estimated from deconvolution methods. Then, groups of spots with similar cellular compositions are defined as niches and represent building blocks that appear repeatedly in multiple samples and in distinct contexts.

6. Molecular niches: Similarly as cell type niches, molecular niches are groups of locations or spots in spatial transcriptomics data that share a specific function. These features complement the structural niches based on cell type compositions since they allow to differentiate locations in spatial transcriptomics with identical organizations but distinct functions. For both the

cell type and molecular niches, the definition of the number of niches is completely empirical.

7. Spatial interactions: In spatial transcriptomics for every measured feature (eg. gene expression, cell type composition, cellular function) it is possible to estimate its spatial relationship with the rest. One example of this type of feature is the correlation of the expression of 2 genes across a slide or region of interest. However, this type of interactions are limited to spatial dependencies within a spot, ignoring the whole tissue organization. Definition of spatial interactions in distinct spatial contexts have been presented in the work of [104] and extended in [105]. In the latter research paper I explored the estimation of spatial interactions between signaling pathway activities and ligand or receptor expressions, demonstrating that tissue organization is not limited to its structural characteristics, but also is reflected in the spatial distributions of distinct cellular functions.

B) **Link molecular variability and cellular functions to tissue structures**

One of the advantages of spatial transcriptomics is the possibility to contextualize expression variability of locations in terms of the organization of cells. The assumption is that cellular phenotypes and responses to stimuli (approximated from gene expression) are influenced by intrinsic and extrinsic processes with distinct importances [162]–[165]. Intrinsic processes relate to the intracellular relations (eg. dependency of signaling pathways, volume of the cell, etc.), while extrinsic processes capture intercellular dependencies and environmental effects. An interplay between the effects of the surroundings and the current cell state allows cells to process information and maintain the functions of tissues.

In this work I leveraged the methodology developed by [105] and spatial transcriptomics data to create spatially contextualized models of signaling pathway activities and cell state function. For both tasks, I evaluated the importance that the cellular composition had on the spatial distribution of signaling pathway activities and cellular states inferred from gene expression. The results of these models provide distributions of spatial relationships (see previous section) that allow to evaluate if

cellular functions are constrained to a specific “cellular layout”, or if emergent cellular functions converge in distinct contexts.

C) Contrasting the molecular and structural components of tissues

The final objective of the proposed computational strategy focuses on contrasting the distributions of the collection of the tissue features between the distinct sample groups to identify conserved disease mechanisms associated with clinical covariates. I applied mainly classic non-parametric statistical population comparison hypothesis tests. Although exhaustive, this strategy has the limitation of evaluating each feature independently, disregarding the dependencies between them.

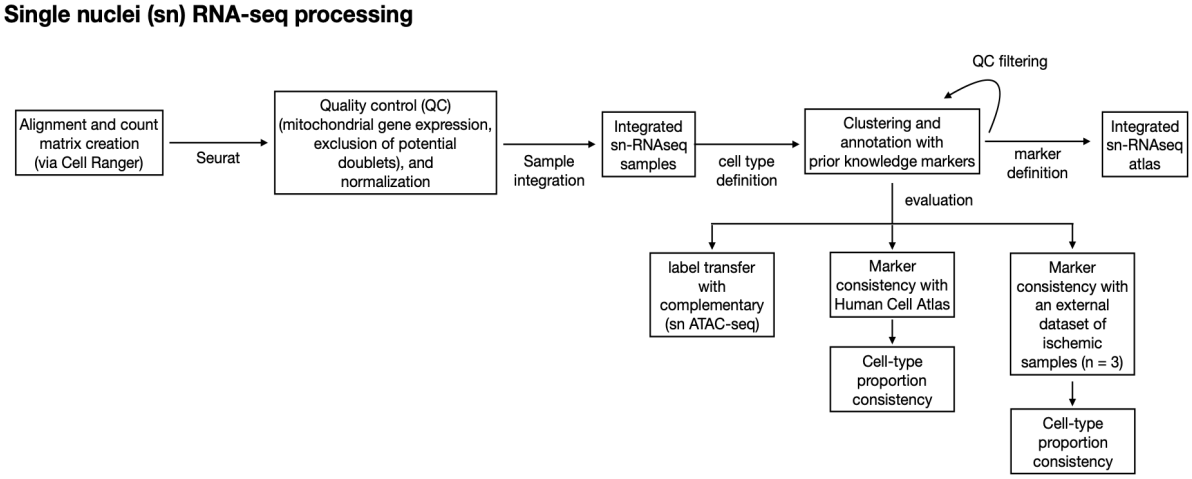

Figure CH3-2. Schematic of the methodology used to analyze single cell nuclei data.
Reprinted from [155]

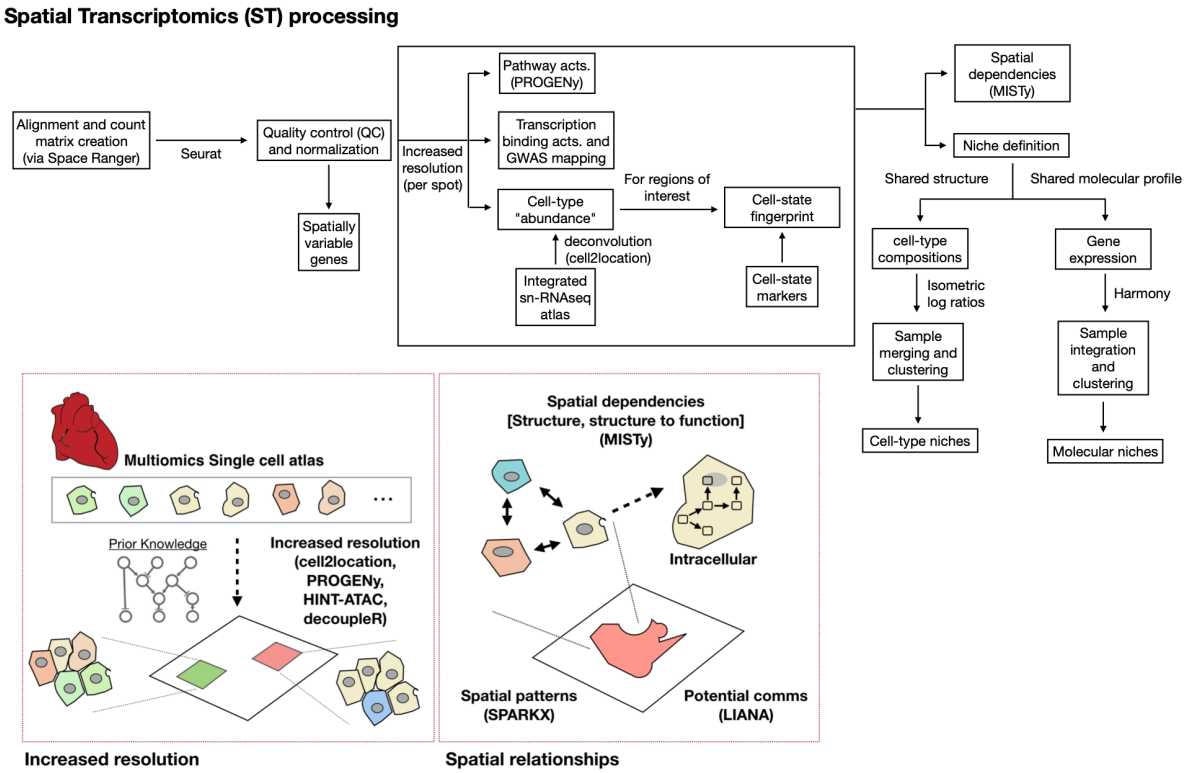

Figure CH3-3. Schematic of the methodology to analyze spatial transcriptomics in this work.
Reprinted and adapted from [155].

3.2.1 Challenges in multi-scale data analysis and integration

My proposed computational strategy was also constrained by multiple technical challenges that are described below:

1. Multi-omics and multi-scale molecular profiles acquired from distinct samples are heterogeneous in multiple levels:

Spatial and single nuclei transcriptomics data obtained from distinct regions and patients come from 10µm cryo-sections of each cardiac specimen and adjacent regions. By experimental design, single nucleus profiles only approximate the molecular profiles of spatial transcriptomics of adjacent regions. For the same reason, while it is expected to observe similar cellular compositions in single cell data, the level of mismatch is unknown.

Sample acquisition occurred in distinct tissue banks and sequencing was performed in distinct locations which adds a known technical confounder. Additionally, samples in distinct physiological zones acquired from the same patient are confounded by the individual. While

technical effects are expected to affect all samples, individual effects are limited to acute myocardial infarction samples, since they are the only type of heart specimens where distinct physiological areas can be located.

Finally, the fact that tissue selection can not be ensured to contain comparable cellular compositions and tissue structures across patients and conditions adds variability that may be greater than the one expected from the disease.

2. No common coordinate system can be defined before data sampling:

As mentioned in the last point, tissue sampling is expected to be heterogeneous at the compositional and organizational level, not only because of the diversity in clinical profiles, but also because of the selection of sampling areas guided by pathologists and clinicians. Even though morphological zones can be defined to limit the sampling area of tissues (eg. left ventricle or apex), the final captured area used to generate single cell and spatial data can not be aligned given the distinct distribution of tissue structures such as vasculature, scars, etc. This non-trivial registration challenge constrains the type of comparisons between different samples to general changes in the tissues rather than changes in exact locations.

3. Non-probabilistic sampling:

Our study characterizes rare specimens of acute myocardial infarction that are not routinely sampled in the clinic. For the same reason, experimental design is limited to tissue banks with selective bias towards specific clinical profiles and populations. This in practice affects the effective “sample size” and the statistical power.

3.2.2 Implementation of the computational strategy

The methodology used to analyze the single nuclei multi-omics and spatial transcriptomics data is described in detail below. I have omitted the computational work related to the inference of transcriptional regulatory processes based on the integration of snRNAseq and snATACseq data that was led by Zhijian Li. The text below was entirely written by me and it is taken from [155], with some necessary adaptations where Zhijian Li and Christoph Kuppe contributed. Code is available in the following repository: https://github.com/saezlab/visium_heart.

snRNA-seq data processing

To identify the major lineages representative of all of our specimens, I created a single nuclei atlas analyzing and integrating each snRNA-seq dataset obtained from the CellRanger software (v6.0.2) using Seurat (v4.0.1) [166].

Each single data set went through identical quality control processing. I discarded nuclei (i) in the top 1% in terms of the number of genes, (ii), with less than 300 genes and less than 500 UMIs, (iii) or more than 5% of mitochondrial gene expression, and (iv) doublets as estimated using *scDblFinder* (v1.4.0) [73] with default parameters. Count matrices were log normalized for downstream analyses using a scaling factor of 10,000. I calculated a dissociation score for each cell using Seurat's module score functions with a gene set provided by O'Flanagan, et. al. [71] and discarded the nuclei that belonged to the top 1%. To generate an integrated atlas of all samples, log normalized expression matrices were merged and dimensionality reduction was performed on the collection of the top 3,000 most variable genes that were shared with most of the samples using principal component analysis (PCA). To select the collection of shared variable genes between samples, first I estimated the top 3,000 most variable genes per sample and then selected the top 3,000 most recurrent genes from them across all samples. PCA correction was performed with *harmony* [84] (v1.0) using as covariates the patient, sample, and batch labels. A shared nearest neighbor (SNN) graph was built with the first 30 principal components (PCs) using Seurat's *FindNeighbors*, and the cells were clustered with a Louvain algorithm with *FindClusters.* A high resolution (1) was selected to generate a large collection of nuclei clusters to capture representative major cell lineages, even if present in low proportions. Cluster markers were identified with Wilcoxon tests as implemented in Seurat's *FindAllMarkers* function. Final assignment of cells to major cell lineages was based on literature marker genes. I filtered out small clusters (median number of nuclei across filtered clusters = 269) with low gene count distributions (median counts across filtered clusters = 756) or feature recovery (median number of genes across filtered clusters = 695), with marker genes that couldn't be assigned to known cell-types of the heart. To visualize all nuclei in a two dimensional embedding, an Uniform Manifold Approximation and Projection (UMAP) was created with Seurat's *RunUMAP* function using the first 30 principal components of harmony's PCA correction embedding.

Major cell-type markers were estimated by performing differential expression analysis of cell-type and patient-specific pseudo bulk profiles. Pseudo bulk profiles were calculated by

summing up the counts of all cells belonging to the same cell-type and patient. Profiles coming from less than 10 cells or profiles from which the maximum gene expression was of less than 1,000 counts per library were discarded. Differentially expressed genes were calculated by fitting a quasi-likelihood negative binomial generalized log-linear model as implemented in *edgeR* (v3.32.1) [47] (FDR < 0.15). Each cell type was compared against the rest.

Comparison with independent healthy and ischemic human heart cell atlases

I compared our generated atlas with another reference human single nuclei RNA-seq atlas at the molecular and compositional level. The counts matrix was downloaded directly from https://www.heartcellatlas.org and I selected the data coming from single nuclei or left-ventricle samples. Nuclei with less than 200 genes, and genes expressed in less than 3 nuclei were excluded. Log normalization with a scaling factor of 10,000 was performed with scanpy's [167] (v1.7.0) *normalize_total* function.

To evaluate our major cell-type annotation, I calculated the enrichment of the top 200 marker genes based on log fold change of each cell-type defined in the reference atlas in the list of the top 200 marker genes of each of our defined cell-types with hypergeometric tests. Marker genes of the reference atlas were calculated with Wilcoxon tests as implemented in scanpy's [167] (v1.7.0) *rank_genes_groups* (adj. p-value < 0.01). Each cell type was compared against the rest. To evaluate the compositional stability of our control samples, I calculated the Pearson correlation between the median proportion of each shared cell-type of the reference atlas and our control, border zone, and remote zone samples.

Similarly, I compared our atlas to an independent collection of human heart nuclei derived from three ischemic specimens. First, I analyzed and integrated the smaller collection of samples using identical procedures as the ones used in our provided atlas. After nuclei clustering, I assigned each cluster to a cell-type using literature markers. Cell-type markers were calculated with Wilcoxon tests (adj. p-value < 0.01) and the top 200 genes based on log fold change were selected. Marker overlap and compositional stability comparison with ischemic specimens from our atlas were performed as described previously.

Cell-type-specific transcription factor binding and regulon activity

To estimate transcription factor (TF) binding activity for each major cell type identified from snATAC-seq data, Zhijian Li generated a naive gene regulatory network by linking

transcription factor binding sites to nearest genes and estimated transcription factor binding activities using footprinting methods (see [155]). The number of ATAC-seq reads in the region with 100 base pairs up-stream and down-stream of the the TF binding site were used to indicate how strong the interaction was: each TF-gene interaction was weighted as the ratio between the number of ATAC-seq reads around the TF binding site associated with that gene and the maximum number of reads observed in any binding site of the transcription factor. All interactions with a weight larger than 0.3 were considered in downstream analysis. This generated weighted and filtered cell-type specific regulons. To infer a transcription factor regulon activity score, I estimated the mean expression of the target genes in each cell-type-specific regulon. Cell-type pseudo bulk profiles were filtered to contain only genes with at least 10 counts in 5% of the samples, before the estimation of normalized weighted means using decoupleR's [52] (v1.1.0) *wmean* function with 1000 permutations. Regulon activities were standardized and correlated with TF binding activities using Spearman correlations. The minimum correlation of 0.5 was used as threshold and the top 5 TFs per cell type were selected for visualization.

Cell-state definitions

Cell-state definition of major cell-types was done by integrating cell-type specific snRNA-seq and snATAC-seq data. For each major cell-type, Zhijian Li first transformed single nuclei chromatin accessibility data into gene scores that allowed for a diagonal integration procedure. Details can be consulted in [155].

Characterization of spatial transcriptomics data sets

Single-slide processing

Filtered feature-barcode expression matrices from *SpaceRanger* (v1.3.2) were used as initial input for the spatial transcriptomics analysis using Seurat (v4.0.1). Spots with less than 300 measured genes and less than 500 unique molecular identifiers (UMIs) were filtered out. Ribosomal and mitochondrial genes were excluded from this analysis. Individual count matrices were normalized with *sctransform* [168], and additional log normalized (size factor = 10, 000) and scaled matrices were calculated for comparative analyses using default settings.

Cell-type compositions were calculated for each spot using cell2location [98] (v0.05). Reference expression signatures of major cell-types were estimated using regularized negative binomial regressions and our integrated snRNA-seq atlas. I fitted a model in six downsampled

iterations of our snRNA-seq atlas (30%) and generated a final reference matrix by taking the mean estimation. Each slide was later deconvoluted using hierarchical bayesian models as implemented in *run_cell2location.* I provided the following hyperparameters: 8 cells per spot, 4 factors per spot, and 2 combinations per spot. Additionally, for each spot I calculated cell-type proportions using the cell-type specific abundance estimations. Cell-type compositions of the complete slide were calculated adding the estimated number of cells of each type across all spots.

To compare the stability of estimated cell compositions between our different data modalities, I calculated Spearman correlations between the estimated cell type proportions of each slide and the observed cell type proportions in its corresponding snRNA-seq and snATAC-seq dataset.

Estimation of functional information from spatial data

For each spot, I estimated signaling pathway activities with PROGENy's [56] [149] (v1.12.0) model matrix using the top 1,000 genes of each transcriptional footprint and the sctransform normalized data. Spatially variable genes were calculated with SPARKX [161] (v1.1.1) using log normalized data (FDR < 0.001). To obtain overrepresented biological processes from each list of spatially variable genes, I performed hypergeometric tests using the set of canonical pathways provided by MSigDB [146] (FDR < 0.05).

Estimation of cell-death molecular footprints from spatial data

To associate the differences in nucleus capture in snRNA-seq between the different samples to cell-death processes, I leveraged the information from spatial transcriptomics to estimate the general expression of genes associated to cell death for each sample. For each unfiltered slide I estimated per spot the normalized gene expression of BioCarta's [146] "Death Pathway" and Reactome's [169] "Regulated Necrosis Pathway" using decoupleR's (v1.1.0) *wmean* method and the sctransform normalized data. To have a final pathway score per slide, I calculated for each slide the mean "pathway expression" across all spots.

Cell-state spatial mapping

To map the functional states of each cell type into spatial locations, I leveraged the deconvolution results of each slide and the set of differentially expressed genes of each recovered cell state. Given the continuous nature of cell-states, I assumed that the collection of

up and downregulated genes of a cell-state represented its transcriptional fingerprint and could be summarized in a continuous score in locations where I could reliably identify the major cell-type from which the state was derived. For a given major cell type of interest *k*, I identified spots where its inferred abundance was of at least 10%. To estimate state scores associated with cell-type *k,* I used decoupleR's (v1.1.0) normalized weighted mean method (*wmean*) and the set of the up-regulated genes of each state defined with snRNA-seq and snATAC-seq (log fold change > 0; Wilcoxon tests, FDR < 0.05). The log fold change of each selected gene was used as the weight in the *wmean* function.

Analysis of ion channel related genes

I related the expression of ion channel related gene sets to the different cardiomyocyte cell states and their location in spatial transcriptomics. First I selected two different gene sets containing ion channel related genes: 1) Reactome's [169] "Ion Channel Transport" and a curated list of transmural ion channels from Grant et al. [170]. I calculated gene set scores for each spatial transcriptomics spot using decoupleR's wmean function. Then I correlated these gene set scores to the spatial mapping of cardiomyocyte cell-states in regions where I observed at least 10% of cardiomyocytes. Additionally, I evaluated if any of the genes belonging to these gene sets were differentially expressed between the vCM1 and failing vCM3 population using Wilcoxon tests as implemented in scran's (v1.18.5) *findMarkers* function (AUC < 0.4, AUC > 0.6, FDR < 0.05).

Spatial map of cell dependencies

I used MISTy's [105] implementation in mistyR (v1.2.1) to estimate the importance of the abundance of each major cell-type in explaining the abundance of the other major cell-types. Cell-type cell2location estimations of all slides were modeled in a multiview model using three different spatial contexts: 1) An intrinsic view that measures the relationships between the deconvolution estimations within a spot, 2) a juxta view that sums the observed deconvolution estimations of immediate neighbors (largest distance threshold = 5), and 3) a para view that weights the deconvolution estimations of more distant neighbors of each cell-type (effective radius = 15 spots). The aggregated estimated importances (median) of each view of all slides were interpreted as cell-type dependencies in different spatial contexts, such as colocalization or mutual exclusion. Nevertheless, the reported interactions don't imply any causal relation. Before aggregation, I excluded the importances of all predictors of target cell-types whose R2 was less than 10% for each slide.

To associate tissue structures with tissue functions, I fitted a MISTy model to explain the distribution of PROGENy's pathway activities standardized scores. The multi view model consisted of the following predictors: 1) An intrinsic view to model pathway crosstalk within a spot, 2) a juxta view to model pathway crosstalk between neighboring spots (largest distance threshold = 5), 3) a para view estimating pathway relations in larger tissue structures (effective radius = 15), and 4) an intrinsic view and 5) a paraview containing cell2location estimations (effective radius = 15). These last two views model explicitly the relations between cell-type compositions of spots and pathway activities. Cycling cells and TNFa were not included in the described analyses. Before aggregation, I excluded the importances of all predictors of target pathway activities whose R2 was less than 10% for each slide.

Niche definitions from spatial transcriptomics data

To identify groups of spots in the different samples that shared similar cell-type compositions, I transformed the estimated cell-type proportions of each spatial transcriptomics spot and slide into isometric log ratios (ILR) [171], and clustered spots into groups. These niches represent groups of spots that are similar in cell composition and represent potential shared structural building blocks of our different slides; I refer to these groups of spots as cell-type niches. Louvain clustering of spots was performed by first creating a shared nearest neighbor graph with *k* different number of neighbors (10, 20, 50) using scran's [172] (v1.18.5) *buildSNNGraph* function. Then, I estimated the clustering resolution that maximized the mean silhouette score of each cluster. I assigned overrepresented cell-types in each structure by comparing the distribution of cell-type compositions within a cell-type niche versus the rest using Wilcoxon tests (FDR < 0.05). I tested if a given cell-state was more representative of a cell-type niche by performing Wilcoxon tests between each niche and the rest (FDR < 0.05). Only positive state scores were considered in this analysis.

Additionally, to complement the repertoire of niches identified with cell-type compositions, I integrated and clustered the Visium spots of all slides using their gene expression. I called these clusters molecular niches. Integration and clustering of spots was performed with the same methodology as the one I used to create the snRNA-Seq atlas. A low resolution was used (0.2) to have a similar number of molecular niches as cell-type niches. Cell-type and cell-state enrichment was performed as mentioned before.

Differential expression analysis of molecular niches enriched with cardiomyocytes

Differential expression analysis between molecular niches enriched in cardiomyocytes (niche-0, niche-1, niche-3) was performed using the log-transformed expression of all spots belonging to a given niche. Wilcoxon tests were performed with scran's [172] (v1.18.5) *findMarkers* function. Genes with a summary area under the curve (AUC) > 0.55 and FDR < 0.05 were considered upregulated genes.

Differential molecular profiles of the molecular niche 10 enriched with capillary endothelial cells

Differential expression analysis between ischemic, fibrotic and myogenic enriched spatial transcriptomic spots was performed with Wilcoxon tests as implemented in scran's [172] (v1.18.5) *findMarkers* function. To obtain overrepresented biological processes from upregulated genes, I performed hypergeometric tests using the set of hallmark pathways provided by MSigDB [146]. Normalized PROGENy's pathway activities for each spot were calculated using decoupleR's *wsum* method with 100 permutations on log-transformed data. Mean normalized pathway scores were calculated per slide and comparisons between groups were performed with Wilcoxon tests. Reported p-values were adjusted for multiple testing using the Benjamini-Hochberg procedure.

General differences in tissue organization

I annotated the different spatial transcriptomic slides into three groups based on histological differences with the help of pathologists: myogenic-enriched, fibrotic-enriched, and ischemic enriched. A general comparison of the sampled patient specimens was performed at the compositional and molecular level.

Hierarchical clustering, with euclidean distances and Ward's algorithm, was used to cluster the pseudo bulk profiles of the spatial transcriptomics datasets (replicates where merged, n = 27). Genes with less than 100 counts in 85% of the sample size were excluded for this analysis. Log-normalization (scale factor = 10, 000) was performed. To visualize the general molecular differences between our samples, log normalized pseudo bulk profiles of the spatial transcriptomics datasets were projected in an UMAP embedding.

To identify compositional differences between our sample groups, I compared cell-type and niche compositions. To identify cell-type composition changes associated to the sample groups, mean cell-type compositions across single cell and spatial datasets were compared with Kruskal-Wallis tests (FDR < 0.1). Pairwise comparisons of sample groups were performed with the Wilcoxon test. Additionally, to test which cell-type and molecular niches had different distributions between our group samples, I performed Kruskal-Wallis tests over the compositions of cell-type or molecular niches (FDR < 0.1). Additional pairwise comparisons were performed with Wilcoxon tests. For this, I only consider slides where no single niche represents more than 80% of the spots. Also, I only considered niches representing more than 1% of the composition of at least 5 slides.

To identify differences between the structurally similar tissues captured in the myogenic enriched group, I separated the samples into remote, border, and control zones and repeated the niche composition comparison described previously.

To identify patterns of tissue organization associated with a sample group, I tested if differential cell dependencies were captured by the MISTy models used to predict cell-type abundances (see methods section "Spatial map of cell dependencies"). First, I filtered the standardized importance matrices of each sample's MISTy model fitted to predict major cell-type abundances to contain only the values of target cell-types predicted with an R2 greater than 0.05. Then, for each slide I created a spatial dependency vector where each element contains the importance of each possible pair of target and predicted cell-types. Finally, I tested which cell interactions had higher importances in one of the sample groups compared to the rest using Wilcoxon tests (FDR < 0.25). To prioritize interactions, I only performed pairwise comparisons between sample groups for cell-type dependencies from which the maximum median importance across all groups was greater than 0.

Estimation of the impacts of the spatial context in gene expression

I used mistyR (v1.2.1) to find the associations between the tissue organization and the spatial distribution of failing cardiomyocytes and the different endothelial, myeloid, and fibroblasts cell-states. I hypothesized that the distribution of specific cell-states in the spatial transcriptomics slides could be modeled by the cell-type composition or cell-state presence of individual spots and their neighborhood.

For a given collection of cell-states of interest, I first defined regions of interest in every single slide as the collection of spots where the inferred abundance of the cell-type from which the cell-state was derived was at least 10%. These regions limit the target spots used in the MISTy model, however the whole slide is used to spatially contextualize the predictors. I used as predictors the abundances of cell-types estimated with cell2location or cell-states scores. To account only for the effects of the activation of a cell-state, the state scores of predictor cell-states were masked to 0 whenever their score was lower than 0. In all models I included two classes of spatially contextualized predictive views: an intrinsic (intra) and a local neighborhood view (para, effective radius = 5).

Specifically I fitted the following models to answer four questions:

1) What are the main cell-types whose abundance within a spot or in the local neighborhood predict the “stressed” vCM3?

vCM3 ~ intra(cell-type abundance) + para(cell-type abundance)

2) What are the main cell-types whose abundance within a spot or in the local neighborhood predict the endothelial subtypes? How do the different subtypes relate to each other?

ECsubtypes ~ intra(ECsubtypes) + para(ECsubtypes) + intra(cell-type abundance) + para(cell-type abundance)

3) What are the myeloid cell-states within a spot or in the local neighborhood that better predict fibroblasts cell-states? How do fibroblasts cell-states relate to each other?

FibroblastStates ~ intra(FibroblastStates) + para(FibroblastStates) + intra(MyeloidStates) + para(MyeloidStates)

4) What are the main cell-types whose abundance within a spot or in the local neighborhood predict the myeloid cell-states? How do the different states relate to each other?

MyeloidStates ~ intra(MyeloidStates) + para(MyeloidStates) + intra(cell-type abundance) + para(cell-type abundance)

Specific view importances were compared between patient groups as described previously with an R2 filter of 0.1.

3.3 Single cell atlas creation and cellular mapping

We obtained 10µm cryo-sections of each cardiac specimen and isolated nuclei from the remaining specimen directly adjacent to the cryo-section with subsequent fluorescent activated nuclei sorting (FANS) for snRNA-seq and snATAC-seq. For details on the analysis of snATAC-seq data consult [155]. After filtering low-quality nuclei, I obtained in total 191,497 nuclei from all samples for snRNA-seq, with an average of 2020.543 genes per nucleus. The spatial transcriptomics datasets contained an average of 3389.519 spots per specimen and 2001.694 genes per spot with a large variability mainly due to the underlying biological process with cell-death. On average, I quantified 4.48 nuclei per spot (standard deviation = 3.24). Moreover, IZ samples demonstrated the lowest abundance in nuclei and an enriched cell death and regulated necrosis gene set enrichment hinting towards necrotic cell death particularly in this area.

I first independently established a map of human heart cell types using the snRNA-seq data. I clustered the cells based on the integrated snRNA-seq data from all samples after batch correction and I annotated the clusters with curated marker genes from literature [157], [173], [174]. In total, 11 cell types were identified: vCMs (cardiomyocytes), Fibs (fibroblasts), vSMCs (vascular smooth muscle cells), PCs (pericytes), Endos (endothelial), Adipos (adipocytes), neuronal, mast, lymphoid, myeloid, and cycling cells (Figure CH3-4a). This annotation is largely in line with the recent literature on healthy human hearts [157] and an independent dataset of ischemic hearts in terms of molecular and cellular composition. Similar procedures were done to the snATAC-seq data. To explore regulatory information provided by the snATAC-seq, we performed transcription factor (TF) footprinting analysis using cell-type-specific pseudo-bulk ATAC-seq profiles. This revealed footprinting-based TF binding activity events confirmed by expression of their target genes in snRNA-seq data (Figure CH3-4b).

I next investigated whether our spatial transcriptomics datasets reflected known biological processes of human myocardial infarction. To this end,I identified spatially variable gene expression across samples with SPARKX [161] and identified overrepresented biological processes using hypergeometric tests (Figure CH3-4c). This analysis revealed functional and organizational differences consistent with the underlying biological conditions. In the acute

myocardial infarction ischemic zone, I observed an enrichment of spatially variable gene expression associated with the innate immune system, neutrophil degranulation and programmed cell-death, and a depletion of fibrotic and muscle contraction processes. Consistently, the chronic remodeled heart (late stage after myocardial infarction) showed enrichment of spatially variable genes associated with extracellular matrix (ECM) proteoglycans, glycoproteins and other matrisome components in line with the expected fibrotic processes captured in these specimens. The borderzone specimens showed an enrichment of genes associated with mitochondrial complex I biogenesis and pyruvate metabolism/citric acid TCA cycle, both confirming the response to injury and potentially altered redox state and metabolism of this area. In the control and remote zone specimens, I observed an enrichment of spatially variable genes associated with muscle contraction linked to an overrepresentation of healthy cardiomyocytes in these samples. Overall, this analysis confirms that the spatial data clearly reflects known zones of biological processes after myocardial infarction.

Since 10x Visium profiles the transcriptome of spatially resolved mini-bulk samples of cell communities rather than single cells, I aimed to increase its resolution by estimating the cellular compositions of each spot and describing cellular processes occurring in specific locations. To estimate the abundance and location of cell-types in each slide, I deconvoluted each spatial transcriptomics spot using as reference the integrated and annotated snRNA-seq data [98] (Figure CH3-4d). The estimated cell-compositions from spatial transcriptomics of each patient generally agreed with their respective observed compositions in snRNA-seq and snATAC-seq data (median Spearman correlation = 0.891 and 0.817, respectively). To extract mechanistic knowledge from each spot I estimated signaling pathway activities from gene expression data (Figure CH3-4d). In combination, my analyses allowed us to link structural information of spatial transcriptomics to cellular functions for each slide, e.g. I observed in areas with an abundance of fibroblasts an increased activity of TGFb and in necrotic regions the location of immune cells coupled to a higher activity of NFkB.

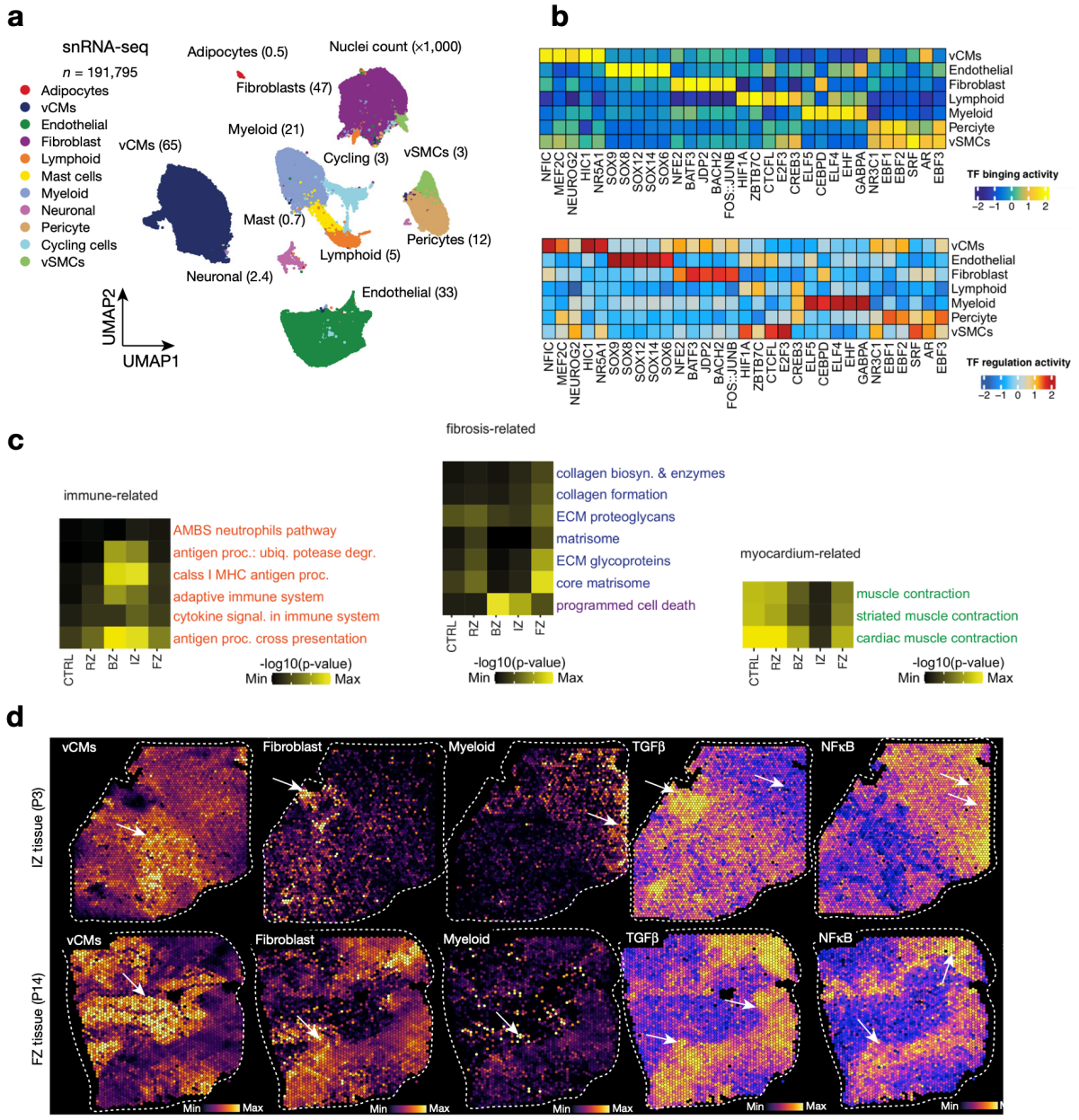

Figure CH3-4. Generation of single cell atlas and spatial mapping of processes and structures.
a, UMAP embedding of snRNA-seq data from all samples (n=191,795) including 11 major cell-types. Number of cells per cell-type (x1000). **b,** Transcription factor (TF) binding and TF target expression per major cell-type based on the snATAC- seq and snRNA-seq data. **c,** Overrepresented spatially variable biological processes (myocardium related, immune related and fibrosis related) across regions. Each cell contains the mean adjusted p-value of a hypergeometric test across all spatial transcriptomics samples belonging to each region. **d,** Representative characterisation of spatial transcriptomics data using cell-type deconvolution (cell2location) and pathway activity (PROGENy). Cell-type deconvolution and histology of these slides confirm the presence of vCMs, fibroblasts and myeloid cells (arrows). Pathway activity in these areas was enriched for TGFβ and NFκB. Reprinted and adapted from [155]

3.4 Modeling global tissue structure with spatial transcriptomics

To explore the general organization principles of the tissues I analyzed, I estimated spatial dependencies between cell-types and molecular processes using spatial transcriptomics. First,

I classified spatial transcriptomics spots from all samples into nine different classes of tissue structures based on their cell-type compositions using the deconvolution results and clustering analysis (Figure CH3-5a). I hypothesized that these tissue structures represent repetitive classes of cell communities or niches and describe in a low resolution the general structural building blocks shared between different specimens (Figure CH3-5b). To describe the recovered tissue structures, I identified the differentially abundant cell-types in each one of them. I observed that four tissue structures, niche 1, niche 7, niche 8, and niche 9 were enriched with cardiomyocytes, endothelial cells, and pericytes (Wilcoxon test, adj. p-value < 0.05, Figure CH3-5c). While these niches recovered the myogenic enriched regions of the tissues, they also contained differential compositions of fibroblasts and immune cells. Additionally, I recovered a fibrogenesis and an inflammatory niche that contained differential compositions of fibroblasts and immune cells (niche 5 and 4 respectively). Finally, I observed niches associated with the lowly abundant cells such as vSMCs (niche 3 and 6), adipocytes, immune and proliferating cells (niche 2). My results align with expected morphological features of the heart and provide a description of cellular colocalization events.

To estimate the dependencies between the abundances of cell-types in each spatial transcriptomics spot as well as their relationships in the immediate or distant neighborhood I fitted spatially contextualized models to the deconvolution scores of all spots and slides. First, I calculated the cell-type dependencies within spots (Figure CH3-5d), representing strong colocalization or mutual exclusion events. Endothelial cells were the most predictive of the abundance of vSMCs, pericytes, adipocytes, and cardiomyocytes, likely reflecting vasculature dependencies. Lymphoid and myeloid cells showed strong dependencies between them, as well as myeloid and fibroblasts, representing the inflammatory and fibrogenesis tissue structure recovered previously. Cardiomyocytes and pericytes showed strong dependencies capturing similar compositional trends as the ones observed in myogenic enriched tissue structures. Then, I calculated the cell-type dependencies occurring between immediate neighboring spots (Figure CH3-5e). Abundance dependencies were similar as the ones observed within spots, suggesting that tissue structures expand in larger areas of the tissue. By looking at immediate neighbors I captured mostly vasculature related interactions that included vSMCs, endothelial cells, pericytes, cardiomyocytes and fibroblasts. Finally, I explored cell dependencies in the distant neighborhood focused in a radius of 15 spots (Figure CH3-5e). The strongest differential interaction that I observed compared to the other descriptions was between vSMCs and endothelial cells. My different spatially contextualized descriptions estimated from spatial

transcriptomics provide a structural reference that aligns with the expected morphological organization of these specimens and demonstrate that proper integration was achieved.

Finally, I identified dependencies between signaling pathways and cell-types in space to link tissue organization to function (Figure CH3-6). Local and extended neighborhood model importances captured relationships between PI3K and p53 signaling which showed a mutually exclusive spatial distribution while both pathways are predicted by cardiomyocytes (Figure CH3-6a,c). PI3K signaling in cardiomyocytes controls the hypertrophic response to preserve cardiac functions [175], while p53 is known to act as a master regulator in cardiac homeostasis [176]. Spatial segregation of these CM-related pathways points towards functional CM heterogeneity. I observed colocalized and extended neighborhood relationships of known key pathways in fibrosis including TGFβ and NFkB predicted by fibroblasts (Figure CH3-6). Overall, CMs were the best predictor cell-types of the activities of the estimated pathways from PROGENy. Hypoxia and WNT pathway showed a stronger colocalization to CMs in ischemic specimens (Figure CH3-6) highly consistent with the cardiomyocyte differentiation events occurring after myocardial infarction [177], [178]. My results compile tissue organization principles of the human heart that relate to coordinated cellular processes and provide a basis for comparative analysis in this atlas.

3.5 Patient comparison between conditions with spatial transcriptomics

To identify general tissue differences during remodeling after myocardial infarction, I compared the specimens of the distinct regions and patients at the molecular and compositional level in different scales. First, I regrouped our samples in three major classes based on structural and histological differences: myogenic-enriched, fibrotic-enriched, and ischemic-enriched. I evaluated the extent in which the molecular information of the visium slides could recover the histological differences. I generated pseudobulk expression profiles of each Visium slide and projected them in low-dimensional UMAP space (Figure CH3-7A). I observed that samples grouped mainly by histological differences, as expected. To unbiasedly group samples, I performed hierarchical clustering of the pseudobulk profiles and observed that except from one ischemic sample, all slides of the same group clustered together. These results suggest that molecular information captured the histological differences of our samples.

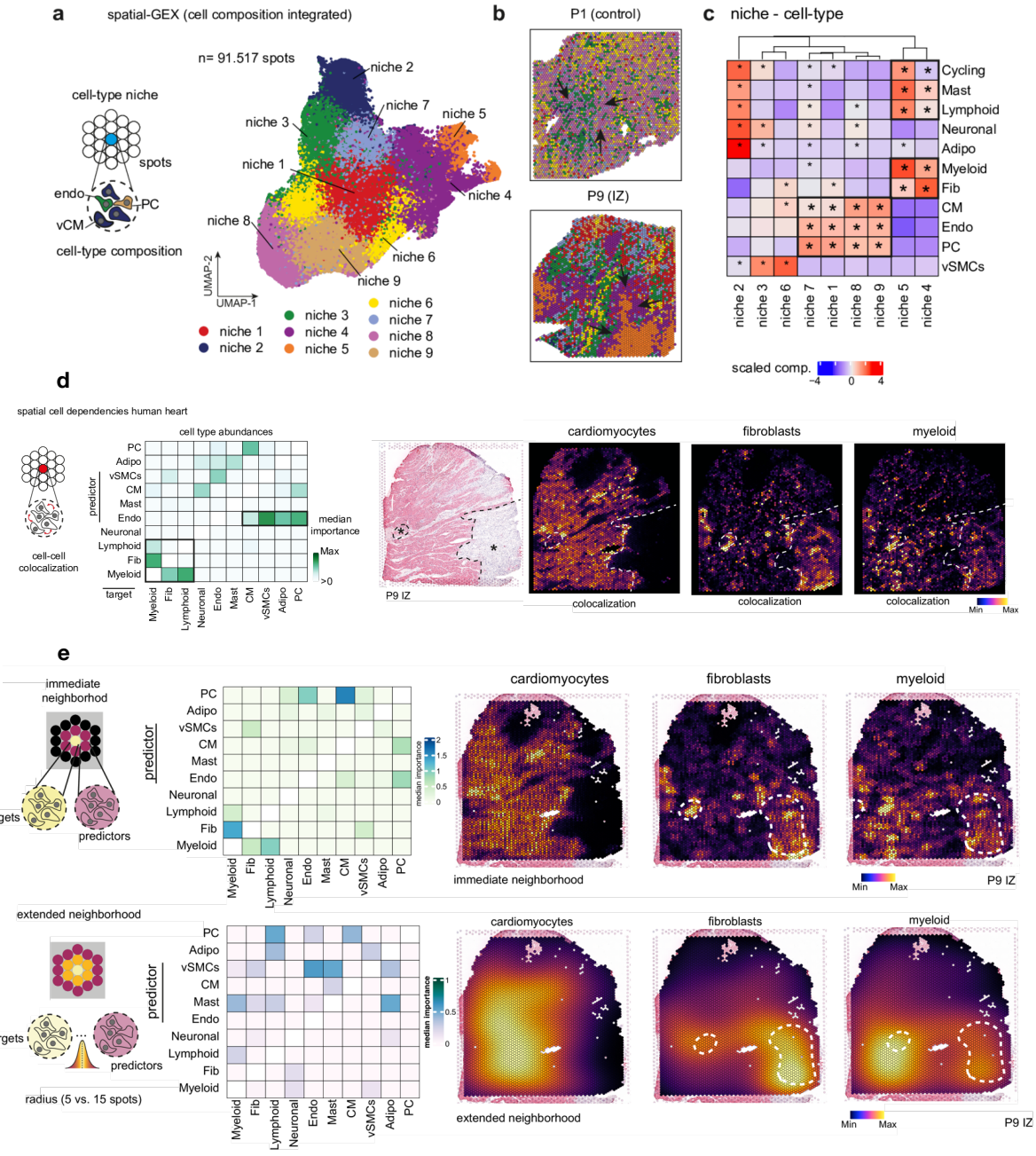

Figure CH3-5. Structural hallmarks of cardiac tissue.
a, Schematic of cell-type niche definition, UMAP embedding and clustering of spatial transcriptomics (ST) spots based on major cell-type compositions transformed into isometric log ratios. **b**, Spatial mapping of cell-type niches in a control and an ischemic (IZ) sample. Arrows show niche 8 areas in the control slide. Arrows in the IZ sample show niche 4 and how it surrounds niche 5. **c,** Scaled median cell-type compositions within each niche. *= reflects increased composition of a cell-type in a niche compared to other niches (one-sided Wilcoxon Rank Sum test, adj. p-value < 0.05). **d**, Median standardized importances (> 0) of cell-type abundances in the prediction of other cell-types within a spot inferred from spatially contextualized models. **e**, Median standardized importances (> 0) of cell-type abundances in the prediction of other cell-types within the immediate neighborhood (upper part) and the extended neighborhood (effective radius of 15 spots) (lower part) inferred from spatially contextualized models. Cell-type abundances of the immediate (upper panels) and extended neighborhood of cardiomyocytes, fibroblasts and myeloid cells. Reprinted and adapted from [155]

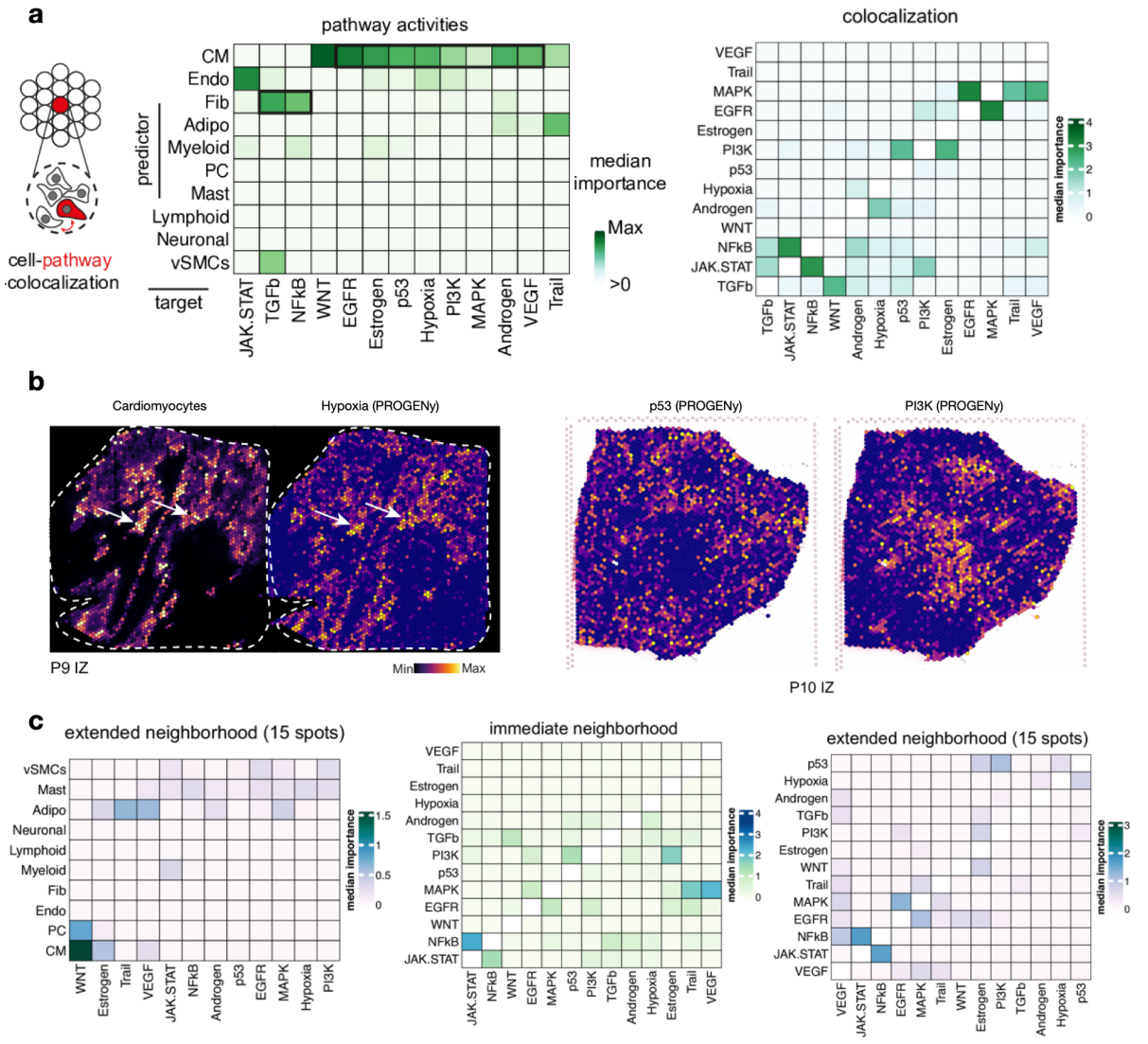

Figure CH3-6. Links between structures and functions.
a, Median standardized importances (> 0) of cell-type abundances (left) and pathway activities (right) in the prediction of signaling pathway activities within a spot inferred from spatially contextualized models. **b**, Spatial distribution of cardiomyocytes and hypoxia, p53 and PI3K signaling pathway activities in ischemic tissue slides. **c**, Median standardized importances (> 0) of cell-type abundances in the extended neighborhood (effective radius = 15), and PROGENy pathway activities in the immediate or extended neighborhood on the prediction of pathway activities inferred from spatially contextualized models. Reprinted and adapted from [155].

Since the pseudobulk profile of each Visium slide summarizes the information of many cell-types, I then evaluated to what extent the differences in cellular composition explained the observed molecular and histological differences (Figure CH3-7B). I estimated for each multi-omic sample a mean composition of all major cell-types and identified that the distributions of compositions of eight cell-types were different between patient groups (Kruskall-Wallis test adj-p < 0.10). Cardiomyocytes were the most abundant cells in myogenic-enriched samples, while fibroblasts and vSMCs were more representative in fibrotic-enriched samples. Ischemic-enriched samples showed a larger composition of myeloid, lymphoid and cell-cycling cells,

with low proportions of cardiomyocytes representing the events expected after myocardial infarction. My comparisons of cell-compositions across multiple modalities aligned with the structural differences observed in histology.

Given the observed differences in cellular compositions, I then investigated if these compositional changes were also reflected on how the tissue organized (Figure CH3-7C). I leveraged the spatial information of Visium and compared the cell-type spatial dependencies of each slide estimated with MISTy. For each slide, I concatenated the standardized importances of each spatially contextualized model to create a cell-type spatial dependency signature. UMAP projection of cell dependencies in all spatial contexts showed that in general, samples were distributed based on their histological group. Hierarchical clustering of model importances showed similar results. To identify specific cell-type dependencies that change between the different patient groups, I made pairwise comparisons of the distribution of importances of each pair of cell-types using Wilcoxon tests. The largest differences between myogenic and fibrotic enriched groups were observed between the dependencies that vSMCs and pericytes had with other cells. I identified differential dependencies between vSMCs and fibroblasts cells, both in the immediate and larger neighborhood, and between cardiomyocytes and pericytes, within the same spot as well as in the larger spatial context. These observations align with the compositional analysis, where I observed slight enrichment of vSMCs and depletion of pericytes in the fibrotic group.

Similarly, observed differences between ischemic and myogenic enriched groups related to endothelial cells, vSMCs, pericytes and cardiomyocytes. I observed changes in dependencies between myeloid and lymphoid cells in the immediate neighborhood, and between fibroblasts and myeloid cells within a spot, which could be associated with immediate inflammatory and fibrotic responses after myocardial infarction. Aligned with my compositional analysis, differential interactions between cardiomyocytes and endothelial cells were observed. Differences between ischemic and fibrotic enriched samples included interactions with pericytes, cardiomyocytes, endothelial and neuronal cells. My analysis showed that cell-dependencies within a histologically defined group are stable and are related to changes in abundance of cell-types.

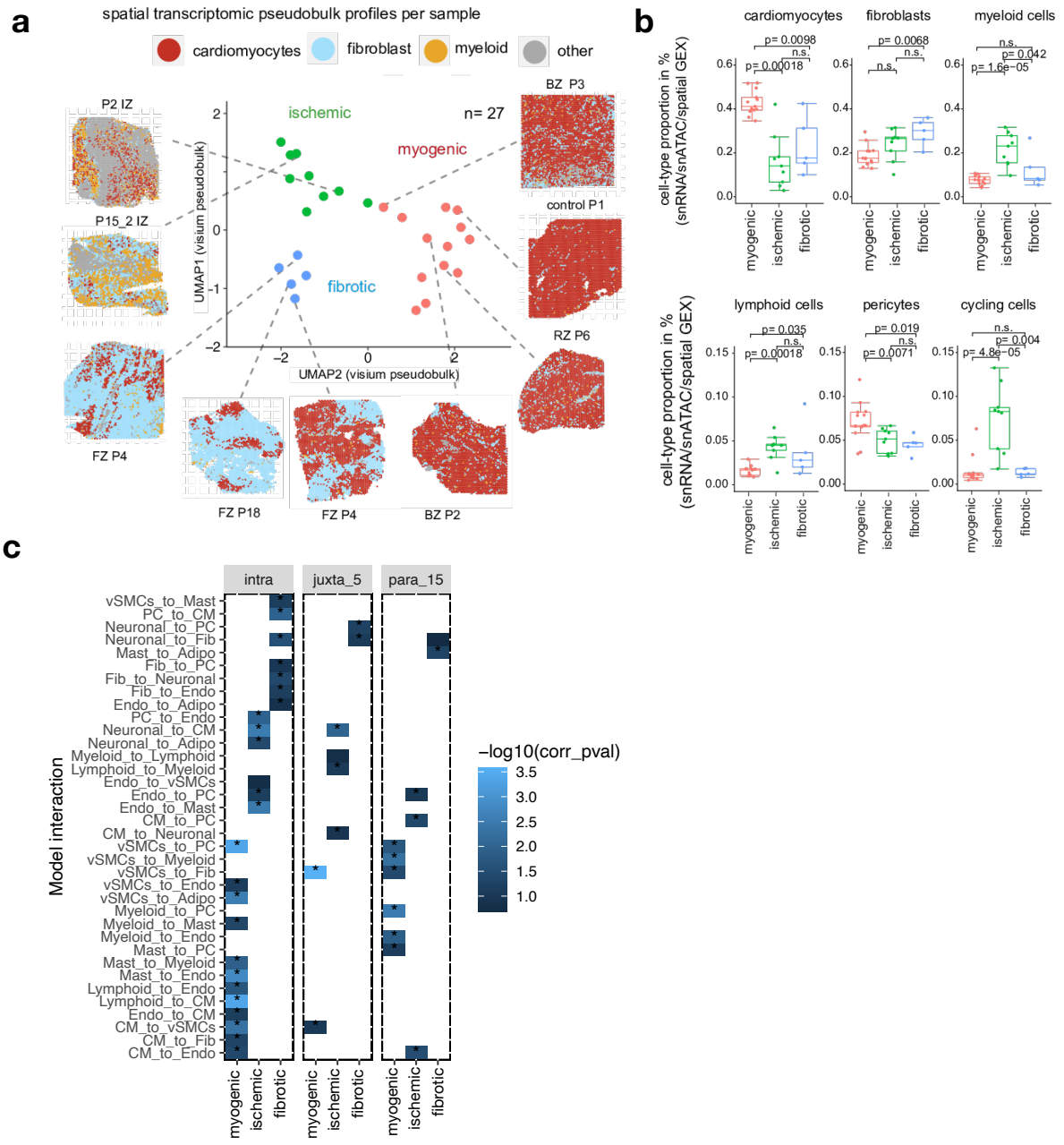

Figure CH3-7. Sample differences at the molecular, compositional and organizational levels.
a, UMAP embedding of the pseudobulk transcriptional profile of all patient samples from spatial transcriptomics (ST) and visualization of major cell-type abundance in three predefined groups; myogenic group (control, RZ, BZ), ischemic group (IZ) and fibrotic group (FZ). Each dot represents a ST patient sample. In the ST slides each spot was labeled by the cell-type with the largest proportion. **b**, Comparisons of patients mean major cell-type proportions (inferred from snRNA-seq, snATAC-seq, and ST data) between myogenic, ischemic and fibrotic groups. P-values are estimated by Wilcoxon rank sum test (unpaired, two-sided). Each dot represents a patient sample (n=13 for myogenic group, n=9 for ischemic group, n=5 for fibrotic group). **c**, Pairwise comparison between patient groups (one vs rest) of the standardized importances of cell-type abundances within the same spot (intra), and in the immediate (juxta) or extended neighborhood (para) to predict other cell-types' abundances (Wilcoxon rank sum test, testing for greater values as alternative hypothesis). * = adj. p-value <= 0.15. Reprinted and adapted from [155].

I hypothesized that the differences in cell composition and organization observed between the samples could be also represented as changes in major tissue structures or cell communities. To identify changes in abundance of tissue structures between our different samples, I tested if there were differences in the compositions of the cellular niches that I found by clustering the Visium spots of all of the samples. Principal component analysis and hierarchical clustering of the isometric log ratio-transformed compositions of cellular niches showed that most of the samples grouped based on their histological groups (Figure CH3-8a). In six out of nine niches I observed differences in compositions of niches (Kruskal-Wallis test, adj. p-value < 0.1). Niche-4 mainly associated with fibrotic structures was observed in higher proportions in fibrotic-enriched slides, while niche-5 associated with pre-fibrotic structures was mainly present in ischemic-slides (Figure CH3-8b). These results agree with the differences in dependencies between myeloid and Fib cells described previously. Niche-8 and niche-9, mostly representing muscle structures, were more present in myogenic and fibrotic enriched samples compared to ischemic samples (Figure CH3-8b). vSMCs-rich niche-6 showed similar patterns. My results suggest that changes in high-level tissue structures are associated with the molecular and compositional differences of the profiled samples.

To fully exploit the molecular resolution of spatial transcriptomics I assessed if I was able to capture tissue structures that couldn't be captured by histology alone and were different between sample groups. I separated myogenic-enriched samples into three groups based on the region and patient condition (CTRL, BZ, and RZ). These tissues, as previously described, share similar cellular compositions and organization in general, however given their origin sampling site may capture distinct molecular patterns, specially given their interaction with ischemic regions. I performed pairwise comparisons of changes in compositions of cellular niches and observed no clear difference between niches (Wilcoxon test, adj. p-value < 0.1). A subtle increase in proportions of niche-1 and niche-7 were observed in RZs. These niches represent muscular structures, but have an overrepresentation of the stressed CM phenotype. Similarly, a subtle increase of proportions of the fibrotic niche-4 were observed in the RZs (Figure CH3-8c). To circumvent the limitations of summarizing the molecular profile of Visium spots with cell-type compositions and complement the catalog of tissue structures, I generated a set of molecular niches based on the integration and clustering of Visium spots using gene expression (Figure CH3-9a). This data representation could allow us to identify molecular differences of structurally similar tissue structures with a better definition. The compositions of the 12 new molecular niches were able to separate the ischemic-enriched samples from the rest, however

differences between fibrotic and myogenic enriched samples were not captured based on hierarchical clustering and principal component analysis (Figure CH3-9b). Molecular niches capturing muscle and myeloid-rich structures showed compositional differences across conditions (Kruskal-Wallis test, adj $p < 0.1$). I observed molecular niches that were able to differentiate samples within the myogenic-enriched group. Niche-1 and niche-2 both represent muscle-rich structures, but the latter contains an enrichment of stressed-CM phenotype markers. I observed an enrichment of niche-2 and a depletion of niche-1 in BZs compared to controls, suggesting that differences in cell-type phenotypes are also relevant to differentiate the sample groups (Figure CH3-9c).

Finally, I compared the molecular profiles of the major cell-types between the different histological groups. I calculated pseudobulk expression profiles of each combination of cell-types and samples from the snRNAseq data. Then, for each cell-type, I calculated sample distances using modified Jensen-Shannon divergences. A full sample divergence matrix was computed by summing all cell-type specific sample divergences. Samples grouped mainly by histological group in an multidimensional scaling projection suggesting that cell-type specific molecular differences are also associated with tissue organization and composition (Figure CH3-9d). In summary, my analyses highlight that specimens of similar histological groups share a stable composition of tissue structures that impact the expression profile of specific cell-types. Moreover, I demonstrated that spatial transcriptomics can help to identify molecular differences in tissues with similar cell-type compositions or structures such as the distinch fibrotic processes in ischemic and fibrotic samples, or the cardiomyocyte phenotypes in the RZ and BZ.

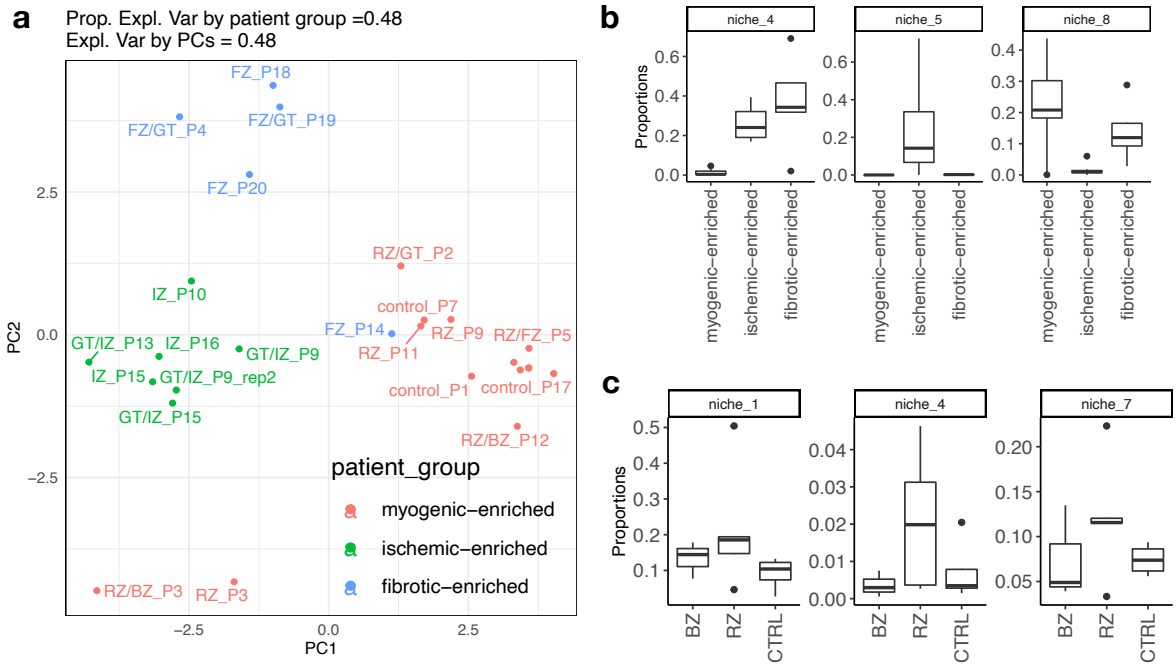

Figure CH3-8. Sample differences using compositional niches.
a, Principal component analysis of the isometric log ratio compositions of cell-type based niches. **b,** Distribution of the compositions of the niches 4, 5 and 8 in distinct patient groups. **c,** Distribution of the compositions of the niches 1, 4 and 7 in distinct groups of myogenic enriched samples. Reprinted and adapted from [155].

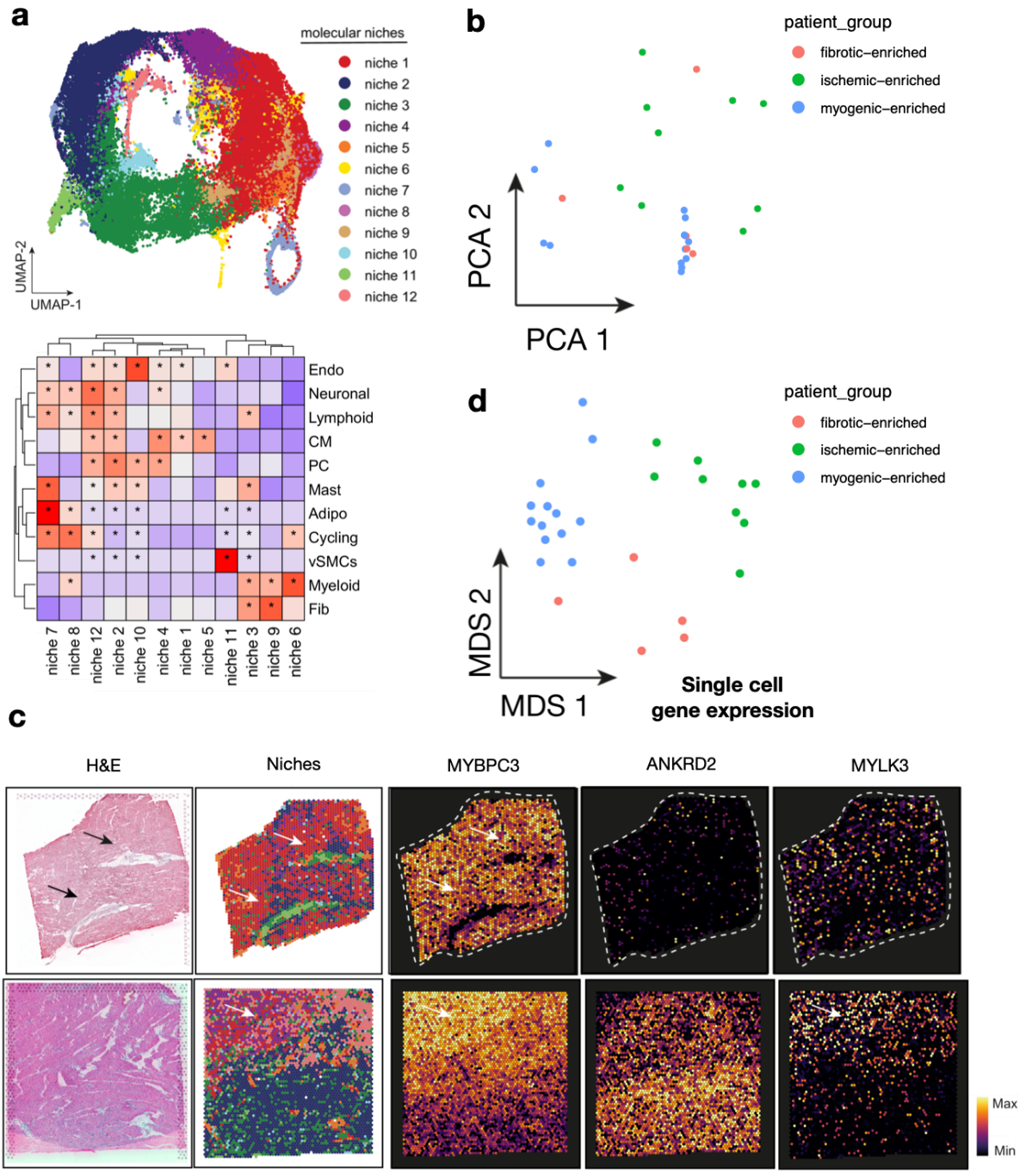

Figure CH3-9. Sample differences using molecular niches and single cell gene expression.
a, UMAP embedding and clustering of spatial transcriptomics (ST) spots based on log normalized gene expression. Scaled median cell-type compositions within each molecular niche. *= reflect increased compositions of a cell-type in a niche compared to other niches (one-sided Wilcoxon rank sum test, adj. p-value < 0.05). **b,** Principal component analysis of the isometric log ratio compositions of molecular niches. **c,** H&E, visualization of molecular niches 1,2, and 4 and gene expression (*MYBPC3*, *ANKRD2*, and *MYLK3*) of a control slide (upper) and a border zone (lower). Note that molecular niche 2 (blue), is present in both the control and borderzone slide, however an increased abundance in the borderzone is observed in the lower part of the slide. **d,** Multidimensional scaling calculated from the molecular distances of samples across cell-types. Reprinted and adapted from [155].

3.5 Spatial contextualization of cell-states

In this section I will describe the analyses I performed to estimate the influence that the tissue structure and organization had in explaining the spatial distribution and variance of cell-states of particular importance in myocardial infarction. The text was written by me and adapted from [155] when neccessary.

3.5.1 Cardiomyocytes

To further investigate distinct cardiomyocyte (CM) states, we aimed to understand the molecular heterogeneity of cardiomyocytes after myocardial infarction (Figure CH3-9). Zhijian Li co-embedded the snRNA-seq and snATAC-seq data from cardiomyocytes into a common low-dimensional space and clustered the cells. This uncovered five cell-states of cardiomyocytes (vCM1-5), spanning multiple samples and modalities. Molecular differences between the distinct clusters revealed that CM states represent distinct cellular stress states within the acute myocardial infarction phase. (i.e., vCM1; “non-stressed”, vCM2 “pre-stressed” and “stressed” vCM3) (Figure CH3-10a,b). For details consult [155].

I next estimated the cell dependencies of the stressed cardiomyocyte state (vCM3) with other cell types within each spatial spot and its local neighborhood (radius of 5 spots) between sample groups (Figure CH3-10c). I observed that the importance of vSMCs in predicting vCM3 within a spot was higher in myogenic and ischemic samples, while the importance of fibroblasts and myeloid cells increased in fibrotic samples (Figure CH3-10c-e). The local neighborhood modeling of vCM3 revealed that the abundance of fibroblasts better explained vCM3 in myogenic enriched samples compared to fibrotic samples. Overall, this demonstrates that the “stressed” CM-state vCM3 occurs in the perivascular niche of larger blood vessels, highlighting the interaction of mesenchymal cells of the perivascular niche with stressed cardiomyocytes in this tissue area. Furthermore, I noticed that when comparing RZ with control samples, stressed vCM3s are best predicted by myeloid cells (Figure CH3-10f). This underlines the importance of immune-CM interactions that could additionally explain the increased arrhythmia susceptibility in the remote regions of the post-infarct heart, since it has been shown that cardiac macrophages influence normal and aberrant cardiac conduction. My results showed that the “stressed”-CM-vCM3 can be found in distinct spatial cell-type neighborhoods enriched by different compositions of vSMCs, fibroblasts, adipocytes or myeloid cells.

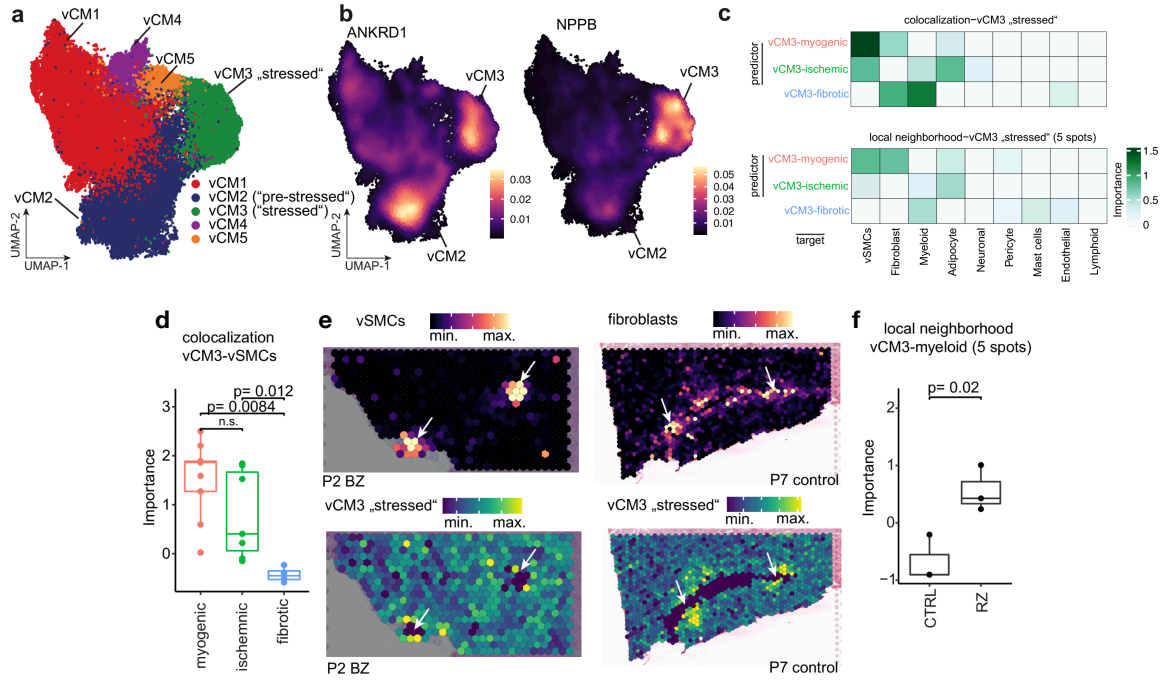

Figure CH3-10. Characterization of cardiomyocyte states.
a, UMAP embedding of sub-clusters of human cardiomyocytes using integrated snATAC-seq and snRNA-seq data. Colors refer to different CM-states. **b,** Gene expression of *ANKRD1* and *NPPB*. Colors refer to gene-weighted kernel density as estimated by using R package *Nebulosa*. **c,** Mean standardized importances of the abundances of major cell-types within a spot and the local neighborhood (effective radius of 5 spots) in the prediction of vCM3 in spatial transcriptomics. **d**, Comparison of standardized importances of vSMCs' abundances within a spot to predict vCM3 between myogenic, ischemic, and fibrotic groups (two-sided Wilcoxon rank sum test, Box-Whisker plots showing median and IQR. Maximum and minimum values as described previously). Each dot represents a sample (n = 9 for myogenic group, n = 7 for ischemic group, n = 4 for fibrotic group). **e,** Visualization examples of the abundances of vSMCs or fibroblasts and vCM3 state scores in a BZ (left) and a control human heart slide (right). **f,** Comparison of standardized importances of Myeloid cells' abundances in the local neighborhood to predict vCM3 between control and remote zone samples (two-sided t-test, Box-Whisker plots showing median and IQR. Maximum and minimum values as described previously). Each dot represents a sample (n = 3 for controls, n = 3 for remote zones). BZ: border zone, RZ: remote zone. Reprinted and adapted from [155].

3.5.1 Endothelial cells

Co-embedding of snRNA- and snATAC-seq data identified five subtypes of endothelial cells from all major vascular beds, namely capillary endothelial cells, arterial endothelial cells, venous endothelial cells, lymphatic and endocardial endothelial cells (Figure CH3-11a,b). I modeled the association of the different endothelial cell subtypes with the abundances of the other major cell-types in spatial transcriptomics (Figure CH3-11c). I observed that the markers of arterial endothelial cells were best predicted by vSMCs within a spot and in the local neighborhood (radius of 5 spots) reflecting the anatomy of arterioles in the heart. Moreover, the expression of markers of capillary endothelial cells were best predicted by the presence of pericytes in the tissue in line with the known presence and role of pericytes in direct contact to

capillary endothelium (Figure CH3-11d). The rest of the endothelial sub-types were mainly predicted by the presence of fibroblasts within a spot and in the local neighborhood. Additionally, I observed that the abundance of myeloid cells correlated with the expression of markers of lymphatic endothelial cells.

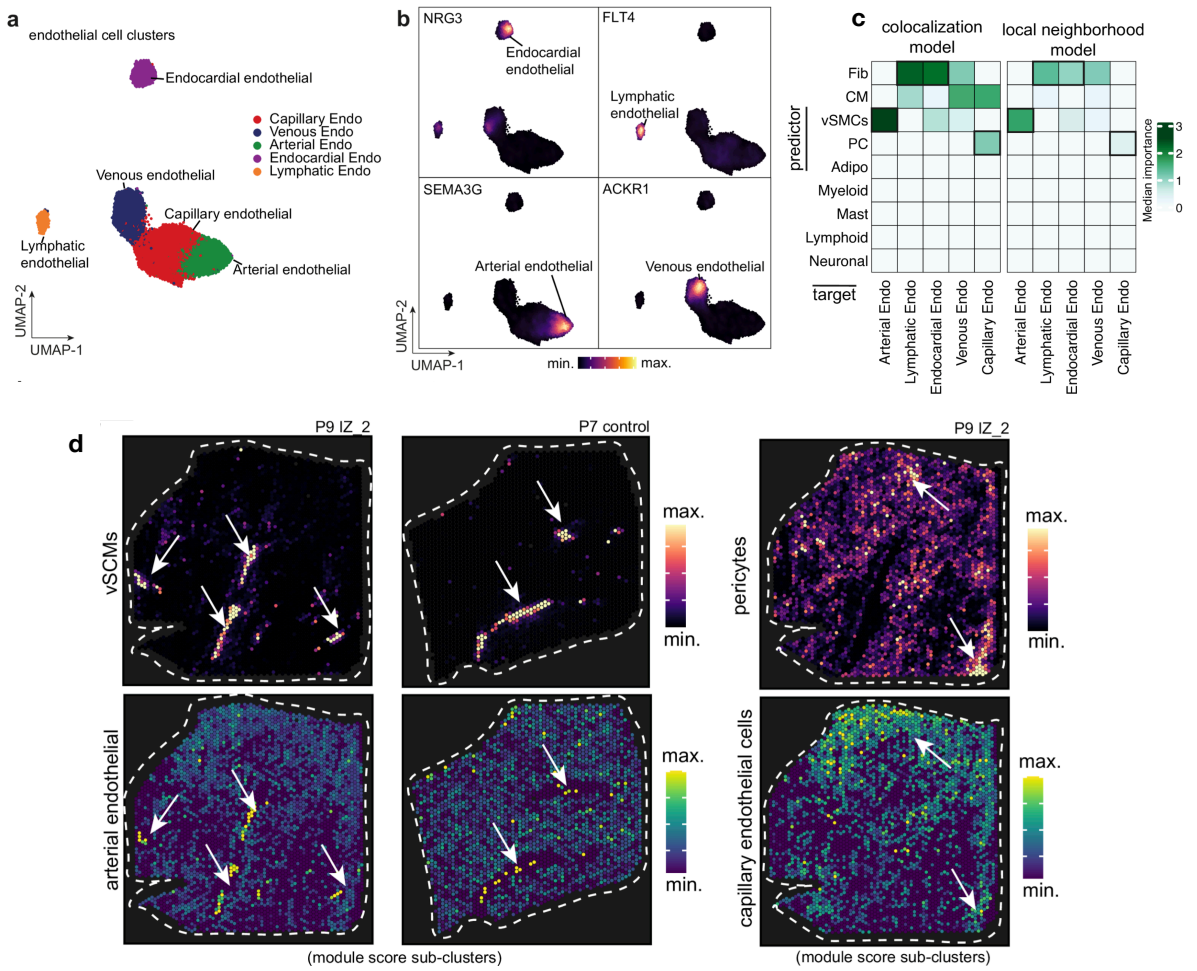

Figure CH3-11. Characterization of endothelial states.
a, UMAP embedding of sub-clusters of human endothelial cells using the integrated snRNA-seq and snATAC-seq data. Colors refer to different endothelial cell populations. **b,** Gene expression of *NRG3*, *FLT4*, *SEMA3G* and *ACKR1*. Colors refer to gene-weighted kernel density as estimated by using R package *Nebulosa*. c, Median standardized importances (>0) of the abundances of major cell-types within a spot (left) and the local neighborhood (effective radius of 5 spots) (right) in the prediction of endothelial cell-state scores in spatial transcriptomics. **d**, Visualization of the spatial distribution of the abundances of vSMCs and the state score of arterial endothelial cells in an ischemic (left) and control (right) sample. Arrows point at colocalization events. Visualization of the spatial distribution of the abundances of pericytes and the state score of capillary endothelial cells sample. Arrows point at colocalization events. Reprinted and adapted from [155].

3.5.1 Fibroblasts and myeloid cells

To dissect molecular and cellular mechanisms of fibrogenesis in the human heart, we clustered all fibroblasts using the integrated snRNA-seq and snATAC-seq data and identified four sub-

clusters (Fib1-4). Molecular characterization of these clusters revealed a functional continuum capturing the increased expression of extracellular matrix related genes which associates with myofibroblast differentiation (Figure CH3-12a-b). Myeloid-derived cells have been reported to be key players in cardiac remodeling following myocardial infarction. To understand their heterogeneity, we sub-clustered them using the multiomic data and identified five sub-clusters across all myocardial infarction samples (Figure CH3-12c). We observed that two clusters showed marker expression of resident myeloid cells (LYVE+ and FOLR+ expressing myeloid cluster), as well as a CCL18 and SPP1 expressing macrophage cluster and a monocyte/cDC cluster (Figure CH3-12d). To further gain insights about the spatial dependencies of the myeloid and fibroblasts states, I modeled their marker expression using the spatial transcriptomics data. I observed that the presence of SPP1+ macrophages better predicted all fibroblasts states compared to other myeloid cell states, with a higher importance for myofibroblasts within a spot and in the local neighborhood. Myofibroblasts marker expression aligned with a gradient of expression of the markers of SPP1+ macrophages (Figure CH3-12e-f). This pattern was also recovered by our cell-type niche definition, where the inflammatory niche 5 was surrounded by the fibrotic-rich niche 4, which I could confirm by a higher expression of SPP1+ macrophages and myofibroblast marker genes in niche 5 compared to niche 4.

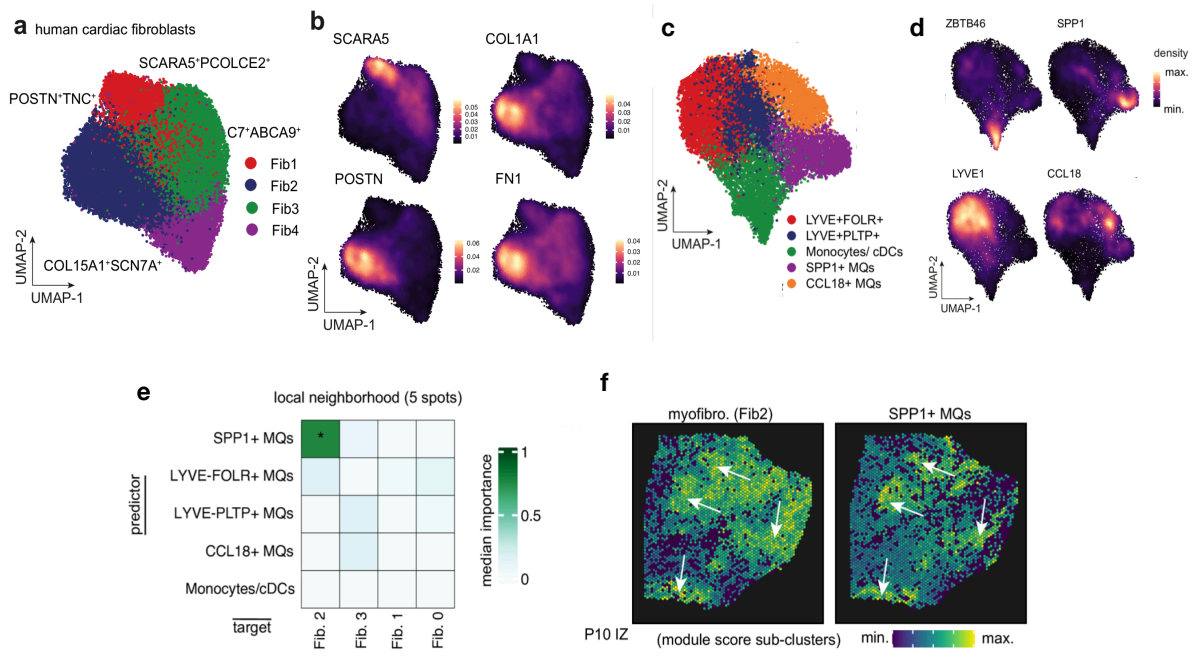

Figure CH3-12. Characterization of fibroblast and myeloid states.
a, UMAP embedding of sub-clusters of human cardiac fibroblasts using the integrated snRNA-seq and snATAC-seq. Colors refer to different populations. b, Gene expression of *SCARA5*, *COL1A1*, *POSTN*, and *FN1*. Colors refer to gene-weighted kernel density as estimated by using R package *Nebulosa*. c, UMAP embedding of sub-

clusters of human cardiac myeloid cells using the integrated snRNA-seq and snATAC-seq. Colors refer to different populations. d, Gene expression of *LYVE1*, *CCL18*, *ZBTB46*, and *SPP1.* Colors refer to gene-weighted kernel density as estimated by using R package *Nebulosa*. e, Median standardized importances (>0) of the cell-state scores of myeloid cells in the local neighborhood (effective radius of 5 spots) in the prediction of fibroblast cell-state scores in spatial transcriptomics. f, Visualisation of cell-state scores of myofibroblasts (Fib2) and SPP1+ MQs in a RZ sample. Arrows point to regions where there's an observed colocalization. Reprinted and adapted from [155].

3.6 Discussion and future perspectives

In this chapter I described the work performed to build the first multi-omic atlas of human myocardial infarction and more general of a human disease, including spatial transcriptomic, single-cell gene expression (snRNA) and chromatin accessibility (snATAC) data. This represents a first step towards a multi-scale understanding of tissue function and transformations during disease.

I presented the fundamental objectives of multi-scale modeling that relate to the proper definition of the molecular and structural characteristics of the tissue. Despite the challenges associated with data quality and generalization, in this work I proposed a computational strategy that allowed us to spatially contextualize the variability observed in single cell transcriptomics. My computational analysis increased the resolution of spatial transcriptomics by estimating for each location the major cell-type compositions and increased the interpretability of their molecular profiles by estimating pathway activities and mapping transcription factor binding activities. These different layers of biological information allowed me to link the organization of human heart tissue specimens of different histomorphological regions, time-points after myocardial infarction and different individuals to cellular functions. I described cardiac "niches", representing structural building blocks that are shared between different slides and could facilitate patient comparison. These niches were generated by clustering spatial transcriptomics spots using cell-type compositions or gene expression profiles that capture both structural and molecular diversity in tissues. Analysis of the integrated snRNA- and snATAC-seq data identified different cell states and sub-types for cardiomyocytes, endothelial cells, fibroblasts, and myeloid cells.

Moreover, I proposed an initial computational strategy for case-control comparisons that incorporate tissue- and cell-centric measurements. I contrasted the compositions of each major cell-type across patient groups with information provided by all omic layers, then I contrasted the compositions of niches from spatial transcriptomics and demonstrated consistent

remodeling events characteristic of fibrotic and ischemic samples. Finally, I identified molecular differences in myogenic regions that differentiated control samples from border zones and remote zones of myocardial infarction patients.

Given the limitations of inferring cell communication events from transcriptomics data and the difficulties of disentangling patterning events from communication events in spatial transcriptomics, spatial modeling was exclusively used to study the spatial organization of cell-types and cell-states in the spatial transcriptomics slides. I estimated the importance of the abundance of each major cell-type in explaining the abundance of the other major cell-types in different spatial contexts. This analysis explicitly focused on tissue patterning and allowed for the identification of differential cell-type dependencies between patient groups. I estimated the relationship between the tissue structure and cell functions, encoded as signaling pathway activities per spot estimated from gene expression. In this model, I predicted spatial relationships between pathway activities and the importances of cell-type abundances in predicting the patterns of pathway activities. Finally, I estimated associations between the tissue organization and the spatial distribution of failing cardiomyocytes and the different endothelial and fibroblast cell-states. I hypothesized that the distribution of specific cell-states in the spatial transcriptomics slides could be modeled by the cell-type composition or cell-state presence of individual spots and their neighborhood. With these models I was able to capture known co-localizations of venous endothelial cells, vascular smooth muscle cells, and capillary endothelial cells and pericytes. Moreover, I describe differential spatial correlations of the transcriptional signature of the “stressed” cardiomyocyte state vCM3 and the abundance of other cell-types between different patient groups. This analysis suggested that the location of this cardiomyocyte state can occur in different tissue regions depending on the time point and region after MI. In relation to fibroblast and myeloid cell-state populations, I observed a spatial gradient of myofibroblasts that aligned with a gradient of SPP1+ macrophages.

The combination of spatial technologies with single cell data represented an opportunity to study how cardiac cell states are influenced by their tissue microenvironment. The inferred interactions between cell-types largely reflect the spatial organization of the tissue and, while many other factors are involved, these interactions provide hypotheses for further analysis. My analyses revealed multiple novel insights into molecular and spatial biology of human myocardial infarction at different disease stages and scales that provide novel insights into the important mechanisms of cardiac injury repair and fibrosis and a valuable resource for the field

and future drug target discovery studies. Additionally, I devised a novel computational strategy for data integration across multiple experiments and modalities, which serves as a framework for future datasets and analyses.

Chapter 4

Multicellular factor analysis for a tissue-centric understanding of disease

In this chapter I introduce a tissue-centric framework to assess the variability of samples within a patient cohort profiled with single cell transcriptomics. By combining the information of single cell, spatial, and bulk transcriptomics, I found that inflammatory and reparative responses associated with myocardium remodeling are dominated by global coordinated transcriptional responses across distinct cell lineages. The results described in this chapter represent a unification of the knowledge generated by my research previously presented in the last two chapters and support the possibility of building holistic multi-scale molecular descriptions of disease.

4.1 Multicellular analysis of single cell omics data

The availability of single cell transcriptomics atlases profiling the pathological state of different tissues and organs in humans has increased during the last years and will continue to expand in different areas of the biomedical field. In these studies a common objective is to generate cross-conditions comparisons of the molecular profiles of major cell lineages, called cell types, by contrasting the gene expression between two or more groups of samples. The differentially expressed genes across sample groups of each cell type then are interpreted as a consequence of the emergence of cell states, groups of cells sharing a function, or as a consequence of a global transcriptional response of the majority of cells to a disease microenvironment. However, cross-condition comparisons are generally performed in each cell type separately, ignoring the fact that cells in tissues are organized in communities with shared functions. This cell-centric approach treats each transcriptional response to disease independent from each other, missing coordinated multicellular changes, where particular gene expression changes may be shared across cell types.

A family of novel tissue-centric methods have been proposed, such as DIALOGUE [90], scITD [91], and tensorcell2cell [92], in which the main objective is to approximate multicellular responses by inferring from a collection of “cell type views" a latent space that captures the shared variability across features of distinct cell types. In these models, cell type views can represent various molecular readouts of a cell type of interest that are quantified across distinct samples. DIALOGUE and scITD model pseudobulk expression profiles of individual cell types, while tensorcell2cell uses communication scores between pairs of cell types. These methods differ in their statistical foundations (Table CH4-1), however all of them allow to interpret each latent variable as a multicellular higher-order transcriptional response, or program, that describes certain aspect of the variability of a sample cohort (eg. clinical condition, technical effects, etc.) facilitating exploratory and unsupervised analysis of highly nested and dimensional single cell data. scITD and tensorcell2cell additionally report reconstruction error metrics of the latent space that quantify how much of the variability of a specific cell type view can be recovered by the latent variables. Moreover, since all of the latent variables are linear combinations of the features of distinct cell type views, it is possible to evaluate shared and cell type specific responses associated with a technical or clinical aspect of the sample cohort.

While these novel tissue-centric methods are helpful in the unsupervised analysis of cross-condition single cell atlases, multiomics integration methods, such as Multi-Omics Factor Analysis (MOFA) [34], [179], can also be repurposed for the inference of multicellular responses, since they share modeling objectives and treat similar multi-table data representations. MOFA is a flexible method that overcomes the limitation of data completeness that multicellular integration methods enforce and provides the possibility of jointly analyzing independent groups of samples with various classes of cell type views, representing a generalization of the aforementioned recently developed methods.

Although all of these methods have the possibility to infer coordinated gene expression events across cell types, no current framework has been proposed to contextualize these multicellular programs estimated from single cell data to other scales such as spatial and bulk omics. Spatial mapping of coordinated transcriptional processes across cell types in tissues could delineate to what extent these are the consequence of a global response to the general tissue microenvironment or the consequence of specific cellular interactions. Complementary, multicellular coordinated changes could be used to reinterpret the collection of transcriptional signatures derived from bulk transcriptomics studies.

Here I propose a cross-condition multicellular analysis from single cell transcriptomics data with MOFA using as case study a collection of large public atlases of human myocardial infarction and end-stage heart failure. I show that MOFA can be used for an unsupervised analysis of the variability of patients in single cell data sets and for the prioritization of coordinated transcriptional changes across cell types associated with clinical aspects. I dissect from these programs the general pathological cell type responses from expression fingerprints of cell state emergence. I show that upon myocardial infarction most of the cell type specific gene expression deregulation is the product of a global tissue level response, rather than only the consequence of new functional populations of cells. Moreover, I explore the expression of these coordinated programs in tissues by mapping their distribution in spatial transcriptomics data and suggest that pathological programs distribute in larger areas of diseased tissues and that this distribution is not constrained to the colocalization of specific cell types. Finally, I show how multicellular factor analysis can be used to generate a global comparison of single cell atlases that can facilitate the meta-analysis and integration of single cell data coming from multiple patient cohorts. I show that the estimated shared multicellular programs can be used to better deconvolute disease signals from bulk transcriptomics data that are not related to compositional changes of tissues. My analyses presented in this work represent a multi-scale and multi-cellular framework that integrates single cell, spatial and bulk transcriptomics for the analysis of cross-condition comparisons with the objective of understanding tissue responses in disease contexts (Figure CH4-1).

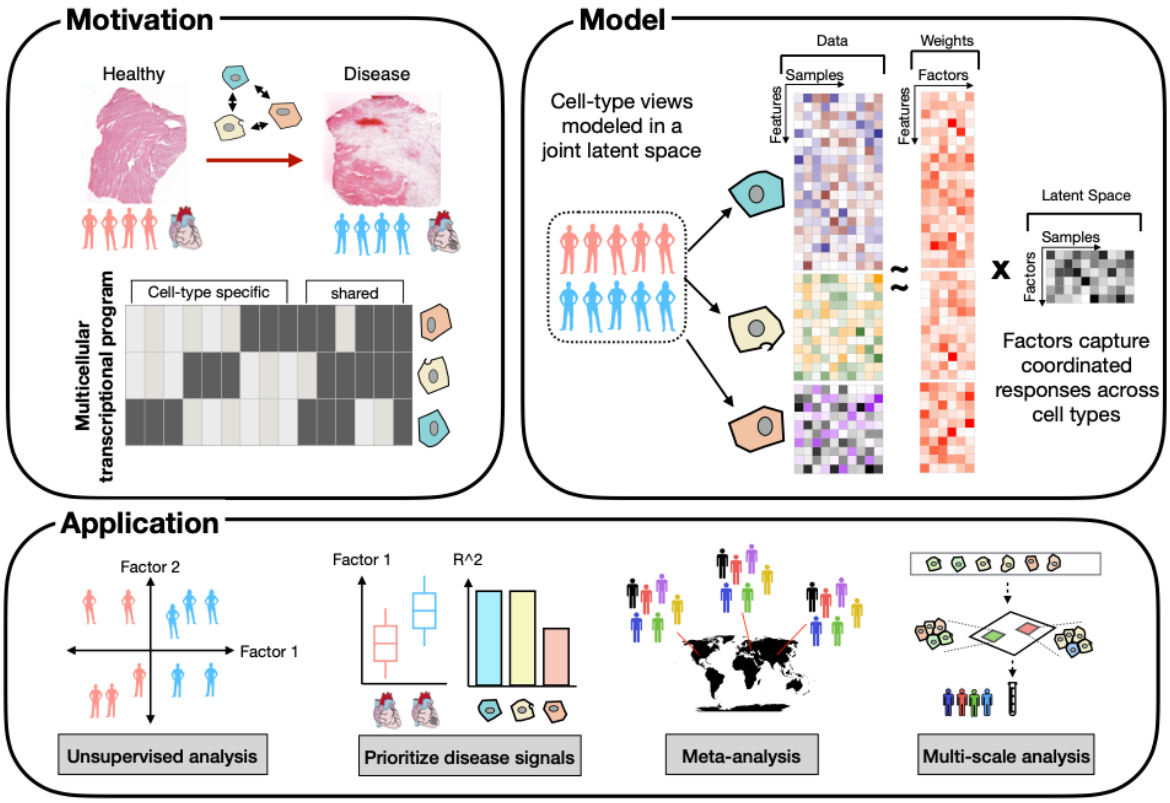

Figure CH4-1. Multicellular factor analysis using MOFA.
Single cell omics data across conditions are generated with the objective to characterize disease processes and mechanisms. A tissue-centric analysis framework of cross-condition single cell atlases focuses on the inference of coordinated molecular events across cell types, in this chapter called multicellular programs. Here I propose to repurpose multiomics factor analysis (MOFA) [34], [179] to simultaneously decompose the variability of multiple cell types (eg. represented as gene expression matrices of pseudobulk profiles) and create a latent space that recovers multicellular transcriptional changes. Throughout this chapter, four applications are presented to show how this analysis can be used for an unsupervised analysis of single cell data of multiple samples and conditions, for a prioritization of disease signals using the inferred latent space, and for a combined analysis of multiple studies across different scales (eg. bulk or spatial transcriptomics).

Method	Statistical method	Input	Output							
			Latent variable scores per sample	Multicellular programs encoded as a loading matrix	Explained variance associated to each factor	Explained variance associated to each view and factor	Robust to non-overlapping feature set	Integrates multiple studies	Input flexibility	Linear
DIALOGUE [90]	Penalized matrix decomposition followed by multi-level modeling step	Expression matrix or reduced matrix (PCA) of each cell	NO	YES	NO	NO	NO	NO	NO	YES
scITD [91]	Tucker decomposition	Tensor of pseudobulk expression profiles per cell type	YES	YES	YES	YES	NO	NO	NO	YES
Tensor cell2cell [92]	Tensor Component Analysis	Coexpression of ligand and receptors of pairs of sender and receiving cell types	YES	YES* (view agnostic)	NO	NO	NO	NO	NO	YES
MOFA [179]	**Probabilistic group factor analysis**	**Collection of pseudobulk expression profiles per cell type or any type**	**YES**	**YES**	**YES**	**YES**	**YES**	**YES**	**YES**	**YES**

Table CH4-1. Comparison of methods for multicellular analysis

4.2 Results

4.2.1 Multicellular factor analysis for an unsupervised evaluation of the variability of samples in single cell cohorts

One of the main challenges of analyzing large cohorts of single cell transcriptomics data is to assess in an unsupervised fashion the variability in gene expression across multiple samples for all cell types. Data-driven approaches are required to describe the complexity of single cell data with a reduced number of variables and to quantify the amount of variance across cell types that can be associated with biological, rather than technical factors. In addition to evaluating the influence of technical or confounding factors, this type of analysis is particularly important for single cell studies designed to study disease processes in tissues where reliable clinical patient groupings are not obvious.

A proposed solution to this task is to identify a shared latent space across cell types where the variability between samples is captured simultaneously. In this latent space, samples are defined by a reduced collection of factors usually built from a linear combination of the features of the original data. In the case of single cell transcriptomics data, the latent space can be inferred from a collection of cell type views representing, for example, the pseudobulk expression profile of each sample across cell types (Figure CH4-1). Ideally, each dimension or factor of the shared latent space is interpretable by 1) quantifying the contribution of each gene for each cell type in its definition and by 2) quantifying how much of the variability of a given view is captured by the factor. From the available methods designed explicitly to capture the multicellular variability of different samples in single cell experiments, scITD [91] is the only methodology that covers these requirements by performing Tucker decomposition [180] to a tensor formed by pseudobulk expression profiles of different cell types. However, one of the limitations of this approach is that it constrains the tensor for data completeness, in other words, the model requires an identical set of genes to be measured across cell types and requires all samples to contain complete profiles for all cell types. Given the differences in cell capture across samples and the expected non-overlapping set of genes expressed in different lineages, this modeling constraint is not always applicable to single cell data sets.

The generation of a latent space that captures the variability of distinct views across samples is a task that has been addressed by state-of-the-art multiomics integration methods established

for bulk data. Hence, it is plausible that these integration methods, such as multi-omics factor analysis (MOFA) [34], [179], could be used for the multicellular analysis of single cell data. Based on probabilistic group factor analysis, MOFA aims at integrating multiple views measured for multiple samples in an unsupervised manner. Compared to scITD, MOFA allows for a more flexible definition of multi view integration, since it allows for the usage of incomplete data with missing values both in views and in features per view. Additionally, MOFA models are completely interpretable allowing to measure the contribution of each view in the reconstruction of the latent space and to associate the global variability of the data to multiple covariates of the samples.

To demonstrate that MOFA can be repurposed to perform an unsupervised multicellular analysis of samples profiled with single cell RNA-seq, I fitted a MOFA model to a collection of 27 left-ventricle heart human samples profiling human myocardial infarction and control myocardium [155]. This data set contains samples of three major disease groups across 11 major cell types: myogenic (n = 13), fibrotic (n = 5) and ischemic (n = 9). Pseudobulk expression profiles were generated for each major cell type of each sample described in the public atlas by summing up the gene counts of all cells belonging to the combination of sample and cell type. Profiles generated with less than 5 cells were excluded. Data was normalized using the trimmed-mean of M values (TMM) method in edgeR [47] with a scale factor of 1 million and log-transformed. For each cell type expression matrix, highly variable genes were selected using scITD's adaptation of PAGODA2's method [91] before fitting a MOFA model with six factors where each cell type represented an independent view. Feature-wise sparsity was not forced in the model to obtain the greatest number of genes per cell type associated with each factor. Lowly abundant cell types that were not captured in all samples ("Mast", "Neuronal", "proliferating", and "Adipocytes") were excluded from the analysis. I associated the factor scores of each patient to disease group and batch labels to evaluate both biological and technical variability in the dataset using analysis of variance (ANOVA). P-values were corrected using the Benjamini-Hochberg procedure.

The latent space returned by the MOFA model fitted to the single cell atlas (Figure CH4-2A) explained on average the 68.3% of the variability of gene expression of the highly variable genes across cell types. Hierarchical clustering of the samples based on their factor scores clearly separated ischemic and fibrotic specimens from myogenic-enriched samples, however I observed that only three out of five fibrotic samples were similar in their factor scores. From

six recovered factors, Factor 1 was the only one associated with the tissue condition labels (adj. p-value < 0.05, mean percentage of explained variance across cell types of 37.15%, Figure CH4-2B), and Factor 2 and Factor 5 were associated with the technical label (adj. p-value = 0.051, mean percentage of explained variance across cell types of 19%). Fibroblasts (Fib), myeloid cells and vascular smooth muscle cells (vSMCs) were the cell types with the largest explained variance associated with the technical variable (> 22%). To easily visualize the biological variability of samples, I performed an Uniform Manifold Approximation and Projection (UMAP) to non-batch related factors (Figure CH4-2C). My results suggest that MOFA can be applied to cross-condition single cell atlases for exploratory unsupervised analysis that allows to detect and prioritize biological signals.

I created a multicellular latent space of six dimensions with scITD to compare the performance of these two methods using the same collection of highly variable genes. The latent variables of scITD recovered in total the 57% of the variability of the expression across cell types, 11% less than the one recovered by the MOFA model. Two recovered factors associated with the tissue condition labels (Factor 2 and 4), with a mean percentage of explained variance across cell types of 9.5%. A single factor (Factor 3) associated with the technical label with a percentage of explained variance across cell types of 7.01%. The mean absolute Pearson correlation between the factors associated with the condition label across methods was of 0.51, with a maximum absolute correlation of 0.75 between MOFA's Factor 1 and scITD's Factor 2. The mean Pearson absolute correlation between the loadings of these two factors across cell types was 0.66, suggesting a high concordance in the latent space reconstruction between the two methods.

4.2.2 Multicellular coordinated programs are related to global responses during myocardial infarction

To characterize the multicellular coordinated gene programs associated with cross-condition differences between sample groups identified with MOFA, I explored the cell type specific gene loadings of Factor 1. First, from the collection of 5209 highly variable genes used in the model, I considered genes to be associated with the factor if the Pearson correlation between their expression in at least one cell type and the factor score across samples was more or equal to 0.5. After filtering genes based on this classification, I observed that the median number of genes associated to the factor per cell type was 593, being lymphoid and vascular smooth

muscle cells the cell types with the least amount of associated genes (Figure CH4-3A). I observed that the majority of the genes correlated to Factor 1 (62%) were associated with more than a single cell type, suggesting that certain disease responses may be shared between cell types (Figure CH4-3B). Then, for each cell type I separated their genes associated with the factor into two groups based on their loadings (positive > 0 and negative < 0). Based on the sample scores of Factor 1, disease and healthy transcriptional programs are represented by the positive and negative gene sets, respectively. To quantify the overlap between these two classes of programs across cell-types I calculated pairwise Jaccard indexes (Figure CH4-3C). The mean Jaccard index across pairwise comparisons of disease and healthy signatures was of 0.33 and 0.23, respectively, suggesting that both, in healthy and disease tissue contexts, cell types activate both shared and specific transcriptional programs.

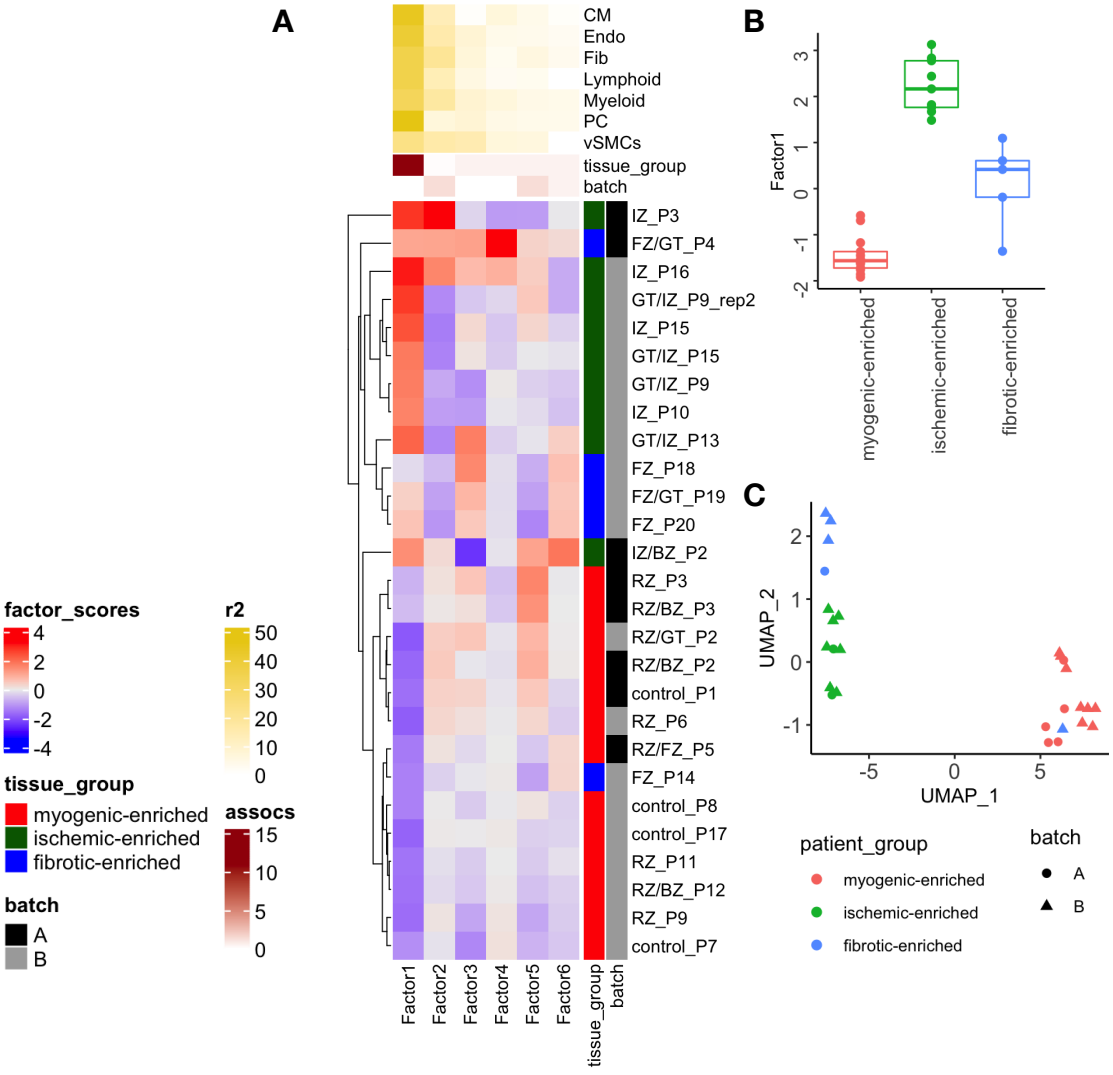

Figure CH4-2 Multicellular factor analysis using MOFA to a single cell atlas of myocardial infarction. **A**, The lower panel shows the hierarchical clustering of the factor scores of the 27 samples inferred by the MOFA model. Next to the sample name, its histological classification and technical batch labels are shown. The middle panel shows the -log10(adj. p-values) of testing for associations between the factor scores and the patient group or batch label. The upper panel shows the percentage of explained variance of each cell type expression matrix recovered by the factor. **B,** Distribution of the scores of Factor 1 across different patient groups. **C,** Uniform Manifold Approximation and Projection of the scores of factors not related to the technical batch.

The MOFA model was fitted to a collection of pseudobulk expression matrices of major cell type lineages under the assumption that transcriptional signatures captured by the gene loadings of the model can recapitulate the emergence of functional cell states together with global responses within a cell lineage. Thus, I next tested for each cell type to what extent the genes associated with Factor 1, and thus associated with changes upon myocardial infarction, agreed with the proposed cell-states presented in my previous work [155], based on the analysis of single cells. Within each cell type, I tested for the overrepresentation of cell state markers in disease and healthy programs recovered from the MOFA factor loadings using hypergeometric tests. I observed across cell-types that disease programs had a greater overrepresentation (adj. p-value < 0.05) of marker genes of cell states that increased in composition in ischemic and fibrotic samples, whilst healthy programs were overrepresented by marker genes of cell-states with greater compositions in healthy samples (Figure CH4-4A, for an example in cardiomyocytes). These results aligned with the expected effect of pseudobulking groups of cells, where the gene expression signal of the most abundant cells will be prioritized. Overall, the gene loadings across cell types of Factor 1 captured transcriptional shifts that relate to the change in compositions of functional cell states as a consequence of the disease context.

Next, I questioned if a global transcriptional response across cells within a lineage could be recovered from the model loadings of genes associated with the factor. I hypothesized that while the emergence of cell-states is a valid abstraction of the molecular processes related to disease, there may be transcriptional responses that are independent from cell-states and represent a global response of cells within the diseased tissue triggered by a larger microenvironment. To quantify the proportion of variance that could be explained by the disease and the cell-state classes within each major cell-type lineage, I performed independent analyses of variance (ANOVA) to the expression of each gene associated to Factor 1 of my MOFA model and by extension to responses upon myocardial infarction. The ANOVAs were fitted to pseudobulk expression profiles of cell-states across samples within each major lineage (cardiomyocytes, fibroblasts, endothelial and myeloid cells). Profiles generated with less than

20 cells were excluded. Normalization of data was performed as previously described. In each ANOVA either the cell-state class or the patient group was used in the test. Eta-squared values of the grouping variable per gene were used to quantify the amount of variance explained by cell-states or patient grouping.

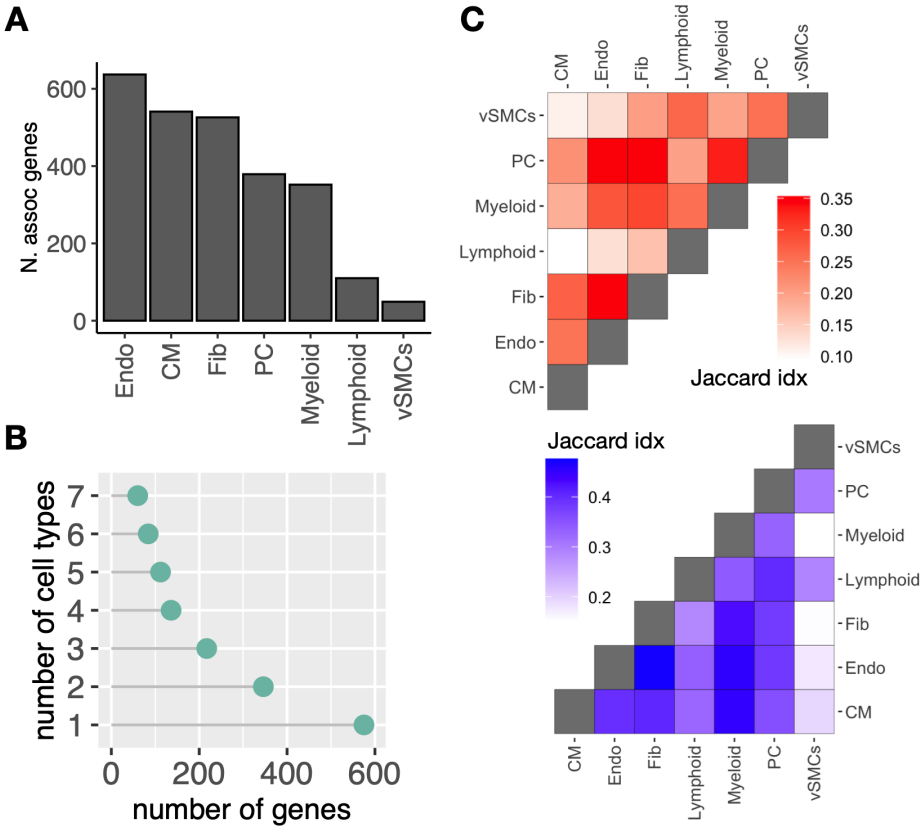

Figure CH4-3 Description of the coordinated responses upon myocardial infarction captured by Factor 1. **A,** Number of genes per cell type associated with Factor 1. **B,** Number of genes associated with Factor 1 that are expressed in a single or multiple cell types. **C,** Pairwise Jaccard index between the healthy and disease signatures between cell types.

Within each major cell type lineage I observed genes associated with Factor 1 that were independent of cell-states. In Figure CH4-4B I showed an example of this type of association for FATC4, a gene that was part of the transcriptional shift of cardiomyocytes. FATC4 is a gene involved in a global response to ischemia in cardiomyocytes (adj. P-value < 0.05), however it is similarly expressed across cell-states. I observed that across cell types, the variability in expression of genes associated with factor 1 were better explained by the disease condition rather than cell-states with evidence at the significance level (Figure CH4-4C, for cardiomyocytes) and the log-ratios of explained variances (Figure CH4-4D). Altogether, my results demonstrate that the genes defining the multicellular latent variable associated with

disease response recovered by the MOFA model capture both, cell-state dependent and independent transcriptional shifts. Moreover, the high ratios between the explained variance associated with disease and cell-states across different lineages suggest that while certain cell-states are more abundant during myocardial infarction, cells within a tissue and lineage partake in a shared global response that could be associated to adaptations to the changing tissue.

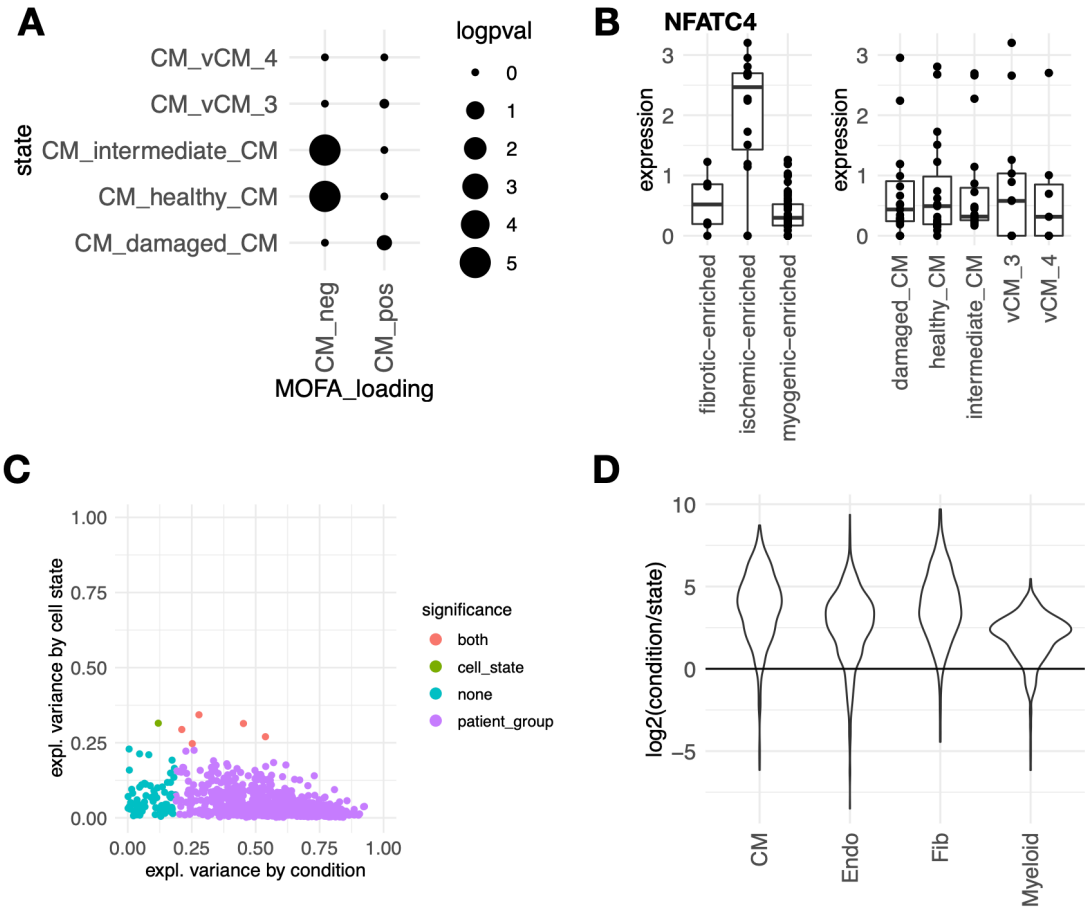

Figure CH4-4 Cell-state dependent and independent transcriptional responses upon myocardial infarction. **A,** Overrepresentation of cell-state markers in healthy and disease signatures estimated by MOFA. Size of the dots represent -log10(adj. p-value). **B**, Distribution of the expression of NFATC4 in cardiomyocytes across patient groups or cell-states. Each dot in both panels is the pseudobulk log expression of each cell state and sample. **C**, Proportion of variance of cardiomyocyte genes associated with Factor 1 explained by the patient group or the cell state. Colors represent significance of the association (adj. p-value < 0.01). **D,** Distribution of the log ratios across major cell types between the proportion of variance explained by the patient group and cell-states. Each point in each distribution is the ratio of a gene of the cell type associated with Factor 1.

4.2.3 Spatial mapping of multicellular coordinated programs upon myocardial infarction

To better understand the multicellular coordination of the disease transcriptional programs identified by the MOFA model, I mapped the previously defined positive and negative gene

sets of each cell type to the collection of paired spatial transcriptomics samples (10X Visium) that were generated together with the single cell data used in the previous sections [155]. To map each specific gene set of each cell type, I first identified regions per slide where the major cell types were present. I classified each spot of each slide to contain a cell type if the proportion of cells within the spot was greater or equal to 0.10. Proportions were obtained from the deconvolution analysis of the Visium slides from my previous work [155]. I then calculated normalized spot level scores for each positive and negative gene set using *decoupler-py*'s weighted mean method using the absolute loading score as weights with 100 permutations [52]. Cell type scores were only inferred in spots where the cell type was present. I assumed that the normalized gene set scores capture the distribution of cells expressing a disease or healthy transcriptional program, based on the positive and negative gene sets, respectively.

The scores of Factor 1 from the MOFA model clearly separated the three main sample groups (myogenic, ischemic and fibrotic) available in this dataset (Figure CH4-2B). I reasoned that this sample separation could be explained, among other processes, by the distribution of cell type disease programs in larger areas in ischemic and fibrotic tissues. To estimate the relative areas across cell types and samples where disease and healthy transcriptional programs were expressed, I first assumed that the effective area of a program for a specific cell type was defined by the number of spots where the cell type was present (see above). Then, for each cell type program I counted in how many spots the inferred gene set score was greater than 2 (representing the number of standard deviations from the mean of the distribution of scores from random gene sets). Finally, the relative area of activation was calculated as the ratio between the spots with active programs and the effective area.

I observed that across cell-types, the expression of disease programs occur in larger areas in ischemic and fibrotic samples compared to myogenic samples, encompassing over 50% of their effective location area except for pericytes and endothelial cells (Figure CH4-5A-B for cardiomyocytes and fibroblasts, Wilcoxon Test adj. p-value < 0.05). Ischemic samples contained larger relative areas of activation of disease programs of cardiomyocytes and fibroblasts compared to fibrotic samples (Figure CH4-5A-B, Wilcoxon Test adj. p-value < 0.05), reflecting the gradual reparative responses of myocardium following infarction. While disease programs showed a clear difference in distribution across conditions, coordinated programs associated to healthy myocardium were expressed almost uniformly across conditions (Figure CH4-5A), suggesting that upon myocardial infarction, major cell type

lineages activate complementary transcriptional programs with the purpose of maintaining tissue homeostasis without losing their usual gene programs. To avoid ambiguous definitions of relative active areas of disease and healthy transcriptional programs, I quantified the mean expression of each program and cell type across the different samples. The comparison of expression scores across conditions confirmed the reported differences in size of relative active areas (Figure CH4-5C, Wilcoxon Test adj. p-value < 0.05). My analyses demonstrated that the scores from the main factor explaining the variability between sample groups from the MOFA model capture the size of the areas where the expression of disease transcriptional programs are activated. The fact that in ischemic and fibrotic slides I observed the activation of disease programs in over 50% of the tissue area provides further evidence to suggest that upon myocardial infarction, global responses within each cell type dominate the molecular reparative responses of the myocardium.

I then estimated the spatial dependencies of the cell type disease transcriptional programs activated in fibrotic and ischemic samples using spatially contextualized models to understand their multicellular coordination. I used MISTy [105] to predict the spatial distribution of all cell type disease programs with two different spatial contexts: 1) An intrinsic view that measures the relationships between the disease programs within a spot, and 2) a para view that weights the program estimations of more distant neighbors of each cell-type (effective radius = 5 spots). With this model I measure if the activation of a cell type disease program in a spot can be explained by the activation of other disease programs within the same spot or in an extended neighborhood. The results of all models were aggregated to estimate distributions of performance and mean spatial dependencies between programs (Figure CH4-5D). My models showed that the spatial distribution of the disease programs of fibroblasts (median $R^2 = 0.62$), cardiomyocytes (median $R^2 = 0.4$) and myeloid cells (median $R^2 = 0.4$) could be explained by the co-localization of disease programs of other cell types. The spatial organization of more distant neighbors of cardiomyocytes contributed 74% to the prediction of their disease programs, while for fibroblasts and myeloid cells it was 40% and 43% respectively. Moreover, I observed a spatial dependency between the distribution of the programs of these three cell types, being fibroblasts the best predictors within the spot and myeloid cells the most important predictors in the local neighborhood (5 spots). To visualize hotspots of co-activation of the disease programs of cardiomyocytes, myeloid and fibroblasts across samples, I encoded the expression of each program in the RGB space (Figure CH4-5E). In this color space, brighter and darker colors represent a high and low expression of a disease program respectively and

the color combination differentiates different events of co-activation of programs between these cell types. As predicted by the MISTy model, I observed that areas activating the disease program of fibroblasts also co-activated programs of either myeloid cells or cardiomyocytes. Although I could identify areas in ischemic samples where the activation of myeloid cells and cardiomyocytes disease programs co-occurred, I did not observe large regions in tissues where the programs of all of the three cell types were activated. The distinct types of disease program "hotspots" suggest that the cell type disease programs identified by the MOFA model are activated in distinct cellular organizations. These observations indicate that while distinct cell type colocalization events may be reflecting tissue structures with specific tasks (eg. myeloid cells and fibroblasts regulate scar formation), in the disease context their molecular profile is coordinated with other tissue structures creating a higher order response that is characteristic of a tissue ecosystem. In summary, my results suggest that the coordinated disease transcriptional responses of cardiomyocytes, fibroblasts and myeloid cells estimated from the MOFA model are spatially organized in fibrotic and ischemic slides and thus likely reflect a multicellular response. However these coordinated responses are not constrained by a fixed local organization of cells within a region of the tissue but rather a global coordination of processes.

4.2.4 Multicellular factor analysis for the meta analysis of single cell atlases of heart failure

I have demonstrated that multicellular factor analysis using MOFA can be performed to a collection of single cell samples to identify coordinated global transcriptional responses in tissues that associate with clinical or technical covariates. I then tested if the model extension of MOFA+ [179], that jointly models independent groups of samples, could be used to perform meta-analysis across multiple case-control patient cohorts. I reasoned that the shared latent space would capture both conserved or specific transcriptional responses across distinct disease contexts.

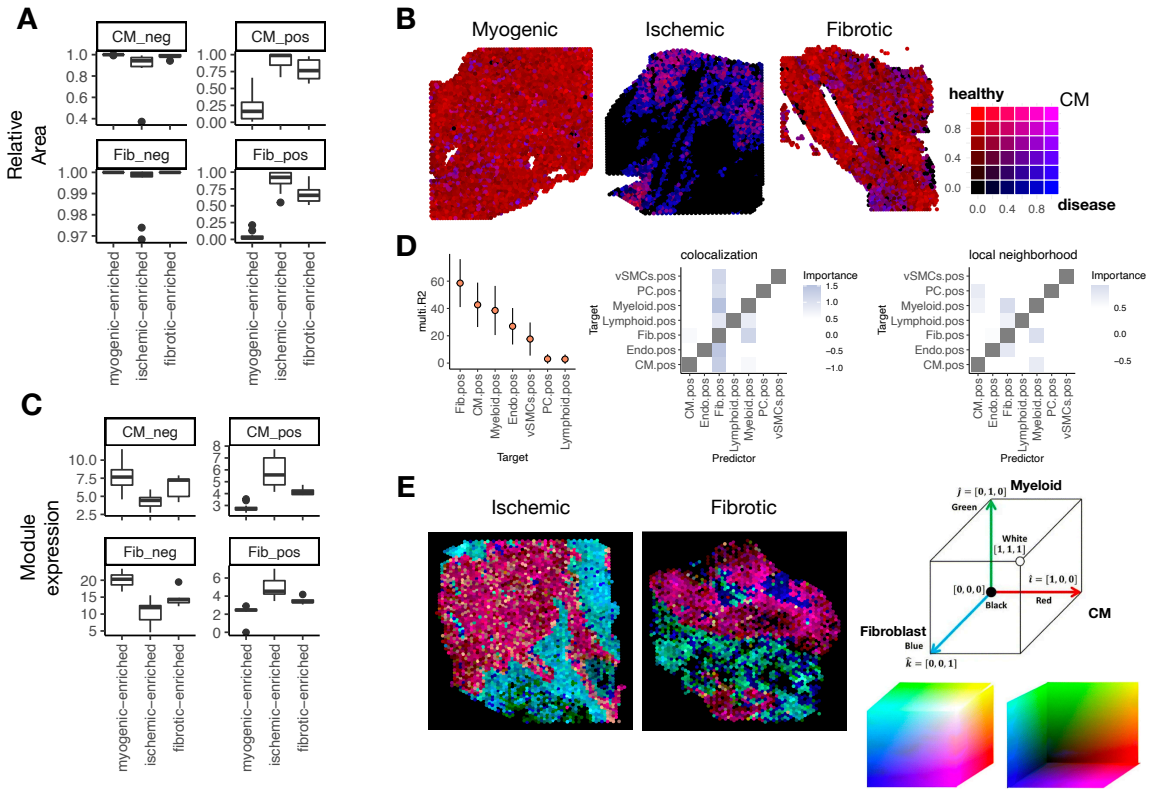

Figure CH4-5 Cell-state dependent and independent transcriptional responses upon myocardial infarction. **A,** Quantification of the relative area of the expression of healthy and disease programs of cardiomyocytes (CM) and fibroblasts (Fib) in spatial transcriptomics slides across patient groups (n = 28). **B,** Spatial mapping of the activation of cardiomyocyte specific healthy and disease programs in three representative examples of the conditions analyzed. The color of each spot represents the combination of the expression of those programs mapped to the RGB space (blue = disease program, red = healthy program). **C,** Quantification of the mean expression of healthy and disease programs of cardiomyocytes (CM) and fibroblasts (Fib) in spatial transcriptomics slides across patient groups (n = 28). **D,** Summary statistics of the spatially contextualized models of the expression of disease signatures fitted to spatial transcriptomics data. Left panel shows the distribution of the variance captured by the model for each cell type specific signature. The rest of the panels show the spatial interactions between signatures within a spot (middle) or in the local neighborhood (right). **E,** Spatial mapping of the activation of cardiomyocyte, fibroblast and myeloid disease programs in two representative examples of fibrotic and ischemic samples. The color of each spot represents the combination of the expression of those programs mapped to the RGB space (red = Cardiomyocytes, green = Myeloid cells ,blue = Fibroblasts). Legend adapted from [181].

I created pseudobulk expression tensors of major cell type lineages in two different single nuclei studies of heart failure. The first study (Chaffin2022) encompassed 42 single nuclei cardiac samples profiling healthy myocardium (n = 16) and end-stage heart failure both from dilated (n = 11) and hypertrophic cardiomyopathies (n = 15) [158]. The second study (Reichart2022) profiled 79 cardiac samples of healthy myocardium (n = 18), together with samples of dilated (n = 52), non-compaction (n = 1), and arrhythmogenic right ventricular (n = 8) cardiomyopathy [159]. After homogenizing the major cell type annotations, pseudobulk profiles of samples and cell types with more than 25 cells were calculated and normalized as

previously mentioned. I identified highly variable genes per cell type using scran's [172] *modelGeneVar* function with a biological variance threshold of 0. To perform the joint MOFA model between datasets, I subsetted the variable genes per cell type of Reichart2022 with the ones calculated from Chaffin2022. The grouped MOFA model was built as previously described with 6 factors, however groups were scaled before fitting the model.

To establish a baseline for each study, I first fitted a MOFA model to each study independently. I observed a mean total amount of explained variance across cell types of 38% for Chaffin2022, from which 21% could be associated with differences between healthy donors and heart failure patients. For Reichart2022 the model captured a mean total amount of explained variance of 37%, from which 19% associated with the distinct patient groups. In contrast, the joint MOFA model (Figure CH4-6A) captured a mean total amount of explained variance of 30 % and 35% for Chaffin2022 and Reichart2022, respectively. Despite the reduction in total explained variance, the joint model captured slightly higher amounts of variance explained by the clinical groups, with 23% for Chaffin2022 and 24% for Reichart2022. I did not find associations between the factors and the study labels, suggesting a proper integration of datasets. Additionally, I could not associate any factor with the sex of the patients (Figure CH4-6A). I visualized the distribution of samples across studies using the scores of the first two factors (Figure CH4-6B), which were associated with the distinct patient groups. I observed a clear distinction between non-failing and failing hearts, however etiologies formed two distinct groups of patients. In one group I distinguished dilated cardiomyopathy patients coming from different studies together with hypertrophic patients. In the other cluster, a second group of dilated cardiomyopathy patients exclusively from Reichart2022 clustered together with the rest of the failing patients. My results show that a multicellular factor analysis can be applied to samples coming from distinct patient cohorts for an unsupervised analysis of their variability and that the latent space together with the cell type specific loadings can be potentially used for the meta-analysis of the transcriptional coordinated responses of a tissue in distinct disease contexts.

4.2.5 Deconvolution of cell type specific transcriptional shifts of heart failure from bulk transcriptomics

Finally, to show that the global transcriptional responses to disease contexts estimated from my proposed multicellular factor analysis model could be used to deconvolute cell type

responses from bulk transcriptomics, I mapped cell type specific heart failure signatures estimated from my previously described meta-model to an independent collection of 16 bulk heart failure transcriptomics studies (ReHeaT) [112]. I hypothesized that a bulk expression profile captures the coordinated multicellular response of all cell types in the tissue and that the contribution of each cell type in this profile could be approximated by quantifying the enrichment of cell type specific signatures. My assumptions are supported by my previous observation that the summarized expression of a disease signature in a tissue is related to the size of the area where it is expressed (see Section 4.2.3).

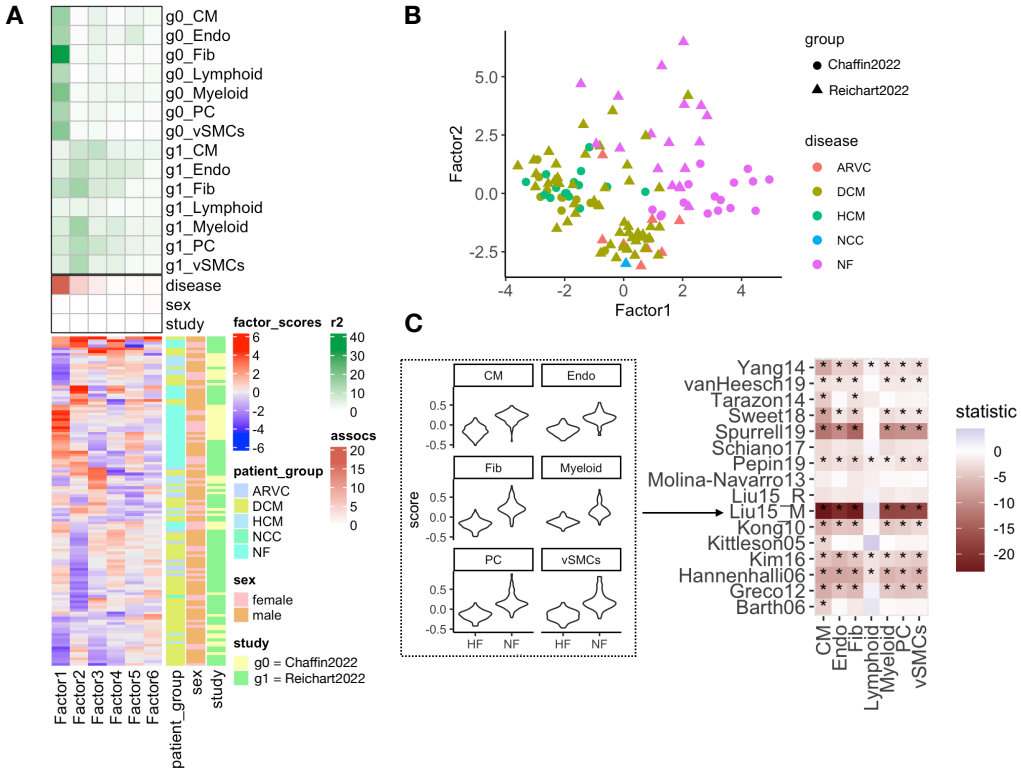

Figure CH4-6 Multicellular factor analysis for the meta-analysis of patient cohorts within and across scales. **A,** Summary statistics of the grouped multicellular factor analysis model. The lower panel shows the hierarchical clustering of the factor scores of the 121 samples inferred by the MOFA model. The patient group class, sex, and study labels are indicated next to the sample name. The middle panel shows the -log10(adj. p-values) of testing for associations between the factor scores and the patient group, sex, or the study label. The upper panel shows the percentage of explained variance of each cell type expression matrix recovered by the factor for each study separately. ARVC = arrhythmogenic right ventricular cardiomyopathy, DCM = dilated cardiomyopathy, HCM = hypertrophic cardiomyopathy, NCC = non-compaction cardiomyopathy , NF = non-failing. **B,** Distribution of the patient samples across studies based on the scores of the first two factors of the MOFA model. **C,** Quantification of the mean expression of cell-type specific programs in a collection of bulk transcriptomics studies. The left panel shows the distribution of cell-type programs from Factor 1 across failing (HF) and non-failing (NF) hearts. Given that high Factor 1 scores are associated with healthy patients, high signature scores are expected in NF samples. The right panel shows the t-values obtained from the comparison of HF vs NF samples across 16 independent bulk studies from ReHeaT. Stars highlight differences in the congruent direction (adj. p-value < 0.2). CM = cardiomyocytes, Fib = fibroblasts, Endo = Endothelial, vSMCs = vascular smooth muscle cells, PC = pericytes.

To calculate the enrichment of cell type specific disease programs in a bulk expression profile, I calculated directed weighted means of each signature using decoupleR's *wmean* function [52]. As inputs for each enrichment run I provided the scaled vector of bulk gene expression and the gene loadings of each cell type from a factor of interest as gene sets. Given the directed quality of the enrichment, the sign of the enrichment scores relate to its association with disease, eg. if positive factor scores are associated with the disease in the multicellular factor analysis model, then positive enrichment scores in the bulk data reflect a higher expression of the disease program. I estimated the enrichment of cell type disease signatures from Factor 1 of the meta-MOFA model fitted to Chaffin2022 and Reichart2022 in all samples available in ReHeaT, a compendium of 16 independent bulk transcriptomics datasets of heart failure [112]. Based on the distribution of the Factor 1 scores, I expected negative enrichment scores to associate with failing hearts samples in the bulk studies. To estimate if the enrichment scores of each cell type signature differentiated failing from non-failing hearts in the right direction across studies, I performed t-tests in which the defined alternative hypothesis was that the mean of enrichment scores in failing heart samples was less than the one from non-failing hearts (Figure CH4-6C).

I observed in 13 of the 16 bulk studies a congruent difference in the expression of cardiomyocyte disease signatures between failing and not failing hearts (adj. p-value < 0.2, Figure CH4-6C). Additionally, in 10 of these studies I could differentiate failing and not failing hearts using the signatures of the rest of the cell types used in the MOFA model, except for lymphoid cells (adj. p-value < 0.2, Figure CH4-6C). These results suggest that the global multicellular responses associated with heart failure estimated from single cell data are transferable to different patient cohorts and data modalities. Moreover, the decomposition of gene expression in terms of cell type responses provides an alternative and complementary understanding of bulk data beyond cell type compositions, given that the effect of lowly abundant cells (eg. pericytes or vascular smooth muscle cells) was still traceable from expression profiles.

4.3 Discussion

Despite the high costs of single cell technologies, it is expected that in the next few years, single cell datasets encompassing hundreds of patients will be generated to better characterize the molecular responses during disease. Consequently, there is a need for tissue-centric frameworks that on the one hand allow for an unsupervised analysis of samples across cell types and on the other provide estimations of coordinated molecular responses that better reflect the multicellular nature of organs.

In this study, I proposed to repurpose the statistical framework of multi-omics factor analysis (MOFA) to estimate cross-condition multicellular responses from single cell transcriptomics data. I demonstrated that the application of MOFA to collections of pseudobulk expression matrices of major cell types can generate a latent space that captures technical and biological variability of whole tissue specimens independent of cell type compositional changes. The interpretability of the model allows it to prioritize shared coordinated transcriptional changes between cell types, without losing the possibility of identifying cell type specific responses. The reconstruction metrics provided by the model can be used to identify subsets of cell types whose gene expression variability associates more with specific clinical covariates of samples. I argued that in comparison to novel methods explicitly built for the modeling of multicellular responses, MOFA has two main advantages: 1) Data completeness is not enforced which allows to better characterize cell-type specific responses and deals with the technical limitations of cell capture, and 2) group modeling can be defined, which facilitates the integration and comparison of multiple patient cohorts. Additionally, MOFA can model a flexible set of cell type views that potentially could capture complex multicellular responses, such as cell communication via ligand-receptor co-expression.

The application of multicellular factor analysis with MOFA to a collection of public single cell atlases of myocardial infarction and heart failure allowed me to show that tissue centric approaches open the door for holistic analysis of the responses of cells in tissues during disease. I showed cell-state independent transcriptional responses of cell types upon myocardial infarction, which may suggest that while certain functional states within a lineage are more favored in a disease context, most of the cells have a shared global response. By combining the results of multicellular factor analysis with spatial transcriptomics datasets, I mapped these cell type global responses and showed that the factor scores provided by our model, associate to

the relative area where they are expressed in tissues. Moreover, I showed that the coordination of cell type global responses in tissues is not constrained to a specific configuration of local cellular interactions, but rather represented the coordinated response of various tissue structures. Given my observations on global responses upon myocardial infarction, I meta-analyzed single cell samples of healthy and heart failure patients with multiple cardiomyopathies from two distinct studies and I showed that despite technical and clinical variability between patients, a conserved transcriptional response across cell types is observed in failing hearts. Importantly, I was able to find traces of these cell type responses in independent bulk data sets providing an alternative and complementary understanding of bulk data beyond cell type compositions. Altogether, I contributed with a framework that allows the community to integrate the measurements of independent single cell, spatial, and bulk datasets to contextualize cell type responses in disease.

My proposed framework is still dependent on the summarization of gene expression per cell-type as pseudobulk profiles, which may potentiate the technical effects of background expression of single cells and confound the estimated multicellular responses. Deep generative models have been proposed to take advantage of single cell measurements to estimate sample-level heterogeneity [182], however my observations on the conservation of global responses across scales provide evidence to suggest that current pseudobulk approaches still provide a meaningful understanding of tissue function. Although this study was focused on the application of MOFA for the understanding of multicellular responses in tissues, my results support the application of models such as MEFISTO [183] to analyze complex time-course experimental designs and MuVI [184] to generate biological interpretable latent spaces. Nevertheless, inherent challenges of factor analysis such as the linear constraints to generate the latent space still need to be addressed in future work.

In summary, I contributed with a framework that allows the integration of the measurements of independent single cell, spatial, and bulk datasets to contextualize multicellular responses in disease. My proposed tissue-centric analysis opens new opportunities to restudy publicly available datasets of precious patient samples as new technologies developed.

Concluding remarks

High-throughput genomic technologies have been applied in biomedicine to link the pathophysiology of disease to molecular and cellular mechanisms in organs. In the study of cardiovascular disease, one of the most prevalent conditions in the human population, the measurement of gene expression changes through different stages of cardiac remodeling aims to better understand the deregulation of the compensatory mechanisms that the heart activates upon damage. This objective has motivated the generation of transcriptomic profiles of human healthy and damaged myocardium across distinct disease conditions and scales using bulk, single cell, and spatial technologies.

This thesis aimed to provide a comprehensive description of the transcriptomic changes upon human myocardial infarction and heart failure. First, I demonstrated that it is possible to identify and prioritize generalizable gene expression changes associated with heart failure from a collection of independent patient cohorts profiled with bulk transcriptomics. Despite technical and clinical differences, I showed convergent global tissue responses that generalize to distinct heart failure etiologies. Moreover, motivated by the need of ever-growing molecular references of disease, I provided a computational framework that allowed the building of a consensus transcriptional signature of heart failure that is easy to update and correct. Knowing the limitations of bulk transcriptomics to delineate cell type specific responses in tissues, I presented afterwards a spatial and single cell atlas of the molecular processes happening after myocardial infarction focusing on distinct physiological zones and time points of cardiac remodeling. I described a multi-omics and multi-scale integration methodology that provided spatially resolved intercellular processes and facilitated the comparison of tissue samples at the organizational, compositional, and molecular level. My results showed that it is possible to identify conserved tissue structures (cellular neighborhoods) that emerge as a consequence of the disease context. Additionally, I demonstrated that the functional variability of cell lineages can be associated with these structures, nevertheless, I also showed data that suggested that the functional variability of cell lineages is inherent in tissues and not triggered by disease. Finally, I repurposed the statistical framework for multi-omics integration to generate multicellular descriptions of tissues from single cell data. My tissue-centric approach demonstrated to be useful to perform unsupervised analysis of large patient cohorts profiled with single cell data, to meta-analyze the cell type responses upon distinct disease contexts from multiple studies,

and to integrate the knowledge across the single cell, spatial, and bulk scales. The multi-scale integration of the transcriptional responses occurring during heart failure suggest that while cell type specific responses can be traced and localized in specific tissue structures, a global coordinated response dominates the molecular landscape of tissues during cardiac remodeling.

The consequences of this observation may impact the way the community prioritizes the molecular readouts provided by multi-scale omics technologies for the study of heart failure. With this work, I aimed to show that while exploring the cell type specific responses associated with heart failure and its etiologies will be valuable to create a catalog of the possible and finite functional transformations of each cell lineage, intercellular transcriptional responses are the hallmarks of the molecular description of disease. These coordinated responses across cell types in tissues are the footprints of the metabolic and organizational state of the tissue, and represent the reaction towards the maintenance of homeostasis of the tissues and organs in a changing environment. If tissues are understood as ecosystems that are defined by the biotic and abiotic components and their interactions within a limited area, the description of cell type responses and their location is limited to the study of the biotic component, ignoring the interdependence between these and the environment.

In my opinion for the cardiovascular field (and potentially other biomedical areas) to completely benefit from the generation of highly resolved and multiplexed data from single cell and spatial technologies, two main research objectives should be prioritized:

1) Identification of the conserved global intercellular responses in tissues, which requires an ever-growing interest in the integration of multiple datasets of distinct patient cohorts across scales and anatomical regions of the heart. My work aims to suggest a possible computational roadmap to achieve this, nevertheless many challenges lie ahead such as the size of the capture area profiled with omics technologies, the lack of a common spatial coordinate system for unorganized tissues, the definition of a cell lineage ontology, the lack of quality standards of experimental protocols and clinical reports, heterogeneous patient representation and centralized data hosting and integration.

2) **Reconciliation of the single cell omics and evolutionary ecology fields**, to better understand the functional heterogeneity of cell lineages and the interdependence of distinct tissue structures within a tissue or an organ. Organizational constraints should exist in tissues

to ensure the functions of organs from development until the death of a human, and it may be possible that the functional variability of cell lineages may ensure homeostasis by creating a pool of possible functions that can be favored in specific microenvironments. It is very likely that in the future more tissue-centric hypotheses will be tested with advanced dynamic profiling technologies of cell cultures and organoids.

In summary, the multi-scale transcriptomics landscape of heart failure described in this work contributed to the study of the impacts of tissue organization in cell function, and vice versa, while integrating the knowledge generated from the analysis of multiple independent patient cohorts across distinct technologies.

References

[1] B. Ziaeian and G. C. Fonarow, “Epidemiology and aetiology of heart failure.,” *Nat. Rev. Cardiol.*, vol. 13, no. 6, pp. 368–378, Jun. 2016, doi: 10.1038/nrcardio.2016.25.

[2] B. Bozkurt *et al.*, “Universal definition and classification of heart failure: a report of the Heart Failure Society of America, Heart Failure Association of the European Society of Cardiology, Japanese Heart Failure Society and Writing Committee of the Universal Definition of Heart Failure: Endorsed by the Canadian Heart Failure Society, Heart Failure Association of India, Cardiac Society of Australia and New Zealand, and Chinese Heart Failure Association.,” *Eur. J. Heart Fail.*, vol. 23, no. 3, pp. 352–380, Mar. 2021, doi: 10.1002/ejhf.2115.

[3] E. Arbustini *et al.*, “The MOGE(S) classification of cardiomyopathy for clinicians.,” *J. Am. Coll. Cardiol.*, vol. 64, no. 3, pp. 304–318, Jul. 2014, doi: 10.1016/j.jacc.2014.05.027.

[4] C.-C. Liew and V. J. Dzau, “Molecular genetics and genomics of heart failure.,” *Nat. Rev. Genet.*, vol. 5, no. 11, pp. 811–825, Nov. 2004, doi: 10.1038/nrg1470.

[5] D. J. Lips, L. J. deWindt, D. J. W. van Kraaij, and P. A. Doevendans, “Molecular determinants of myocardial hypertrophy and failure: alternative pathways for beneficial and maladaptive hypertrophy.,” *Eur. Heart J.*, vol. 24, no. 10, pp. 883–896, May 2003, doi: 10.1016/s0195-668x(02)00829-1.

[6] S. Cresci *et al.*, “Heart failure in the era of precision medicine: A scientific statement from the american heart association.,” *Circ. Genom. Precis. Med.*, vol. 12, no. 10, pp. 458–485, Oct. 2019, doi: 10.1161/HCG.0000000000000058.

[7] M. A. Peterzan, C. A. Lygate, S. Neubauer, and O. J. Rider, “Metabolic remodeling in hypertrophied and failing myocardium: a review.,” *Am. J. Physiol. Heart Circ. Physiol.*, vol. 313, no. 3, pp. H597–H616, Sep. 2017, doi: 10.1152/ajpheart.00731.2016.

[8] D. L. Mann and M. R. Bristow, “Mechanisms and models in heart failure: the biomechanical model and beyond.,” *Circulation*, vol. 111, no. 21, pp. 2837–2849, May 2005, doi: 10.1161/CIRCULATIONAHA.104.500546.

[9] A. C. Campos de Carvalho, T. H. Kasai-Brunswick, and A. Bastos Carvalho, “Cell-Based Therapies for Heart Failure.,” *Front. Pharmacol.*, vol. 12, p. 641116, Apr. 2021, doi: 10.3389/fphar.2021.641116.

[10] A. Iorio, A. Pozzi, and M. Senni, “Addressing the heterogeneity of heart failure in future randomized trials.,” *Curr. Heart Fail. Rep.*, vol. 14, no. 3, pp. 197–202, Jun. 2017, doi: 10.1007/s11897-017-0332-1.

[11] S. J. Shah *et al.*, “Research priorities for heart failure with preserved ejection fraction: national heart, lung, and blood institute working group summary.,” *Circulation*, vol. 141, no. 12, pp. 1001–1026, Mar. 2020, doi: 10.1161/CIRCULATIONAHA.119.041886.

[12] M. Colvin *et al.*, “Heart Failure in Non-Caucasians, Women, and Older Adults: A White Paper on Special Populations From the Heart Failure Society of America Guideline Committee.,” *J. Card. Fail.*, vol. 21, no. 8, pp. 674–693, Aug. 2015, doi: 10.1016/j.cardfail.2015.05.013.

[13] D. Tirziu, F. J. Giordano, and M. Simons, “Cell communications in the heart.,” *Circulation*, vol. 122, no. 9, pp. 928–937, Aug. 2010, doi: 10.1161/CIRCULATIONAHA.108.847731.

[14] K. Fountoulaki, “Cellular Communications in the Heart | CFR Journal,” Sep. 2015.

[15] F. Perbellini, S. A. Watson, I. Bardi, and C. M. Terracciano, “Heterocellularity and Cellular Cross-Talk in the Cardiovascular System.,” *Front. Cardiovasc. Med.*, vol. 5, p. 143, Nov.

2018, doi: 10.3389/fcvm.2018.00143.
[16] C. M. Howard and T. A. Baudino, “Dynamic cell-cell and cell-ECM interactions in the heart.,” *J. Mol. Cell. Cardiol.*, vol. 70, pp. 19–26, May 2014, doi: 10.1016/j.yjmcc.2013.10.006.
[17] N. Takeda and I. Manabe, “Cellular Interplay between Cardiomyocytes and Nonmyocytes in Cardiac Remodeling.,” *Int. J. Inflam.*, vol. 2011, p. 535241, Sep. 2011, doi: 10.4061/2011/535241.
[18] V. Talman and R. Kivelä, “Cardiomyocyte-Endothelial Cell Interactions in Cardiac Remodeling and Regeneration.,” *Front. Cardiovasc. Med.*, vol. 5, p. 101, Jul. 2018, doi: 10.3389/fcvm.2018.00101.
[19] F. C. Simões and P. R. Riley, “Immune cells in cardiac repair and regeneration.,” *Development*, vol. 149, no. 8, Apr. 2022, doi: 10.1242/dev.199906.
[20] J. G. Travers, F. A. Kamal, J. Robbins, K. E. Yutzey, and B. C. Blaxall, “Cardiac fibrosis: the fibroblast awakens.,” *Circ. Res.*, vol. 118, no. 6, pp. 1021–1040, Mar. 2016, doi: 10.1161/CIRCRESAHA.115.306565.
[21] P. Kong, P. Christia, and N. G. Frangogiannis, “The pathogenesis of cardiac fibrosis.,” *Cell. Mol. Life Sci.*, vol. 71, no. 4, pp. 549–574, Feb. 2014, doi: 10.1007/s00018-013-1349-6.
[22] S. Steffens, M. Nahrendorf, and R. Madonna, “Immune cells in cardiac homeostasis and disease: emerging insights from novel technologies.,” *Eur. Heart J.*, vol. 43, no. 16, pp. 1533–1541, Apr. 2022, doi: 10.1093/eurheartj/ehab842.
[23] J. G. Rurik, H. Aghajanian, and J. A. Epstein, “Immune cells and immunotherapy for cardiac injury and repair.,” *Circ. Res.*, vol. 128, no. 11, pp. 1766–1779, May 2021, doi: 10.1161/CIRCRESAHA.121.318005.
[24] T. Oka, H. Akazawa, A. T. Naito, and I. Komuro, “Angiogenesis and cardiac hypertrophy: maintenance of cardiac function and causative roles in heart failure.,” *Circ. Res.*, vol. 114, no. 3, pp. 565–571, Jan. 2014, doi: 10.1161/CIRCRESAHA.114.300507.
[25] A. R. Joyce and B. Ø. Palsson, “The model organism as a system: integrating ‘omics’ data sets.,” *Nat. Rev. Mol. Cell Biol.*, vol. 7, no. 3, pp. 198–210, Mar. 2006, doi: 10.1038/nrm1857.
[26] K. J. Karczewski and M. P. Snyder, “Integrative omics for health and disease.,” *Nat. Rev. Genet.*, vol. 19, no. 5, pp. 299–310, May 2018, doi: 10.1038/nrg.2018.4.
[27] Y. Hasin, M. Seldin, and A. Lusis, “Multi-omics approaches to disease.,” *Genome Biol.*, vol. 18, no. 1, p. 83, May 2017, doi: 10.1186/s13059-017-1215-1.
[28] X. Chen, S. A. Teichmann, and K. B. Meyer, “From Tissues to Cell Types and Back: Single-Cell Gene Expression Analysis of Tissue Architecture,” *Annu. Rev. Biomed. Data Sci.*, vol. 1, no. 1, pp. 29–51, Jul. 2018, doi: 10.1146/annurev-biodatasci-080917-013452.
[29] G. Palla, D. S. Fischer, A. Regev, and F. J. Theis, “Spatial components of molecular tissue biology.,” *Nat. Biotechnol.*, vol. 40, no. 3, pp. 308–318, Mar. 2022, doi: 10.1038/s41587-021-01182-1.
[30] M. P. Nagle, G. S. Tam, E. Maltz, Z. Hemminger, and R. Wollman, “Bridging scales: From cell biology to physiology using in situ single-cell technologies.,” *Cell Syst.*, vol. 12, no. 5, pp. 388–400, May 2021, doi: 10.1016/j.cels.2021.03.002.
[31] A. Dugourd *et al.*, “Causal integration of multi-omics data with prior knowledge to generate mechanistic hypotheses.,” *Mol. Syst. Biol.*, vol. 17, no. 1, p. e9730, Jan. 2021, doi: 10.15252/msb.20209730.
[32] N. C. Duarte *et al.*, “Global reconstruction of the human metabolic network based on genomic and bibliomic data.,” *Proc Natl Acad Sci USA*, vol. 104, no. 6, pp. 1777–1782, Feb. 2007, doi: 10.1073/pnas.0610772104.

[33] K. Kamimoto, C. M. Hoffmann, and S. A. Morris, “CellOracle: Dissecting cell identity via network inference and in silico gene perturbation,” *BioRxiv*, Feb. 2020, doi: 10.1101/2020.02.17.947416.

[34] R. Argelaguet *et al.*, “Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets.,” *Mol. Syst. Biol.*, vol. 14, no. 6, p. e8124, Jun. 2018, doi: 10.15252/msb.20178124.

[35] A. Qoku and F. Buettner, “Encoding Domain Knowledge in Multi-view Latent Variable Models: A Bayesian Approach with Structured Sparsity,” *arXiv*, 2022, doi: 10.48550/arxiv.2204.06242.

[36] T. Maier, M. Güell, and L. Serrano, “Correlation of mRNA and protein in complex biological samples.,” *FEBS Lett.*, vol. 583, no. 24, pp. 3966–3973, Dec. 2009, doi: 10.1016/j.febslet.2009.10.036.

[37] L. C. R. Fraser, R. J. Dikdan, S. Dey, A. Singh, and S. Tyagi, “Reduction in gene expression noise by targeted increase in accessibility at gene loci.,” *Proc Natl Acad Sci USA*, vol. 118, no. 42, Oct. 2021, doi: 10.1073/pnas.2018640118.

[38] L. Isbel, R. S. Grand, and D. Schübeler, “Generating specificity in genome regulation through transcription factor sensitivity to chromatin.,” *Nat. Rev. Genet.*, Jul. 2022, doi: 10.1038/s41576-022-00512-6.

[39] K. Shah and S. Tyagi, “Barriers to transmission of transcriptional noise in a c-fos c-jun pathway.,” *Mol. Syst. Biol.*, vol. 9, p. 687, 2013, doi: 10.1038/msb.2013.45.

[40] B. S. Zhao, I. A. Roundtree, and C. He, “Post-transcriptional gene regulation by mRNA modifications.,” *Nat. Rev. Mol. Cell Biol.*, vol. 18, no. 1, pp. 31–42, Jan. 2017, doi: 10.1038/nrm.2016.132.

[41] O. N. Jensen, “Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry.,” *Curr. Opin. Chem. Biol.*, vol. 8, no. 1, pp. 33–41, Feb. 2004, doi: 10.1016/j.cbpa.2003.12.009.

[42] R. Stark, M. Grzelak, and J. Hadfield, “RNA sequencing: the teenage years.,” *Nat. Rev. Genet.*, vol. 20, no. 11, pp. 631–656, Nov. 2019, doi: 10.1038/s41576-019-0150-2.

[43] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, “Quantitative monitoring of gene expression patterns with a complementary DNA microarray.,” *Science*, vol. 270, no. 5235, pp. 467–470, Oct. 1995, doi: 10.1126/science.270.5235.467.

[44] Z. Wang, M. Gerstein, and M. Snyder, “RNA-Seq: a revolutionary tool for transcriptomics.,” *Nat. Rev. Genet.*, vol. 10, no. 1, pp. 57–63, Jan. 2009, doi: 10.1038/nrg2484.

[45] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, “Mapping and quantifying mammalian transcriptomes by RNA-Seq.,” *Nat. Methods*, vol. 5, no. 7, pp. 621–628, Jul. 2008, doi: 10.1038/nmeth.1226.

[46] M. E. Ritchie *et al.*, “limma powers differential expression analyses for RNA-sequencing and microarray studies.,” *Nucleic Acids Res.*, vol. 43, no. 7, p. e47, Apr. 2015, doi: 10.1093/nar/gkv007.

[47] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.,” *Bioinformatics*, vol. 26, no. 1, pp. 139–140, Jan. 2010, doi: 10.1093/bioinformatics/btp616.

[48] M. I. Love, W. Huber, and S. Anders, “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.,” *Genome Biol.*, vol. 15, no. 12, p. 550, 2014, doi: 10.1186/s13059-014-0550-8.

[49] W. T. Barry, A. B. Nobel, and F. A. Wright, “Significance analysis of functional categories in gene expression studies: a structured permutation approach.,” *Bioinformatics*, vol. 21, no. 9, pp. 1943–1949, May 2005, doi: 10.1093/bioinformatics/bti260.

[50] L. Väremo, J. Nielsen, and I. Nookaew, “Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods.,” *Nucleic Acids Res.*, vol. 41, no. 8, pp. 4378–4391, Apr. 2013, doi: 10.1093/nar/gkt111.

[51] R. Mathur, D. Rotroff, J. Ma, A. Shojaie, and A. Motsinger-Reif, “Gene set analysis methods: a systematic comparison.,” *BioData Min.*, vol. 11, p. 8, May 2018, doi: 10.1186/s13040-018-0166-8.

[52] P. Badia-i-Mompel *et al.*, “decoupleR: Ensemble of computational methods to infer biological activities from omics data,” *BioRxiv*, Nov. 2021, doi: 10.1101/2021.11.04.467271.

[53] K. Korthauer *et al.*, “A practical guide to methods controlling false discoveries in computational biology.,” *Genome Biol.*, vol. 20, no. 1, p. 118, Jun. 2019, doi: 10.1186/s13059-019-1716-1.

[54] B. Szalai and J. Saez-Rodriguez, “Why do pathway methods work better than they should?,” *FEBS Lett.*, vol. 594, no. 24, pp. 4189–4200, Dec. 2020, doi: 10.1002/1873-3468.14011.

[55] A. Dugourd and J. Saez-Rodriguez, “Footprint-based functional analysis of multiomic data.,” *Current Opinion in Systems Biology*, vol. 15, pp. 82–90, Jun. 2019, doi: 10.1016/j.coisb.2019.04.002.

[56] M. Schubert *et al.*, “Perturbation-response genes reveal signaling footprints in cancer gene expression.,” *Nat. Commun.*, vol. 9, no. 1, p. 20, Jan. 2018, doi: 10.1038/s41467-017-02391-6.

[57] L. Garcia-Alonso *et al.*, “Transcription factor activities enhance markers of drug sensitivity in cancer.,” *Cancer Res.*, vol. 78, no. 3, pp. 769–780, Feb. 2018, doi: 10.1158/0008-5472.CAN-17-1679.

[58] P. Jiang *et al.*, “Systematic investigation of cytokine signaling activity at the tissue and single-cell levels.,” *Nat. Methods*, vol. 18, no. 10, pp. 1181–1191, Oct. 2021, doi: 10.1038/s41592-021-01274-5.

[59] L. Garcia-Alonso, C. H. Holland, M. M. Ibrahim, D. Turei, and J. Saez-Rodriguez, “Benchmark and integration of resources for the estimation of human transcription factor activities.,” *Genome Res.*, vol. 29, no. 8, pp. 1363–1375, Aug. 2019, doi: 10.1101/gr.240663.118.

[60] A. Dixit *et al.*, “Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens.,” *Cell*, vol. 167, no. 7, pp. 1853-1866.e17, Dec. 2016, doi: 10.1016/j.cell.2016.11.038.

[61] R. Elmentaite, C. Domínguez Conde, L. Yang, and S. A. Teichmann, “Single-cell atlases: shared and tissue-specific cell types across human organs.,” *Nat. Rev. Genet.*, vol. 23, no. 7, pp. 395–410, Jul. 2022, doi: 10.1038/s41576-022-00449-w.

[62] M. Uhlén *et al.*, “Tissue-based map of the human proteome.,” *Science*, vol. 347, no. 6220, p. 1260419, Jan. 2015, doi: 10.1126/science.1260419.

[63] E. Mereu *et al.*, “Benchmarking single-cell RNA-sequencing protocols for cell atlas projects.,” *Nat. Biotechnol.*, vol. 38, no. 6, pp. 747–755, Jun. 2020, doi: 10.1038/s41587-020-0469-4.

[64] G. X. Y. Zheng *et al.*, “Massively parallel digital transcriptional profiling of single cells.,” *Nat. Commun.*, vol. 8, p. 14049, Jan. 2017, doi: 10.1038/ncomms14049.

[65] J. Fischer and T. Ayers, “Single nucleus RNA-sequencing: how it’s done, applications and limitations.,” *Emerg. Top. Life Sci.*, vol. 5, no. 5, pp. 687–690, Nov. 2021, doi: 10.1042/ETLS20210074.

[66] B. B. Lake *et al.*, “A comparative strategy for single-nucleus and single-cell transcriptomes confirms accuracy in predicted cell-type expression from nuclear RNA.,” *Sci. Rep.*, vol. 7, no.

1, p. 6031, Jul. 2017, doi: 10.1038/s41598-017-04426-w.
[67] R. S. Brüning, L. Tombor, M. H. Schulz, S. Dimmeler, and D. John, “Comparative analysis of common alignment tools for single-cell RNA sequencing.,” *Gigascience*, vol. 11, Jan. 2022, doi: 10.1093/gigascience/giac001.
[68] M. D. Young and S. Behjati, “SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data.,” *Gigascience*, vol. 9, no. 12, Dec. 2020, doi: 10.1093/gigascience/giaa151.
[69] S. J. Fleming *et al.*, “Unsupervised removal of systematic background noise from droplet-based single-cell experiments using CellBender,” *BioRxiv*, Oct. 2019, doi: 10.1101/791699.
[70] M. D. Luecken and F. J. Theis, “Current best practices in single-cell RNA-seq analysis: a tutorial.,” *Mol. Syst. Biol.*, vol. 15, no. 6, p. e8746, Jun. 2019, doi: 10.15252/msb.20188746.
[71] C. H. O’Flanagan *et al.*, “Dissociation of solid tumor tissues with cold active protease for single-cell RNA-seq minimizes conserved collagenase-associated stress responses.,” *Genome Biol.*, vol. 20, no. 1, p. 210, Oct. 2019, doi: 10.1186/s13059-019-1830-0.
[72] N. M. Xi and J. J. Li, “Benchmarking Computational Doublet-Detection Methods for Single-Cell RNA Sequencing Data.,” *Cell Syst.*, vol. 12, no. 2, pp. 176-194.e6, Feb. 2021, doi: 10.1016/j.cels.2020.11.008.
[73] P.-L. Germain, A. Lun, C. Garcia Meixide, W. Macnair, and M. D. Robinson, “Doublet identification in single-cell sequencing data using scDblFinder.,” *F1000Res.*, vol. 10, p. 979, Sep. 2021, doi: 10.12688/f1000research.73600.2.
[74] C. Ahlmann-Eltze and W. Huber, “Transformation and Preprocessing of Single-Cell RNA-Seq Data,” *BioRxiv*, Jun. 2021, doi: 10.1101/2021.06.24.449781.
[75] Z. Miao, P. Moreno, N. Huang, I. Papatheodorou, A. Brazma, and S. A. Teichmann, “Putative cell type discovery from single-cell gene expression data.,” *Nat. Methods*, vol. 17, no. 6, pp. 621–628, Jun. 2020, doi: 10.1038/s41592-020-0825-9.
[76] C. H. Holland *et al.*, “Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data.,” *Genome Biol.*, vol. 21, no. 1, p. 36, Feb. 2020, doi: 10.1186/s13059-020-1949-z.
[77] S. Aibar *et al.*, “SCENIC: single-cell regulatory network inference and clustering.,” *Nat. Methods*, vol. 14, no. 11, pp. 1083–1086, Nov. 2017, doi: 10.1038/nmeth.4463.
[78] V. Marot-Lassauzaie, B. J. Bouman, F. D. Donaghy, and L. Haghverdi, “Towards reliable quantification of cell state velocities,” *BioRxiv*, Mar. 2022, doi: 10.1101/2022.03.17.484754.
[79] V. Bergen, R. A. Soldatov, P. V. Kharchenko, and F. J. Theis, “RNA velocity-current challenges and future perspectives.,” *Mol. Syst. Biol.*, vol. 17, no. 8, p. e10282, Aug. 2021, doi: 10.15252/msb.202110282.
[80] E. Armingol, A. Officer, O. Harismendy, and N. E. Lewis, “Deciphering cell-cell interactions and communication from gene expression.,” *Nat. Rev. Genet.*, vol. 22, no. 2, pp. 71–88, 2021, doi: 10.1038/s41576-020-00292-x.
[81] D. Dimitrov *et al.*, “Comparison of methods and resources for cell-cell communication inference from single-cell RNA-Seq data.,” *Nat. Commun.*, vol. 13, no. 1, p. 3224, Jun. 2022, doi: 10.1038/s41467-022-30755-0.
[82] H. T. N. Tran *et al.*, “A benchmark of batch-effect correction methods for single-cell RNA sequencing data.,” *Genome Biol.*, vol. 21, no. 1, p. 12, Jan. 2020, doi: 10.1186/s13059-019-1850-9.
[83] M. D. Luecken *et al.*, “Benchmarking atlas-level data integration in single-cell genomics.,” *Nat. Methods*, vol. 19, no. 1, pp. 41–50, Jan. 2022, doi: 10.1038/s41592-021-01336-8.
[84] I. Korsunsky *et al.*, “Fast, sensitive and accurate integration of single-cell data with Harmony.,” *Nat. Methods*, vol. 16, no. 12, pp. 1289–1296, Dec. 2019, doi: 10.1038/s41592-

019-0619-0.
[85] C. Xu, R. Lopez, E. Mehlman, J. Regier, M. I. Jordan, and N. Yosef, “Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models.,” *Mol. Syst. Biol.*, vol. 17, no. 1, p. e9620, 2021, doi: 10.15252/msb.20209620.
[86] V. Petukhov *et al.*, “Case-control analysis of single-cell RNA-seq studies,” *bioRxiv*, Jan. 2022.
[87] M. Büttner, J. Ostner, C. L. Müller, F. J. Theis, and B. Schubert, “scCODA is a Bayesian model for compositional single-cell data analysis.,” *Nat. Commun.*, vol. 12, no. 1, p. 6876, Nov. 2021, doi: 10.1038/s41467-021-27150-6.
[88] H. L. Crowell *et al.*, “muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data.,” *Nat. Commun.*, vol. 11, no. 1, p. 6077, Nov. 2020, doi: 10.1038/s41467-020-19894-4.
[89] J. W. Squair *et al.*, “Confronting false discoveries in single-cell differential expression.,” *Nat. Commun.*, vol. 12, no. 1, p. 5692, Sep. 2021, doi: 10.1038/s41467-021-25960-2.
[90] L. Jerby-Arnon and A. Regev, “DIALOGUE maps multicellular programs in tissue from single-cell or spatial transcriptomics data.,” *Nat. Biotechnol.*, vol. 40, no. 10, pp. 1467–1477, Oct. 2022, doi: 10.1038/s41587-022-01288-0.
[91] J. Mitchel *et al.*, “Tensor decomposition reveals coordinated multicellular patterns of transcriptional variation that distinguish and stratify disease individuals,” *BioRxiv*, Feb. 2022, doi: 10.1101/2022.02.16.480703.
[92] E. Armingol *et al.*, “Context-aware deconvolution of cell-cell communication with Tensor-cell2cell.,” *Nat. Commun.*, vol. 13, no. 1, p. 3665, Jun. 2022, doi: 10.1038/s41467-022-31369-2.
[93] M. Asp, J. Bergenstråhle, and J. Lundeberg, “Spatially Resolved Transcriptomes-Next Generation Tools for Tissue Exploration.,” *Bioessays*, vol. 42, no. 10, p. e1900221, Oct. 2020, doi: 10.1002/bies.201900221.
[94] L. Moses and L. Pachter, “Museum of spatial transcriptomics.,” *Nat. Methods*, vol. 19, no. 5, pp. 534–546, May 2022, doi: 10.1038/s41592-022-01409-2.
[95] F. Salmén *et al.*, “Barcoded solid-phase RNA capture for Spatial Transcriptomics profiling in mammalian tissue sections.,” *Nat. Protoc.*, vol. 13, no. 11, pp. 2501–2534, Nov. 2018, doi: 10.1038/s41596-018-0045-2.
[96] Z. Ni *et al.*, “SpotClean adjusts for spot swapping in spatial transcriptomics data.,” *Nat. Commun.*, vol. 13, no. 1, p. 2971, May 2022, doi: 10.1038/s41467-022-30587-y.
[97] A. Andersson *et al.*, “Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography.,” *Commun. Biol.*, vol. 3, no. 1, p. 565, Oct. 2020, doi: 10.1038/s42003-020-01247-y.
[98] V. Kleshchevnikov *et al.*, “Cell2location maps fine-grained cell types in spatial transcriptomics.,” *Nat. Biotechnol.*, vol. 40, no. 5, pp. 661–671, May 2022, doi: 10.1038/s41587-021-01139-4.
[99] M. Elosua-Bayes, P. Nieto, E. Mereu, I. Gut, and H. Heyn, “SPOTlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes.,” *Nucleic Acids Res.*, vol. 49, no. 9, p. e50, May 2021, doi: 10.1093/nar/gkab043.
[100] R. Lopez *et al.*, “DestVI identifies continuums of cell types in spatial transcriptomics data,” *Nature Biotechnology*, Apr. 2022.
[101] T. Biancalani *et al.*, “Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram.,” *Nat. Methods*, vol. 18, no. 11, pp. 1352–1362, Nov. 2021, doi: 10.1038/s41592-021-01264-7.
[102] K. Li *et al.*, “Computational elucidation of spatial gene expression variation from spatially

resolved transcriptomics data.," *Mol. Ther. Nucleic Acids*, vol. 27, pp. 404–411, Mar. 2022, doi: 10.1016/j.omtn.2021.12.009.
[103] D. Schapiro *et al.*, "histoCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data.," *Nat. Methods*, vol. 14, no. 9, pp. 873–876, Sep. 2017, doi: 10.1038/nmeth.4391.
[104] D. Arnol, D. Schapiro, B. Bodenmiller, J. Saez-Rodriguez, and O. Stegle, "Modeling Cell-Cell Interactions from Spatial Molecular Data with Spatial Variance Component Analysis.," *Cell Rep.*, vol. 29, no. 1, pp. 202-211.e6, Oct. 2019, doi: 10.1016/j.celrep.2019.08.077.
[105] J. Tanevski, R. O. R. Flores, A. Gabor, D. Schapiro, and J. Saez-Rodriguez, "Explainable multiview framework for dissecting spatial relationships from highly multiplexed data.," *Genome Biol.*, vol. 23, no. 1, p. 97, Apr. 2022, doi: 10.1186/s13059-022-02663-5.
[106] D. S. Fischer, A. C. Schaar, and F. J. Theis, "Learning cell communication from spatial graphs of cells," *BioRxiv*, Jul. 2021, doi: 10.1101/2021.07.11.451750.
[107] D. Li, J. Ding, and Z. Bar-Joseph, "Identifying signaling genes in spatial single-cell expression data.," *Bioinformatics*, vol. 37, no. 7, pp. 968–975, May 2021, doi: 10.1093/bioinformatics/btaa769.
[108] Z. Cang and Q. Nie, "Inferring spatial and signaling relationships between cells from single cell transcriptomic data.," *Nat. Commun.*, vol. 11, no. 1, p. 2084, Apr. 2020, doi: 10.1038/s41467-020-15968-5.
[109] X.-L. Meng, "Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election," *Ann. Appl. Stat.*, vol. 12, no. 2, pp. 685–726, Jun. 2018, doi: 10.1214/18-AOAS1161SF.
[110] E. M. DeFilippis *et al.*, "Improving enrollment of underrepresented racial and ethnic populations in heart failure trials: A call to action from the heart failure collaboratory.," *JAMA Cardiol.*, vol. 7, no. 5, pp. 540–548, May 2022, doi: 10.1001/jamacardio.2022.0161.
[111] A. R. Bentley, S. Callier, and C. N. Rotimi, "Diversity and inclusion in genomic research: why the uneven progress?," *J. Community Genet.*, vol. 8, no. 4, pp. 255–266, Oct. 2017, doi: 10.1007/s12687-017-0316-6.
[112] R. O. Ramirez Flores *et al.*, "Consensus Transcriptional Landscape of Human End-Stage Heart Failure.," *J. Am. Heart Assoc.*, vol. 10, no. 7, p. e019667, Apr. 2021, doi: 10.1161/JAHA.120.019667.
[113] J. Yang *et al.*, "Decreased SLIM1 expression and increased gelsolin expression in failing human hearts measured by high-density oligonucleotide arrays.," *Circulation*, vol. 102, no. 25, pp. 3046–3052, Dec. 2000, doi: 10.1161/01.cir.102.25.3046.
[114] R. Raghow, "An 'omics' perspective on cardiomyopathies and heart failure.," *Trends Mol. Med.*, vol. 22, no. 9, pp. 813–827, Sep. 2016, doi: 10.1016/j.molmed.2016.07.007.
[115] G. H. Kim, N. Uriel, and D. Burkhoff, "Reverse remodelling and myocardial recovery in heart failure.," *Nat. Rev. Cardiol.*, vol. 15, no. 2, pp. 83–96, Feb. 2018, doi: 10.1038/nrcardio.2017.139.
[116] A. Alimadadi, P. B. Munroe, B. Joe, and X. Cheng, "Meta-Analysis of Dilated Cardiomyopathy Using Cardiac RNA-Seq Transcriptomic Datasets.," *Genes (Basel)*, vol. 11, no. 1, Jan. 2020, doi: 10.3390/genes11010060.
[117] M. Asakura and M. Kitakaze, "Global gene expression profiling in the failing myocardium.," *Circ. J.*, vol. 73, no. 9, pp. 1568–1576, Sep. 2009.
[118] U. C. Sharma, S. Pokharel, C. T. A. Evelo, and J. G. Maessen, "A systematic review of large scale and heterogeneous gene array data in heart failure.," *J. Mol. Cell. Cardiol.*, vol. 38, no. 3, pp. 425–432, Mar. 2005, doi: 10.1016/j.yjmcc.2004.12.016.
[119] A. S. Barth, A. Kumordzie, C. Frangakis, K. B. Margulies, T. P. Cappola, and G. F.

Tomaselli, “Reciprocal transcriptional regulation of metabolic and signaling pathways correlates with disease severity in heart failure.,” *Circ. Cardiovasc. Genet.*, vol. 4, no. 5, pp. 475–483, Oct. 2011, doi: 10.1161/CIRCGENETICS.110.957571.

[120] Y. Liu *et al.*, “RNA-Seq identifies novel myocardial gene expression signatures of heart failure.,” *Genomics*, vol. 105, no. 2, pp. 83–89, Feb. 2015, doi: 10.1016/j.ygeno.2014.12.002.

[121] S. Hannenhalli *et al.*, “Transcriptional genomics associates FOX transcription factors with human heart failure.,” *Circulation*, vol. 114, no. 12, pp. 1269–1276, Sep. 2006, doi: 10.1161/CIRCULATIONAHA.106.632430.

[122] S. van Heesch *et al.*, “The translational landscape of the human heart.,” *Cell*, vol. 178, no. 1, pp. 242-260.e29, Jun. 2019, doi: 10.1016/j.cell.2019.05.010.

[123] M. E. Sweet *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*, vol. 19, no. 1, p. 812, Nov. 2018, doi: 10.1186/s12864-018-5213-9.

[124] M. M. Kittleson *et al.*, “Gene expression analysis of ischemic and nonischemic cardiomyopathy: shared and distinct genes in the development of heart failure.,” *Physiol. Genomics*, vol. 21, no. 3, pp. 299–307, May 2005, doi: 10.1152/physiolgenomics.00255.2004.

[125] E. Tarazón *et al.*, “RNA sequencing analysis and atrial natriuretic peptide production in patients with dilated and ischemic cardiomyopathy.,” *PLoS ONE*, vol. 9, no. 3, p. e90157, Mar. 2014, doi: 10.1371/journal.pone.0090157.

[126] C. H. Spurrell *et al.*, “Genome-Wide Fetalization of Enhancer Architecture in Heart Disease,” *BioRxiv*, Apr. 2019, doi: 10.1101/591362.

[127] S. W. Kong *et al.*, “Heart failure-associated changes in RNA splicing of sarcomere genes.,” *Circ. Cardiovasc. Genet.*, vol. 3, no. 2, pp. 138–146, Apr. 2010, doi: 10.1161/CIRCGENETICS.109.904698.

[128] M. M. Molina-Navarro *et al.*, “Differential gene expression of cardiac ion channels in human dilated cardiomyopathy.,” *PLoS ONE*, vol. 8, no. 12, p. e79792, Dec. 2013, doi: 10.1371/journal.pone.0079792.

[129] S. Greco *et al.*, “MicroRNA dysregulation in diabetic ischemic heart failure patients.,” *Diabetes*, vol. 61, no. 6, pp. 1633–1641, Jun. 2012, doi: 10.2337/db11-0952.

[130] K.-C. Yang *et al.*, “Deep RNA sequencing reveals dynamic regulation of myocardial noncoding RNAs in failing human heart and remodeling with mechanical circulatory support.,” *Circulation*, vol. 129, no. 9, pp. 1009–1021, Mar. 2014, doi: 10.1161/CIRCULATIONAHA.113.003863.

[131] A. S. Barth *et al.*, “Identification of a common gene expression signature in dilated cardiomyopathy across independent microarray studies.,” *J. Am. Coll. Cardiol.*, vol. 48, no. 8, pp. 1610–1617, Oct. 2006, doi: 10.1016/j.jacc.2006.07.026.

[132] M. E. Pepin *et al.*, “DNA methylation reprograms cardiac metabolic gene expression in end-stage human heart failure.,” *Am. J. Physiol. Heart Circ. Physiol.*, vol. 317, no. 4, pp. H674–H684, Oct. 2019, doi: 10.1152/ajpheart.00016.2019.

[133] E. H. Kim *et al.*, “Differential protein expression and basal lamina remodeling in human heart failure.,” *Proteomics Clin. Appl.*, vol. 10, no. 5, pp. 585–596, May 2016, doi: 10.1002/prca.201500099.

[134] C. Schiano *et al.*, “Heart failure: Pilot transcriptomic analysis of cardiac tissue by RNA-sequencing.,” *Cardiol. J.*, vol. 24, no. 5, pp. 539–553, May 2017, doi: 10.5603/CJ.a2017.0052.

[135] E. Petretto *et al.*, “Integrated genomic approaches implicate osteoglycin (Ogn) in the regulation of left ventricular mass.,” *Nat. Genet.*, vol. 40, no. 5, pp. 546–552, May 2008, doi:

10.1038/ng.134.
[136] F. Wittchen *et al.*, “Genomic expression profiling of human inflammatory cardiomyopathy (DCMi) suggests novel therapeutic targets.,” *J. Mol. Med.*, vol. 85, no. 3, pp. 257–271, Mar. 2007, doi: 10.1007/s00109-006-0122-9.
[137] L. Laugier *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.*, vol. 65, no. 7, pp. 1103–1111, Oct. 2017, doi: 10.1093/cid/cix506.
[138] K. M. Akat *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*, vol. 111, no. 30, pp. 11151–11156, Jul. 2014, doi: 10.1073/pnas.1401724111.
[139] P. W. Angel *et al.*, “A simple, scalable approach to building a cross-platform transcriptome atlas.,” *PLoS Comput. Biol.*, vol. 16, no. 9, p. e1008219, Sep. 2020, doi: 10.1371/journal.pcbi.1008219.
[140] B. S. Carvalho and R. A. Irizarry, “A framework for oligonucleotide microarray preprocessing.,” *Bioinformatics*, vol. 26, no. 19, pp. 2363–2367, Oct. 2010, doi: 10.1093/bioinformatics/btq431.
[141] A. Lachmann *et al.*, “Massive mining of publicly available RNA-seq data from human and mouse.,” *Nat. Commun.*, vol. 9, no. 1, p. 1366, Apr. 2018, doi: 10.1038/s41467-018-03751-6.
[142] Y. Liao, G. K. Smyth, and W. Shi, “The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads.,” *Nucleic Acids Res.*, vol. 47, no. 8, p. e47, May 2019, doi: 10.1093/nar/gkz114.
[143] H. Choi, R. Shen, A. M. Chinnaiyan, and D. Ghosh, “A latent variable approach for meta-analysis of gene expression data from multiple microarray experiments.,” *BMC Bioinformatics*, vol. 8, p. 364, Sep. 2007, doi: 10.1186/1471-2105-8-364.
[144] A. Subramanian *et al.*, “Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.,” *Proc Natl Acad Sci USA*, vol. 102, no. 43, pp. 15545–15550, Oct. 2005, doi: 10.1073/pnas.0506580102.
[145] L.-C. Chang, H.-M. Lin, E. Sibille, and G. C. Tseng, “Meta-analysis methods for combining multiple expression profiles: comparisons, statistical characterization and an application guideline.,” *BMC Bioinformatics*, vol. 14, p. 368, Dec. 2013, doi: 10.1186/1471-2105-14-368.
[146] A. Liberzon, C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. P. Mesirov, and P. Tamayo, “The Molecular Signatures Database (MSigDB) hallmark gene set collection.,” *Cell Syst.*, vol. 1, no. 6, pp. 417–425, Dec. 2015, doi: 10.1016/j.cels.2015.12.004.
[147] G. Korotkevich, V. Sukhov, N. Budin, B. Shpak, M. N. Artyomov, and A. Sergushichev, “Fast gene set enrichment analysis,” *BioRxiv*, Jun. 2016, doi: 10.1101/060012.
[148] M. J. Alvarez *et al.*, “Functional characterization of somatic mutations in cancer using network-based inference of protein activity.,” *Nat. Genet.*, vol. 48, no. 8, pp. 838–847, Aug. 2016, doi: 10.1038/ng.3593.
[149] C. H. Holland, B. Szalai, and J. Saez-Rodriguez, “Transfer of regulatory knowledge from human to mouse for functional genomics analysis.,” *Biochim. Biophys. Acta Gene Regul. Mech.*, vol. 1863, no. 6, p. 194431, Jun. 2020, doi: 10.1016/j.bbagrm.2019.194431.
[150] A. Egerstedt *et al.*, “Profiling of the plasma proteome across different stages of human heart failure.,” *Nat. Commun.*, vol. 10, no. 1, p. 5830, Dec. 2019, doi: 10.1038/s41467-019-13306-y.
[151] A. Athar *et al.*, “ArrayExpress update - from bulk to single-cell expression data.,” *Nucleic Acids Res.*, vol. 47, no. D1, pp. D711–D715, Jan. 2019, doi: 10.1093/nar/gky964.

[152] T. Barrett *et al.*, “NCBI GEO: archive for functional genomics data sets--update.,” *Nucleic Acids Res.*, vol. 41, no. Database issue, pp. D991-5, Jan. 2013, doi: 10.1093/nar/gks1193.
[153] C. Cummins *et al.*, “The european nucleotide archive in 2021.,” *Nucleic Acids Res.*, vol. 50, no. D1, pp. D106–D110, Jan. 2022, doi: 10.1093/nar/gkab1051.
[154] L. Collado-Torres *et al.*, “Reproducible RNA-seq analysis using recount2.,” *Nat. Biotechnol.*, vol. 35, no. 4, pp. 319–321, Apr. 2017, doi: 10.1038/nbt.3838.
[155] C. Kuppe *et al.*, “Spatial multi-omic map of human myocardial infarction.,” *Nature*, vol. 608, no. 7924, pp. 766–777, Aug. 2022, doi: 10.1038/s41586-022-05060-x.
[156] T. J. Cahill and R. K. Kharbanda, “Heart failure after myocardial infarction in the era of primary percutaneous coronary intervention: Mechanisms, incidence and identification of patients at risk.,” *World J. Cardiol.*, vol. 9, no. 5, pp. 407–415, May 2017, doi: 10.4330/wjc.v9.i5.407.
[157] M. Litviňuková *et al.*, “Cells of the adult human heart.,” *Nature*, vol. 588, no. 7838, pp. 466–472, Dec. 2020, doi: 10.1038/s41586-020-2797-4.
[158] M. Chaffin *et al.*, “Single-nucleus profiling of human dilated and hypertrophic cardiomyopathy.,” *Nature*, vol. 608, no. 7921, pp. 174–180, Aug. 2022, doi: 10.1038/s41586-022-04817-8.
[159] D. Reichart *et al.*, “Pathogenic variants damage cell composition and single cell transcription in cardiomyopathies.,” *Science*, vol. 377, no. 6606, p. eabo1984, Aug. 2022, doi: 10.1126/science.abo1984.
[160] L. S. Tombor *et al.*, “Single cell sequencing reveals endothelial plasticity with transient mesenchymal activation after myocardial infarction.,” *Nat. Commun.*, vol. 12, no. 1, p. 681, Jan. 2021, doi: 10.1038/s41467-021-20905-1.
[161] J. Zhu, S. Sun, and X. Zhou, “SPARK-X: non-parametric modeling enables scalable and robust detection of spatial expression patterns for large spatial transcriptomic studies.,” *Genome Biol.*, vol. 22, no. 1, p. 184, Jun. 2021, doi: 10.1186/s13059-021-02404-0.
[162] R. Foreman and R. Wollman, “Mammalian gene expression variability is explained by underlying cell state.,” *Mol. Syst. Biol.*, vol. 16, no. 2, p. e9146, Feb. 2020, doi: 10.15252/msb.20199146.
[163] O. Symmons and A. Raj, “What’s Luck Got to Do with It: Single Cells, Multiple Fates, and Biological Nondeterminism.,” *Mol. Cell*, vol. 62, no. 5, pp. 788–802, Jun. 2016, doi: 10.1016/j.molcel.2016.05.023.
[164] B. A. Kramer and L. Pelkmans, “Cellular state determines the multimodal signaling response of single cells,” *BioRxiv*, Dec. 2019, doi: 10.1101/2019.12.18.880930.
[165] B. A. Kramer, J. Sarabia Del Castillo, and L. Pelkmans, “Multimodal perception links cellular state to decision-making in single cells.,” *Science*, vol. 377, no. 6606, pp. 642–648, Aug. 2022, doi: 10.1126/science.abf4062.
[166] Y. Hao *et al.*, “Integrated analysis of multimodal single-cell data,” *BioRxiv*, Oct. 2020, doi: 10.1101/2020.10.12.335331.
[167] F. A. Wolf, P. Angerer, and F. J. Theis, “SCANPY: large-scale single-cell gene expression data analysis,” *Genome Biol.*, vol. 19, no. 1, p. 15, Feb. 2018, doi: 10.1186/s13059-017-1382-0.
[168] C. Hafemeister and R. Satija, “Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression.,” *Genome Biol.*, vol. 20, no. 1, p. 296, Dec. 2019, doi: 10.1186/s13059-019-1874-1.
[169] M. Gillespie *et al.*, “The reactome pathway knowledgebase 2022.,” *Nucleic Acids Res.*, vol. 50, no. D1, pp. D687–D692, Jan. 2022, doi: 10.1093/nar/gkab1028.
[170] A. O. Grant, “Cardiac ion channels.,” *Circ. Arrhythm. Electrophysiol.*, vol. 2, no. 2, pp. 185–

194, Apr. 2009, doi: 10.1161/CIRCEP.108.789081.
[171] V. Pawlowsky-Glahn and A. Buccianti. “Compositional Data Analysis: Theory and Applications”. *John Wiley & Sons*, Sept. 2011. doi: 10.1002/9781119976462
[172] A. T. L. Lun, D. J. McCarthy, and J. C. Marioni, “A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. [version 2; peer review: 3 approved, 2 approved with reservations],” *F1000Res.*, vol. 5, p. 2122, Aug. 2016, doi: 10.12688/f1000research.9501.2.
[173] L. Wang *et al.*, “Single-cell reconstruction of the adult human heart during heart failure and recovery reveals the cellular landscape underlying cardiac function.,” *Nat. Cell Biol.*, vol. 22, no. 1, pp. 108–119, Jan. 2020, doi: 10.1038/s41556-019-0446-7.
[174] N. R. Tucker *et al.*, “Transcriptional and cellular diversity of the human heart.,” *Circulation*, vol. 142, no. 5, pp. 466–482, Aug. 2020, doi: 10.1161/CIRCULATIONAHA.119.045401.
[175] T. Aoyagi and T. Matsui, “Phosphoinositide-3 kinase signaling in cardiac hypertrophy and heart failure.,” *Curr. Pharm. Des.*, vol. 17, no. 18, pp. 1818–1824, 2011, doi: 10.2174/138161211796390976.
[176] T. W. Mak, L. Hauck, D. Grothe, and F. Billia, “p53 regulates the cardiac transcriptome.,” *Proc Natl Acad Sci USA*, vol. 114, no. 9, pp. 2331–2336, Feb. 2017, doi: 10.1073/pnas.1621436114.
[177] E. P. Daskalopoulos, K. C. M. Hermans, B. J. A. Janssen, and W. Matthijs Blankesteijn, “Targeting the Wnt/frizzled signaling pathway after myocardial infarction: a new tool in the therapeutic toolbox?,” *Trends Cardiovasc. Med.*, vol. 23, no. 4, pp. 121–127, May 2013, doi: 10.1016/j.tcm.2012.09.010.
[178] M. W. Bergmann, “WNT signaling in adult cardiac hypertrophy and remodeling: lessons learned from cardiac development.,” *Circ. Res.*, vol. 107, no. 10, pp. 1198–1208, Nov. 2010, doi: 10.1161/CIRCRESAHA.110.223768.
[179] R. Argelaguet *et al.*, “MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data.,” *Genome Biol.*, vol. 21, no. 1, p. 111, May 2020, doi: 10.1186/s13059-020-02015-1.
[180] L. R. Tucker, “Some mathematical notes on three-mode factor analysis,” *Psychometrika*, vol. 31, no. 3, pp. 279–311, Sep. 1966, doi: 10.1007/BF02289464.
[181] F. García-Lamont, J. Cervantes, A. López-Chau, and S. Ruiz-Castilla, “Color image segmentation using saturated RGB colors and decoupling the intensity from the hue,” *Multimed. Tools Appl.*, vol. 79, no. 1–2, pp. 1555–1584, Jan. 2020, doi: 10.1007/s11042-019-08278-6.
[182] P. Boyeau, J. Hong, A. Gayoso, M. Jordan, E. Azizi, and N. Yosef, “Deep generative modeling for quantifying sample-level heterogeneity in single-cell omics,” *BioRxiv*, Oct. 2022, doi: 10.1101/2022.10.04.510898.
[183] B. Velten *et al.*, “Identifying temporal and spatial patterns of variation from multimodal data using MEFISTO.,” *Nat. Methods*, vol. 19, no. 2, pp. 179–186, Feb. 2022, doi: 10.1038/s41592-021-01343-9.
[184] A. Qoku and F. Buettner, “Encoding Domain Knowledge in Multi-view Latent Variable Models: A Bayesian Approach with Structured Sparsity,” *arXiv*, Apr. 2022.