TUM School of Computation,
Information and Technology

TECHNISCHE UNIVERSITÄT MÜNCHEN

Dissertation

Multi-omics integration for Atrial Fibrillation

Ines Marion Assum

April 2022

TUM School of Computation, Information and Technology

TECHNISCHE UNIVERSITÄT MÜNCHEN

Multi-omics integration for Atrial Fibrillation

Ines Marion Assum

Vollständiger Abdruck der von der TUM School of Computation, Information and Technology der Technischen Universität München zur Erlangung des akademischen Grades einer

Doktorin der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzende:	Prof. Dr. Julia Schnabel
Prüfer der Dissertation:	1. TUM Junior Fellow Dr. Matthias Heinig 2. Prof. Dr. Julien Gagneur

Die Dissertation wurde am 06.04.2022 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 29.07.2022 angenommen.

Acknowledgment

I would like to thank

... my supervisor Matthias Heinig for his continuous support, his vision and ideas, for all the discussions and for the great research topic that allowed me to learn so much over the past few years.

... the Heinig Lab, especially Kathi, Hans, Toray, Barbara and Simon, for being such an awesome group, for all the feedback on half-baked presentations, the great input and advice and all the discussions over coffee and tea.

... Julia Krause, the best collaborator, for the great teamwork, for always being understanding, answering all my questions, for all the hours in the lab creating the data and for even more hours discussing the results.

... Tanja Zeller and Renate Schnabel for their dedication to great interdisciplinary work, welcoming me into the team and giving me the opportunity to work on such a unique cohort.

... all the people at ICB that have become more than colleagues to me, especially Paul and Maren and the old office with Valle, Hans and Turid.

... Anna Sacher, the best science manager ever and all the office staff making our lives easier.

... Fabian Theis for growing this amazing institute with such a great work environment.

... Joseph Högel and Herbert Heinz for giving me the opportunity of an internship and a bachelor thesis in such a great research environment, and for teaching me all the basics I needed for my PhD.

... Michael Menden and the Menden Lab for all their support in the last phase of my thesis.

Finally, I would like to thank my family. Thank you, Brigitte and Günter, for taking care of us, for providing a safe heaven, and for being there whenever we needed it most. And most of all, I would like to thank my husband Johannes. Thank you for all the good times, for always being understanding even when I was difficult. Without you, your kindness, your patience and support, I could not have made it through the hard times.

Abstract

Genome-wide association studies (GWAS) for atrial fibrillation (AF) have uncovered numerous disease-associated variants. Their underlying molecular mechanisms, especially consequences for mRNA and protein expression remain largely elusive. Thus, novel multi-omics approaches are needed for deciphering the underlying molecular networks.

Making use of the deeply-phenotyped AFHRI-B cohort, we integrated genomics, transcriptomics and proteomics to assess genome-wide consequences of common genetic variation for transcript and protein abundance of nearby genes by evaluating *cis* expression quantitative trait loci (eQTLs) and *cis* protein quantitative trait loci (pQTLs).
By the integration of multiple omics-modalities we identified three functional *cis* QTL categories representing shared regulation affecting both transcripts and proteins, regulation of only the mRNA level (independent *cis* eQTLs) and variation that is only observed on protein level (independent *cis* pQTLs). Additionally, we validated our findings with classical colocalization analyses. Using public annotations, we assessed overrepresented regulatory elements to infer possible mechanisms underlying the genetic regulation.
Finally, we confirmed the relevance of the discovered QTLs by an overrepresentation of *cis* eQTLs and pQTLs for GWAS loci associated with general cardiovascular disease, arrhythmias, and specifically, atrial fibrillation.

Current research suggests, however, that *cis*-genetic variation only accounts for a modest fraction of disease heritability compared to much larger contributions of *trans*-genetic effects. As part of the omnigenic model, it has been hypothesized that those *trans* effects can accumulate on few genes that are directly linked to the phenotype in question. According to this model, we developed a pathway enrichment approach to identify candidates satisfying the core gene properties in order to narrow down the search space for *trans* QTL testing.
Specifically, we used a genome-wide polygenic risk score (PRS) as a proxy for *trans*-genetic effects. While correcting for the strongest *cis* effects, we assessed the correlation between PRS and transcript and protein expression (eQTS/pQTS) to rank genes accordingly. As potential core genes are expected to share molecular function, we used Gene Set Enrichment Analysis (GSEA) together with the Gene Ontology gene set annotations in order to reconstruct possible biological networks.The resulting leading edge genes of significantly enriched pathways were subsequently selected as candidates. The genetic disease link was then evaluated by *trans* eQTL/pQTL testing with AF GWAS hits for only this small subset.

Using this novel approach, we identified two *trans* eQTLs and five *trans* pQTLs. One of the *trans* eQTLs was the transcription factor (TF) NKX2-5, where we did extensive follow-up analyses including the replication of the corresponding pQTL on protein level, analyzing the TF network and its functional targets.
For many of our putative core genes we were able to find further literature support of the disease link to AF or a related cardiovascular disease.

Finally, we took a closer look at molecular pathophysiology of AF via traditional differential transcriptome and proteome analysis. Due to generally low effect sizes, we further investigated possibilities of multi-omics pathway enrichment analysis in order to make full use of this dataset by accumulating information about regulated processes across different genes and modalities.
Since most pathway enrichment methods focus on only one data type, we conducted a simulation study, which compared different multi-omics pathway enrichment approaches as well as ad-hoc combinations of analyses of single data modalities. While using multi-omics data in general, as well as including direction of effect greatly improved model performance, methods that directly integrate multiple omics only slightly outperformed the ad-hoc combination in the case of similar pathway activations across the different omics.
Using those insights, we identified a broad range of mechanisms, including cardiac, metabolic and immune pathways, for different AF subtypes which were in line with previous literature. Additionally, we observed regulation which was specific to either transcriptomics or proteomics.
In order to make similar analyses more accessible, we also provided the EnrichmentNodes plugin for KNIME, an interactive graphical workflow development framework for combining different analysis in a user-friendly way without any actual programming.

Zusammenfassung

Genomweite Assoziationsstudien (GWAS) für Vorhofflimmern haben bereits zahlreiche krankheitsassoziierte Varianten aufgedeckt. Die zugrunde liegenden molekularen Mechanismen, insbesondere die Auswirkungen auf die mRNA- und Proteinexpression, sind jedoch noch weitgehend ungeklärt. Daher sind neue Multi-omics-Ansätze erforderlich, um die zugrunde liegenden molekularen Netzwerke zu entschlüsseln.

Basierend auf den Daten der vielfältig charakterisierten AFHRI-B-Kohorte konnten Genomik, Transkriptomik und Proteomik kombiniert werden, um die genomweiten Auswirkungen genetischer Varianten auf die Transkript- und Proteinvariation nahe gelegener Gene durch die Auswertung von *cis* eQTLs und *cis* pQTLs zu bewerten. Die Integration mehrerer Datentypen ermöglichte die Identifikation dreier funktioneller *cis*-QTL-Kategorien, die entweder eine gemeinsame Regulation über Transkriptom und Proteom hinweg darstellen oder eine Regulation, welche nur auf mRNA-Ebene (unabhängige *cis* eQTLs) beziehungsweise nur auf Proteinebene (unabhängige *cis* pQTLs) beobachtet wird. Anschließend wurden die Ergebnisse mit klassischen Kolokalisierungsanalysen validiert. Anhand von öffentlichen Annotationen identifizierten wir überrepräsentierte regulatorische Elemente, um mögliche Mechanismen abzuleiten, die der genetischen Regulation zugrunde liegen.
Die Relevanz der entdeckten QTLs konnte schließlich mit einer Überrepräsentation von *cis* eQTLs und pQTLs für solche GWAS-Loci bestätigt werden, die mit allgemeinen Herz-Kreislauf-Erkrankungen, Herzrhythmusstörungen und insbesondere Vorhofflimmern in Verbindung stehen.

Aktuelle Forschungsergebnisse deuten jedoch darauf hin, dass die *cis*-genetische Variation nur einen relativ kleinen Teil zur Erblichkeit von Krankheiten beiträgt, während *trans*-genetische Effekte einen wesentlich höheren Anteil erklären. Im Rahmen des *omnigenic model* wurde die Hypothese aufgestellt, dass sich genetische *trans*-Effekte auf wenigen Genen akkumulieren können, die direkte Auswirkungen auf den betreffenden Phänotyp haben. Auf Grundlage dieses Modells entwickelten wir einen Ansatz, um Kandidaten basierend auf diesen zentralen Eigenschaften zu priorisieren und auf diese Weise den Suchraum für *trans*-QTL-Tests einzugrenzen.
Konkret verwendeten wir einen genomweiten polygenen Risikoscore (PRS), der die *trans*-genetischen Effekte repräsentiert. Indem wir zudem für die stärksten *cis*-Effekte korrigierten, konnte die Korrelation des PRS mit Transkript- sowie Proteinexpression (eQTS bzw. pQTS) genutzt werden, um die Gene entsprechend einzustufen. Des Weiteren ist zu erwarten, dass entsprechende Gene im Zentrum molekularer Prozesse eingebunden sind. Deshalb wurde die Gene Set Enrichment Analyse (GSEA) auf den

so gewonnenen Rangfolgen zusammen mit Gene ontology (GO) Genannotationen genutzt, um mögliche biologische Netzwerke zu rekonstruieren. Gene mit dem größten Beitrag wurden sodann als potenzielle Kandidaten ausgewählt und der jeweilige Krankheitszusammenhang schließlich über *trans* eQTL/pQTL-Analysen mit AF-assoziierten Genorten verifiziert.
Mit diesem neuartigen Ansatz identifizierten wir zwei *trans* eQTLs und fünf *trans* pQTLs. Bei einem der *trans* eQTLs handelte es sich um den Transkriptionsfaktor (TF) NKX2-5, dessen zugehöriges pQTL auf Proteinebene repliziert sowie das TF-Netzwerk und dessen funktionelle Zielgene analysiert wurden.
Für viele unserer potentiellen *core genes* konnten wir in der Literatur weitere Belege für einen Zusammenhang mit Vorhofflimmern oder einer verwandten Herz-Kreislauf-Erkrankung finden.

Schließlich untersuchten wir die molekulare Pathophysiologie von Vorhofflimmern mithilfe traditioneller differenzieller Transkriptom- und Proteomanalysen. Aufgrund der allgemein geringen Effektgrößen nahmen wir mögliche Veränderungen der zugehörigen Signalwege in den Blick um Effekte ähnlicher Mechanismen über verschiedene Gene und Modalitäten hinweg zu integrieren.
Da sich die meisten Pathway Enrichment Methoden nur auf eine molekulare Modalität beschränken, verglichen wir mit Hilfe einer Simulationsstudie verschiedene Ansätze zur direkten Integration oder ad-hoc Kombination mehrerer Datentypen. Die besten Ergebnisse wurden dabei unter Einbeziehung mehrer Datentypen und speziell auch der Effektrichtung erzielt. Die direkte Integration mehrerer Modalitäten zeigte gegenüber der ad-hoc-Kombination nur im Falle ähnlicher Aktivierungen in den verschiedenen Omics einen leichten Vorteil.
Mithilfe dieser Erkenntnisse konnten wir ein breites Spektrum von regulierten kardialen, metabolischen und immunologischen Prozessen für beide Subtypen des Vorhofflimmers identifizieren. Entsprechende Mechanismen stimmten in hohem Maße mit existierender Literatur überein und beinhalteten jeweils Prozesse, die spezifisch nur im Transkriptom oder Proteom zu beobachten waren.
Wir stellten außerdem eine Erweiterung EnrichmentNodes des KNIME Frameworks zur Verfügung, welche die Integration ähnlicher Analyseschritte in einer interaktiven, grafischen Entwicklungsumgebung von Arbeitsabläufen möglich macht ohne dass hierfür Programmierkenntnisse notwendig sind.

Contents

1 Introduction

Atrial fibrillation (AF) is the most common cardiac arrhythmia which affects more than 33 million individuals worldwide [Chugh et al., 2014]. For individuals of age 55 and older, the life time risk of AF is estimated at one out of three individuals [Magnussen et al., 2017, Staerk et al., 2017].

AF is characterized by a distorted electrophysiological signal transduction in the human heart, leading to an irregular and often fast rhythm. The disorganized excitation of the atrium prohibits proper propagation to the ventricle, which leads to constraint contractions (Figure 1.1). Together with other comorbidities such as heart failure and myocardial infarction, AF significantly increases the risk of stroke and death [Staerk et al., 2017].

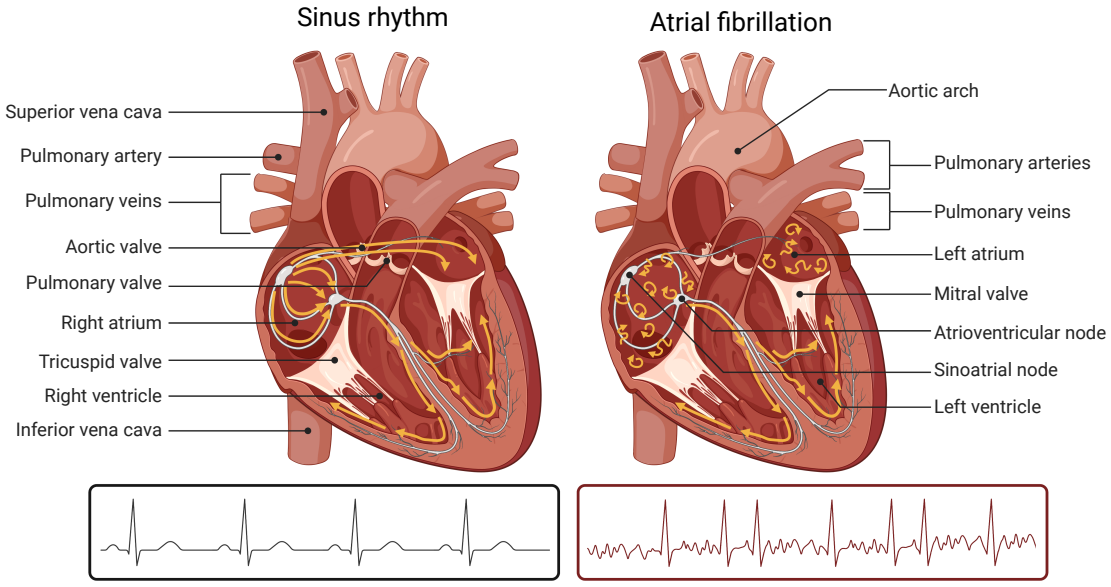


Figure 1.1: Disrupted signal transduction in atrial fibrillation.
Schematic representation of the electrical signal conduction from the sinoatrial node through the atria and atrioventricular node to the ventricles. Asynchronous electrical signals prevent proper stimulation of the ventricles, resulting in an irregular and rapid rhythm as depicted in the symbolic electrocardiographic readings. Figure created with BioRender.com using the template "Heart Anatomy" (Jean-Francis Berry).

Important risk factors for AF are age, sex, genetics and ethnicity as well as common comorbidities such as hypertension, diabetes mellitus, coronary artery disease, chronic kidney disease and obesity [Magnussen et al., 2017, Staerk et al., 2017, Hindricks et al., 2021].

1.1 Systems biology of atrial fibrillation

Extensive research [Brundel et al., 2002, Corradi et al., 2008, Dobrev and Nattel, 2011, Lin et al., 2014, Opacic et al., 2016, Sigurdsson et al., 2017, Magnussen et al., 2017, Kanaan et al., 2019, Dobrev et al., 2019, Thomas et al., 2019, Nattel et al., 2020, Van Ouwerkerk et al., 2020, Hindricks et al., 2021] investigating the molecular mechanisms underlying AF as well as corresponding risk factors has been carried out over the last decades. Due to the complex nature of human disease, interdisciplinary research integrating clinical, molecular and genetic data with computational approaches are needed to further our understanding of the pathophysiology of AF.

More than 100 genetic loci have been identified. Their functional mechanisms, however, remain largely unknown [Roselli et al., 2018]. Downstream changes are in general highly specific to the tissue of question. Left and right atrial appendage tissue has been studied, which revealed de-regulated molecular processes such as structural remodeling of the atrial substrate, changes in metabolism and inflammatory reactions including fibrosis [Hindricks et al., 2021, Staerk et al., 2017, Hu et al., 2015, Tu et al., 2014].

Still, we are lacking a conclusive understanding of comprehensive disease mechanisms, which would be highly relevant to developing and improving treatment options and preventative measures [Hindricks et al., 2021].

1.1.1 Deeply phenotyped atrial fibrillation cohort

Disease-relevant molecular investigations are strongly restricted by the accessibility of relevant heart tissue which leads to an inherently small sample size. With the exception of post-mortem specimens, donor tissue can only be procured from individuals with other cardiac conditions requiring heart surgery.

In order to assess relevant mechanisms at different steps of gene regulation, multiple intermediate molecular data types need to be taken into account. Following this line of action, extensive phenotyping was carried out on right atrial appendage tissue samples from donors undergoing coronary artery bypass surgery. Together with clinical data on AF and covariates, this represents a unique cohort integrating genomics, transcriptomics, proteomics and metabolomics data (Figure 1.2).

1.1.2 Tissue-specific gene regulation and omics measurement techniques

Genetic variants can influence biological processes in various ways. First of all, variants can change the amino acid sequence of a protein and therefore influence functionality. However, only a very small fraction of the human genome contains protein-coding information and also variants outside of these regions, also referred to as non-coding, have been found to play a vital role in gene regulation on diverse levels.

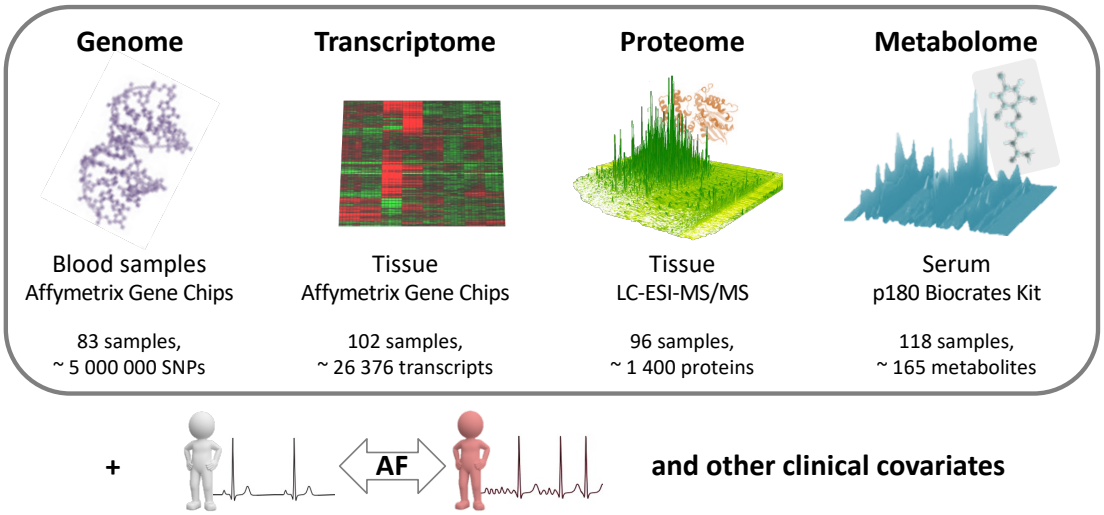


Figure 1.2: Overview of the deeply phenotyped AFHRI-B cohort.
Heart atrial appendage samples were retrieved during coronary artery bypass surgery and used for transcriptomics, proteomics and metabolomics profiling. Baseline blood samples were used for genotyping and serum metabolomics. Figure created by Ines Assum and Julia Krause.

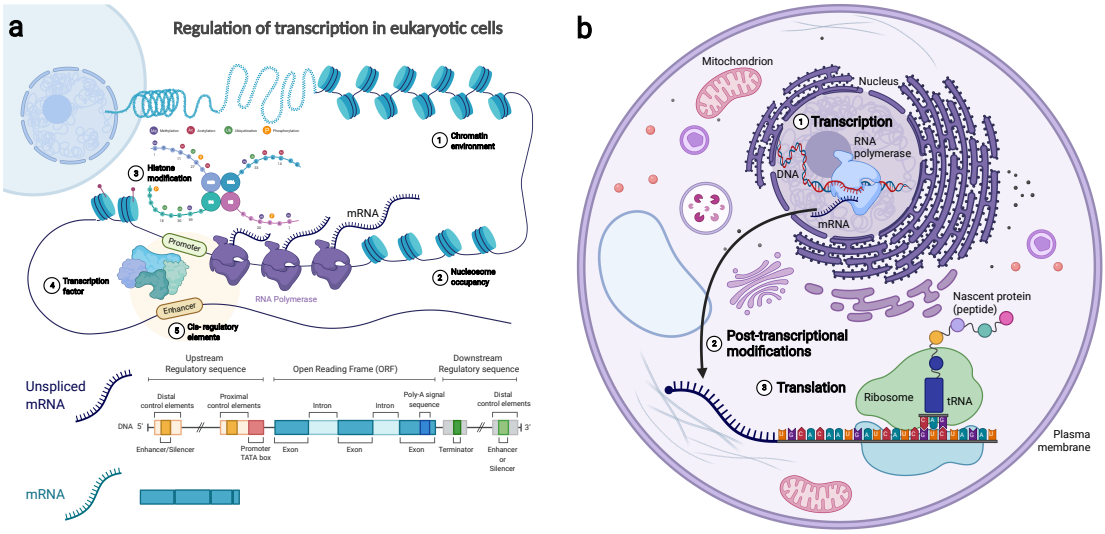


Figure 1.3: Gene regulation in human cells.
a: Schematic representation of transcriptional regulation.
b: Protein biosynthesis, where DNA is transcribed into RNA, processed with post-transcriptional regulation and translated into proteins.
Figures created with BioRender.com using the templates "Regulation of Transcription in Eukaryotic Cells" (Biljana Atanasovska, PhD), "Histone Modification, Eukaryotic and Prokaryotic Gene Structure" and "Structural Overview of an Animal Cell".

In order to investigate how genetic variants influence gene regulation, it is important to understand the different steps of protein bio synthesis as summarized in Figure 1.3. Vital steps in a human eukaryotic cell are transcription, i.e. reading the genetic code from the deoxyribonucleic acid (DNA) to ribonucleic acid (RNA), and translation, i.e. translating mature mRNA into proteins. In a very brief and simplified description, chromatin consist of DNA which is condensed and wrapped around histones in the nucleus. Epigenetic modifications of those histones can, amongst other factors, influence accessibility of the DNA. To transcribe the RNA from the DNA of a gene, protein complexes called the RNA polymerase bind the DNA in specific sequence regions called promoters to initiate transcription. A typical human protein coding gene is divided into multiple functional regions, starting at the promoter and transcription start site (TSS), followed by the 5′ untranslated region (5′ UTR), a sequence of exons that contain the protein-coding information and introns with non-coding sections completed by the poly-A tail, a terminator sequence and finally the 3′ untranslated region (3′ UTR) (Figure 1.3a).
After transcription, the unspliced mRNA is transported out of the nucleus into the cytoplasm, processed by splicing, i.e. keeping only the exons, and potentially modified post-transcriptionally. This product is then referred to as mature mRNA or transcript. Finally, ribosomes translate mRNA by adding reading out base triplets and accordingly, adding the corresponding amino-acid to a nascent protein chain, so called peptides. Most of the time, multiple subunits with specific processing steps, modifications and folding for each of them are needed for a functional protein (Figure 1.3b).
All of these processes are very complex and highly regulated by diverse genetic and epigenetic mechanisms, and many of them are still poorly understood [Buccitelli and Selbach, 2020]. Regulatory elements, such as binding sites, enhancers and silencers, can be placed in the UTR or upstream and downstream of the gene.

Therefore, new measurement techniques have been introduced to observe and investigate intermediate molecular species. Genetic variation is often considered by evaluating single-nucleotide-polymorphism (SNP) which are regions in the genome, that differ only by one specific base. SNP genotypes and transcriptomics can be measured by specific microarrays. Microarrays consists of probes that omit a fluorescent signal if they are bound by their specific counterparts. Compared to that, RNA-sequencing (RNA-seq) enables actual sequencing of the RNA or also DNA fragments in form of complementary DNA (cDNA) [Wang et al., 2009].
To evaluate the interaction of proteins and DNA, DNA and associated proteins are cross-linked and fragmented. Using specific antibodies, these fragments are selected and enriched for a protein of interest. Massively parallel sequencing is performed to then identify the DNA regions bound by the protein of interest. The method is denoted by the name chromatin immunoprecipitation followed by sequencing (ChIP-seq) [Park, 2009]. Following a similar idea, promoter-capture Hi-C is used to map DNA-DNA interactions by ligating and then sequencing interacting DNA regions [Davies et al., 2017].
Finally, untargeted and targeted mass spectrometry methods can be used to quantify proteomics [Westont and Hood, 2004] and metabolomics [Baharum and Azizan, 2018].

1.2 Systems genetics approach to human disease

1.2.1 Genetics, genome-wide associations studies and complex traits

Genetic information in humans is organized as DNA on chromosomes. Since humans are diploid organisms, they retain two copies or chromosomes for most of the genetic information. For reproduction, one copy is randomly chosen in the context of a process called meiosis to be passed on to the offspring. Due to crossover, recombinational and mutational events, genetic variation is introduced at different steps of meiosis as well as during replication of genetic information for the division of cells (mitosis).
With the rise of array measurement technologies, it became possible to evaluate a large variety of genetic markers. Multiple variations of the same genetic sequence can occur as we just introduced, and are denoted as alleles. The combination of the two observed alleles for the same locus or gene of one individuum - or two single bases in the case of SNPs - are also described by the term genotype. One key characteristic of such a SNP is how often alternative alleles are observed compared to a reference allele in a population of many individuals, also referred to as minor allele frequency (MAF). SNPs with a MAF > 5 % are often described as common variants as opposed to rare variants, that occur in only a small percentage of the population (e.g. < 1 % or < 5 %). Within the 46 chromosomes, human genetic code is organized in haplotypes, i.e. genetic regions that are often inherited together with a lower rate of recombination. The corresponding correlation structure, where certain variants are observed together more often than expected only based on their allele frequency is referred to as linkage disequilibrium (LD). By genotyping approximately one half to one million of those SNPs and imputing further variants using reference panels including known haplotypes, it has become possible to efficiently genotype millions of variants in ever growing cohorts.
In genome-wide association studies (GWAS), each individual SNP can further be evaluated for an association with a certain trait. These kind of studies have discovered thousands of disease-associated loci and improved our understanding of genetic and phenotypic relationships [Wang and Wang, 2018], especially if a trait is influenced by a multitude of variants across many genes. Such traits can be quantitative (e.g. height) or categorical (e.g. disease phenotypes) and they are also referred to as complex traits.
AF is a typical example for such a polygenic, complex trait.

1.2.2 Quantitative trait locus analyses

More than 95 % of GWAS variants are localized in non-coding regions [Roselli et al., 2018], most likely affecting expression through changes in regulatory elements rather than changing a specific protein. Of course, in these cases, elucidating the biological mechanisms is much more challenging and remains often elusive. Brem et al. [2002] were one of the first to systematically evaluate the consequences of specific genetic variants on the expression of particular genes in budding yeast, considering *cis* expression quantitative trait loci (eQTL), where the gene and variant are in relative close proximity

to each other with respect to genetic position and *trans* eQTLs, mapping association where the variant is located further away or even on a different chromosome. Since then, larger studies in lymphoblastoid cell lines (LCL) investigated functional genetic variation in humans [Lappalainen et al., 2013]. Westra et al. [2013] used *trans* eQTLs in non-transformed peripheral blood samples to identify downstream effects of disease-associated variants from genome-wide association studies. Since then, QTL analyses in LCLs have been extended to proteomics level [Hause et al., 2014] and were used to compare genetic consequences on mRNA, ribosomal occupancy and proteins by Battle et al. [2015].
Proteins as end products of gene expression, directly interacting with our environment, are especially interesting to study. Due to the easy accessibility Suhre et al. [2017], Sun et al. [2018], Yao et al. [2018] and Ferkingstad et al. [2021] studied the plasma proteome to map genetic risk to disease end points and identify putative causal genes and pathways, with the latest study including over 35 000 individuals.
Regulation of gene expression in the disease context can be investigated in large cohorts using widely accessible bio material like blood samples. Whole blood *cis* eQTLs were used to study genetic effects in coronary artery disease [Joehanes et al., 2017]. However, gene expression varies strongly across different tissues. Therefore, disease-relevant tissues should be evaluated as for example left ventricle tissue samples in the case dilated cardiomyopathy [Heinig et al., 2017]. Instead of focusing on only one specific tissue, the GTEx consortium [Gamazon et al., 2018] systematically evaluated *cis*-regulatory patterns across 54 human tissues in non-diseased donors.
The most comprehensive description of non-coding variants associated with AF so far have been given by van Ouwerkerk et al. [2019, 2020].

1.2.3 Genetic architecture of complex polygenic traits

Even though important insights have been gained by QTL studies, a large fraction of heritability remains elusive [Manolio et al., 2009]. As opposed to Mendelian diseases, where the disorders are caused by specific mutations in single genes, most complex traits are influenced by a multitude of genes which is why they are also denoted as polygenic traits.
While mostly rare monogenic effects also exist in AF, additive polygenic effects of many common variants explain a much higher proportion disease heritability [Choi et al., 2020].
However, it remains difficult to quantify contributions to overall heritability due to *cis* and *trans* effects, different effect sizes and allele frequency coupled with partly spurious correlations in the genetic make-up and linkage disequilibrium.

1.2.3.1 The omnigenic model

To quantify the contribution of *cis* and *trans* effects on complex traits, Boyle et al. [2017] created the theoretical omnigenic model that particularly is able to explain the following

characteristics of complex trait heritability:

- *Cis* effects can only explain a modest fraction of heritability.
- Common variants account for a large proportion of the total heritability, but the individual effect sizes of most variants are very small.
- A modest increase in per-SNP heritability in tissue-specific compared to broadly active regulatory elements indicates that regulation is driven by factors across the whole genome and diverse gene functional categories.
- Few rare variants with large effect sizes contribute much less to the total heritability compared to the many common variants with very small effect sizes.
- Due to the incomparable sizes of protein-coding and non-coding regions in the genome which both contribute to variation of a complex trait, heritability is dominated by non-coding variants.
- While GWAS hits are strongly enriched for *cis* and *trans* eQTLs, many disease-associated loci without an eQTL exist.

From these observations, Boyle et al. [2017] derived that the easiest model which would explain such a behavior would be that genes can actually be divided into two classes: few central core genes and many peripheral genes, where core genes accumulate genetic perturbations from many different peripheral genes and master regulators. While genetic variation can indirectly influence core gene expression by changes of peripheral genes involved in shared gene regulatory networks, core gene function should rather be conserved due to their vital role in biological processes. Therefore, they are expected to directly and functionally relate to a phenotype. A more severe or loss-of-function mutations would then be observed as a Mendelian disease.
Based on this omnigenic idea of genetic architecture, Liu et al. [2019] developed a quantitative model describing *cis* and *trans* contributions to complex trait heritability. They estimate, that based on actual *cis* and *trans* eQTL effect sizes, at least 70 % of disease heritability was explained by *trans* effects propagated through gene regulatory networks. Within these networks, multiple *trans* effects can accumulate on just a few central genes. Identifying these *trans* regulations remains challenging, as the effect size of each individual *trans* locus is very small [Westra et al., 2013, Võsa et al., 2021] compared to the large multiple testing burden.

Since many genetic variants contribute to disease risk, scores that summarize the individual overall genetic predisposition have been created [Kalsto et al., 2019, Khera et al., 2018]. A short introduction into different classes of genetic risk scores is given in the corresponding method section 3.2.6.

1.2.3.2 Quantitative trait scores (QTS)

Since genetic risk scores summarize genetic contributions to disease risk, it can act as a proxy of the accumulation of *trans* effects [Võsa et al., 2021]. Therefore, genes that are highly influenced by the propagation of multiple *trans* effects should be able to be identified by a higher correlation of gene expression with the risk score. This concept

of systematically assessing the correlation of transcriptomics with an (omnigenic risk) score, has been termed expression quantitative trait score (eQTS) [Võsa et al., 2021]. The same can be applied to protein measurements which might be even more relevant for phenotypic consequences, and is denoted as protein quantitative trait score (pQTS).

1.3 Multi-omics pathway analyses

The recent improvement of measurement technologies has lead to the evaluation of many differential transcript or protein expression experiments. However, deriving new insights into de-regulated biological processes from comparing differences in expression levels between multiple groups is not always straight forward. Depending on the effect sizes, genes appearing on the top of the differential expression result list might only be those with very large effects, missing many relevant genes or only identifying those which are already known. This is especially challenging in the context of small effects limited sample sizes. Also, simple summary statistics give information about the differences of a specific gene in varying settings, it does however not add any information about the biological processes involved. Particularly, too few or too many differential findings as well as the choice of a strict cutoff dividing genes in significant and non-significant ones challenge the interpretation of such analysis results [Subramanian et al., 2005].
Next to a single, strongly regulated gene, even smaller coordinated regulations of multiple genes involved in the same biological function or signaling pathway might be much more important. Therefore, techniques have been developed to infer regulated processes from differential expression results. Such approaches often denoted as pathway or gene set enrichment analysis make use of prior knowledge of co-regulation of genes involved in similar biological processes and look for patterns of common regulation of genes belonging to the same group.

Such prior knowledge can be derived from cataloging general biological knowledge about genes and genetic interactions. This approach was taken by the Gene Ontology (GO) Consortium building a formal representation about biological processes, cellular components and molecular function [Carbon et al., 2019, Ashburner et al., 2000]. It is introduced in more detail in the material section 2.2.1.
Similarly, Reactome annotations consist of a relational database for signaling and metabolic molecules with their relations organized into biological pathways and processes [Gillespie et al., 2022], the STRING [Szklarczyk et al., 2021] database provides information about protein-protein interactions and the Kyoto Encyclopedia of Genes and Genomes[1] (KEGG) describes pathways for higher-level functions and utilities of the biological system with a stronger focus on disease mechanisms just to mention a few.

A general approach to make use of this knowledge is to evaluate how many significant

[1] `https://www.kegg.jp/`

genes from the analysis are overlapping with a specific set of genes of interest using the aforementioned resources. The so-called overrepresentation analysis (ORA) evaluates overrepresentation of significant genes in a gene set using Fisher's exact test or a hypergeometric distribution to evaluate statistical significance.

This comes with common challenges, such as picking the right significance threshold. Subramanian et al. [2005] introduced gene set enrichment analysis (GSEA). Instead of dividing genes into discrete groups, they are ranked according to their correlation with a phenotype of interest, or more concrete, based their association evaluated through common summary statistics. Next, for each gene set of interest, the accumulation of genes at the top or bottom of the list - representing up- or down-regulation - are evaluated by an enrichment score.

However, this approach leads to many false positives in case of strongly correlated or even hierarchical annotations such as the GO categories. Therefore, going back to significantly and not-significantly observed genes, Bauer et al. [2010] introduced their model-based gene set analysis (MGSA) to estimate posterior probabilities for each annotation term based on a Bayesian network model incorporating the annotation topology.

As demonstrated by Buccitelli and Selbach [2020], regulatory mechanisms can take effect at different steps of gene regulation and therefore, affect different omics level. To make use of the growing availability of multi-omics data, single-omic methods can be integrated ad-hoc on pathway level.
Alternatively, Sass et al. [2013] developed the multi-level ontology analysis (MONA) approach that directly integrates multiple omics in one model by extending the Bayesian network proposed by MGSA.

While the Bayesian models outperform GSEA in general in a single-omic scenario, the impact of different factors such as the coverage of measured genes, specifically with respect to multi-omics integration remains elusive. Also, little is known about how the different significance cutoffs affect model performance of MGSA and MONA, and instead of comparing direct integration in the case of MONA to an ad-hoc integration of multiple omics on pathway level, only one omic was considered for GSEA and MGSA. Furthermore, the correlation between omics and the similarity or differences of pathway activations in different omics might significantly affect the choice of method.

1.4 Aims and contributions

In summary, omics technologies opened up new opportunities to study complex human diseases. However, new approaches are needed to integrate the different data types and derive insights into underlying molecular mechanisms.

In particular,

1. ... gene regulation is a complex process, which is very specific to the tissue of question. However, downstream consequences of genetic variation, which would be vital to understand the underlying mechanisms, are mostly unknown.
2. ... more than one hundred genetic loci have been associated with AF. However, their function and affected genes often remain elusive.
3. ... *trans*-genetic effects have been shown to majorly contribute to common disease. However, current analyses are largely impaired due to the vast search space in combination with limited sample sizes.
4. ... the omnigenic model proposes the existence of core genes. Due to their direct link to the phenotype, these core genes could be a key component of understanding more complex disease mechanisms. However, so far it has been challenging to identify core genes due to complex interactions and small effect sizes.
5. ... current measurement technologies enable the evaluation of multiple molecular data types. However, it is difficult to leverage the full potential of multi-omics data to derive de-regulated biological processes in AF.

In this thesis, we want to leverage the full potential of multi-omics data to better understand underlying mechanisms of complex diseases - and specifically AF - by the following contributions:

1. By integrating genomics, transcriptomics and proteomics data, we were able to derive a comprehensive map of genome-wide *cis*-regulatory mechanisms highlighting differences in regulation on transcript and protein level.
2. The integration of *cis* eQTLs and pQTLs in combination with a multitude of annotations makes it possible to evaluate the context of each GWAS hit individually. Our functional *cis* QTL categories aid in shortlisting possible underlying mechanisms or causal factors. Specifically, also non-significant associations contribute valuable information.
3. We propose a PRS-based candidate selection approach to make targeted *trans* QTL analyses possible in a relatively small clinical cohort.
4. The resulting two *trans* eQTLs and five *trans* pQTLs, as well as the analysis of the NKX2-5 TF network with 13 identified targets result in overall 20 putative core genes for AF.
5. We have extended existing methods for multi-omics gene set enrichment analysis and made them accessible to a broader audience in the graphical workflow framework KNIME. Our simulation study warranted insights on the performance of different methods under varying underlying pathway activations. By leveraging

the gained knowledge, we were able to identify common molecular mechanisms underlying AF even in the case of extremely small effect sizes.

These contributions are part of manuscripts, which have either already been published in peer-reviewed journals or are currently in preparation. Therefore, parts of this thesis are similar to the following publications:

1. Ines Assum†, Julia Krause†, Markus O. Scheinhardt, Elke Hammer, Christian Müller, Christin S. Börschel, Uwe Völker, Lenard Conradi, Bastiaan Geelhoed, Tanja Zeller*, Renate Schnabel* and Matthias Heinig*. **Tissue-specific multi-omics analysis of atrial fibrillation.** *Nature Communications* **13**, 441 (2021).
2. Ines Assum†, Julia Krause†, Renate Schnabel, Tanja Zeller and Matthias Heinig. **Multi-omics pathway analysis in atrial fibrillation.** (*Manuscript in preparation*).

On top of these, I lead or contribute to the following articles:

1. Ines Assum, ..., Matthias Heinig and Silke Szymczak. **Benchmarking of gene set enrichment methods across species.** (*Manuscript in preparation*).
2. Julia Krause, ..., Ines Assum, Matthias Heinig, ..., Justus Stenzig*, Tanja Zeller* **An arrhythmogenic metabolite in atrial fibrillation.** (*Manuscript in revision*).

† authors contributed equally (shared first author), * authors contributed equally (shared last authors);

Besides contributions to the articles mentioned above, I have been leading the development and maintenance of the following extension to the KNIME framework:

1. Ines Assum and Kristof Gilicze **EnrichmentNodes** `https://github.com/InesAssum/EnrichmentNodes`

1.5 Outline

The general topic of this thesis is the integration of multi-omics data to infer disease-specific gene-regulatory mechanisms.

The introduction starts with the general motivation, biological background and the challenges of the field addressed by this thesis.

This is followed by the materials chapter with the detailed presentation of the AF multi-omics cohort as well as some relevant, public resources.
In the methods chapter, an overview over the commonly used concepts and approaches is given. This is then used to describe to the specific methods needed for the evaluation of atrial specific *cis*- and trans-regulatory mechanisms followed by the descriptions for multi-omics gene set enrichment analyses.

Subsequently, we present new research results and corresponding discussion on three three topics:
In the fourth chapter, we present downstream consequences of common genetic variation on transcript and protein abundance including *cis* eQTL and pQTL integration in form of functional *cis* QTL categories, enrichment of regulatory elements and their overlap with known GWAS hits.

Next, we evaluate omnigenic contributions to AF resulting in the identification of two *trans* eQTLs and five *trans* pQTLs using our novel targeted *trans* QTL approach. We further investigate the NKX2-5 transcription factor network including its potential targets, ending up with identifying overall twenty putative core genes for AF.

To assess the performance of different multi-omics gene set enrichment methods, we conducted a simulation study including a variety of multi-omics specific factors and simulation scenarios. We apply the results to our AF multi-omics results to identify common mechanisms of AF and present a KNIME plugin enabling the application of all evaluated enrichment methods as part of an interactive, graphical workflow development framework.

Lastly, we summarize and discuss project over-arching findings including current limitations and give an outlook to future projects and further research.

2 Materials

2.1 Atrial fibrillation in high risk individuals-biopsy (AFHRI-B) cohort

The AFHRI-B (Atrial fibrillation in high risk individuals-biopsy) is an epidemiological, prospective, single center cohort study for the improvement of atrial fibrillation risk stratification in high risk individuals. Multi-omics data used in this thesis was obtained from 118 patients undergoing open heart coronary artery bypass surgery at the Department of Cardiology and Department of Cardiovascular Surgery, University Medical Center Hamburg-Eppendorf in the years 2012 to 2015 (see Figure 1.2). The observational cohort study was approved by the Ethikkommission Ärztekammer Hamburg (PV3982) and was performed in compliance with the Declaration of Helsinki.
The aim of this study was to increase our understanding with regard to the pathogenesis of AF and to find new potential therapeutic targets. Systemic protein expression is considered as an intermediate biochemical marker of genetic variability. Multi-omics data shall be used to identify gene expression patterns and proteins specific to AF in atrial cardiac tissue.

In the following, we will give an overview over data collection, processing and basic characterization necessary for downstream analyses. These descriptions will be highly similar to the Materials and Methods sections described in previous work [Assum et al., 2022a].

The cohort consisted of patients older than 18 years who were scheduled to undergo open heart coronary artery bypass surgery. Patients with other or additional surgeries, e.g. valve surgery, were excluded. During surgery, right atrial appendage tissue remnants were collected when the extracorporeal circulation was started and shock frozen immediately. Omics data for tissue samples, blood plasma and genotypes were measured in multiple batches depending on the amount and quality of the material and resources available. If measurements for the same omic were performed in multiple batches, there were equal numbers of cases and controls in each run. For the current analyses, N = 118 patients with omics data were available.

2.1.1 Cardiovascular phenotypes and risk factors

Information on classical cardiovascular risk factors and potential confounders (age, sex, body mass index, systolic and diastolic blood pressure, hypertension, hypertension medication, diabetes, diabetes medication, history of myocardial infarction, smoking) [Magnussen et al., 2017] was collected by questionnaire and from medical records. Additionally, C-reactive protein (CRP) and N-terminal pro-brain natriuretic peptide (NTproBNP) were available from routine pre-surgical work-up. Follow-up for AF and other cardiovascular disease outcomes was done by questionnaire, telephone interview and medical chart review.

Atrial fibrillation subtypes Based on patient history and routine cardiology work-up, non-valvular prevalent AF was the clinical diagnosis used as an outcome in our analyses. Dependent on specific parts of this thesis, we distinguished specific subtypes of AF patients: Prevalent AF patients presented with either current AF episodes or a prevalent AF diagnosis before the bypass surgery. In contrast, patients who only developed AF after surgery were labeled post-operative or incident AF. Analyses which considered prevalent and incident AF phenotypes compared only cases of this subgroup to controls and excluded cases of incident and prevalent AF respectively, while overall AF merged prevalent and incident AF cases.

Clinical baseline Baseline characteristics of the cohort stratified by analysis type can be found in Table 2.1 and respective blood samples were aliquoted and stored prior to surgery.

Table 2.1: AFHRI-B cohort baseline table.
AF, atrial fibrillation; PRS, genome-wide polygenic score; BMI, body mass index; BP, blood pressure; conc., concentration;

Variable	All samples	eQTL/eQTS	pQTL/pQTS	All omics
Samples measured, n (%)	118 (100)	75 (64)	74 (63)	66 (56)
Women, n (%)	13 (11)	5 (7)	4 (5)	3 (5)
Prevalent AF, n (%)	15 (13)	13 (17)	9 (12)	9 (14)
AF PRS, median (IQR)	32.40 (32.33-32.48)	32.40 (32.33-32.48)	32.40 (32.33-32.49)	32.39 (32.33-32.48)
Age, median (IQR), y	66.8 (59.5-73.5)	68.4 (60.6-73.8)	67.2 (59.7-73.7)	67.7 (60.6-73.7)
BMI, median (IQR), kg/m^2	27.8 (24.8-30.4)	28.3 (25.0-30.5)	27.6 (24.8-30.5)	28.0 (25.0-30.5)
Systolic BP, median (IQR), mmHg	135 (122-145)	137 (123-146)	136 (122-145)	136 (123-145)
Diastolic BP, median (IQR), mmHg	76 (70-82)	76 (70-81)	76 (70-80)	76 (70-80)
Hypertension, n (%)	105 (89)	68 (91)	66 (89)	60 (91)
Hypertension medication, n (%)	98 (83)	64 (85)	61 (82)	56 (85)
Diabetes, n (%)	36 (31)	25 (33)	23 (31)	22 (33)
Diabetes medication, n (%)	33 (28)	24 (32)	22 (30)	21 (32)
Myocardial infarction, n (%)	45 (38)	29 (39)	30 (41)	27 (41)
Smoking, n (%)	28 (24)	14 (19)	15 (20)	13 (20)
Fibroblast-score, median (IQR)	80.43 (79.52-81.87)	80.42 (79.12-82.81)	80.26 (79.08-81.87)	80.26 (79.08-82.06)
RIN-score, median (IQR)	7.7 (7.1-8.1)	7.6 (7.1-8.1)	7.6 (7.1-8.1)	7.6 (7.1-8.1)
Protein conc., median (IQR), $\mu g/\mu l$	0.87 (0.45-1.31)	0.80 (0.45-1.30)	0.80 (0.45-1.32)	0.80 (0.45-1.30)

2.1.2 Omics data

Overview: Analyses were performed in all samples with respective omics data which passed appropriate quality control as stated in the preprocessing steps. This resulted in different samples being available for different analyses, as visualized in Figure 2.1.

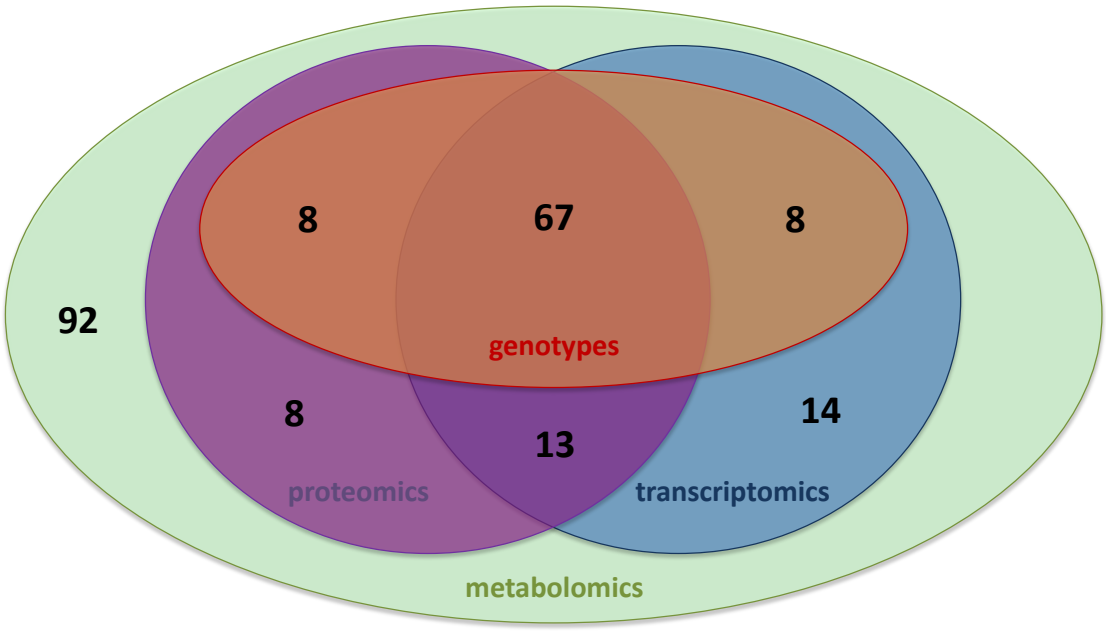


Figure 2.1: Overview of overlapping omics data for different samples.
Overview of metabolomics, transcriptomics, proteomics and genotype data for the samples in the AFHRI-B. Shown are only measurements which have passed quality control. Transcriptomics and proteomics were derived from atrial tissue.

2.1.2.1 Genotypes

Genotype data for 83 blood samples was generated by Julia Krause and Tim Hartmann using the Affymetrix GeneChips Genome-Wide Human SNP Array 6.0. Markus Scheinhardt performed SNP-calling with the Birdseed v2 algorithm to turn intensities from the CEL-files into genotypes and carried out standard quality control procedures as described in Anderson et al. [2010] using the PLINK 1.9[1] software. Only SNPs with a MAF > 0.01, a call rate > 98 % and a HWE exact test P > 10^{-6} passed quality control, resulting in 749 272 SNPs. Genotypes were kindly imputed by Matthias Heinig with IMPUTE2 [Howie et al., 2009] based on the 1000 Genomes Phase 3 genotypes [Auton et al., 2015, Sudmant et al., 2015] (per SNP: confident genotype calls with genotype probability > 95 %, percentage of confident genotype calls across samples > 95 %) and included only variants with HWE P > 10^{-4} resulting in 5 050 128 SNPs.
To evaluate close or spurious relatedness, identity-by-descent (IBD) was investigated in PLINK via the $\hat{\pi}$ estimates. Maximum $\hat{\pi}$ was 0.04, which was lower than the estimate

[1] `https://www.cog-genomics.org/plink/1.9`

of $\hat{\pi} = 0.0625$ for fourth-degree relatives and well below recommended and applied threshold of 0.1875 representing relatedness between third- and second-degree relatives [Anderson et al., 2010]. In order to evaluate ancestry and population structure, PCA was performed on the variance-standardized relationship matrix of ancestry-informative SNPs [Anderson et al., 2010] from the merged genotype data of AFHRI-B and 1000 Genomes samples. All samples from the AFHRI-B cohort were of central European ancestry and showed no population substructure (see Figure 2.2).

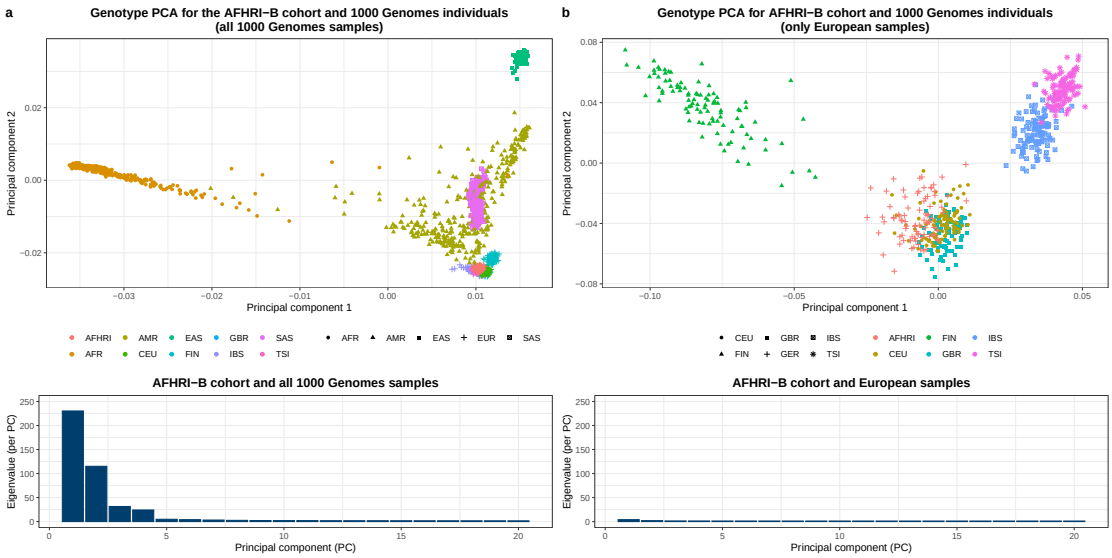


Figure 2.2: Evaluation of the genetic population structure of the AFHRI-B cohort compared to the 1000 Genomes individuals.
a: PCA plot on the covariance matrix of ancestry-informative SNPs over all samples including all populations. Clear differentiation between different populations with the AFHRI-B samples clustering in the middle of the European populations.
b: PCA plot on the covariance matrix of ancestry-informative SNPs over all European individuals. Eigenvalues are much smaller compared to the previous analysis in (a) and the AFHRI-B cohort samples still cluster together homogeneously.
PCA, principal component analysis; AFR, African ancestry; AMR, American ancestry; EAS, East Asian ancestry; EUR, European ancestry; SAS, South Asian ancestry;
AFHRI, AFHRI-B cohort; CEU, Utah residents with Northern and Western European ancestry; FIN, Finnish in Finland; GBR, British in England and Scotland; IBS, Iberian populations in Spain; TSI, Toscani in Italy;

2.1.2.2 Genetic risk score for AF

A genetic risk score was calculated for the AFHRI-B cohort using the genome-wide polygenic score for AF derived from the LDpred algorithm [Vilhjálmsson et al., 2015] and published by Khera et al. [2018]. Since genetic risk scores are most meaningful when considering their distribution across the general population, we calculated risk score values also for unrelated, European (CEU) individuals from the 1000 Genomes Projects [Auton et al., 2015, Sudmant et al., 2015].
For 6 730 540 variants out of 6 730 541 in the score, we merged the Phase 3 genotypes

of the 1000 Genomes individuals (407 samples) with the genotypes of the AFHRI-B cohort (83 samples). The Plink 1.9[2] function *score* was used to compute risk score values. Percentiles across all 490 individuals were then used for all further analyses.

2.1.2.3 Transcriptomics in human heart tissue

The mRNA data was generated from human heart atrial appendage tissue samples which were obtained by Lenard Conradi during coronary artery bypass surgery. Tissue samples were directly frozen in liquid nitrogen and pulverized for further analysis. RNA isolation, RNA assessment using the RNA integrity number (RIN), which assigns a quality score between 1 (poor) and 10 (best), and gene expression quantification using the HuGene 2.0 ST Arrays (Affymetrix® GeneChip WT Plus Reagent Kit) were performed by Julia Krause.
The R bioconductor package oligo [Carvalho and Irizarry, 2010] was used to create expression sets from the raw CEL files, to perform standard microarray background correction, quantile-normalization per sample as well as to log-transform the data. For all further analyses, left atrial appendage tissues and samples with a RIN-score smaller than 6 were excluded. Microarray measurements were performed in two batches, resulting in few biological replicates. In this case, only the one with the highest RIN-score was kept. In order to derive gene level expression values, the mean of multiple transcript clusters annotated to the same gene symbol was computed, resulting in expression data for 26 376 genes in 102 right atrial appendage samples. Assessing the expression patterns in a PCA plot (see Figure 2.3a), there was no apparent separation of the three disease groups.

2.1.2.4 Untargeted proteomics in human heart tissue

Proteomics quantifications were kindly performed by Elke Hammer at the Proteomics facility Greifswald: "To measure the protein concentrations of 97 right atrial appendage samples, the tissues were homogenized using a micro dismembrator (Braun, Melsungen, Germany) at 2 600 rpm for 2 minutes in 100 μl of 8M urea/2M thiourea (UT). Then homogenates were resuspended in 300 μl of UT. Nucleic acid fragmentation was gained by sonication on ice three times for 5 s each with nine cycles at 80 % energy using a Sonoplus (Bandelin, Berlin, Germany). The homogenates were centrifuged at 16 000 x g for one hour at 4 °C. After that, protein concentration was determined by Bradford with BSA as standard (SE). 3 μg protein were reduced and alkylated and digested with LysC (1:100) for 3 h followed by tryptic digestion overnight both at 37 °C. Subsequently peptide solutions were desalted on C18 material (μ ZipTip). Finally mass spectrometry analysis was performed on a LC-ESI-MS/MS machine (LTQ Orbitrap Velos). One sample was excluded due to irregularities in the chromatographic pattern. The Rosetta Elucidator 3.3 workflow was used to extract feature intensity and derive protein intensities by summing of all isotope groups with the same peptide annotation for all peptides

[2] `https://www.cog-genomics.org/plink/1.9`

annotated to one protein (further parameters: Uniprot_Sprot_human_rel. 2016_05: static modification: carbamidomethylation at Cys, variable modification: oxidation at methionine, 2 missed cleavages, fully tryptic, filtered for peptides with FDR < 0.05 corr. to Peptide Teller probability > 0.94 and shared peptides were excluded). Intensities for 1 419 proteins with one or more peptides (877 with 2 or more peptides) were quantified for 96 samples, median-normalized and log10-transformed" [Assum et al., 2022a]. Although the same amount of protein (3 μg) was used for quantifications, confounding by the original protein concentration of the sample was observed when investigating PCA plots as depicted in Figure 2.3b. Therefore, this was used as a technical covariate in all further analyses.

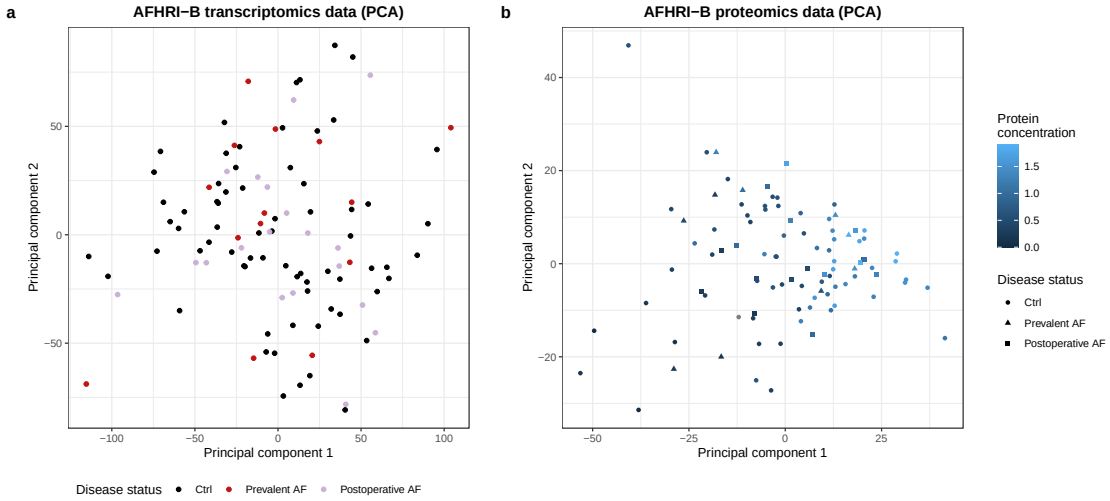


Figure 2.3: PCA plots of the transcriptomics and proteomics data.
a: Expression data does not show a separation of the disease groups in a PCA.
b: Protein concentration of the original sample is highly correlating with the first two principle components.
PCA, principal component analysis; Ctrl, control;

2.2 Public data

2.2.1 Pathway annotations

Depending on the context and discipline, the term pathway can have different meanings. Clinical pathways are defined as step by step procedures for the treatment of medical conditions, while a pathway in molecular biology often stands for a partial or simplified representation of complex physiological mechanisms. In particular, metabolic pathways often refer to specific chemical reactions in the cell and signaling pathways map the transduction of signals by interactions modifying gene and protein expression or regulation. A more general approach which comes closest to the meaning of pathways in this thesis are genetic pathways as gene-regulatory networks. They are described by groups of interacting and regulating genes on DNA, RNA, protein or metabolite level. Pathway annotations are used to connect groups of genes or metabolites to biological processes and pathway enrichment analysis tries to infer molecular or physiological changes from individual patient measurements or differential expression results. This prior knowledge is often encoded in a graph structure which can be exported into a list of genes annotated to terms that represent specific aspects of biological processes.

2.2.1.1 Kyoto Encyclopedia of Genes and Genomes (KEGG)

KEGG[3] is a database resource which maps complex biological systems, ecological systems and molecular mechanisms of the cell to general pathways. Information is provided on different levels, ranging from exact mechanistic equations to general mapping of genes to biological processes. Large-scale molecular datasets, such as genome sequencing and other high-throughput experimental technologies, were used to create this curated database. In this thesis, we mainly focus on the human disease pathways in the form of gene lists associated to a pathway term.

2.2.1.2 The Gene Ontology resource (GO)

The Gene Ontology (GO) resource[4] [Ashburner et al., 2000, Carbon et al., 2019] provides structured, computable knowledge regarding the function of genes and gene products. It contains three distinct but related ontologies with annotations about the function of a particular gene. The "Molecular Function" ontology focuses on the molecular activities of individual gene products or molecular complexes. For "Cellular Component", classes refer to cellular anatomy rather than processes and describe the structures, where gene products perform their function. Finally, "Biological Process" describes larger processes which involve multiple molecular activities.

[3] `https://www.genome.jp/kegg/`
[4] `http://geneontology.org/`

The Gene Ontology (GO) knowledge base is structured hierarchically, with smaller, more specific child terms being a subset of a more general parent term. While GO provides species-agnostic annotations, in this thesis we mainly focus on annotations of human genes by extracting lists of genes annotated to different terms of the biological process ontology.

2.2.2 The 1000 Genomes Project

The goal of the 1000 Genomes Project [Auton et al., 2015, Sudmant et al., 2015] was to create a reference map for genetic variants with an allele frequency of at least 1 % in different populations.

Due to advances in sequencing technology and reduced cost connected to those, sequencing larger amounts of complete genomes became feasible. The 1000 Genomes Project provides a comprehensive resource on human genetic variation and the genetic data was made freely accessible for researchers all over the world.

Even though deeply sequencing all samples still remained too expensive, sequencing all samples to four times the genome coverage allowed an efficient detection of most variants with a minor allele frequency larger than 1 %. This was done by combining information across samples while leveraging the limited number of haplotypes. Sequencing depth was sufficient to assign high confidence genotypes to all variant sites discovered in this project for all the 2 504 samples.

In the meantime, ongoing progress in enhancing sequencing techniques has led to even larger and more diverse panels for genome annotations. However, to this date, it is one of the most commonly used genetic references.

2.2.3 The Genotype Tissue Expression consortium

Extending genome annotations further to the transcript level, the aim of the Genotype-Tissue Expression (GTEx) project [Gamazon et al., 2018] was to build a comprehensive public resource to study tissue-specific gene expression and regulation.
Molecular assays including whole genome sequencing (WGS), whole exome sequencing (WES) and RNA sequencing (RNA-seq) were performed on samples from nearly 1 000 individuals and 54 non-disesased tissue sites to assess general and tissue-specific impact of genetic variation on gene expression. The GTEx Portal[5] provides open access to data including gene expression, QTLs and histology images.

For this thesis, tissue-specific transcript expression (GTEx v8 data for right atrium and left ventricle) as well as summary statistics for atrial appendage eQTLs (GTEx v7) were used.

[5] `https://www.gtexportal.org/home/`

Furthermore, matched tissue-specific proteomics data was made available as part of the publication from Jiang et al. [2020].

Gene expression as well as its regulation by genetic variants differ tremendously between tissues. In an exemplary fashion, this is visualized in Figure 2.4 by considering gene expression for marker genes of important cell types in the human heart as reported in the human heart atlas [Litviňuková et al., 2020] across all tissues available in GTEx.

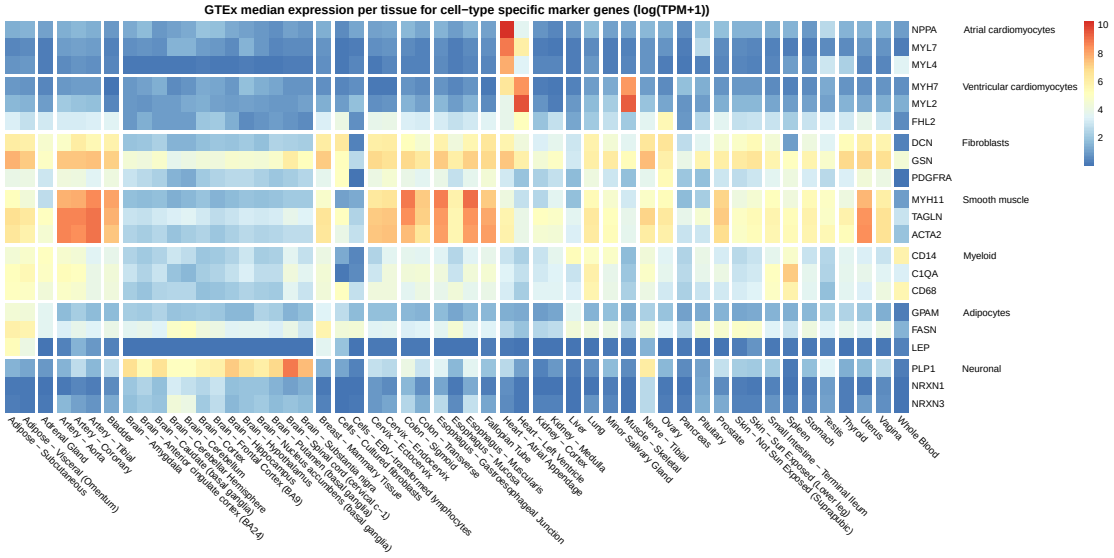


Figure 2.4: GTEx tissue-specific expression patterns.
Gene expression per GTEx tissue for marker genes of cell types relevant in human cardiac tissues. GTEx, genotype-tissue expression;

3 Methods

3.1 Preprocessing procedures

3.1.1 Normalisation procedures

Biological data can be diverse and normally follow very various distributions. Due to constraints of the applicability of statistical methods as well as improving the comparability between different samples, normalization and harmonization might be necessary.

In this regard, direct biological measurements are often distributed exponentially, while many methods investigate linear relations in the data. Therefore, logarithmic transformation is most commonly used as a first step of preprocessing.

Definition 3.1 (Log-transformation) *Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be a data matrix with n observations (e.g. samples) of p features (e.g. genes).*
Then the logarithmic transformation of base b, $\mathbf{X}' = x'_{ij}$, of $\mathbf{X} = x_{ij}$ is defined as

$$x'_{ij} = log_b(x_{ij}) \qquad \forall x_{ij} \in \mathbf{X}. \tag{3.1}$$

In case of negative or 0 values in the data, a pseudo count a to derive strictly positive values can be added to the data before transformation, i.e. $x'_{ij} = log_b(x_{ij} + a)$.

Additionally, values are often centered to mean 0 or scaled to variance 1 per row or column depending on the context.

Furthermore, distributions can be transformed to following a normal distribution. This is specifically important for data which did not follow a log-normal distribution and therefore could not be transformed to a normal distribution by applying the logarithmic transformation as defined in Definition 3.1.

Definition 3.2 (Quantile normalization) *Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be a data matrix with n observations (e.g. samples) of p features (e.g. genes).*
Let $\mathbf{r_k} = (r_{k_1}, \ldots, r_{k_p})$ be the corresponding vector containing the rank of feature values for observation k across all features with $\mathbf{x_k} = x_{kj}$ for all $j \in 1, \ldots, p$ and let $\mathbf{r_l} = (r_{l_1}, \ldots, r_{l_n})$ be the corresponding vector containing the rank of observation values for feature l across all observations with $\mathbf{x_l} = x_{il}$ for all $l \in 1, \ldots, n$ of the data matrix $\mathbf{X} = x_{ij}$. Then the quantile

normalization to a normal distribution per observation $\mathbf{x_k}' = x'_{k\cdot}$ *or feature* $\mathbf{x_l}' = x'_{\cdot l}$ *is defined as*

$$x'_{kj} = q_{\mathcal{N}(\mu,\sigma^2)}\left(\frac{r_{k_j}}{p+1}\right) \qquad \forall j \in 1 \dots p, \qquad \textit{and} \tag{3.2}$$

$$x'_{il} = q_{\mathcal{N}(\mu,\sigma^2)}\left(\frac{r_{l_i}}{n+1}\right) \qquad \forall i \in 1 \dots n, \tag{3.3}$$

with $q_{\mathcal{N}(\mu,\sigma^2)}$ *being the quantile function of a normal distribution.*

In practice, the standard normal distribution $\mathcal{N}(0,1)$ is most often used.

3.1.2 Imputation

With any large-scale molecular experiments, the increasing number of features measured poses a great challenge on evaluating every feature for every sample. While many preprocessing procedures filter for appropriate quality, few missing observations remain in many contexts. However, many downstream analyses require fully observed data. For these cases, a k-nearest neighbor (KNN) approach was used to impute missing observations as described in Troyanskaya et al. [2001].

Definition 3.3 (K-nearest neighbor imputation) *Let* $\mathbf{X} \in \mathbb{R}^{n\times p}$ *be a data matrix with n observations (e.g. samples) of p features (e.g. genes).*
Then for every missing value $x_{i^*j^*}$ *and similarity measure* $\rho_{jl} = \rho(\mathbf{x_{\cdot j}}, \mathbf{x_{\cdot l}})$ *for* $l \neq j$, $x_{i^*j^*}$ *is estimated by the mean of k features with the highest similarity to* j^* *observed in* i^**, i.e.*

$$\hat{x}_{i^*j^*} = \frac{1}{k}\sum_{j=l_1}^{l_k} x_{i^*j} \qquad \textit{with } \rho_{j^*l_1} \geq \rho_{j^*l_2} \geq \dots \geq \rho_{j^*l_k} \geq \rho_{j^*l_{k+1}} \geq \dots \geq \rho_{j^*l_{p-1}}. \tag{3.4}$$

After reasonable normalization (e.g. log-transformation for expression data), Euclidean distance proved to be a sufficiently accurate norm to be used as similarity measure [Troyanskaya et al., 2001].

3.1.3 Probabilistic estimation of expression residuals (PEER)

Population based gene expression data is often confounded by unwanted variation, such as batch effects or unknown biological or environmental factors. Probabilistic estimation of expression residuals (PEER) implements a Bayesian approach based on factor analysis to infer hidden determinants across samples [Stegle et al., 2012]. Identifying and including such factors as covariates greatly improves statistical power, e.g. in eQTL analyses.

In that case, gene expression data is modeled as a general additive model for different sources of variation including known and unknown factors [Stegle et al., 2010]:

$$\mathbf{X} = \mathbf{X}^{(1)}(\mathbf{S}) + \mathbf{X}^{(2)}(\mathbf{F}) + \mathbf{X}^{(3)}(\mathbf{H}) + \cdots + \boldsymbol{\varepsilon} \tag{3.5}$$

S denotes the transcriptional state which might be dependent on *cis*-genetic variation. **F** represents factors which are known, such as age, sex or known experimental or environmental conditions. On the contrary, **H** models hidden factors, i.e. confounders which are unknown. Further epistatic or environmental interactions, non-linear effects and a noise term ε complete the model.
Based on a predefined number of hidden factors, all contributions are inferred in an iterative approach as a joint model where gene expression follows a normal distribution assuming we consider a gene expression matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ with n samples and p genes:

$$P(x_{ij}|x_{ij}^{(1)}, x_{ij}^{(2)}, x_{ij}^{(3)}, \tau_j) = \mathcal{N}(x_{ij}|x_{ij}^{(1)}, x_{ij}^{(2)}, x_{ij}^{(3)}, \frac{1}{\tau_j}) \tag{3.6}$$

with a gamma prior on the noise precisions $P(\tau_j) \sim \Gamma(\tau_j|a_\tau, b_\tau)$.

Interestingly, PEER is specifically designed to be used for determining hidden factors in *cis* eQTL analysis. Such hidden factors, however, can also represent *trans* regulation such as coordinated changes by a transcription factor. Therefore, PEER can be used to infer *trans* effect in combination with transcription factor activity inference, but it is not suited to account for hidden confounders in *trans* QTL analyses [Stegle et al., 2012].

3.1.4 Correction for cell type composition

Tissue biopsies contain a variety of cell types. In atrial tissue, the most prominent ones are atrial cardiomyocytes, fibroblasts, pericytes, endothelial cells, myeloid cells, smooth muscle cells, lymphoid cells, adipocytes, neuronal cells and mesothelial cells in descending order [Litviňuková et al., 2020]. The cell type composition differs between individuals and particularly, the contribution of fibroblasts and cardiomyocytes, one of the functionally most relevant cell types in primary atrial appendage tissue, is therefore confounding the convoluted expression profile.
In order to correct for that, we used a fibroblast-score based on genes up-regulated in fibroblasts compared to cardiomyocytes in rats to account for the amount of fibroblasts in the sample AFAP1L2, ARHGAP20, CILP, CLEC3B, COL14A1, CPXM2, DCDC2, ELN, FCRL2, FGF10, FOSB, FRAS1, ITGBL1, JAG1, KIAA1199, NOV, NRG1 and SCN7A [Heinig et al., 2017]. For each tissue sample, the fibroblast-score was computed by adding up expression values for the 14 genes measured by transcriptomics (AFAP1L2, ARHGAP20, CILP, CLEC3B, DCDC2, ELN, FCRL2, FGF10, FOSB, FRAS1, ITGBL1, NOV, NRG1, SCN7A). Since only CLEC3B and COL14A1 were measured on protein level, the fibroblast-score for tissue samples with only proteomics but no transcriptomics was imputed based on the existing transcriptomics-derived fibroblast-score values and all proteomics data. Cell type correction was inferred by fibroblast- rather than by cardiomyocyte-specific gene signatures in order to prevent interference with changes

that are connected to structural remodeling, a common mechanism in progressing AF [Corradi et al., 2008] which affects cardiomyocyte gene expression profiles as well as abundance. Even though the difference was not significant, we did see a trend of higher fibroblast scores in AF cases as visualized in Figure 3.1a.
As a validation of the fibroblast-score in human atrial tissue, we considered GTEx gene expression from the right atrial appendage and matched score values with annotations of fibrosis in the histology of the corresponding tissue samples. Even with this rather coarse-grained labels, we saw higher fibroblast-scores for samples with reported fibrosis as shown in Figure 3.1b.

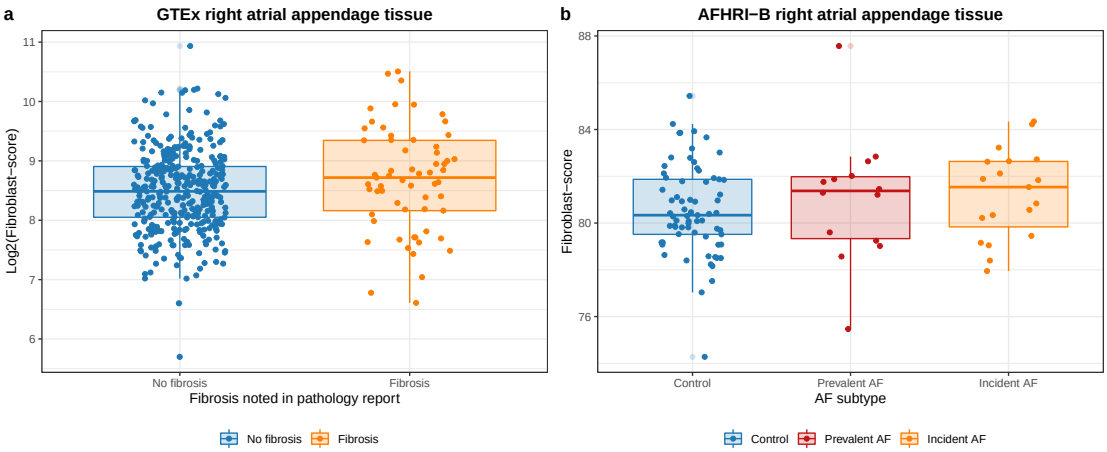


Figure 3.1: Comparison of fibroblast-score values for atrial tissue samples.
a: Fibroblast-score computed based on RNA-seq data from GTEx right atrial appendage tissues. Score values are compared for tissue samples without and with fibrosis noted in the pathology report.
b: Fibroblast-score computed based on micro-array transcriptomics data from the AFHRI-B cohort right atrial appendage tissues. Score values are compared for tissue samples of controls and prevalent as well as incident AF cases.
GTEx, genotype tissue expression; AF, atrial fibrillation;

3.1.5 Principal component analysis (PCA)

One of the most commonly used dimensionality reduction techniques, the Principal Component Analysis (PCA), was first introduced by Pearson [1901]. A lower dimensional representation of the data matrix $\mathbf{X}$ can be derived by identifying orthogonal projections which maximize the variance explained by the chosen directions. This is based on the empirical estimation of the covariance matrix followed by an eigenvector decomposition.

Definition 3.4 (Principal Component Analysis) *Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be a data matrix with n observations (e.g. samples) of p features (e.g. genes).*
Assume that the corresponding covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$ has full rank. Then it is symmetric and positive semidefinite and has an eigenvalue decomposition $\boldsymbol{\Sigma}$

$$\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}} = \boldsymbol{\Gamma}^T \boldsymbol{\Sigma} \boldsymbol{\Gamma} \tag{3.7}$$

with eigenvectors $\mathbf{w}_l$ and eigenvalues λ_l.
Then the Principle Component Analysis (PCA) $\mathbf{X}'$ of $\mathbf{X}$, based on the first k principal components, is defined as

$$\mathbf{X}' = (\boldsymbol{\Gamma}'^T \mathbf{X}^T)^T \qquad with \quad \boldsymbol{\Gamma}' = (\mathbf{w}_1, \ldots, \mathbf{w}_k) \tag{3.8}$$

for the k largest eigenvalues λ_l, $l \in 1, \ldots, k$ and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$.

Principal components can be derived for samples or features. In practice and also for rank deficient matrices, a singular value decomposition (SVD) is often used to derive the corresponding principal components.

3.2 Association analysis

3.2.1 Linear regression

Seal [1967] first introduced the concept of linear regression analysis. Here, we explain a variable $\mathbf{y}$ given several predictors $\mathbf{x}$ while assuming an additive linear relation.

Definition 3.5 (Multipe linear regression) *Let $\mathbf{X} = x_{ij} \in \mathbb{R}^{n \times p+1}$ be a data matrix with n observations (samples) $i \in 1, \ldots, n$ and p features (covariates) $j \in 0, \ldots, p$.*

Then the multiple linear regression model is defined as

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \\ &= \mathbf{x}_i \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \ldots, n, \\ with\ \mathbf{x}_i &= (1, x_{i1}, \ldots, x_{ip})\ and\ \boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T. \end{aligned} \tag{3.9}$$

This is often written in matrix-vector-notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{3.10}$$

with

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ 1 & x_{21} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} \quad and \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}. \tag{3.11}$$

3.2.2 Logistic regression

Similarly, we can also consider logistic regression for binary variables $\mathbf{y}$.

Definition 3.6 (Logistic regression) *Let $\mathbf{X} = x_{ij} \in \mathbb{R}^{n \times p+1}$ be a data matrix with n observations (samples) $i \in 1, \ldots, n$ and p features (covariates) $j \in 0, \ldots, p$.*
Based on the notation introduced in the multivariate linear regression model (Definition 3.5), the logistic regression model is defined as

$$P(y_i = 1) = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}, \qquad i \in 1, \ldots, n \quad \textit{or} \tag{3.12}$$

$$P(\mathbf{y} = 1) = \frac{1}{1 + \exp(-\mathbf{X}\boldsymbol{\beta})} \qquad \textit{with} \tag{3.13}$$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ 1 & x_{21} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix} \quad \textit{and} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}. \tag{3.14}$$

For the corresponding log odds $\log\left(\frac{P(\mathbf{y}=1)}{P(\mathbf{y}=0)}\right)$ *we then find*

$$\log\left(\frac{P(\mathbf{y}=1)}{P(\mathbf{y}=0)}\right) = \ln\left(\frac{P(\mathbf{y}=1)}{1 - P(\mathbf{y}=1)}\right) = \mathbf{X}\boldsymbol{\beta}. \tag{3.15}$$

Both types of regression models are evaluated in R using a maximum likelihood approach (least squares estimator).

3.2.3 Hypothesis testing

For inference of a linear regression model, we generally assume a Gaussian error, i.e.

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2), \qquad i \in 1, \ldots, n, \tag{3.16}$$

follow a normal distribution and are independent and identically distributed (i.i.d.) with mean zero and unknown variance σ^2. For more details on linear and logistic regression models, please refer to Fahrmeir et al. [2013].

Definition 3.7 (Likelihood of a linear regression model) *Let* $\mathbf{X} = x_{ij} \in \mathbb{R}^{n \times p+1}$ *be a data matrix with n observations (samples)* $i \in 1, \ldots, n$ *and p features (covariates)* $j \in 0, \ldots, p$ *and let us assume a Gaussian error model.*

Given a linear relationship between a response variable $\mathbf{y}$ *and* $\mathbf{X}$*, the same variance* σ^2 *of the distribution of* $\mathbf{y}$ *for each* $\mathbf{X}$ *and a normal distribution of* $\mathbf{y}$ *for each* $\mathbf{X}$*, the likelihood of the multiple linear regression model*

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{3.17}$$

with

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ 1 & x_{21} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix},$$

$$\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad \textit{and} \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \quad i \in 1, \ldots, n \quad \textit{(i.i.d.)}, \tag{3.18}$$

is defined as

$$\mathcal{L}(\boldsymbol{\theta}; \mathbf{X}) := P(\mathbf{y}|\boldsymbol{\theta}) = P(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2) \tag{3.19}$$

and together with the Gaussian error assumption, the identity matrix $\mathbf{I}$ *and* $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ip})$*:*

$$\mathcal{L}(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X}) \sim \mathcal{N}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}) = \frac{1}{\sqrt{2\pi\sigma^2}}\, exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta})^2\right). \tag{3.20}$$

The corresponding Maximum Likelihood and least squares estimator are both given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{3.21}$$

Definition 3.8 (Hypothesis testing for the coefficients of a linear regression model) *Let us assume a multiple linear regression model as defined in Definition 3.5 and Definition 3.7 with the likelihood* $\mathcal{L}(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X})$ *and* $\mathbf{X} \in \mathbb{R}^{n \times p+1}$*.*
Then for given estimates $\hat{\boldsymbol{\beta}}$ *of* $\boldsymbol{\beta}$ *and* $\hat{\sigma}^2$ *of* σ^2 *of the multiple linear regression model* $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ *and null hypothesis* $H_0 : \beta_i = 0$ *as well as alternative hypothesis* $H_1 : \beta_i \neq 0$*, the i-th regression coefficient* $\hat{\beta}_i$ *divided by its standard error* $se(\hat{\beta}_i)$ *follows a Student's t-distribution, i.e.*

$$\frac{\hat{\beta}_i}{se(\hat{\beta}_i)} \sim t_{n-p-1}, \qquad i \in 1, \ldots, n. \tag{3.22}$$

Definition 3.9 (Hypothesis testing for the coefficients of a logistic regression model) *Let us assume a logistic regression model with* $\ln\left(\frac{P(\mathbf{y}=1)}{1-P(\mathbf{y}=1)}\right) = \mathbf{X}\boldsymbol{\beta}$ *and* $\mathbf{X} \in \mathbb{R}^{n\times p+1}$ *as defined in Definition 3.6.*
For given estimates $\hat{\boldsymbol{\beta}}$ *of* $\boldsymbol{\beta}$ *and null hypothesis* $H_0 : \beta_i = 0$ *as well as alternative hypothesis* $H_1 : \beta_i \neq 0$*, the i-th regression coefficient* $\hat{\beta}_i$ *divided by its standard error* $se(\hat{\beta}_i)$ *follows a standard Normal distribution, i.e.*

$$\frac{\hat{\beta}_i}{se(\hat{\beta}_i)} \sim \mathcal{N}(0,1), \qquad i \in 1,\ldots,n. \tag{3.23}$$

3.2.4 Multiple hypothesis testing

With the growing possibilities of measuring biological markers, in particular in bioinformatics, statistical analyses no longer refrains to testing a single hypothesis but to screening a large number of markers (e.g. genes) in parallel [Farcomeni, 2008]. This led to the need of new methods to control false positive discoveries, which accumulate by testing thousands or even millions of hypotheses at the same time. Table 3.1 introduces common notation used in the definitions of the following paragraphs.

Table 3.1: Confusion matrix for (multiple) hypothesis testing.
Outcomes in testing m hypotheses;

	H_0 rejected	**H_0 accepted**	**Total**
H_0 false	True positives TP = $N_{1\|1}$	False negatives (Type II error) FN = $N_{1\|0}$	M_1
H_0 true	False positives (Type I error) FP = $N_{0\|1}$	True negatives TN = $N_{0\|0}$	M_0
Total	R (rejections)	$m - R$	m

Performance metrics

Definition 3.10 (Commonly used performance metrics) *Let us consider a hypothesis test for an experiment as described in Table 3.1 with the corresponding counts of True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN) to define the following performance metrics:*

$$\textit{Accuracy:} \quad ACC = \frac{TP + TN}{TP + FN + FP + TN} \tag{3.24}$$

$$\textit{Sensitivity:} \quad SENS = \frac{TP}{TP + FN} \tag{3.25}$$

$$\textit{Specificity:} \quad SPEC = \frac{TN}{TN + FP} \tag{3.26}$$

$$\textit{F1 score:} \quad F1 = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{3.27}$$

$$\textit{Precision:} \quad PREC = \frac{TP}{TP + FP} \tag{3.28}$$

$$\textit{Recall:} \quad REC = \frac{TP}{TP + FN} \tag{3.29}$$

$$\textit{False positive rate:} \quad FPR = \frac{FP}{FP + TN} \tag{3.30}$$

$$\textit{False negative rate:} \quad FNR = \frac{FN}{FN + TP} \tag{3.31}$$

$$\textit{False discovery rate:} \quad FDR = \frac{FP}{TP + FP} \tag{3.32}$$

Note that sensitivity and recall are actually identical. However, depending on the context, the terms sensitivity and specificity or precision and recall are often considered together. To keep in line with common notations, both are used and defined.

Multiple hypothesis testing correction When performing a larger number of m tests, the probability of obtaining at least one false positive test drastically increases.

Definition 3.11 (Family-wise error rate (FWER)) *Let us consider a multiple hypothesis testing experiment as introduced in Table 3.1.*
To control the type I error ($N_{1|0}$) in m experiments, the family-wise error rate (FWER) is defined as

$$FWER := P(N_{1|0} \geq 1). \tag{3.33}$$

For m independent tests with a significance threshold α, we can then derive the FWER by

$$FWER = P(N_{1|0} \geq 1) = 1 - (1 - \alpha)^m. \tag{3.34}$$

Definition 3.12 (Bonferroni multiple hypothesis testing correction) *Let us consider a multiple hypothesis testing experiment as introduced in Table 3.1.*
Then the Bonferroni P value correction for a P value p_k as part of m experiments is defined as

$$p^* := p_k \cdot m \tag{3.35}$$

or, equivalently, a corrected α threshold controlling the type I error is defined as

$$\alpha^* = \frac{\alpha}{m}. \tag{3.36}$$

Definition 3.13 (False discovery proportion (FDP) and false discovery rate (FDR)) *Let us consider a multiple hypothesis testing experiment as introduced in Table 3.1.*
When performing m tests with a threshold α controlling the type I error ($N_{1|0}$) the false discovery proportion (FDR) is defined as

$$FDP := \begin{cases} \frac{N_{1|0}}{R}, & \text{if } R > 0 \\ 0, & \text{if } R = 0 \end{cases}, \tag{3.37}$$

the false discovery rate (FDR) is defined as

$$FDR := \mathbf{E}[FDP] \tag{3.38}$$

and the positive false discovery rate (pFDR) is defined as

$$pFDR = q := \mathbf{E}[FDP|R > 0]. \tag{3.39}$$

Definition 3.14 (Step-up Benjamini-Hochberg FDR control) *Let us consider a multiple hypothesis testing experiment as introduced in Table 3.1.*
Then for ordered P values $0 \leq p_1 \leq \cdots \leq p_m \leq 1$ from m experiments consider

$$q_k = p_k \frac{m}{k}. \tag{3.40}$$

Then for the largest k^*, such that $p_{k^*} \leq q_{k^*}$, the adjusted P values p_k^* are defined iteratively as

$$p_k^* := \begin{cases} q_k = p_k \frac{m}{k}, & \text{if } k \in 1, \ldots, k^* \\ \min(q_k, q_{k+1}), & \text{if } k \in k^* + 1, \ldots, m \end{cases}. \tag{3.41}$$

Then a hypothesis test is considered significant with respect to the FDR significance level α if $p_k^* < \alpha$.

Definition 3.15 (Storey's q-value method) *Let us consider a multiple hypothesis testing experiment as introduced in Table 3.1.*
Then for a significance cutoff α with $0 < \alpha < 1$, the FDR can be approximated as

$$FDR(\alpha) \approx \frac{\mathbf{E}[N_{1|0}(\alpha)]}{\mathbf{E}[R(\alpha)]} \approx \frac{M_0 \cdot \alpha}{R(\alpha)} \approx \frac{\textit{False positives for } \alpha}{\textit{Number of significant tests for } \alpha}. \tag{3.42}$$

Accordingly, we can define $\pi_0 \equiv \frac{M_0}{m}$ as the fraction of true nulls and derive an estimate depending on a tuning parameter λ as

$$\hat{\pi}_0(\lambda) = \frac{\#\{p_j > \lambda \quad \textit{for } j \in 1, \ldots, m\}}{m(1-\lambda)}. \tag{3.43}$$

The estimate of π_0, $\hat{\pi}_0$, can be used to derive q values controlling the FDR [Storey and Tibshirani, 2003]. Furthermore, $(1 - \hat{\pi}_0)$ can also be used to evaluate replication rates when comparing matched P value distributions of two experiments. For reasonable

distributions of P values, $\hat{\pi}_0$ estimates should get more reliable for higher parameters of λ.
Leveraging that, the R Bioconductor package qvalue[1] automatically fits a natural cubic spline with three degrees of freedom to several $\hat{\pi}_0(\lambda)$ and returns a $\hat{\pi}_0$ estimate based on the limit of λ towards 1.

3.2.5 MatrixEQTL

QTL analyses evaluate the association of genetic variants with a quantitative trait using linear regression or ANOVA models. In a standard eQTL experiment, many variants and genes are evaluated adding up to an extensive amount of regressions to be computed resulting in a heavy computational burden. Therefore, more efficient implementations to derive results were needed.

In this thesis, we will make use of the R package MatrixEQTL[2] [Shabalin, 2012]. Using special preprocessing techniques and expressing the most intense computational parts in terms of large matrix operations as summarized in the following, they successfully achieved a speed-up of two to three orders of magnitude compared to using the base R implementation of linear regression analysis.

Test statistics highly depend on sample correlations. Applying certain normalization procedures once to every SNP and gene, they significantly reduce the complexity for deriving further computations.
Computing a test statistic is relatively cheap compared to estimating the corresponding P value. Therefore, MatrixEQTL uses a pre-defined significance threshold set by the user and translated to the matched test statistic, to only derive P values for associations of interest based on test statistic cutoffs.
In order to make huge computations feasible with respect to memory requirements, results are evaluated in blocks of $10^4 \times 10^4$ variables.
Similarly, a slightly modified definition of the Benjamini-Hochberg procedure is used, summarizing the ranks of P values which were not computed:

Let $p_1 < p_2 < \cdots < p_K$ be the P values which passed the significance threshold and m be the total number of test performed. Then

$$q_K = \frac{m}{K} p_K \qquad \text{and} \tag{3.44}$$

$$q_k = min\left(\frac{m}{k} p_k, q_k + 1\right) \qquad \text{for k = 1, ..., K-1.} \tag{3.45}$$

[1] `http://github.com/jdstorey/qvalue`
[2] `https://github.com/andreyshabalin/MatrixEQTL`

3.2.6 Genetic risk scores

With the growing number of GWAS studies, more and more variants associated with disease have been discovered. Causal variants are often obscured by e.g. LD leading to large numbers of highly correlated features.

However, it is of great interest to summarize genetic predisposition in a more dense form. Therefore, genetic risk scores (GRS) combining multiple risk variants have been introduced.

Definition 3.16 (Classical genetic risk score) *Let $X = x_{ij} \in \mathbb{R}^{n \times p}$ be a data matrix with n observations (samples) $i \in 1, \ldots, n$ and p features (genetic variants) $j \in 1, \ldots, p$ coded as absolute number of risk alleles per variant.*
Then the weighted genetic risk score GRS for sample i is defined as

$$GRS_i = \sum_{j=1}^{p} \beta_j \cdot x_{ij} \qquad \text{with weights } \beta_j, \text{ for } j \in 1 \ldots p. \tag{3.46}$$

While this is a very simple concept, the selection of variants and deriving the weights β is not.

3.2.6.1 Regression based approaches

Most algorithms tried to solve this problems using classical variable selection methods such as LASSO [Tibshirani, 1996] or other penalized regression models, restricting the number of variants taken into account.

Another class of commonly used approaches is summarized by the thresholding and pruning term. For existing GWAS summary statistics, considering a specific significance cutoff, variants are ordered based on their association with the trait of question. Starting with the most significant variant, only this variant is kept while all variants in close proximity and in high LD are removed. Finally, only one representative SNP per LD clump is selected for the GRS across the whole genome. Corresponding weights in the risk score are often derived from the original GWAS summary statistics.

Of course, many more variations and extensions exist.

3.2.6.2 LDpred based genome-wide polygenic risk scores

In contrast, a new class of genome-wide polygenic risk scores (PRS) derived using the LDpred algorithm [Vilhjálmsson et al., 2015] have been recently used with great success [Khera et al., 2018]. The general idea is to rather consider only a fraction of variants, denoted by ρ, to be actual causal and adjust the corresponding weights

of correlated variants rather than excluding them from the analysis. These PRS are based on a Bayesian approach modeling the posterior mean effect using a point-normal mixture prior which is dependent on LD information and the tuning parameter ρ.
For details, please refer to the original publication [Vilhjálmsson et al., 2015].

3.3 Pathway enrichment methods

3.3.1 Gene set enrichment analysis (GSEA)

Subramanian et al. [2005] and Mootha et al. [2003] introduced the concept of gene set enrichment analysis (GSEA), where a list of genes was ranked based on the correlation with a phenotype. Using this ranking, it can be assessed whether a predefined set of genes is accumulating at the top (i.e. positive correlation) or bottom (i.e. negative correlation) as visualized in Figure 3.2. Similarly, a test statistic can be used for the ranking and genes belonging to the same pathway can be defined as one corresponding gene set. In order to assess statistical significance, the original publication proposed applying a weighted Kolmogorov-Smirnov-like statistic described theoretically in Hollander et al. [2015]. The final enrichment score (ES) was then defined as the maximum deviation from zero per gene set when running down the ranked list. Empirical P values were derived by permuting phenotype labels of the original data, recalculating the gene ranking and ES in order to estimate the null distribution of the ESs.
Permuting phenotype labels and recomputing the statistics can be computationally quite demanding. Alternatively, for each gene set of the same size, only gene labels can be permuted instead. As also introduced by Subramanian et al. [2005], empirical P values can then derived by comparing the actual ES values to the random null distribution.

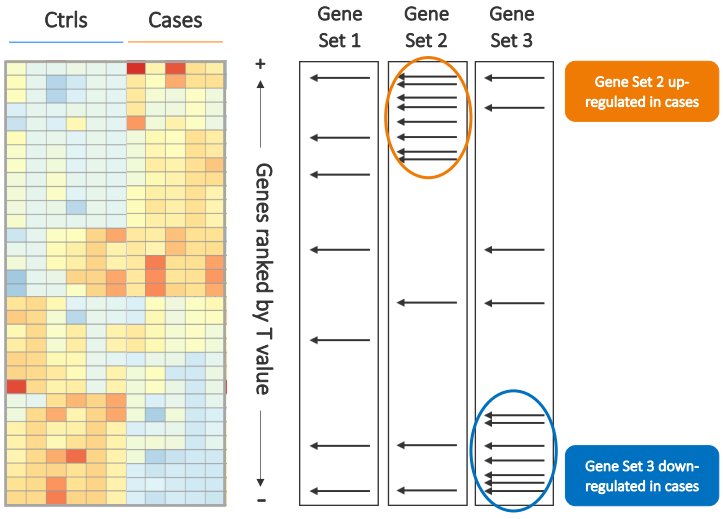


Figure 3.2: Basic idea of gene set enrichment analysis.
Given the summary statistics of a differential expression experiment, genes can be ranked according to their correlation with the phenotype. Different gene sets based on pathway annotations can then be evaluated by assessing if genes belonging to the same gene set accumulate at top or bottom of the gene ranking. Ctrls, controls;

This procedure was extended by Korotkevich et al. [2019] who provided an efficient implementation with the R Bioconductor package fgsea[3]. Let R be a list of gene level statistics with $R_i > R_j$ for $i < j$ for n genes thus, $i \in H = \{1, 2, \ldots, n\}$, and let T be a list of gene sets representing pathway annotations.

[3] https://bioconductor.org/packages/release/bioc/html/fgsea.html

For gene set $t \in T$ with size $|t| = k$ and the summed ranks $NR = \sum_{i \in t} |R_i|$, the enrichment score $ES(t)$ for gene set t is defined based on the running sum $ES_i(t)$

$$ES_i(t) = \begin{cases} 0 & \text{if } i = 0, \\ ES_{i-1}(t) + \frac{1}{NR}|R_i| & \text{if } 1 \leq i \leq n \text{ and } i \in t, \\ ES_{i-1}(t) - \frac{1}{n-k} & \text{if } 1 \leq i \leq n \text{ and } i \notin t \end{cases} \tag{3.47}$$

as

$$ES(t) = ES_{i^*} \quad \text{where } i^* = \arg\max_i |ES_i|. \tag{3.48}$$

Empirical P values for the null distribution of the ES for a gene set of size k can then be derived by permutation analysis.

3.3.2 Bayesian approaches for modeling pathway activations

GSEA in its original formulation as described above has some significant shortcomings with respect to highly overlapping gene sets, as already discussed in the original publication [Subramanian et al., 2005].

Instead of testing each category separately for enrichment, Bayesian approaches can estimate a posterior probability for each term to be active by updating the distribution for a hypothesis based on evidence from a differential expression experiment. This is realized by a Bayesian network, which we will introduce in more detail in the following.

Complementary to GSEA, a significance cutoff is chosen and each gene is therefore observed as either not significant (inactive or off) or significant (active or on).
The core of such a Bayesian network model is based on three layers: the term layer T, the hidden gene layer H and the observation layer O. Nodes between the term and hidden gene layer are connected based on the gene set annotations and activations are modeled with the observation layer. Additionally, we have a prior for each term being active, as well as false positive (α) and false negative (β) probabilities for observing the hidden state of each gene as visualized in Figure 3.3.

Definition 3.17 (Bayesian network base model) *Based on Figure 3.3, the term, hidden gene and observation layer can be modeled as boolean variables that can be active (1, on) or inactive (0, off). For each hidden gene H_i, we can define the set of terms it is annotated to as $T(H_i)$. Accordingly, we consider a hidden node active if any of the terms $T(H_i)$, which H_i is annotated to, is active, i.e.*

$$P(H_i = 1|T) = \begin{cases} 1 & \textit{if } \exists T_j \in T(H_i) \colon T_j = 1 \\ 0 & \textit{otherwise.} \end{cases} \tag{3.49}$$

Depending on the state of a hidden node, we can compare the corresponding observation and

derive error rates α and β:

$$P(O_i = 1|H_i) = \begin{cases} 1-\alpha & \text{if } H_i = 1 \quad \textit{(true positives, states match)} \\ \alpha & \text{if } H_i = 0 \quad \textit{(false positives, hidden node is inactive)} \end{cases} \tag{3.50}$$

$$P(O_i = 0|H_i) = \begin{cases} 1-\beta & \text{if } H_i = 0 \quad \textit{(true negatives, states match)} \\ \beta & \text{if } H_i = 1 \quad \textit{(false negatives, hidden node is active)} \end{cases} \tag{3.51}$$

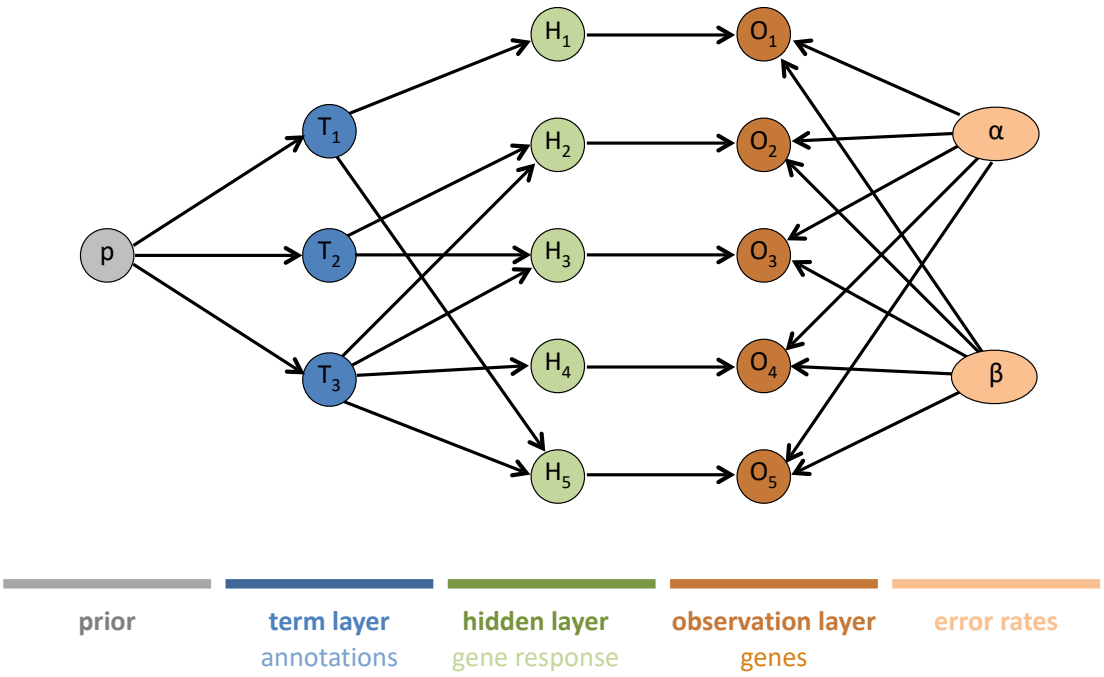


Figure 3.3: Bayesian network modeling pathway activations.
Nodes in the term, hidden and observation layer are modeled as boolean nodes. All terms share a prior p and the edges connecting terms and hidden nodes are defined by e.g. pathway annotations or a similar gene set collection. Hidden nodes are connected to their corresponding observation. All observations share the same false positive (α) and false negative (β) probabilities.

In the following, we will present two implementations including extensions of this base model. However, first let us introduce some basic concepts from statistics:

Definition 3.18 (Conditional probability) *For two events A and B and $P(B) \neq 0$, the conditional probability of A given B is true, $P(A|B)$, is defined as*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{3.52}$$

Definition 3.19 (Statistical independence) *Two events A and B are independent $A \perp B$ if and only if their joint probability equals the product of their probabilities, i.e.*

$$P(A \cap B) = P(A)P(B) \qquad \textit{for } A \neq B. \tag{3.53}$$

Theorem 3.1 (Bayes' theorem) *Let A and B be events and $P(B) \neq 0$. Let $P(A)$ and $P(B)$ be the (marginal) probabilities of observing A and B, let $P(A|B)$ be the conditional probability of event A occurring given that B is true and $P(B|A)$ be the conditional probability of event B occurring given that A is true, respectively. Then*

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad \text{for } A \neq B. \tag{3.54}$$

Bayes' theorem is often used in statistics for Bayesian inference. In this case, a posterior probability - $P(A|B)$ in the above notation - is derived using a prior probability $P(A)$ and additional information, e.g. a likelihood function, $P(B|A)$ from either observed data or a statistical model. The marginal likelihood $P(B)$ is independent of A and therefore does not change the probabilities of different events A.

3.3.3 Model-based gene set analysis (MGSA)

Model-based gene set analysis (MGSA) [Bauer et al., 2010] models the state propagation from terms to nodes using local probability distributions based on the Bayesian network introduced in Definition 3.17 and Figure 3.3.
Following the scheme of the model, for known values of the prior p, false positive rates α and false negative rates β, the joint probability distribution can be derived as

$$\begin{aligned} P(T, H, O) &= P(T)P(H|T)P(O|H) \\ &= P(T)\prod_{i=1}^{n} P(H_i|T)P(O_i|H_i), \end{aligned} \tag{3.55}$$

leveraging the graph structure of the network in Figure 3.3 and the link between term and hidden node activations from Definition 3.17.

Let $n_{oh|T} = |\{i|O_i = o \wedge H_i = h\}|$ be the number of genes observed as true positives, true negatives, false positives and false negatives based on the configurations of combinations of $o, h \in \{0, 1\}$ given a set of active terms T. By using Bernoulli distributions, the probability to observe gene activations given T can be written as

$$\begin{aligned} P(O|T) &= \prod_{i=1}^{n} P(H_i|T)P(O_i|H_i) \\ &= \alpha^{n_{10|T}}(1-\alpha)^{n_{00|T}}(1-\beta)^{n_{11|T}}\beta^{n_{01|T}}. \end{aligned} \tag{3.56}$$

The marginal probabilities (i.e. specifically $P(T)$) cannot be derived analytically, therefore a Metropolis-Hastings algorithm based on a Marcov chain Monte Carlo (MCMC) method is used. By performing a random walk over the term and parameter configurations, the target distribution $P(T|O)$ is sampled asymptotically. Detailed descriptions on the underlying theory and implementations can be found in Andrieu et al. [2003], Diaconis [2009] and Diaconis and Saloff-Coste [1998].

Briefly, new states are sampled based on a proposal density function $Q_T(\cdot|T^t)$ for the current state T^t by considering $N(T^t)$, the number of possible transitions to a neighborhood by changing the state of only one term. If the new configuration (T^u) fits the observed data better, it is accepted as a next step with a higher probability. Taking into account the transition probabilities and Bayes' theorem, the acceptance probability can be simplified and written as

$$P_{accept}(T^t, T^u) = \frac{P(O|T^u)P(T^u)N(T^t)}{P(O|T^t)P(T^t)N(T^u)}. \tag{3.57}$$

Let $C(T_j)$ bet the number of sampled configurations where term T_j was active. Then a posterior for T_j, $P(T_j|O)$, can be derived as the fraction of iterations, where the term T_j was active, i.e.

$$P(T_j|O) \approx \frac{C(T_j)}{L}$$

with l the number of iterations after the burn-in phase.

Now this needs to be adjusted to the case of unknown α and β which must be considered in the joint probability distribution, i.e.

$$P(p, T, H, \alpha, \beta, O) = P(p)P(T|p)P(H|T)P(\alpha)P(\beta)P(O|H, \alpha, \beta) \tag{3.58}$$

with uniform priors $U(0,1)$ on p, α and β.

Details concerning the proposal density depending on state and parameter transitions as well as determination of the neighborhoods specifying the application of the Metropolis-Hastings algorithm as shown in Algorithm 3.1, can be found in the original publication [Bauer et al., 2010].

Algorithm 3.1 (Metropolis-Hastings algorithm to sample posterior term probabilities)
Input: *Observations O, number of steps L.*
Initialize $T^t \leftarrow \underbrace{(0, \ldots, 0)}_{m \text{ times}}$*;*
For $l = 1, \ldots, L$
  $T^u \sim Q_T(\cdot|T^t)$*, i.e. chosse a neighbor candidate by either*
  · toggling a term
  · exchanging an active term with an inactive one
  $a \leftarrow \frac{P(O|T^u)P(T^u)N(T^t)}{P(O|T^t)P(T^t)N(T^u)}$
  $r \sim U(0,1)$
  If $r < a$ ***then***
    $T^t \leftarrow T^u$
  end
end
Return *Approximation of ther posterior term probabilities*
$(P(T_1 = 1|O), \ldots, P(T_m = 1|O)) = \left(\frac{C(T_1)}{L}, \ldots, \frac{C(T_m)}{L}\right)$

3.3.4 Multi-level ontology analysis MONA

Multi-level ontology analysis (MONA)[4] [Sass et al., 2013] is relying on the same base model as described in Definition 3.17 and Figure 3.3.
Compared to MGSA, instead of uniform distributions for priors p, α and β, a Beta-function $\text{Beta}(a, b)$ is chosen.
The MONA is based on a modular framework, where different extensions for the hidden and observation layer can be added on top of the original base model.

3.3.4.1 MONA cooperative model

The authors extend the base model to a two-omics cooperative model suited for the integration of two molecular modalities, e.g. transcriptomics and proteomics.

To this end, they introduce a second set of observations (O_i^I, O_i^{II}) and error rates (α^I, β^I, α^{II}, β^{II}) that still refer to the same hidden node H_i:

$$P(O_i^I = 1|H_i) = \begin{cases} 1-\alpha^I & \text{if } H_i = 1 \quad \text{(true positives, states match)} \\ \alpha^I & \text{if } H_i = 0 \quad \text{(false positives, hidden node is inactive)} \end{cases} \tag{3.59}$$

$$P(O_i^I = 0|H_i) = \begin{cases} 1-\beta^I & \text{if } H_i = 0 \quad \text{(true negatives, states match)} \\ \beta^I & \text{if } H_i = 1 \quad \text{(false negatives, hidden node is active)} \end{cases} \tag{3.60}$$

$$P(O_i^{II} = 1|H_i) = \begin{cases} 1-\alpha^{II} & \text{if } H_i = 1 \quad \text{(true positives, states match)} \\ \alpha^{II} & \text{if } H_i = 0 \quad \text{(false positives, hidden node is inactive)} \end{cases} \tag{3.61}$$

$$P(O_i^{II} = 0|H_i) = \begin{cases} 1-\beta^{II} & \text{if } H_i = 0 \quad \text{(true negatives, states match)} \\ \beta^{II} & \text{if } H_i = 1 \quad \text{(false negatives, hidden node is active)} \end{cases} \tag{3.62}$$

By linking the different observations O_i^I, O_i^{II} to the same hidden node H_i, term activations and both observations layers are coupled directly. However, since observations may heavily depend on the type of measurement technique they were derived from, false positive and false negative error rates are shared for all nodes occurring in one omic, but modeled separately for the different omic types I and II.

[4] https://www.helmholtz-munich.de/icb/research/labs/computational-cell-maps/projects/mona/index.html

3.3.4.2 Bayesian inference implementation

While the MCMC approach asymptotically provides a sampler of the target distribution, it has potentially long runtimes and is very specific to the model design.
To avoid the corresponding convergence issues, the authors made use of expectation propagation (EP) to infer the marginal probabilities. EP approaches approximate a target distribution in a factorized form. The Kullback-Leibler (KL) divergence,

$$\mathrm{KL}(p||q) = \int p(x) \log \frac{p(x)}{q(x)} dx, \tag{3.63}$$

can be used as a measure to assess differences of distributions. Therefore, the factorized marginal probabilities can be estimated by iteratively minimizing the KL divergence.

In the case of the two-omics cooperative model, the posterior depends on the parameter vector

$$\boldsymbol{\theta} = \{p, T, H, \alpha^I, \alpha^{II}, \beta^I, \beta^{II}\}. \tag{3.64}$$

Similarly to the MGSA joint probability (Equation 3.58), the posterior $P(\boldsymbol{\theta}|O)$ can be factorized as

$$P(\boldsymbol{\theta}|O) = \frac{P(T|p)P(O|H, \alpha^I, \alpha^{II}, \beta^I, \beta^{II})P(H|T)P(\alpha^I)P(\alpha^{II})P(\beta^I)P(\beta^{II})P(p)}{P(O)} \tag{3.65}$$

$$= \frac{1}{P(O)} \prod_{k=1}^{K} f_k(\boldsymbol{\theta}), \tag{3.66}$$

with the K different factors $f_k(\boldsymbol{\theta})$ to infer.

Instead of sampling the full posterior distribution, each factor $f_k(\boldsymbol{\theta})$ is approximated based on minimizing the Kullback-Leibler divergence by matching the first two moments resulting in the following algorithm 3.2 [Sass et al., 2013]:

Algorithm 3.2 (Expectation propagation for approximating the MONA model posterior)
Input: *Factorized posterior*

$$P(\boldsymbol{\theta}|O) = \frac{1}{P(O)} \prod_{k=1}^{K} f_k(\boldsymbol{\theta}).$$

Initialize Gaussian term approximations $\tilde{f}_k(\boldsymbol{\theta})$, $k \in 1, \ldots, K$;
Repeat
  For $j = 1, \ldots, K$
    Update $\tilde{f}_j$ *such that*

$$\tilde{f}_j(\boldsymbol{\theta}) \prod_{i \neq j} \tilde{f}_i(\boldsymbol{\theta}) \text{ minimizes KL-divergence from } f_j(\boldsymbol{\theta}) \prod_{i \neq j} \tilde{f}_i(\boldsymbol{\theta})$$

  end
until *all* $\tilde{f}_j(\boldsymbol{\theta})$ *converged;*
Approximate

$$P(O) \approx \tilde{Z} = \int \prod_{k=1}^{K} \tilde{f}_k(\boldsymbol{\theta}) \, \mathrm{d}\, \boldsymbol{\theta};$$

Return *Approximation* $Q(\boldsymbol{\theta}|O)$ *of* $P(\boldsymbol{\theta}|O)$ *with*

$$Q(\boldsymbol{\theta}|O) = \frac{1}{\tilde{Z}} \prod_{k=1}^{K} \tilde{f}_k(\boldsymbol{\theta});$$

Additionally, Sass et al. [2013] also introduce an inhibitory model for the analysis of miRNA interference for mRNAs. Since this variation was not used in this thesis, we chose to focus on the cooperative model only. However, it does show the flexibility of the approach in tailoring the underlying network architecture to the assumed regulatory mechanism.

3.4 Methods for quantitative trait loci analyses

This section is based on and partly identical to the publication by Assum et al. [2022a] and also available as a preprint on *bioRxiv*[5,6]:
Tissue-specific multi-omics analysis of atrial fibrillation[7]
Ines Assum†, Julia Krause†, Markus O. Scheinhardt, Christian Müller, Elke Hammer, Christin S. Börschel, Uwe Völker, Lenard Conradi, Bastiaan Geelhoed, Tanja Zeller*, Renate B. Schnabel* and Matthias Heinig*, *Nature Communications* **13**, 441 (2022). Authors marked with † or * contributed equally to this work.
Code related to this project is available at `https://github.com/heiniglab/symatrial`[8] [Assum and Heinig, 2021].

Unless stated otherwise explicitly, the Benjamini-Hochberg procedure was used to estimate the false discovery rate (FDR) per omic type and to account for multiple hypothesis testing.

Definition 3.20 (Highly variable genes) *Let m_G be the median expression and σ_G^2 be the variance of transcript measurements of gene G across N observations (samples). Then a gene i can be defined as highly variable, if for the variance σ_G^2 and estimates $\hat{\beta}_0$, $\hat{\beta}_1$ of the linear model*

$$\begin{aligned} log(\sigma_G^2) &\sim \beta_0 + \beta_1 \cdot m_G + \varepsilon \quad and \\ \sigma_{G_i}^2 &> exp\left(\hat{\beta}_0 + \hat{\beta}_1 \cdot m_G + 3 \cdot \hat{\sigma}^2\left(\hat{\beta}_1\right)\right). \end{aligned} \tag{3.67}$$

Partial correlations The R package ppcor [Seongho, 2015] was used to evaluate partial correlations.

3.4.1 Genotypes, transcriptomics and proteomics data

For QTL analyses, outliers in the expression data which coincide with rare genotypes can lead to false positive findings. For SNPs with less than three individuals having an homozygous-minor-allele genotype, all samples with homozygous-minor-allele genotype were therefore recoded to heterozygous genotype.
Preprocessing of the transcriptomics and proteomics data was described in sections 2.1.2.3 and 2.1.2.4. The technical covariates RIN-score (RIN) and the protein concentration of the original tissue sample (prot. conc.) were used in further analyses where appropriate.
Matched genotypes and mRNA measurements were available for 75 out of 102 individuals with expression data for 26 376 genes. Similarly, proteomics data were available for 75 individuals. The 1 337 proteins suitable for *cis* QTL analysis contained 62 missing

[5] `https://www.biorxiv.org/content/10.1101/2020.04.06.021527v1`
[6] `https://www.biorxiv.org/content/10.1101/2020.04.06.021527v2`
[7] `https://doi.org/10.1038/s41467-022-27953-1`
[8] `https://doi.org/10.5281/zenodo.5094276`

values (0.06 % of all values). As PEER factors can only be inferred from datasets without missing values, proteomics were imputed using the KNN-method implemented in the R bioconductor package impute[9].

3.4.1.1 Protein analysis using Western blot

Transcription factors show much lower abundance compared to other structural proteins in atrial tissue. Therefore, the TF NKX2-5 was specifically assessed by Julia Krause using Western blot analysis: "Human atrial tissue samples (15 mg each) were pulverized in liquid nitrogen and lyzed with M-PER Mammalian Protein Extraction Reagent (Thermo Scientific) supplemented with protease inhibitor. Protein concentrations were measured using a BCA assay (Thermo Scientific). The same amount of protein for each sample was heated at 95 °C for 10 minutes in 1x Laemmli. Proteins were separated on a 10 % SDS-PAGE gel and transferred to nitrocellulose membranes. Membranes were blocked with 5 % skim milk in TBS-T for 1 hour. Staining with the primary antibody was performed overnight at 4 °C, and secondary antibody staining for 1 hour at room temperature. The following primary antibodies were used: NKX2-5 (ab205263, 1:1 000), alpha actinin (CST #3134, 1:1 000), GAPDH (CST #3683, 1:2 000). The following HRP-conjugated secondary antibody was used: goat anti-rabbit IgG (PI-1000-1, 1:10 000). The antibodies were visualized with enhanced chemiluminescence (ECL) detection reagent (Bio-Rad #1705060) or the SuperSignal West Pico PLUS chemiluminescent substrate (Thermo Scientific #34579). The membranes were reprobed with GAPDH antibody after incubation with stripping buffer (Thermo Scientific #46430) for 4 minutes, washing and blocking with 5 % skim milk in TBS-T. Antibody detection was performed with a chemiluminescence imaging system (FUSION Solo S). Blot analyses were achieved with the Image Lab software (Bio-Rad 6.1)." [Assum et al., 2022a].

3.4.1.2 Protein-per-mRNA ratios

mRNA and protein measurements were per-sample quantile-normalized and log-transformed in the course of standard data preprocessing. In order to assess individual relative protein-per-mRNA ratios, measurements were additionally quantile-normalized per gene before computing the difference of the (log-transformed) values.

3.4.1.3 Residuals

For each gene, we evaluated variation shared between omic levels across the individuals. Therefore, residuals were computed using linear regression analysis on the per-sample quantile-normalized, log-transformed mRNA and protein values. Residuals were then

[9] `https://bioconductor.org/packages/release/bioc/html/impute.html`

derived from evaluating the following models as illustrated in Figure 3.4:

$$\text{mRNA} \sim \beta_0 + \beta_1 \cdot \text{protein} + \varepsilon \quad \text{for } \textbf{mRNA residuals} \text{ and} \tag{3.68}$$

$$\text{protein} \sim \beta_0 + \beta_1 \cdot \text{mRNA} + \varepsilon \quad \text{for } \textbf{protein residuals}. \tag{3.69}$$

Covariates were not included in the computation of residuals but used for further analyses.

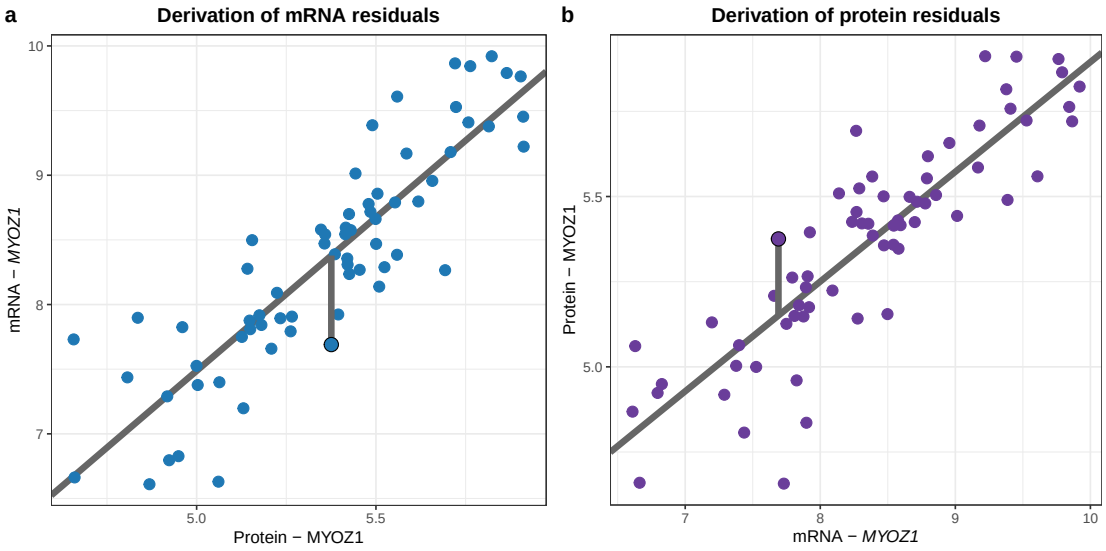


Figure 3.4: mRNA and protein residual derivation for the gene MYOZ1.
a: Transcript-protein correlation and visualization of the definition of mRNA residuals based on model 3.68.
b: Protein-transcript correlation and visualization of the definition of protein residuals based on model 3.69.

3.4.2 Annotations

3.4.2.1 Genome annotations

Ensembl BioMart [Kinsella et al., 2011] GRCh37.p13 hg19 annotations were used as genome annotations. Specifically, transcript start and end, transcription start site (TSS) and exon annotations per gene were downloaded[10]. Additionally, promoter regions were derived from the Gencode [Frankish et al., 2019] v31lift37[11] basic and long non-coding RNA transcript start annotations.

[10] `http://feb2014.archive.ensembl.org/biomart/martview/`
[11] `https://www.gencodegenes.org/human/release_19.html`

Table 3.2: Ensembl variant effect prediction (VEP) consequences of variants.
Table source data taken from
`http://www.ensembl.org/info/genome/variation/prediction/predicted_data.html`.

Display term	SO description	SO accession	IMPACT
Transcript ablation	A feature ablation whereby the deleted region includes a transcript feature	SO:0001893	HIGH
Splice acceptor variant	A splice variant that changes the 2 base region at the 3′ end of an intron	SO:0001574	HIGH
Splice donor variant	A splice variant that changes the 2 base region at the 5′ end of an intron	SO:0001575	HIGH
Stop gained	A sequence variant whereby at least one base of a codon is changed, resulting in a premature stop codon, leading to a shortened transcript	SO:0001587	HIGH
Frameshift variant	A sequence variant which causes a disruption of the translational reading frame, because the number of nucleotides inserted or deleted is not a multiple of three	SO:0001589	HIGH
Stop lost	A sequence variant where at least one base of the terminator codon (stop) is changed, resulting in an elongated transcript	SO:0001578	HIGH
Start lost	A codon variant that changes at least one base of the canonical start codon	SO:0002012	HIGH
Transcript amplification	A feature amplification of a region containing a transcript	SO:0001889	HIGH
Inframe insertion	An inframe non synonymous variant that inserts bases into in the coding sequence	SO:0001821	MODERATE
Inframe deletion	An inframe non synonymous variant that deletes bases from the coding sequence	SO:0001822	MODERATE
Missense variant	A sequence variant, that changes one or more bases, resulting in a different amino acid sequence but where the length is preserved	SO:0001583	MODERATE
Protein altering variant	A sequence_variant which is predicted to change the protein encoded in the coding sequence	SO:0001818	MODERATE
Splice region variant	A sequence variant in which a change has occurred within the region of the splice site, either within 1-3 bases of the exon or 3-8 bases of the intron	SO:0001630	LOW
Incomplete terminal codon variant	A sequence variant where at least one base of the final codon of an incompletely annotated transcript is changed	SO:0001626	LOW
Start retained variant	A sequence variant where at least one base in the start codon is changed, but the start remains	SO:0002019	LOW
Stop retained variant	A sequence variant where at least one base in the terminator codon is changed, but the terminator remains	SO:0001567	LOW
Synonymous variant	A sequence variant where there is no resulting change to the encoded amino acid	SO:0001819	LOW
Coding sequence variant	A sequence variant that changes the coding sequence	SO:0001580	MODIFIER
Mature miRNA variant	A transcript variant located with the sequence of the mature miRNA	SO:0001620	MODIFIER
5 prime UTR variant	A UTR variant of the 5′ UTR	SO:0001623	MODIFIER
3 prime UTR variant	A UTR variant of the 3′ UTR	SO:0001624	MODIFIER
Non coding transcript exon variant	A sequence variant that changes non-coding exon sequence in a non-coding transcript	SO:0001792	MODIFIER
Intron variant	A transcript variant occurring within an intron	SO:0001627	MODIFIER
NMD transcript variant	A variant in a transcript that is the target of NMD	SO:0001621	MODIFIER
Non coding transcript variant	A transcript variant of a non coding RNA gene	SO:0001619	MODIFIER
Upstream gene variant	A sequence variant located 5′ of a gene	SO:0001631	MODIFIER
Downstream gene variant	A sequence variant located 3′ of a gene	SO:0001632	MODIFIER
TFBS ablation	A feature ablation whereby the deleted region includes a transcription factor binding site	SO:0001895	MODIFIER
TFBS amplification	A feature amplification of a region containing a transcription factor binding site	SO:0001892	MODIFIER
TF binding site variant	A sequence variant located within a transcription factor binding site	SO:0001782	MODIFIER
Regulatory region ablation	A feature ablation whereby the deleted region includes a regulatory region	SO:0001894	MODERATE
Regulatory region amplification	A feature amplification of a region containing a regulatory region	SO:0001891	MODIFIER
Feature elongation	A sequence variant that causes the extension of a genomic feature, with regard to the reference sequence	SO:0001907	MODIFIER
Regulatory region variant	A sequence variant located within a regulatory region	SO:0001566	MODIFIER
Feature truncation	A sequence variant that causes the reduction of a genomic feature, with regard to the reference sequence	SO:0001906	MODIFIER
Intergenic variant	A sequence variant located in the intergenic region, between genes	SO:0001628	MODIFIER

3.4.2.2 Variant Effect Predictions

Depending on the type and localization of variants in protein-coding or non-coding regions, different consequences of genetic variation for transcripts and proteins can be observed. Ensembl Variant Effect Predictions (VEP) [McLaren et al., 2016] catalogue most likely effects of sequence variants in a comprehensive fashion represented by different labels (see Table 3.2).

For our dataset, VEP based on rs-ID-gene-pairs were obtained from the BioMart GRCh37.p13 hg19 download page. VEP were downloaded from the Ensembl Biomart GRCh37.p13 based on SNP rs-IDs. The label "Missense" was used to summarize all possible missense consequences of the variant (gained stop codon, a frameshift/amino-acid altering/protein-altering variant, a lost start/stop codon, an inframe insertion/dele-

tion).

3.4.2.3 GWAS catalog

The GWAS catalog[12] [Buniello et al., 2019] systematically summarizes the results of GWAS across various traits such as diseases, measurements or other phenotypes. Traits are organized in 17 categories in the EFO-mapping[13], of which we were interested in cardiovascular measurements (EFO_0004298) and cardiovascular disease (EFO_0000319). We will refer to those two categories as "cardiovascular traits". We further defined the label "arrhythmias", including the subset of traits annotated to atrial fibrillation, cardiac arrhythmia, sudden cardiac arrest, supraventricular ectopy, early cardiac repolarization measurement, heart rate, heart rate variability measurement, P wave duration, P wave terminal force measurement, PR interval, PR segment, QRS amplitude, QRS complex, QRS duration, QT interval, R wave amplitude, resting heart rate, RR interval, S wave amplitude and T wave amplitude. Finally, we also considered the specific group "AF" of terms connected to AF based on the traits Atrial fibrillation and QT interval.

Table 3.3: ChromHMM chromatin states for the 15 state model.
Table source data taken from
`https://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html`.

State no.	Mnemonic	Description	Open chromatin
1	TssA	Active TSS	1
2	TssAFlnk	Flanking Active TSS	1
3	TxFlnk	Transcr. at gene 5′ and 3′	0
4	Tx	Strong transcription	0
5	TxWk	Weak transcription	0
6	EnhG	Genic enhancers	1
7	Enh	Enhancers	1
8	ZNF/Rpts	ZNF genes & repeats	0
9	Het	Heterochromatin	0
10	TssBiv	Bivalent/Poised TSS	1
11	BivFlnk	Flanking Bivalent TSS/Enh	1
12	EnhBiv	Bivalent Enhancer	1
13	ReprPC	Repressed PolyComb	0
14	ReprPCWk	Weak Repressed PolyComb	0
15	Quies	Quiescent/Low	0

3.4.2.4 Chromatin states

Since no epigenetic annotations were available for the AFHRI-B cohort, we relied on public data from the Roadmap Epigenomics project [Roadmap Epigenomics Consortium et al., 2015] for chromatin states. The ChromHMM 15 state model coremarks presented in Table 3.3 trained on human heart right atrial appendage tissue

[12] `https://www.ebi.ac.uk/gwas/`, downloaded 2019-11-26
[13] `https://www.ebi.ac.uk/gwas/api/search/downloads/trait_mappings`, downloaded 2019-11-26

(E104_15_coreMarks_dense.bed) were therefore used. As noted in the "Open chromatin" column in Table 3.3, these states were also used to filter for open chromatin regions with possible active TF binding.

3.4.2.5 Binding sites

Transcription factor binding sites (TF BS) TF BS were based on ChIP-seq data from the ReMap TF database [Chèneby et al., 2018] (ReMap 2018 v1.2[14]) and GSE133833[15], which contained NKX2-5 BS from human iPSC-derived cardiomyocytes [Benaglio et al., 2019]. As those annotations were not tissue-specific, only binding sites for TFs with a median expression of $\log(\text{TPM}+1) \geq 1$ in GTEx right atrial appendage tissue and a minimal overlap of 25 bp with open chromatin regions (see Table 3.3) were considered. Further fine mapping for functional NKX2-5 BS was performed for the transcription factor activity analyses. Promoter were annotated as 2 000 bp upstream and 200 bp downstream of Gencode v31lift37 [Frankish et al., 2019] basic and long non-coding RNA transcript start positions and extended for regions linked to those by promoter-capture HiC data[16] in human iPSC-derived cardiomyocytes [Montefiori et al., 2018] (E-MTAB-6014, capt-CM-replicated-interactions-1kb.bedpe). NKX2-5 BS were considered functional, if they had a 50 bp overlap with open chromatin (see Table 3.3) and promoter regions.

miRNA BS miRNA BS were obtained from TargetScan 7.2[17] [Agarwal et al., 2015] choosing the option of default predictions for conserved target sites of conserved miRNA families.

RNA-binding protein binding sites (RBP BS) Toray Akcan provided RBP BS derived from HepG2 and K562 cell line eCLIP data of the ENCODE[18] Project Consortium [Dunham et al., 2012, Davis et al., 2018] (`https://www.encodeproject.org`). Initial peak calling was based on the ENCODE uniform processing pipeline. Bed-files were then filtered for a positive enrichment as well as a Fisher P value $> -\log10(0.05)$. Finally, overlapping peaks were merged.

[14] `http://pedagogix-tagc.univ-mrs.fr/remap/download/remap2018/hg19/MACS/remap2018_nr_macs2_hg19_v1_2.bed.gz`

[15] `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE133833`

[16] `https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-6014/`

[17] `http://www.targetscan.org/vert_72/vert_72_data_download/Predicted_Target_Locations.default_predictions.hg19.bed.zip`

[18] `https://www.encodeproject.org/report/?type=Experiment&status=released&replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&biosample_ontology.classification=cell+line&assay_title=eCLIP&limit=all`

3.4.3 Evaluation of *cis* QTLs in atrial tissue

3.4.3.1 *Cis* QTL covariates including PEER factors

Expression analysis in human tissues remains challenging due to confounding factors, such as biological differences in cell-type compositions or technical variation of the measurement technique. As many of those confounders are actually not known, methods to estimate these unknown factors have been developed. One of these methods, PEER [Stegle et al., 2012], has been introduced at the beginning of this thesis in section 3.1.3. It has been shown that PEER factors can successfully account for known and unknown confounders in gene expression data when used as covariates for *cis* QTL analyses [Lappalainen et al., 2013, Gamazon et al., 2018]. Moreover, they can either be used in combination or even instead of known covariates [Gamazon et al., 2018] and the number of PEER factors used in the analyses is a hyperparameter which can be used to boost QTL discoveries [Lappalainen et al., 2013]. In this thesis, the package PEER[19] in R 3.5.1 was used for the estimation of PEER factors.
For QTL analyses, the following models were evaluated (for the results see Figure 4.2 and Table 3.4):

1. One to 30 PEER factors without additional covariates.
2. One to 30 PEER factors and the fibroblast-score.
3. One to 30 PEER factors and the first three genotype principal components.
4. One to 30 PEER factors, covariates age, sex, BMI, disease status (overall AF) and the fibroblast-score.
5. One to 30 PEER factors, covariates age, sex, BMI, disease status, fibroblast-score and the first three genotype principal components.

3.4.3.2 *Cis* QTL computation

For QTL computations, the *cis* range was set 10^6 bp. QTLs were evaluated using a linear model with additive (0 for homozygous major allele, 1 for heterozygous, 2 for homozygous minor allele) coding of the genotypes (see models 3.70 and 3.71) and the R package MatrixEQTL[20] [Shabalin, 2012]. Dominant-recessive and ANOVA models were avoided to minimize the discovery of spurious associations. Additionally, all analyses were performed for per-sample quantile-normalized as well as additional per-gene quantile-normalized expression values in order to limit the influence of outliers. The covariate sets and number of PEER factors for the final QTL computations as shown in Table 3.4 were chosen according to the highest number of discoveries of genes with at

[19] `https://github.com/PMBio/peer/`
[20] `https://github.com/andreyshabalin/MatrixEQTL`

least one significant QTL based on a Benjamini-Hochberg FDR < 0.05:

$$\text{expr} \sim \beta_0 + \beta_1 \cdot \text{SNP} + \sum_i \beta_{1+i} \text{PEER}_i + \varepsilon \quad \text{and} \tag{3.70}$$

$$\text{ratio} \sim \beta_0 + \beta_1 \cdot \text{SNP} + \beta_2 \cdot \text{fibroblast-score} + \sum_i \beta_{2+i} \text{PEER}_i + \varepsilon. \tag{3.71}$$

Table 3.4: Number of PEER factors and covariates for QTL computations.
Combination of normalization and covariate options used for the final QTL computations including sample size and degrees of freedom.
QTL, quantitative trait locus; eQTL, expression quantitative trait loci; pQTL, protein quantitative trait loci; res eQTL, expression residual quantitative trait loci; res pQTL, protein residual quantitative trait loci; ratioQTL, ratio quantitative trait loci;

QTL type	Normalization	Covariate set	Number of PEER factors	N	df
eQTL	Per sample and gene quantile-normalization	PEER factors only	12	75	61
pQTL	Per sample and gene quantile-normalization	PEER factors only	10	75	63
res eQTL	Per sample and gene quantile-normalization	PEER factors only	8	66	56
res pQTL	Per sample and gene quantile-normalization	PEER factors only	12	66	52
ratioQTL	Per sample and gene quantile-normalization	PEER factors and fibroblast-score	9	66	54

Due to LD, the SNP correlation structure leads to blocks of linked SNPs all having a significant QTL. In order to assess the number of independent QTL loci per gene and omic type, LD clumping was performed with the software Plink[21] (parameters $R^2 = 0.5$, clump size 250 kb, for FDR based summary statistics: primary FDR cutoff 0.05 and secondary FDR cutoff 0.8, for P value based summary statistics: primary P value cutoff 10^{-5} and secondary P value cutoff 0.05). For each LD clump, we will considered the lead SNP of the clump as a representative of the associated loci.

3.4.3.3 Definition of functional *cis* QTL categories

We used a residual regression approach to assess shared and independent effects of *cis* genetic variants on different omics levels. For residual mRNA and residual protein values (see Equation 3.68-3.69), variation shared between the omics has been removed. Therefore, if a QTL is shared between mRNA and protein, we should observe an eQTL and pQTL, but both the residual eQTL and the residual pQTL should disappear (see also overview Table 3.5). Similarly, for a truly independent eQTL we would expect to see an eQTL as well as residual eQTL in the absence of a pQTL and residual pQTL and respectively, we can define an independent pQTL as a pQTL and residual pQTL without an eQTL or residual eQTL.
The grouping of QTLs by those three distinct categories relied on statistical cutoffs

[21] `https://www.cog-genomics.org/plink/1.9`

(FDR </> 0.05) and was used to identify extreme cases. Mixed forms of the different mechanisms exist and might be missed by this approach (as also shown in Table 3.5). On the other hand, observing an eQTL and pQTL for the same SNP with the same direction of effect does not necessarily implicate a truly shared QTL, as shown in the example in Figure 3.5.

Table 3.5: Definition of functional *cis* QTL categories.
Integration of multi-omics QTL summary statistics to assess functional *cis* QTL categories.
QTL, quantitative trait locus; eQTL, expression quantitative trait loci; pQTL, protein quantitative trait loci; res eQTL, expression residual quantitative trait loci; res pQTL, protein residual quantitative trait loci;

Functional QTL category	eQTL	pQTL	res eQTL	res pQTL
Shared *cis* eQTL/pQTL	1 (FDR < 0.05)	1 (FDR < 0.05)	0 (FDR > 0.05)	0 (FDR > 0.05)
Independent *cis* eQTL	1 (FDR < 0.05)	0 (FDR > 0.05)	1 (FDR < 0.05)	0 (FDR > 0.05)
Independent *cis* pQTL	0 (FDR > 0.05)	1 (FDR < 0.05)	0 (FDR > 0.05)	1 (FDR < 0.05)
Weak shared eQTL/pQTL	1 1	1 1	1 0	0 1
Weak independent eQTL	1 (FDR < 0.05)	0 (FDR > 0.05)	0 (FDR > 0.05)	0 (FDR > 0.05)
Weak independent pQTL	0 (FDR > 0.05)	1 (FDR < 0.05)	0 (FDR > 0.05)	0 (FDR > 0.05)

3.4.3.4 Colocalization analysis

Approximate Bayes Factor analyses can be applied to estimate causal variants underlying eQTL and pQTL summary statistics per gene. In this thesis, the R package coloc [Giambartolomei et al., 2014] with the *coloc.abf()* function was used. The results supplied estimates of the posterior probabilities for the following five hypotheses:

- **H0 (no causal variant):**
  No significant association of the considered SNPs with none of the traits.
- **H1 (causal variant for trait 1 only):**
  One variant which is causal for the signal for the first trait (i.e. expression/eQTL), but not the second (i.e. proteomics/pQTL).
- **H2 (causal variant for trait 2 only):**
  Corresponds to a causal variant only for trait 2 (pQTL) and no association with expression.
- **H3 (two distinct causal variants):**
  Associations with the trait are caused by two different variants.
- **H4 (one common causal variant):**
  Associations with both traits are caused by the same variant.

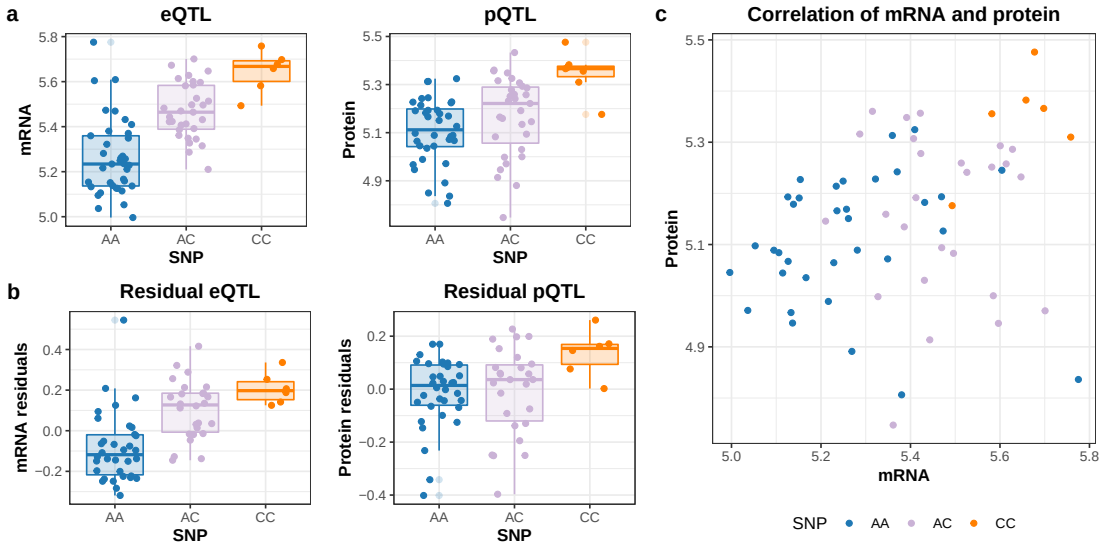


Figure 3.5: Assessing shared and independent effects using residual analysis.
a: Strong QTL on mRNA and protein level.
b: Strong QTL for both mRNA residuals and protein residuals, even though shared effects were removed by the residual computation (based on Equations 3.68 and 3.69).
c: Even though the effects of the same SNP on mRNA and protein are independent, transcript and protein are still highly correlated (two-sided Spearman's rank correlation of $\rho = 0.36$, $P = 0.0031$).
Schematic representation is based on real data of the gene COQ5 and SNP rs12309824.
eQTL, expression quantitative trait locus; pQTL, protein quantitative trait locus; SNP, single-nucleotide polymorphism;

When evaluating posterior probabilities, we considered the sum for H1 and H3 for independent eQTLs as well as the sum for H2 and H3 for independent pQTLs. Posterior probabilities ≥ 0.5 were used as threshold to classify colocalization.

3.4.3.5 Enrichment of functional elements

In order to evaluate whether specific functional elements or a specific chromatin context were enriched in QTLs, we compared annotations of SNP-gene pairs from QTLs to non-QTLs similar to analyses performed by Battle et al. [2015]. The null hypothesis was estimated from non-QTL SNP-gene pairs in a similar context with respect to MAF of the SNP and distance to the TSS. This was carried out using the following steps:

1. The target set was defined as either
   - the top SNP (FDR < 0.05) per gene or
   - the top 5 SNPs (FDR < 0.05) per gene.
2. For each of the SNP-gene pairs defined in step 1, we sampled 100 SNP-gene pairs with
   - a FDR > 0.05,
   - a similar MAF (difference ≤ 0.05) and

- a similar distance of the SNP to the nearest TSS of the corresponding gene (difference $\leq$ 1 000 bp).

For the ranking of SNPs per gene the eQTL FDR was used for eQTLs, the pQTL FDR for pQTLs, the pQTL FDR for shared *cis* eQTLs/pQTLs, the residual eQTL FDR for independent *cis* eQTLs and the residual pQTL FDR for independent *cis* pQTLs. After selecting the target and sampling the background set, annotations were compared for each of the 15 chromatin states as well as genomic regions like exons, splice sites (10 bp around exon start and end positions), 5' and 3' UTRs, TF, RNA-binding protein or miRNA binding sites and whether the SNP mutation was known to have a missense mutation or nonsense-mediated decay as consequence. Fisher's exact test was used to compute odds-ratios and enrichments on the QTL-by-annotation contingency tables.

3.4.3.6 GWAS overlap and enrichments

The web service SNiPA[22] [Arnold et al., 2015] was used to find proxies (EUR population, $R^2 > 0.8$) for all SNPs from the GWAS catalog which were annotated to traits assigned to the EFO terms cardiovascular measurements EFO_0004298, cardiovascular disease EFO_0000319 and rheumatoid arthritis (RA) EFO_0000685. For every original GWAS hit and every gene, we then selected the strongest proxy-gene pair as QTL to annotate the GWAS hit.
For further quantifying a general enrichment of GWAS hits in QTLs, we evaluated for every SNP tested for QTL association if it possessed a significant QTL (FDR < 0.05) or if it was a proxy ($R^2 > 0.8$) for a GWAS hit belonging to the trait categories cardiovascular traits, arrhythmias, AF or RA. Cross tables for the corresponding GWAS annotations and significant eQTLs as well as pQTLs were constructed and tested for enrichment using Fisher's exact test.

[22] `https://snipa.helmholtz-muenchen.de/`

3.4.4 Quantitative trait scores and gene regulation in *trans*

3.4.4.1 eQTS and pQTS rankings

Expression quantitative trait score (eQTS) and protein quantitative trait score (pQTS) rankings were derived through the correlation of transcriptomics and proteomics data with the polygenic risk score (PRS) percentiles by the following linear models:

$$\begin{aligned}\text{mRNA} \sim \beta_0 + \beta_1 \cdot \text{PRS} + \beta_2 \cdot \text{age} + \beta_3 \cdot \text{sex} + \beta_4 \cdot \text{BMI} + \beta_5 \cdot \text{sysBP} + \beta_6 \cdot \text{CRP} + \\ + \beta_7 \cdot \text{NT-proBNP} + \beta_8 \cdot \text{RIN} + \sum_i \beta_{8+i}\text{SNP}_i + \varepsilon \end{aligned} \tag{3.72}$$

$$\begin{aligned}\text{protein} \sim \beta_0 + \beta_1 \cdot \text{PRS} + \beta_2 \cdot \text{age} + \beta_3 \cdot \text{sex} + \beta_4 \cdot \text{BMI} + \beta_5 \cdot \text{sysBP} + \beta_6 \cdot \text{CRP} + \\ + \beta_7 \cdot \text{NT-proBNP} + \beta_8 \cdot \text{protein conc.} + \sum_i \beta_{8+i}\text{SNP}_i + \varepsilon \end{aligned} \tag{3.73}$$

For each gene, $\sum_i \beta_i \text{SNP}_i$ describes the independent *cis* QTL loci represented by the lead SNP identified in LD clumping. They were included as covariates additionally to age, sex, BMI, systolic blood pressure (sysBP), C-reactive protein (CRP) and N-terminal prohormone of brain natriuretic peptide (NT-proBNP) to account for the *cis*-genetic effects possibly contained in the PRS to ensure that eQTS and pQTS rankings are focused on *trans*-genetic effects. Summary statistics, especially T values for the PRS (β_1) in Equations 3.72 and 3.73, were used to define the eQTS and pQTS.

3.4.4.2 Pathway enrichment analysis

Gene set enrichment analysis (GSEA) [Subramanian et al., 2005] was used to identify molecular mechanisms which were driven by *trans*-genetic effects by selecting gene sets enriched for particularly high and low QTS values. To be able to evaluate whether AF-specific mechanisms can be detected, Gene Ontology biological processes represented by the MSigDB v6.1 gene set collection[23] (c5.bp.v6.1.symbols.gmt.txt) [Subramanian et al., 2005, Ashburner et al., 2000, Carbon et al., 2019] which rely on molecular interactions, have been used. In order to prevent circular argumentation and bias, disease-specific annotations like the KEGG human disease pathways were avoided.
GSEA computations were performed with the Bioconductor R package fgsea [Korotkevich et al., 2019] running 100 000 permutations on the eQTS and pQTS T value rankings. Gene sets with at least 15 and a maximum of 500 transcripts as well as at least 15 and a maximum of 500 proteins were considered.
Leading edge genes supplied by the fgsea analysis were considered drivers of the gene set enrichment and used for further analyses.

[23] `https://www.gsea-msigdb.org/gsea/msigdb/genesets.jsp?collection=BP`

3.4.4.3 SNP and gene candidate selection for *trans* analyses

The multiple testing burden of the *trans* analyses were reduced by largely restricting the SNP and gene candidates which were to be tested. To derive the final 108 SNP candidates, all AF GWAS hits annotated with atrial fibrillation in the GWAS catalog [Buniello et al., 2019] were further filtered for a MAF ≥ 0.1 and pruned for selecting only one representative per LD block (highest GWAS P value [Roselli et al., 2018], $R^2 > 0.5$ by SNiPA [Arnold et al., 2015]).
In order to select the gene candidates, we carried out a power analysis for *trans* eQTL detection using the 108 SNPs, a fixed sample size of N = 74 and the strongest *trans* eQTL with 21.8 % of variance explained from the eQTLGen Consortium [van der Wijst et al., 2020]. Power calculations for the F test revealed that, with a Bonferroni-adjusted significance level of 5 % and a power of at least 50 %, 23 genes can be tested in this setting (see Figure 5.7). Therefore, the next step was to select the 23 most promising genes.
Specifically, we used two criteria: First, potential candidate genes needed to be leading edge genes of a significantly enriched gene set (FDR < 0.05) and second, due to the hierarchical structure of the GO biological processes, genes driving the enrichment of multiple gene sets should be prioritized as they are contained in parent terms as well as smaller, more specialized child terms. Accordingly, 1 261 transcripts for 81 significantly enriched gene sets were reduced to 23 genes which appeared in the leading edge of enriched pathways 14 or more times. The same procedure was applied for proteomics, but since there was only one significantly enriched biological process, no further prioritization could be applied and all 152 proteins were kept for *trans* pQTL testing.

3.4.4.4 *Trans* QTL computations

Trans QTLs were calculated with the R package MatrixEQTL [Shabalin, 2012] for 108 SNPs (AF GWAS hits), 23 transcripts and 152 proteins. The additive SNP effect and covariates age, sex, BMI, systolic blood pressure, C-reactive protein, N-terminal prohormone of brain natriuretic peptide, the fibro-score and RIN-score/protein concentration for transcripts/proteins were included in the linear models 3.74 and 3.75. Since PEER factors can in general account for broad changes in gene expression as potentially caused by important TFs, they might overcorrect and remove desired variation. Therefore, it was suggested to use PEER factors only for *cis* but not *trans* analyses [Lappalainen et al., 2013].

$$\begin{aligned}\text{mRNA} \sim \beta_0 + \beta_1 \cdot \text{SNP} + \beta_2 \cdot \text{age} + \beta_3 \cdot \text{sex} + \beta_4 \cdot \text{BMI} + \beta_5 \cdot \text{sysBP} + \beta_6 \cdot \text{CRP} + \\ + \beta_7 \cdot \text{NT-proBNP} + \beta_8 \cdot \text{fibroblast-score} + \beta_9 \cdot \text{RIN} + \varepsilon \quad (3.74)\\ \text{protein} \sim \beta_0 + \beta_1 \cdot \text{SNP} + \beta_2 \cdot \text{age} + \beta_3 \cdot \text{sex} + \beta_4 \cdot \text{BMI} + \beta_5 \cdot \text{sysBP} + \beta_6 \cdot \text{CRP} + \\ + \beta_7 \cdot \text{NT-proBNP} + \beta_8 \cdot \text{fibroblast-score} + \beta_9 \cdot \text{protein conc.} + \varepsilon \quad (3.75)\end{aligned}$$

3.4.4.5 NKX2-5 *trans* pQTL evaluation

The logarithm of NKX2-5 protein intensities normalized to alpha-actinin were used to evaluate the rs9481842-NKX2-5 *trans* pQTL. The same additive linear model and the same covariates as in the original *trans* QTL computations were used with exception of sex, which was dropped because only one female sample was included in this analysis (Equation 3.76).

$$\begin{aligned}\text{NKX2-5 protein} \sim \beta_0 + \beta_1 \cdot \text{rs9481842} + \beta_2 \cdot \text{age} + \beta_3 \cdot \text{BMI} + \beta_4 \cdot \text{sysBP} + \beta_5 \cdot \text{CRP} + \\ + \beta_6 \cdot \text{NT-proBNP} + \beta_7 \cdot \text{fibroblast-score} + \varepsilon \end{aligned} \tag{3.76}$$

3.4.4.6 Definition of NKX2-5 targets

Downstream consequences of the SNP rs9481842 via the TF NKX2-5 were evaluated using genomics, transcriptomics and proteomics data as well as publicly available annotations such ChIP-seq binding sites, promoter-capture HiC and chromatin marks. The data integration and selection procedure is also visualized in Figure 5.12 in Chapter 5.

To define the TF targets, the following linear regression models were evaluated:

- Association of GWAS SNP with target transcript (*trans* eQTL):

$$\text{target transcript} \sim \beta_0 + \beta_1 \cdot \text{rs9481842} + \beta_2 \cdot \text{fibroblast-score} + \beta_3 \cdot \text{RIN} + \varepsilon \tag{3.77}$$

- Independent effects of the SNP on target transcript which are not mediated by the TF transcript:

$$\begin{aligned}\text{target transcript} \sim \beta_0 + \beta_1 \cdot \text{rs9481842} + \beta_2 \cdot \text{TF transcript} + \\ + \beta_3 \cdot \text{fibroblast-score} + \beta_4 \cdot \text{RIN} + \varepsilon \end{aligned} \tag{3.78}$$

- Association of target protein with TF transcript:

$$\begin{aligned}\text{target protein} \sim \beta_0 + \beta_1 \cdot \text{TF transcript} + \beta_2 \cdot \text{fibroblast-score} + \\ + \beta_3 \cdot \text{protein conc.} + \varepsilon \end{aligned} \tag{3.79}$$

- Association of GWAS SNP with target protein (*trans* pQTL) for Table 5.10:

$$\text{target protein} \sim \beta_0 + \beta_1 \cdot \text{SNP} + \beta_2 \cdot \text{fibroblast-score} + \beta_3 \cdot \text{protein conc.} + \varepsilon \tag{3.80}$$

Only genes with both transcriptomics and proteomics data were considered and the following four properties were evaluated to define the NKX2-5 TF targets:

a) The target gene has a functional NKX2-5 binding site with an overlap of a ChIP-seq BS, an open chromatin state and a promoter or promoter interacting region (HiC) as described in section 3.4.2.5.

The effect of the GWAS SNP is mediated by the NKX2-5 TF:

b) The target gene has an *trans* eQTL for the GWAS SNP rs9481842, therefore $\beta_1 < 0$ and $P(\beta_1) < 0.05$ for model 3.77.
c) The association disappears when considering the TF transcript expression, i.e. the eQTL is mediated by NKX2-5 and $P(\beta_1) > 0.2$ for model 3.78.

Finally, the TF should highly influence protein intensities of the target gene:

d) There is a strong positive correlation of the target protein with the NKX2-5 transcript.
We evaluated model 3.79 for all remaining candidates and performed FDR correction on the corresponding P values. All genes with $\text{FDR}(P(\beta_1)) < 0.05$ and $T(\beta_1) > 0$ (for model 3.79) were defined as functional NKX2-5 targets.

3.4.4.7 Differential proteome analysis for AF

Prevalent AF was the phenotype fitting best to the AF definition used in the PRS. Therefore, protein intensities from these cases were compared to controls in the AFHRI-B cohort while excluding post-operative AF cases. Stringent adjustment for confounders and cardiovascular risk factors was carried out by including covariates (= covs) age, sex, BMI, diabetes, sysBP, hypertension medication, myocardial infarction, smoking status, fibroblast-score and protein concentration in model 3.82:

$$\text{protein} \sim \beta_0 + \beta_1 \cdot \text{AF} + \sum_{\text{cov} \in \text{covs}} \beta_{\text{cov}} \cdot \text{cov} + \varepsilon \tag{3.81}$$

3.4.4.8 Replication in independent datasets

Replication in the GSE128188 dataset The GSE128188 dataset[24] [Thomas et al., 2019] contained RNA-seq transcriptomics from atrial appendage tissue of males undergoing coronary artery bypass grafting and/or atrial/mitral valve repair or replacement surgery for ten AF cases as well as ten controls. Processing was performed using the R bioconductor edgeR [Robinson et al., 2010] functions *calcNormFactors()* and *rpkm()* in order to obtain log-transformed TMM-based RPKMs. AF differential expression summary statistics from the original publication were provided by the authors Thomas et al. [2019].
In order to replicate the down-regulation of NKX2-5 targets in this independent dataset, GSEA was performed on the log fold changes while ranking significant genes before non-significant ones. The ten left and right atrial appendage samples as well as the mean of cases and the mean of controls were scaled and centered per gene to produce the comparable values which were visualized.

[24] `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE128188`

Replication in the PXD006675 dataset The PXD006675 repository[25] [Doll et al., 2017] provided deep mass spectrometry proteomics of six left atrial tissue samples (three AF cases as well as three controls). AF differential protein expression summary statistics from the original publication were provided by the authors Doll et al. [2017].
In order to replicate the down-regulation of NKX2-5 targets in this independent dataset, GSEA was performed on the log fold changes while ranking significant genes before non-significant ones. Proteomics data of triplicates for three AF cases and three controls were median normalized per measurement and log-transformed. All samples as well as the mean of the AF cases and the mean of the controls were scaled and centered per gene to produce the comparable values which were visualized.

[25] https://www.ebi.ac.uk/pride/archive/projects/PXD006675

3.5 Methods for multi-omics enrichment analyses

3.5.1 Differential expression analysis for atrial fibrillation

Differential analyses were carried out for transcriptomics and proteomics measurements. In general, we considered three different possible linear regression models as defined in Equation 3.82 adjusting for different sets of risk factors:

- Model A
  covs: age, sex and RIN-score/protein concentration;
- Model B
  covs: age, sex, BMI, diabetes, sysBP, hypertension medication, myocardial infarction, smoking status, RIN-score/protein concentration and fibroblast-score;
- Model C
  covs: age, sex, BMI, diabetes, sysBP, hypertension medication, myocardial infarction, smoking status, CRP, NTproBNT, RIN-score/protein concentration and fibroblast-score;

$$\text{expression} \sim \beta_0 + \beta_1 \cdot \text{AF} + \sum_{\text{cov} \in \text{covs}} \beta_{\text{cov}} \cdot \text{cov} + \varepsilon \tag{3.82}$$

Model A does not take into account important confounders. Model C is very restrictive as discussion is still ongoing if either NTproBNT is a biomarker for independent heart failure or if its increase might also be directly caused by AF. Therefore, model B was used for the differential expression analyses for mRNA and protein abundance.

3.5.2 Extensions to the MONA console app

The original implementation of the MONA framework[26] [Sass et al., 2013] is a Windows executable using the Microsoft .NET Framework 4.0 for the model inference. Instead of the graphical user interface, Andreas Kopf provided an implementation of MONA as a command line tool which is based on C# code compiled into a Windows executable that was run using mono[27] including the .NET framework on unix-based systems.

The following extensions of the MONA console app were realized by either adjusting the input files or adapting the corresponding C# code used to generate the model design.

[26] `https://www.helmholtz-munich.de/icb/research/labs/computational-cell-maps/projects/mona/index.html`
[27] `https://www.mono-project.com/`

3.5.2.1 Adding direction of effect

No modification of the original code is needed to include direction of effect. Only the corresponding assignment matrix, terms, observations and the potential missingness vector need to be modified.

For the corresponding observations for the molecular modality X with measurements $\{X_i\}$ we define

$$O_i^{X,+} = \begin{cases} 1 & \text{if up-regulation observed for } X_i \\ 0 & \text{if down-regulation observed for } X_i \\ 0 & \text{if no-regulation observed for } X_i \end{cases} \quad \text{and} \tag{3.83}$$

$$O_i^{X,-} = \begin{cases} 0 & \text{if up-regulation observed for } X_i \\ 1 & \text{if down-regulation observed for } X_i. \\ 0 & \text{if no-regulation observed for } X_i \end{cases} \tag{3.84}$$

The corresponding assignment matrix also needs to take into account up- and down-regulation of the terms and hidden states. Therefore, the original assignment matrix $\mathbf{A}$ with rows $(H_1, H_2, \ldots, H_p)$ and columns $(T_1, T_2, \ldots, T_k)$ needs to be extended to

$$\mathbf{A}^{+/-} = \begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{A} \end{pmatrix} \tag{3.85}$$

where $\mathbf{A}^{+/-}$ has rows $(H_1^+, H_2^+, \ldots, H_p^+, \quad H_1^-, H_2^-, \ldots, H_p^-)$ and columns $(T_1^+, T_2^+, \ldots, T_p^+, \quad T_1^-, T_2^-, \ldots, T_p^-)$.

Irrespective of whether the observations $O_i^{X,+}/O_i^{X,-}$ are up- or down-regulated, they still share the same error rates α^X and β^X as visualized in Figure 6.2 for the case of the two-omics cooperative model.

Also note, that for the MONA console app not the full assignment matrix is used, but each row is representing a hidden node and the corresponding terms are reported as numbers, i.e. $i-1$ for term T_i and multiple terms separated by commas.

3.5.2.2 Adding a third molecular modality

In order to add a third omic to the cooperative model, the model setup needed to be modified. To make the integration of e.g. transcriptomics and metabolomics possible, in addition to a third observation layer, we also had to adjust the code to take into account missingness for all three omics.

A compiled version of the MONA console app - containing the three-omics cooperative model as algorithm number 8 - is available in the binaries folder of the **enrich** Docker container and in the src/library/binaries folder of the EnrichmentNodes repository `https://github.com/InesAssum/EnrichmentNodes`.
It can be called from a unix-like system including the needed requirements by the following command:

```
mono <path_to/MonaConsoleApp.exe> 8 <path_assignment_matrix>
<path_observations_omic1> <path_term_names> <path_output_file> 1
<path_observations_omic2> <path_observations_omic3>
<path_missings_omic1> <path_missings_omic2> <path_missings_omic3>
```

Additionally, the **MonaConsoleApp2.exe** executable includes as two-omics cooperative model (algorithm number 7) an implementation taking into account two inputs with missingness information for both omics.

3.5.3 A simulation study to benchmark multi-omics enrichment methods

Code used to derive the simulations and evaluate the results is available as part of the EnrichmentNodes repository `https://github.com/InesAssum/EnrichmentNodes` on GitHub at src/R/simulation_study.

All analyses were performed with R version 4.1.1.

3.5.3.1 Simulation of summary statistics

Pathway annotations The MSigDB KEGG pathways c2.cp.kegg.v7.2.symbols.gmt[28] [Subramanian et al., 2005, Ashburner et al., 2000, Carbon et al., 2019] were used as gene set annotations.

Simulation scenarios We evaluated different simulation scenarios with respect to three different properties and all their combinations (overall 18 settings).
For shared pathway activations, six gene sets were randomly sampled to be active in both omics. Only gene sets with more than ten measured genes for the second omic were included. For the independent pathway activations, three gene sets for each omic were sampled independently. Again, only gene sets with more than ten genes measured in the corresponding omic were considered.
The coverage was set to 30 %, to 100 % and to around 10 % when using the actual measured proteins of the AFHRI-B cohort. Only proteins with corresponding transcript measurements were considered.

[28] `https://www.gsea-msigdb.org/gsea/msigdb/genesets.jsp?collection=BP`

As correlation between omics, we considered $\rho \in \{0.2, 0.3, 0.8\}$, with again $\rho = 0.2$ being closest to the actual data.

Sampling scheme and thresholds For each scenario we used the following procedure to simulate two-omics correlated summary statistics:

1. **For each replicate**
   a) **Ground truth:**
   Define the true regulated genes and pathways.
      - Start by selecting gene sets which should be regulated for either one or both omics.
      - Select a sign for up- or down-regulation for each regulated gene set.
      - Assign all regulated genes by following the gene set annotations for each of the omics, taking into account coverages of each omic.
      - Assign the corresponding sign based on the sign of the gene set for each gene. If a gene is occurring in multiple regulated gene sets, then the sign is sampled randomly based on the frequency of the different signs for the corresponding pathway.

   b) **Simulate summary statistics:**
   Based on this ground truth, multi-omics summary statistics will be simulated. While including varying combinations of error rates, they will all be built on the same set of regulated pathways.
      - Simulate the background of unregulated, multi-omics Z scores taking into account the correlation parameter ρ.
      - Simulate Z scores for all genes which are regulated in either one of the omics based on Equations 6.9, 6.10 and 6.11 taking into account the predefined error rates.
        The signal is resampled if the deviation of the assigned error rate from the observed error rate is too large.
        If resampling the signal does not lead to the desired results, also the background sampling is repeated.

   c) **Save data:**
   Save simulated multi-omics summary statistics. Collapse results across all combinations of error rates into one file in order to avoid too many small files.
2. **Diagnostics:**
   Detailed diagnostics about the different simulated datasets for each replicate including preset and actual error rates are collapsed and returned together with visualizations and information about the maximum deviation. A parameter file including all parameters, the used seed and time needed for the simulation is created in the end and added to the result folder.

Above, we have given a more technical representation of the simulation scheme. For a detailed description of the rationale and Equations 6.9, 6.10 and 6.11, refer to the corresponding results section 6.3.

For our simulation study, we evaluated all scenarios with a standard deviation of $\sigma_{BG} = 2$ for the background normal distribution and adjusted the standard deviation for the signal normal distribution dependent on the pre-set false positive rate α as $\sigma_{sig}(\alpha) = 1.5 - \alpha$.

Data was simulated for most of the combinations of false positive and false negative error rates $\alpha, \beta \in \{10^{-6}, 0.01, 0.05, 0.1, 0.15, 0.2, 0.3, 0.4, 0.5, 0.6\}$ with 100 replicates for each of the combinations of error rates for each of the 18 different simulation scenarios.

When simulating the data, we set the thresholds described in Table 3.6 depending on the pre-defined error rates to satisfy accurate actual error rates in the data. If the error rates in the data deviated too much, the signal distribution was resampled 100 times. If the threshold was still not met, the background distribution was resampled up to 50 times. If after that the threshold was still not met, the closest match was chosen.

Table 3.6: Maximum deviation from the desired error rates.
Data was simulated for $\alpha, \beta \in \{10^{-6}, 0.01, 0.05, 0.1, 0.15, 0.2, 0.3, 0.4, 0.5, 0.6\}$, for other combinations of error rates thresholds should be adjusted.

Desired error rate for $\varepsilon \in \{\alpha, \beta\}$	$\varepsilon < 10^{-4}$	$10^{-4} \leq \varepsilon < 0.01$	$0.01 \leq \varepsilon < 0.1$	$\varepsilon \geq 0.1$
Maximum allowed deviation δ	$\delta < 10 \cdot \varepsilon$	$\delta < 2 \cdot \varepsilon$	$\delta < \varepsilon / 3$	$\delta < 0.01$

Data structure of the simulated data The simulated data per replicate is saved into one file consisting of a list with the simulations for each combination of error rates α and β. Data was bundled this way to avoid too many small files. All the computations are optimized to only read in all data per replicate once and then applied to all different error rates.

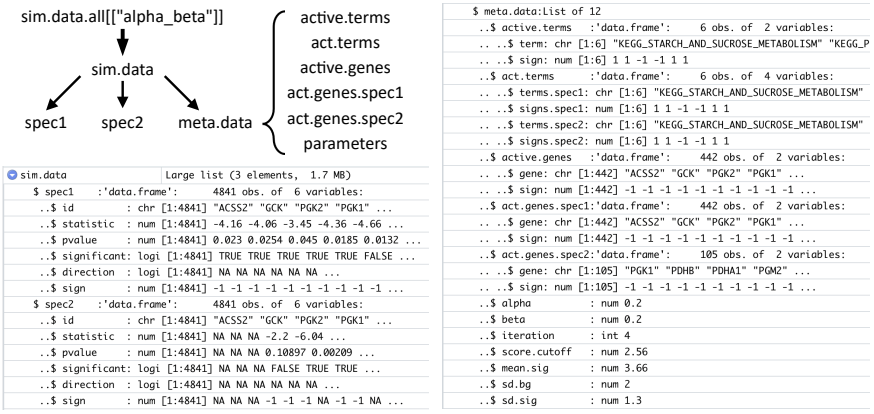


Figure 3.6: Data structures for the simulated correlated multi-omics summary statistics.
Simulation data is structured as nested lists. Each replicate consists of lists for each of the combinations of error rates, where each of those lists contains three slots (a data.frame *spec1*, a data.frame *spec2* and a list *meta.data*). *meta.data* contains five data.frames and seven additional parameters.

For each setting, simulated data is again a list of lists, containing the simulated summary statistics for omic 1, omic 2 and meta data which contain the information about the ground truth and parameters important for deriving thresholds for significance.
More specifically, the columns *id*, *statistic* and *significant* were used in the simulation evaluation.

3.5.3.2 Enrichment analysis

For all methods, all gene sets with a size between ten and 500 were evaluated for the first omic. For the second omic, all gene sets with a size between five and 500 were considered, unless the coverage was 100 %, then the minimal threshold was set at ten in line with the first omic.

Gene set enrichment analysis (GSEA) GSEA was applied using the Bioconductor R package fgsea 1.18.0[29] whose implementation is described in Korotkevich et al. [2019] using the *fgsea* function with parameter *eps = 0*.
For applying GSEA to P values, the absolute simulated Z scores were used together with the parameter *scoreType = "pos"*.

Model-based gene set analysis (MGSA) MGSA was applied using the *mgsa()* function (with default parameters) which is part of the Bioconductor R package mgsa 1.40.0[30] [Bauer et al., 2010].
MGSA including direction of effect was evaluated based on the same transformation to the set of observed genes and the pathway annotations as described in the method section 3.5.2.1.

Single MONA and the MONA cooperative model All MONA models were run using the MonaConsoleApp. The standard option 1 for the first parameter of the Beta distribution was used for priors.

3.5.4 Multi-omics pathway enrichment for atrial fibrillation in human atrial tissue

Just as in the simulation study, MSigDB KEGG pathways c2.cp.kegg.v7.2.symbols.gmt[31] [Subramanian et al., 2005, Ashburner et al., 2000, Carbon et al., 2019] were used as gene set annotations.

[29] `http://bioconductor.org/packages/fgsea/`
[30] `http://bioconductor.org/packages/mgsa`
[31] `https://www.gsea-msigdb.org/gsea/msigdb/genesets.jsp?collection=BP`

All gene sets with at least one available gene were used. To account for interference due to very small gene sets, the exact amount of genes measured for each gene set are visualized with the posterior results in Figure 6.13.

Transcriptomics and proteomics differential expression results according to model B (covariates: age, sex, BMI, diabetes, sysBP, hypertension medication, myocardial infarction, smoking status, RIN-score/protein concentration and fibroblast-score) with a P value significance threshold of 0.25 were used.

3.5.5 EnrichmentNodes - a multi-omics enrichment extension for KNIME

Source code to create our EnrichmentNodes plugin is available on GitHub `https://github.com/InesAssum/EnrichmentNodes`.
The knime folder contains all the specification needed to build the nodes using the Generic KNIME nodes node generator. To run nodes from the EnrichmentNodes plugin, the **enrich** Docker container (with R 4.1.1) is needed which is available on Docker Hub `https://hub.docker.com/repository/docker/inesassum/enrich` and can be easily pulled via command line: docker pull inesassum/enrich:latest
Source files to build the Docker container and all example data for the EnrichmentNodes plugin can be found in the *src* folder.

3.5.5.1 Plugin development using KNIME for developers and GKN

In this section, we describe the development of custom nodes which can be imported as plugins of the Konstanz Information Miner (KNIME) [Berthold et al., 2009] using Generic KNIME nodes (GKN)[32].

Used software versions and dependencies EnrichmentNodes were developed with the KNIME SDK 4.3.4 and Java 8 on macOS Catalina 10.15.7. Updated installation instructions, also for future versions, can be found on GitHub `https://github.com/knime/knime-sdk-setup`.

The EnrichmentNodes plugin was build using the following software versions and dependencies:

- Git[33]
- OpenJDK 8[34] JavaSE-1.8 (AdoptOpenJDK (OpenJ9) 8 [1.8.0_275])
- Eclipse IDE for RCP and RAP Developers[35], Version: 2020-03 (4.15.0), Build id: 20200313-1211

[32] `https://github.com/genericworkflownodes/GenericKnimeNodes`
[33] `https://git-scm.com/downloads`
[34] `https://adoptopenjdk.net/`
[35] `https://www.eclipse.org/downloads/packages/release/2021-03/r/eclipse-ide-rcp-and-rap-developers`

- Docker Desktop 3.3.3 (64133)[36]
- Apache Ant(TM) version 1.10.9

Node development for GKN nodes Generic KNIME nodes are an easy way to turn command line tools into KNIME nodes [Fillbrunn et al., 2017]. All executables can either be supplied in a specific folder or a Docker container can be used to run the commands.

In our case, we implemented the second version. All information needed for the plugin generation have to be supplied in a specific form of the following structure:

- `plugin.properties`
  This is a text file containing information about the plugin package, name and version. For each node, it also specifies the corresponding executable or the Docker container to use.
- `descriptors`
  This folder contains the main information about the graphical interface to set input, output and parameters for a tool. Each node has its own .ctd file mapping the options selected in the graphical interface to the command line tool. Additionally, the `mime.types` text file sets specific file categories for all input and output files, making file handling and display possible.
- `payload`
  Any binaries needed to run the included tools have to be saved here. In case of using a Docker container, this folder can be omitted. Otherwise, subfolders for the different operating systems need to exist.
- `icons`
  If specific icons should be displayed for the extension when starting KNIME or if the symbol representing nodes in the node repository should be adjusted, these icons have to be placed here.
- `contributing-plugins`
  Another optional folder which can be used to place dependencies such as OSGI bundles and Eclipse plugin projects.
- `DESCRIPTION`
  A text file like a Readme with a description of the nodes.
- `LICENSE`
  All licensing information for any used software needs to be placed in this text file.
- `COPYRIGHT`
  This text file contains the copyright information for the project, including e.g. authors and institutions.

Figure 6.14 in the results section 6.5 gives an example for a very simple .ctd file for a "Hello world" node. The easiest way to generate custom plugins is to start using

[36] `https://www.docker.com/products/docker-desktop`

an existing tool, e.g. our EnrichmentNodes[37] or the ImmunoNodes[38] [Schubert et al., 2017].

Protocol to run the GKN node generator Once the node template is prepared, the following steps are necessary to create the plugin.

1. **Build Generic KNIME nodes:**
   Starting with a copy of the GenericKnimeNodes repository on Git Hub, one needs to change into the corresponding folder and run `ant .`, e.g.

   ```
   cd /Users/example/work
   git clone https://github.com/genericworkflownodes/GenericKnimeNodes
   cd GenericKnimeNodes
   ant .
   ```

2. **Build the plugin template:**
   Still at the local copy of the `GenericKnimeNodes` repository, the plugin template has to be built by calling `ant -Dplugin.dir=` on the full path to the plugin template without a trailing slash, e.g.

   ```
   ant -Dplugin.dir=/Users/example/work/EnrichmentNodes/knime
   ```

   This will create a folder `generated_plugins`.
3. **Install the KNIME SDK:**
   - In Eclipse, the git project is imported via:
     `File` → `Import` → `Git` → `Projects from Git File` → `Clone URI`
     using the URI: `https://github.com/knime/knime-sdk-setup`.
   - Next, the `master` branch and the desired `releases` branch needs to be selected.
     EnrichmentNodes were created using the `releases/2020-12` branch, i.e. KNIME SDK 4.3.4.
   - Finally, the project is imported via
     `Import existing Eclipse projects`, selecting all projects and then `Finish`.
4. **Load the KNIME SDK:**
   In the Eclipse Project Explorer, the target platform is loaded by navigating to
   `knime-sdk-setup` → `or.knime.sdk.setup` → `KNIME-AP-complete.target`
   and a double click on `KNIME-AP-complete.target`.
   Resolving the target platform might take quite some time. In case of errors occurring after the target platform was loaded, one might still be able to proceed as most errors will not hinder the node creation.
5. **Import GenericKnimeNodes and the new plugin in Eclipse:**
   - In Eclipse, the GKN are imported by selecting
     `File` → `Open Projects from File System...`.
   - The build of the new plugin is imported as
     `File` →`Import` → `General` → `File System` → `From directory`

[37] `https://github.com/InesAssum/EnrichmentNodes`
[38] `https://github.com/FRED-2/ImmunoNodes/`

by navigating to the corresponding folder, e.g.
`/Users/example/work/GenericKnimeNodes/generated_plugins/`,
selecting all projects and clicking `Finish`.

6. **Run KNIME SDK and the node generator:**
   - In the Eclipse Project Explorer, the KNIME SDK is started with a right click on `KNIME-AP-complete.target` → `Run As` → `Run Configurations...` with the following settings:
   - On the left: `Eclipse Application` → `KNIME Analytics Platform`.
   - On the right: `Main` → `Run a product:` → `org.knime.product.KNIME_PRODUCT` and the matching `Java Runtime Environment` → `Execution environment:`, e.g. `JavaSE-1.8 (AdoptOpenJDK (OpenJ9) 8 [1.8.0_275])`
     and confirming via `Run`.
     The KNIME Analytics Platform should now be starting. Again, the platform might need some time to load.
7. **Export plugin as a .jar file:**
   - For easy installation without the KNIME SDK, plugins can be exported as .jar files by navigating to the plugin in the Eclipse Project Explorer.
   - A right click on the top folder of the plugin opens the dialog `Export` → `Deployable plug-ins and fragments` → `Next >` and after selecting the corresponding plugin, it can be saved by choosing the destination folder via `Directory` → `Browse` and confirming with `Finish`.
   - The plugin will be saved as a .jar file in a folder `plugins` at that destination.

4 Multi-omics analysis of atrial-specific *cis*-regulatory mechanisms

In this chapter, we use *cis* quantitative trait loci (QTL) analyses to better understand consequences of genetic variation to transcript and protein expression with a special focus on the challenges and benefits of integrating multi-omics data.

This chapter is based on and partly identical to the publication by Assum et al. [2022a] and also available as a preprint on *bioRxiv*[1,2]:
Tissue-specific multi-omics analysis of atrial fibrillation[3]
Ines Assum†, Julia Krause†, Markus O. Scheinhardt, Christian Müller, Elke Hammer, Christin S. Börschel, Uwe Völker, Lenard Conradi, Bastiaan Geelhoed, Tanja Zeller*, Renate B. Schnabel* and Matthias Heinig*, *Nature Communications* **13**, 441 (2022). Authors marked with † or * contributed equally to this work.
Code related to this project is available at `https://github.com/heiniglab/symatrial`[4] [Assum and Heinig, 2021].

Genome-wide association studies (GWAS) for atrial fibrillation (AF) have uncovered numerous disease-associated variants. The largest studies to date consisted of more than one million individuals, including more than 60 000 cases [Roselli et al., 2018, Nielsen et al., 2018], enabling investigation of the contribution of common and also rare genetic variant to AF risk. More than 100 distinct genetic loci have been identified to be associated with atrial fibrillation [Nielsen et al., 2018] and the total SNP-heritability of AF estimated by those studies ranged between 11.2 % and 22.4 %. Genome-wide significant common genetic variants explain around 5 % of the total AF variability [Choi et al., 2020]. Despite the strong heritable component, it still remains very challenging to pinpoint underlying genetic mechanisms as 95 % of GWAS hit lie in non-coding regions of the genome [Roselli et al., 2018]. Previous efforts of identifying regulatory elements which are altered by AF-associated variants uncovered possible consequences for target gene expression [van Ouwerkerk et al., 2020]. However, for most AF-associated loci, causal variants, genes and the propagation of effects from transcript to protein level remain largely elusive.
Transcriptional as well as post-transcriptional regulation play a vital role in any biological process. Thus, novel multi-omics approaches which take into account transcriptomics and proteomics are needed to advance our current understanding of

[1] `https://www.biorxiv.org/content/10.1101/2020.04.06.021527v1`
[2] `https://www.biorxiv.org/content/10.1101/2020.04.06.021527v2`
[3] `https://doi.org/10.1038/s41467-022-27953-1`
[4] `https://doi.org/10.5281/zenodo.5094276`

complex molecular networks. Furthermore, vast differences of gene regulation have been observed between different tissues [Gamazon et al., 2018]. Therefore, in order to investigate potential causes of diseases, it is mandatory to study appropriate tissues, which remains particularly challenging for cardiovascular and neurological diseases. The matched genotypes, mRNA and protein measurements available in the AFHRI-B cohort offered a unique opportunity to study consequences of genetic variation on transcript and protein abundance specific for human atrial tissue.

4.1 Downstream consequences of common genetic variants on transcript and protein abundance for nearby genes

Here, we present a multi-omics analysis which uses genomics, transcriptomics and proteomics of human atrial tissue to better understand how genetics are related to molecular changes and further on, cardiovascular phenotypes. The first aim was to systematically integrate omics data and identify genome-wide *cis*-regulatory mechanisms on transcript as well as protein level. A common approach to investigate those mechanisms is to consider tissue-specific *cis*-acting expression quantitative trait locus (eQTL), where genetic variants affect the transcription of nearby genes. *Cis* eQTL in human atrial tissue have already been performed by the GTEx consortium which will be used as a replication for our results. We will further consider the genetic *cis*-regulation of proteins, which has been first described in our publication Assum et al. [2022a] and extend on integrating transcriptomics and proteomics in order to better describe possible regulatory mechanisms.

Microarray transcriptomics and mass spectrometry-based proteomics in human atrial tissue for patients undergoing coronary artery bypass surgery were available for *cis* quantitative trait locus (QTL) analyses.

4.1.1 Correlation between mRNA and protein

As expected from other studies [Civelek and Lusis, 2014, Hause et al., 2014, Battle et al., 2015, Liu et al., 2016, Sun et al., 2018, Eraslan et al., 2019, Jiang et al., 2020], correlation of transcriptomics and proteomics is limited. In this dataset, the 79 matched samples showed a median correlation of 0.15 per gene and 0.21 per sample (see also Figure 4.1). Linear models predicting protein abundance from transcript expression can be used to describe the variance shared across omics by evaluating the R^2 per gene, which was estimated at 0.027 per gene and 0.044 per sample. General variability in gene expression also affects correlation between mRNA and protein, as shown by the higher correlation of highly variable genes (see Definition 3.20) with a median correlation per gene of 0.23.

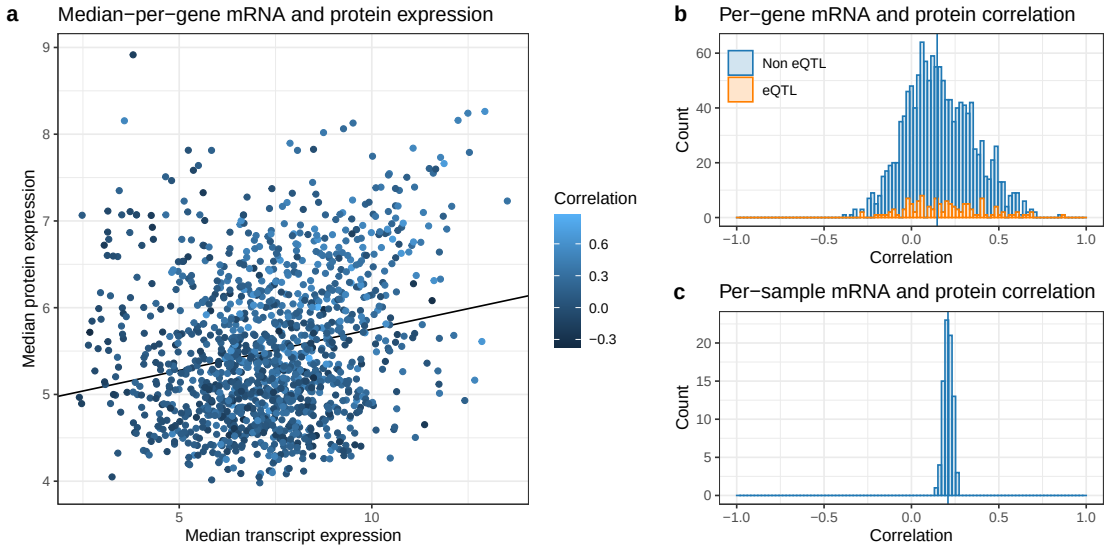


Figure 4.1: Transcript and protein correlations in the AFHRI-B cohort.
a: Median transcript and protein expression and their correlation per gene.
b: Histogram of the mRNA and protein correlation per gene. Genes with a corresponding *cis* eQTL in the AFHRI-B cohort are highlighted.
c: Histogram of the mRNA and protein correlation per sample. Correlations show much less variability.
eQTL, expression quantitative trait locus;

4.1.2 Natural *cis*-genetic variation of the human atrial transcriptome and proteome

Cis expression quantitative trait loci (*cis* eQTLs) were calculated using over 4.8 million genetic variants and expression values for 16 306 genes for 75 individuals. Similarly, *cis* protein quantitative trait loci (*cis* pQTLs) were evaluated for over 2.3 million SNPs in a *cis* range of 1 337 proteins for 75 individuals. Matched genotypes, transcriptomics and proteomics data were available for 66 individuals to integrate data for the computation of *cis* ratio quantitative trait loci (ratioQTLs), *cis* residual expression quantitative trait loci (*cis* res eQTLs) and *cis* residual protein quantitative trait loci (*cis* res pQTLs) on almost 2.5 million variants and 1 243 genes (Table 4.1).
PEER factors [Stegle et al., 2012, Lappalainen et al., 2013] can be used to correct for known as well as unknown confounders when assessing disease-independent effects of *cis*-genetic variants. The numbers of PEER factors were optimized for the maximal amount of discoveries with the results summarized in Figure 4.2 and Table 4.1. For the final set of PEER factors used, we also confirmed that they correlate with the cohort covariates and therefore were able to pick up confounders such as cardiovascular risk factors and technical covariates which is visualized in Figure 4.3.

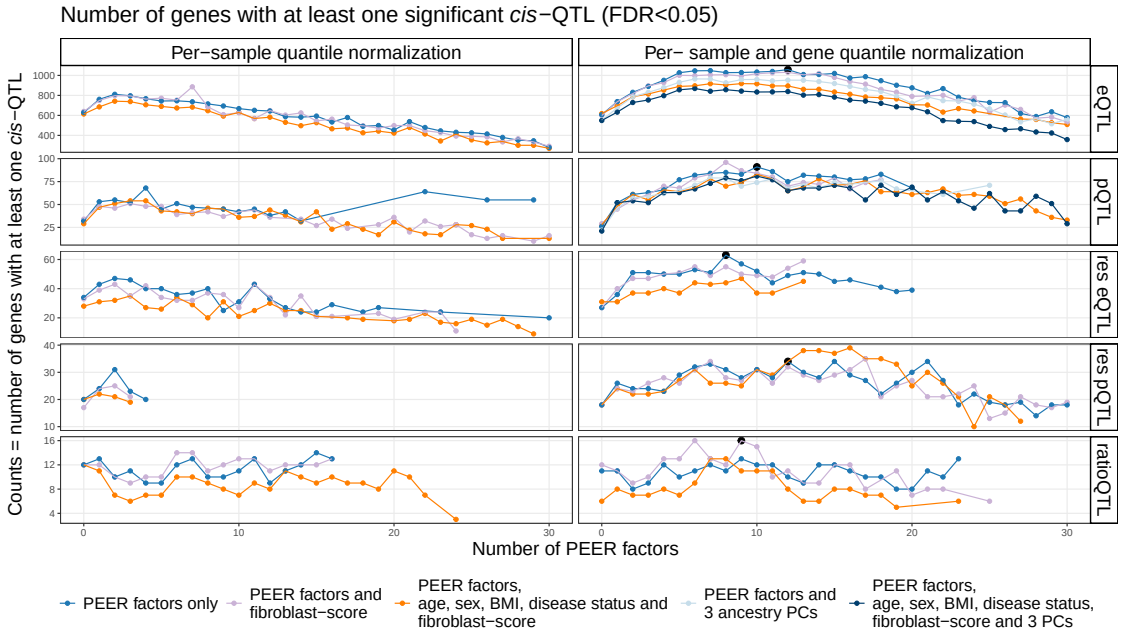


Figure 4.2: *Cis* QTL analysis results for different covariate sets and number of PEER factors.
PEER analysis was performed to account for unknown variation in the data. Displayed are the number of genes with at least one QTL variant with a FDR < 0.05 for different combinations of normalization, number of PEER factors used in the regression and additional covariates. Black dots mark the chosen number of PEER factors and covariates as the maximal number of discovered QTL genes at a FDR < 0.05. QTL, quantitative trait loci; PEER, probabilistic estimation of expression residuals; FDR, false discovery rate; eQTL, expression quantitative trait loci; pQTL, protein quantitative trait loci; res eQTL, expression residual quantitative trait loci; res pQTL, protein residual quantitative trait loci; ratioQTL, ratio quantitative trait loci;
Figure and legend taken from the supplementary material Assum et al. [2022b] [Assum et al., 2022a]`https://doi.org/10.1038/s41467-022-27953-1`.

Table 4.1: Summary of tested data and discovered *cis* quantitative trait loci.
Significant *cis* QTLs in human heart right atrial appendage tissue for mRNA and protein measurements for FDR < 0.05 (according to Benjamini-Hochberg procedure) and P value < 10^{-5}. Loci denote the number of independent loci derived by LD-clumping.
QTL, quantitative trait loci; FDR, false discovery rate; LD, linkage disequilibrium; eQTL, expression quantitative trait loci; pQTL, protein quantitative trait loci; ratioQTL, ratio quantitative trait loci; N, sample size.
Table and legend adapted from Assum et al. [2022a] `https://doi.org/10.1038/s41467-022-27953-1`.

	Tested:			FDR < 0.05:			P < 10^{-5}:			
	SNPs	**Pairs**	**Genes**	**Pairs**	**Genes**	**Loci**	**Pairs**	**Genes**	**Loci**	**N**
Results for all available transcriptomics and proteomics measurements:										
eQTL	4 861 118	56 139 851	16 306	57 403	1 058	1 657	40 267	552	870	75
pQTL	2 323 504	4 508 654	1 337	4 081	91	139	2 543	45	71	75
Results only for genes with both transcriptomics and proteomics measurements:										
eQTL	2 249 758	4 198 168	1 243	4 603	124	201	3 218	64	109	75
pQTL	2 249 758	4 198 168	1 243	3 906	87	133	2 406	42	66	75
ratioQTL	2 249 758	4 198 168	1 243	563	16	23	575	18	27	66
res eQTL	2 249 758	4 198 168	1 243	2 261	63	99	1 504	41	62	66
res pQTL	2 249 758	4 198 168	1 243	1 316	34	45	1 194	29	38	66

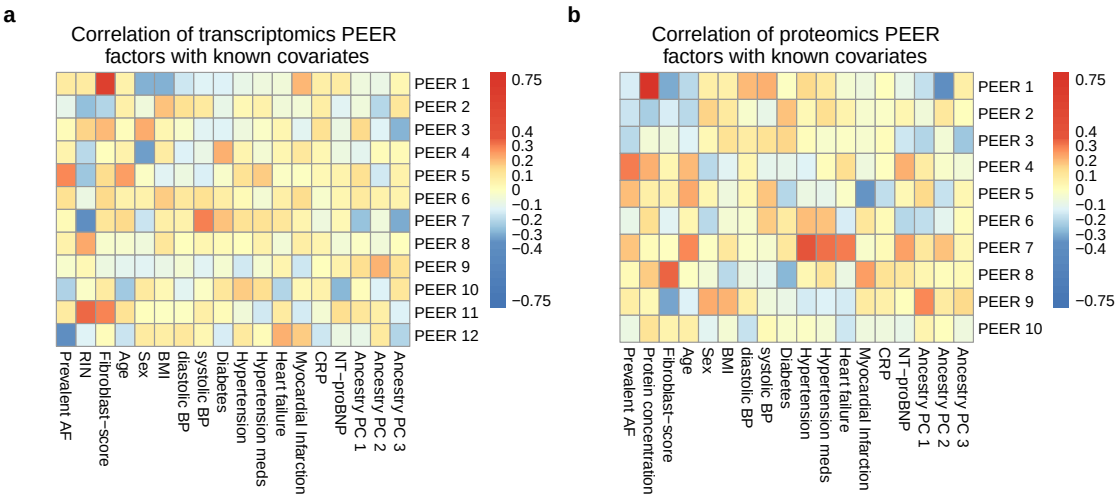


Figure 4.3: Correlation of PEER factors with common risk factors of AF and technical covariates. Pearson correlation of different PEER factors and known risk factors or technical covariates.
a: Transcriptomics analysis: Fibroblast-score and RIN-score highly correlate with PEER factors used in the final *cis* eQTL analysis.
b: Proteomics analysis: Fibroblast-score and original sample protein concentration highly correlate with PEER factors used in the final *cis* pQTL analysis.
PEER, probabilistic estimation of expression residuals; QTL, quantitative trait loci; AF, (prevalent) atrial fibrillation; RIN, RNA integrity number; BMI, body mass index; diasBP, diastolic blood pressure; sysBP, systolic blood pressure; HF, heart failure; MI, myocardial infarction; CRP, C-reactive protein; NT-proBNP, N-terminal prohormone of brain natriuretic peptide;
Figure and legend taken from the supplementary material Assum et al. [2022b] [Assum et al., 2022a] `https://doi.org/10.1038/s41467-022-27953-1`.

4.1.3 Replication

In order to validate our findings, we compared our QTLs for different sets of covariates to the GTEx atrial appendage eQTLs [Gamazon et al., 2018] and across tissues to plasma pQTLs [Sun et al., 2018]. We further used Storey's q-value method which estimates the fraction of true null hypotheses $\hat{\pi}_0$ based on the distribution of matched sets of P values in another dataset. A replication rate can then be determined by considering the corresponding fraction of true discoveries $1 - \hat{\pi}_0$.

4.1.3.1 GTEx replication

We used GTEx v7 [Gamazon et al., 2018] *cis* eQTLs for right atrial appendage tissue to compare to the *cis* QTL results from the AFHRI-B cohort. First, data was filtered for overlapping SNPs and genes, to then compare the most significant marker for each gene with a significant eQTL (significance threshold nominal $P < 1 \times 10^{-5}$) from one dataset to the other as shown in Figure 4.4. None of the eQTLs significant in both cohorts showed discordant effects. Furthermore, of all genes represented in both studies (54 %), 62 % replicated and 87 % showed concordant effects (therefore 25 % had concordant

effects without significance).
Second, we evaluated the general replication rate of our eQTLs in the GTEx data. Effect sizes for the best eQTL showed a very high correlation of 0.81 ($P = 2.3 \times 10^{71}$) and using Storey's q-value method [Storey and Tibshirani, 2003], we estimated a replication rate of 85 %. Conversely, if we checked the top GTEx SNPs per gene, that were also available for the AFHRI-B cohort, 85 % showed concordant allelic effects but only 7.8 % replicated. Even though the large differences in sample size might have lead to shifted significance levels and therefore explain the low number of replicated loci, the high agreement of effect sizes represents a high concordance of the two datasets.

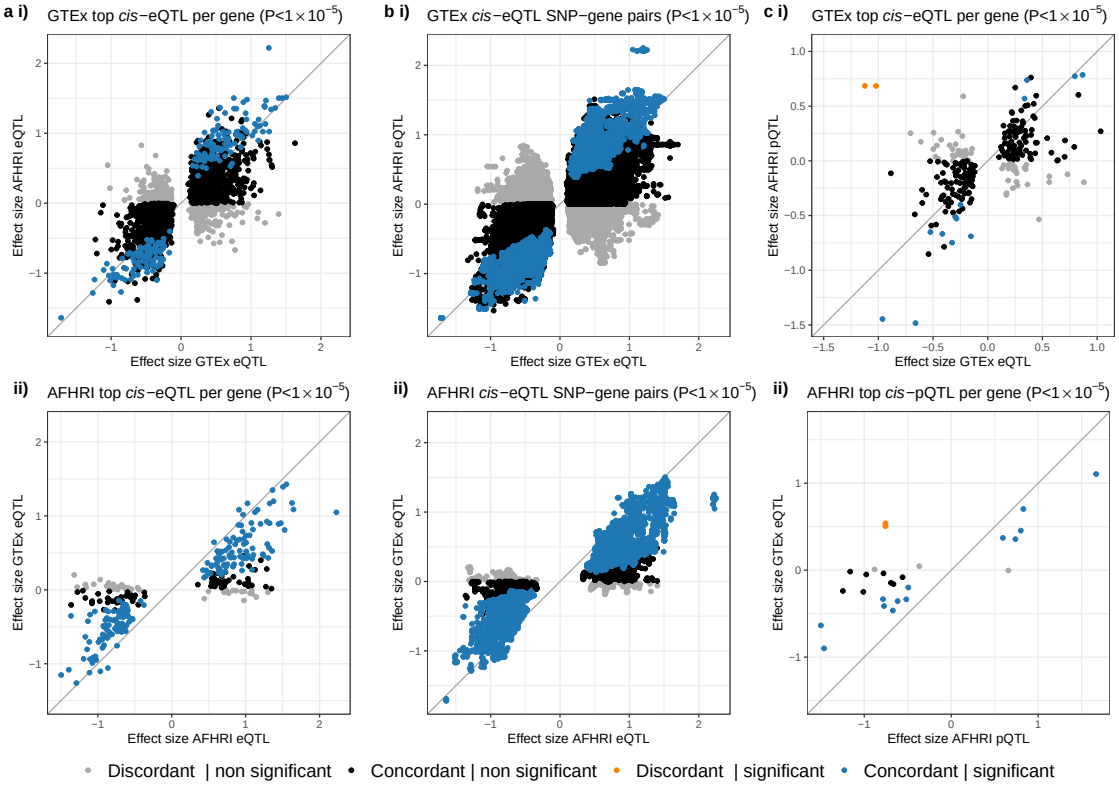


Figure 4.4: Comparison of *cis* eQTL and pQTL results to GTEx *cis* eQTLs in atrial appendage tissue.
a: Comparison of effect sizes of eQTLs in the GTEx and AFHRI cohort for i) the top significant *cis* eQTL per gene in GTEx ($P < 1 \times 10^{-5}$) and ii) the top significant *cis* eQTL per gene in AFHRI ($P < 1 \times 10^{-5}$).
b: Comparison of effect sizes of eQTLs in the GTEx and AFHRI cohort for i) all significant *cis* eQTL SNP-gene pairs in GTEx ($P < 1 \times 10^{-5}$) and ii) all significant *cis* eQTL SNP-gene pairs in AFHRI ($P < 1 \times 10^{-5}$).
c: Comparison of effect sizes of eQTLs in GTEx and pQTLs in the AFHRI cohort i) the top significant *cis* eQTL per gene in GTEx ($P < 1 \times 10^{-5}$) and ii) the top significant *cis* pQTL per gene in AFHRI ($P < 1 \times 10^{-5}$).
All cutoffs refer to uncorrected, nominal P values derived by two-sided tests. Permutation tests were performed for all reported GTEx results, all data from the AFHRI cohort was evaluated by T tests.
eQTL, expression quantitative trait loci; pQTL, protein quantitative trait loci; GTEx, Genotype-Tissue Expression project;
Figure and legend taken from the supplementary material Assum et al. [2022b] [Assum et al., 2022a] `https://doi.org/10.1038/s41467-022-27953-1`.

4.1.3.2 Plasma pQTLs

Plasma pQTLs from the Sun et al. [2018] study were filtered for *cis* regions and overlapping proteins when mapping the aptamer-based multiplex protein assay (SOMAscan) to gene symbols used in the AFHRI-B proteomics. As this was a comparison of blood plasma with heart tissue and different measurement technologies, naturally expression and detection of transcripts and proteins differed significantly. None of our significant *cis* pQTL genes (nominal P value $P < 1 \times 10^{-5}$) were measured in the plasma pQTL study and of all our significant *cis* eQTL genes ($P < 1 \times 10^{-5}$), only AKR1B1 was measured as a plasma protein. Comparing the top significant plasma *cis* pQTLs per gene to the AFHRI-B *cis* pQTLs and *cis* eQTLs, we still observe a clear correlation of effect sizes (Figure 4.5), even though significance varies.

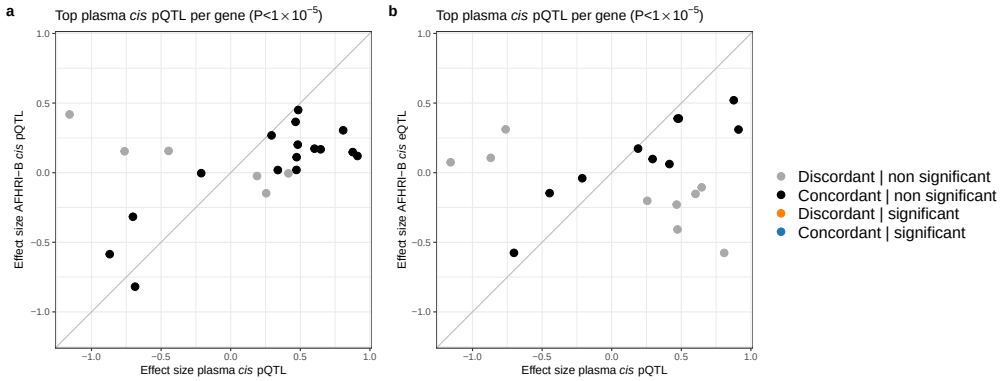


Figure 4.5: Comparison of *cis* eQTL and *cis* pQTL results to plasma *cis* pQTLs.
a: Comparison of effect sizes of the top significant plasma *cis* pQTL per gene ($P < 1 \times 10^{-5}$) to the AFHRI-B *cis* pQTLs.
b: Comparison of effect sizes of the top significant plasma *cis* pQTL per gene ($P < 1 \times 10^{-5}$) to the AFHRI-B *cis* eQTLs.
eQTL, expression quantitative trait loci; pQTL, protein quantitative trait loci;

4.1.3.3 Replication across different covariate sets

The amount of discovered QTLs varied when using different sets of covariates including various numbers of PEER factors. Since covariate sets were optimized for the maximum amount of discoveries, it was necessary to show that there was no inflation by false discoveries.
We therefore compared effect sizes for all SNP-mRNA pairs which were significant with any of the sets of covariates. The corresponding Pearson's correlation ranged between 0.954 and 0.988 with replication rates based on Storey's q-value method of over 99 %. Figure 4.6a is a scatter plot of the effect sizes of all SNP-mRNA pairs which were significant for either only PEER factors as covariates or the most comprehensive set of covariates, including PEER factors, age, sex, BMI, disease status, the fibroblast-score and three genotype principal components.
The same was observed for proteins, where the Pearson's correlation of effect sizes

for all significant SNP-protein pairs (see Figure 4.6b) for any of the sets of covariates ranged between 0.956 and 0.980. Q-value based replication rates were larger than 99 % as well.

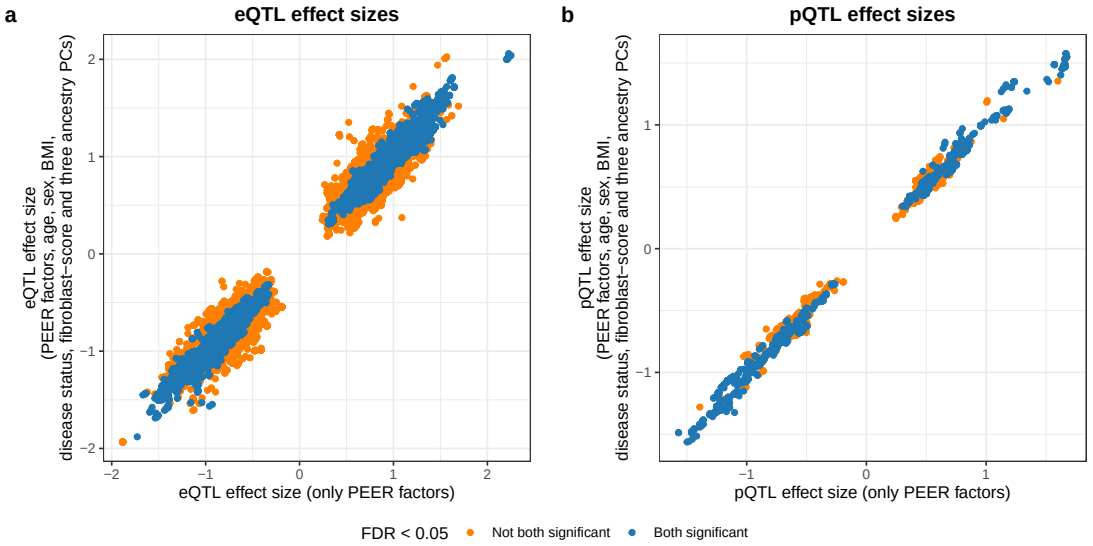


Figure 4.6: Correlation between *cis* eQTL/pQTL effect sizes computed using different sets of covariates.
a: Comparison of *cis* eQTL effect sizes for all SNP-mRNA pairs significant (FDR < 0.05) when using either twelve PEER factors alone or six PEER factors, age, sex, BMI, disease status, the fibroblast-score and three ancestry prinicpal components as covariates.
b: Comparison of *cis* pQTL effect sizes for all SNP-protein pairs significant (FDR < 0.05) when using either ten PEER factors alone or ten PEER factors, age, sex, BMI, disease status, the fibroblast-score and three ancestry prinicpal components as covariates.
eQTL, expression quantitative trait loci; pQTL, protein quantitative trait loci;

Another possibility to assess reliability of results is to consider replication rates for an independent dataset. We therefore compared the *cis* QTLs for the different covariate sets to the GTEx atrial appendage eQTLs and estimated the replication rates (Storey's q-value method) which are summarized in Table 4.2. Replication rates where much higher when comparing all SNP-gene pairs instead of only the top eQTL per gene, but were almost identical when using twelve PEER factors only and when taking into account PEER factors, age, sex, BMI, disease status, fibroblast-score and three ancestry principal components.

4.1.4 Overlap of *cis* eQTLs and pQTLs

Local genetic variation can influence biological processes on various levels, including transcriptional and post-transcriptional regulation [Hause et al., 2014, Battle et al., 2015, Eraslan et al., 2019] and can therefore affect mRNA and protein abundance independently. In the following, we therefore functionally characterize and compare *cis*-genetic effects on transcript and protein level. While many studies still rely heavily on transcriptomics measurements due to differences in effort and feasibility compared

Table 4.2: Replication rates of *cis* eQTLs for the AFHRI-B cohort in the GTEx dataset for different sets of eQTL covariates.
The fraction $\hat{\pi}_0$ of true null hypothesis was estimated from P values using Storey's q-value method. The replication rate between the datasets was then determined as $1 - \hat{\pi}_0$.
eQTL, expression quantitative trait loci; PEER, probabilistic estimation of expression residuals; FDR, false discovery rate; BMI, body mass index; PC, principal component;

Significance cutoff (AFHRI-B cohort)	Considered associations	Twelve PEER factors	Six PEER factors, age, sex, BMI, disease status, fibroblast-score and three ancestry PCs
FDR < 0.05	All SNP-gene pairs	92 % ($\hat{\pi}_0 = 0.08$)	91 % ($\hat{\pi}_0 = 0.09$)
FDR < 0.05	Top eQTL per gene	59 % ($\hat{\pi}_0 = 0.41$)	60 % ($\hat{\pi}_0 = 0.40$)
$P < 1 \times 10^{-5}$	All SNP-gene pairs	95 % ($\hat{\pi}_0 = 0.05$)	94 % ($\hat{\pi}_0 = 0.06$)
$P < 1 \times 10^{-5}$	Top eQTL per gene	85 % ($\hat{\pi}_0 = 0.15$)	85 % ($\hat{\pi}_0 = 0.15$)

to proteomics technologies, proteins abundances have been suggested as more direct determinants for phenotypic consequences of QTLs [Battle et al., 2015].
Therefore, we considered the mRNA and protein correlation as well as the overlap of matched *cis* QTLs. Figure 4.7 visualizes the distribution of *cis* eQTLs and *cis* pQTLs (see also Table 4.1) and its overlap across the genome.

Moreover, we were able to evaluate in detail, which *cis* eQTL and pQTL loci overlap. Therefore, we compared the effect sizes of each top *cis* QTL per gene from one omic to the other as shown in Figure 4.8a. Of all 1 243 genes with transcriptomics and proteomics data, 124 genes had a significant *cis* eQTL (FDR < 0.05). The top eQTL per gene was also a significant pQTL (FDR < 0.05) for only 10 % of the hits, but 74 % showed concordant effects. Using Storey's q-value method, we estimated a replication rate of 32 % and effect sizes showed a correlation of 0.58. Similarly, of 87 genes with transcript and protein measurements and a significant pQTL (FDR < 0.05), only 14 % also had an eQTL for the top pQTL per gene while 22 % showed concordant effects with a q-value based replication rate of 50 % (Figure 4.8b) and a Pearson correlation of effect sizes of 0.66.
Additionally, SNP-gene pairs can be evaluated instead of the top QTL per gene only. Comparing all significant *cis* eQTLs to pQTLs, 14 % (642 out of 3 906) replicated and 74 % had concordant effects. Conversely, 16 % of pQTLs were replicated by eQTLs and 81 % had concordant effects. The correlation of effect sizes ranged between 0.58 and 0.75.

4.1.4.1 *Cis* eQTL/pQTL overlap in other studies

First of all, as this was the first study of pQTLs in human tissue, no comparable dataset was available. However, there have been other pQTL studies working with e.g. plasma [Sun et al., 2018] or lymphoblastoid cell lines (LCL) [Battle et al., 2015].
Sun et al. [2018] compared their pQTLs to eQTLs from the GTEx project [Gamazon et al., 2018]. For 40 % of the identified plasma pQTLs they found a corresponding eQTL in any of the GTEx tissues. Only half of them (19 %) were found in whole blood, the

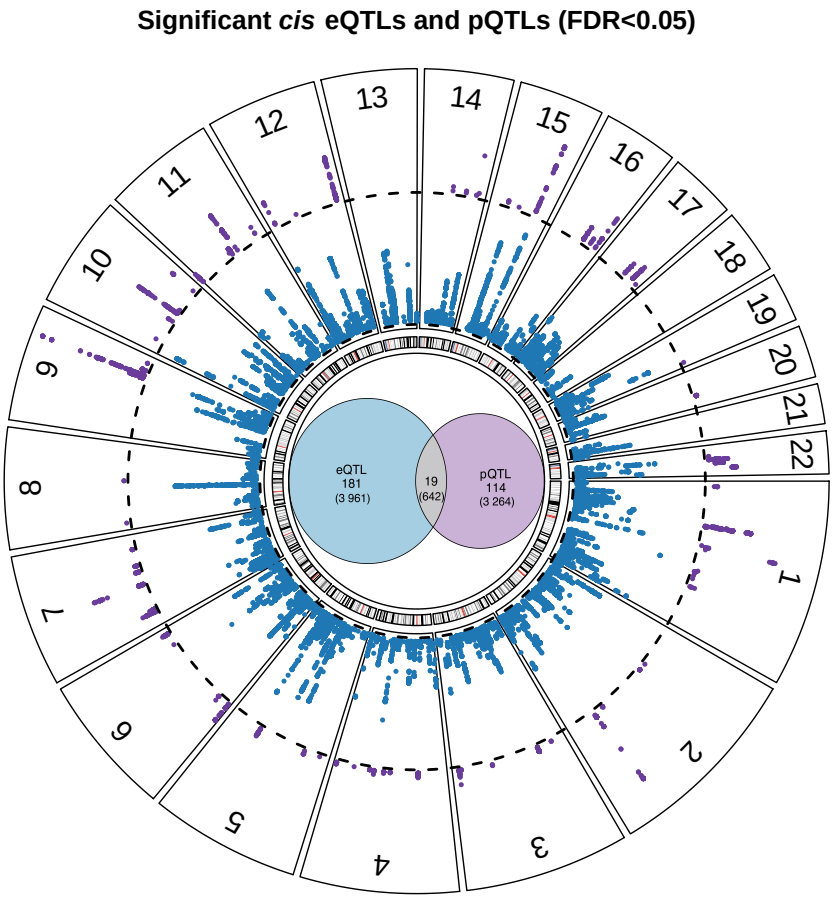


Figure 4.7: Significant *cis* eQTLs, *cis* pQTLs and their overlap.
Circular plot of the significant *cis* eQTLs (blue) and pQTLs (purple) at a FDR cutoff of 0.05 (dotted line, plot created using the R package circlize [Gu et al., 2014]). Considering only genes with both transcriptomics and proteomics measurements, we visualized the overlap of significant eQTLs and pQTLs in the circle center. In total, the lead SNP-gene pair of 200 QTL clumps in 124 genes had a significant eQTL and 133 loci in 87 genes a significant pQTL. Only 19 lead variants (13 genes) had an eQTL and pQTL for the same gene. The numbers in brackets represent the number of significant SNP-gene pairs.
eQTL, expression quantitative trait loci; pQTL, protein quantitative trait loci; QTL, quantitative trait loci; FDR, false discovery rate; LD, linkage disequilibrium; SNP, single nucleotide polymorphism;
Figure and legend adapted from Assum et al. [2022a] `https://www.nature.com/articles/s41467-022-27953-1/figures/1`, licensed under CC BY 4.0 `https://creativecommons.org/licenses/by/4.0/`.

tissue type closest to plasma. Additionally, according to the supplemental information given by the authors, the overlap was even smaller when considering heart tissues. In this case, less than 7.5 % of plasma pQTLs had a matched eQTL in GTEx atrial appendage or left ventricle tissue. Depending on the tissue, 12 % to 21 % of GTEx eQTLs had a matched plasma pQTL which is comparable to 17 % observed for our dataset.
Battle et al. [2015] evaluated consequences of genetic variants in LCLs by considering matched genotypes, transcriptomics, ribosome occupancy and protein abundance.

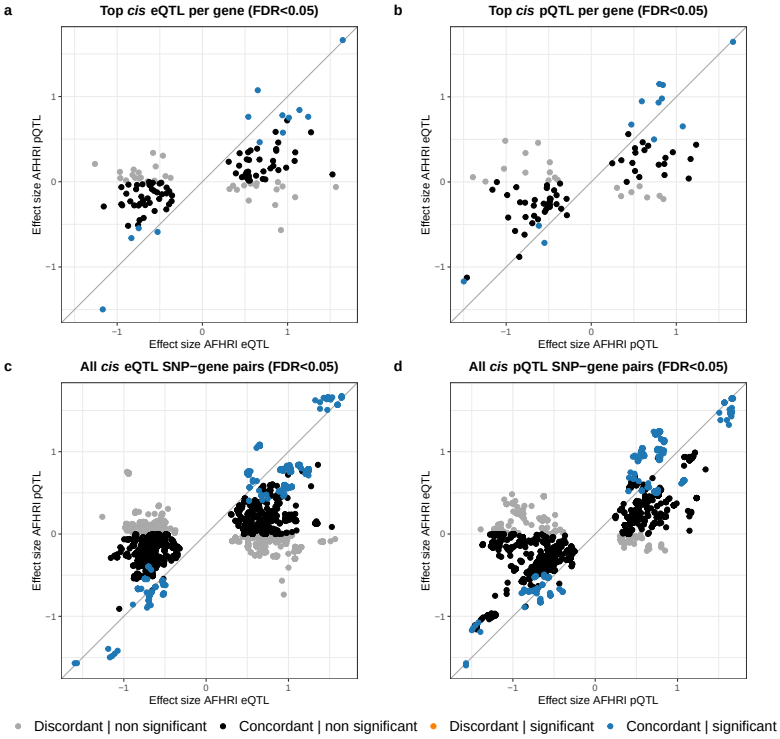


Figure 4.8: Between-omic comparison of *cis* QTL results.
a: Effect size of the top significant *cis* eQTL per gene (FDR < 0.05) to its matched pQTL.
b: Effect size of the top significant *cis* pQTL per gene (FDR < 0.05) to its matched eQTL.
c: Effect sizes of all significant *cis* eQTL SNP-gene pairs (FDR < 0.05) to the matched pQTLs.
d: Effect sizes of all significant *cis* pQTL SNP-gene pairs (FDR < 0.05) to the matched eQTLs.
eQTL, expression quantitative trait loci; pQTL, protein quantitative trait loci; SNP, single-nucleotide polymorphism; FDR, false discovery rate;
Figure and legend adapted from the supplementary material Assum et al. [2022b] [Assum et al., 2022a] `https://doi.org/10.1038/s41467-022-27953-1`.

Considering eQTL SNP-gene pairs, 35 % were also found on protein level and vice versa, 67 % of pQTLs were also eQTLs. Even though these replication rates were much higher than in our study, there were multiple important differences which explain the differing results: First, the coverage of proteins was much higher, as 4 381 proteins were measured. Second, a less stringent cutoff of FDR < 0.1 was chosen and multiple testing adjustment was only applied to the significant SNP-gene pairs in the discovery dataset. Third, transcriptomics as well as proteomics were obtained with different measurement techniques. Finally, a similar study by Hause et al. [2014] also evaluated *cis* eQTLs and pQTLs in LCLs and actually found no overlapping *cis* pQTLs and *cis* eQTLs at a significance threshold of FDR < 0.05.
As also shown in our data, different methods to assess replication rates lead to different results. Significant SNP-gene pairs or lead SNPs per gene can be evaluated. Additionally, LD information can be included by taking the lead SNP per LD regions ($r^2 \geq 0.8$ for Sun et al. [2018]) per gene. Different datasets, replication rates and methods to assess the overlap are also summarized in Table 4.3.

Table 4.3: Overlap of *cis* eQTLs and pQTLs in other studies.
Comparison of *cis* eQTL/pQTL overlap in previously published studies compared to our dataset. Information about plasma pQTL overlap with GTEx eQTLs is either based on the Sun et al. manuscript (marked with "paper") or derived from the Sun et al. supplementary table S8 (marked with "suppl.", [Sun et al., 2018]).
eQTL, expression quantitative trait loci; pQTL, protein quantitative trait loci; FDR, false discovery rate; LCL, lymphoblastoid cell line; GTEx, Genotype-Tissue Expression project;
Table and legend taken from the supplementary material Assum et al. [2022b] [Assum et al., 2022a] `https://doi.org/10.1038/s41467-022-27953-1`.

Cohort / datasets	Overlap of *cis* pQTLs with *cis* eQTL	Overlap of *cis* eQTLs with *cis* pQTL
AFHRI cohort:		
QTL genes	n = 21 (24 %)	n = 21 (17 %)
AFHRI cohort:		
SNP-gene pairs	n = 642 (16 %)	n = 642 (14 %)
LCL Hause et al. [2014]:	none at FDR < 0.05	none at FDR < 0.05
LCL Battle et al. [2015]:		
SNP-gene pairs	67 %	35 %
Plasma pQTL Sun et al. [2018] to all GTEx tissues:		
Lead SNPs for same gene, high LD regions ($r^2 \geq 0.8$)	n = 224 (40 %) (paper)	
All sentinel SNPs listed in the Sun et al. supplements	n = 320 (∼40 %) (suppl.)	
Plasma pQTL Sun et al. [2018] to GTEx whole blood:		*cis* eQTL: P < 1.5×10^{-11}
Lead SNPs for same gene, high LD regions ($r^2 \geq 0.8$)	n = 117 (19 %) (paper)	12.2 % (paper)
All sentinel SNPs listed in the Sun et al. supplements	n = 152 (∼19 %) (suppl.)	
Plasma pQTL Sun et al. [2018] to GTEx heart tissue:		
Lead SNPs for same gene, high LD regions ($r^2 \geq 0.8$)		
All sentinel SNPs listed in the Sun et al. supplements	n = 60 (∼7.5 %) (suppl.)	
Plasma pQTL Sun et al. [2018] to GTEx liver:		*cis* eQTL: P < 1.5×10^{-11}
Lead SNPs for same gene, high LD regions ($r^2 \geq 0.8$)	n = 70 () (paper)	14.8 % (paper)
All sentinel SNPs listed in the Sun et al. supplements	n = 126 (∼16 %) (suppl.)	
Plasma pQTL Sun et al. [2018] to GTEx monocytes:		*cis* eQTL: P < 1.5×10^{-11}
Lead SNPs for same gene, high LD regions ($r^2 \geq 0.8$)	n = 52 () (paper)	14.7 % (paper)
All sentinel SNPs listed in the Sun et al. supplements	n = 94 (∼12 %) (suppl.)	

Additionally, we can also evaluate the correlation of mRNA and protein values across samples for every gene while separating genes without and with an eQTL. As shown in Table 4.4, correlations between mRNA and protein were very similar between the AFHRI-B cohort (AFHRI-B eQTL and GTEx eQTL annotations) and the Hapmap LCL data. The results are also visualized in Figure 4.9 with the stronger separation in panel c being due to a generally lower correlation of mRNA and protein for genes without an eQTL.

Table 4.4: Correlation between mRNA and protein for *cis* eQTL genes.
Comparison of median R^2 and correlation dependent on existing eQTL annotations. Compared are AFHRI genes non *cis* eQTL vs. *Cis* eQTL genes, AFHRI-B genes without an *cis* eQTL in the GTEx atrial appendage data vs. genes with eQTL in GTEx [Gamazon et al., 2018] and Hapmap LCL [Battle et al., 2015] data without and with an eQTL for the LCL computations.
eQTL, expression quantitative trait score; LCL, lymphoblastoid cell line; GTEx, Genotype-Tissue Expression project;
Table taken from the supplementary material Assum et al. [2022b] [Assum et al., 2022a] `https://doi.org/10.1038/s41467-022-27953-1`.

	AFHRI-B cohort		AFHRI-B cohort		Hapmap LCL data	
Measure	**Non *cis* eQTL**	***cis* eQTL**	**Non GTEx *cis* eQTL**	**GTEx *cis* eQTL**	**LCL non *cis* eQTL**	**LCL *cis* eQTL**
Median(R^2)	0.0265	0.0378	0.0243	0.0354	0.0205	0.0376
Mean(R^2)	0.0634	0.0868	0.0604	0.0771	0.0466	0.0798
Median(correlation)	0.145	0.174	0.137	0.175	0.105	0.176
Mean(correlation)	0.163	0.195	0.156	0.188	0.110	0.185

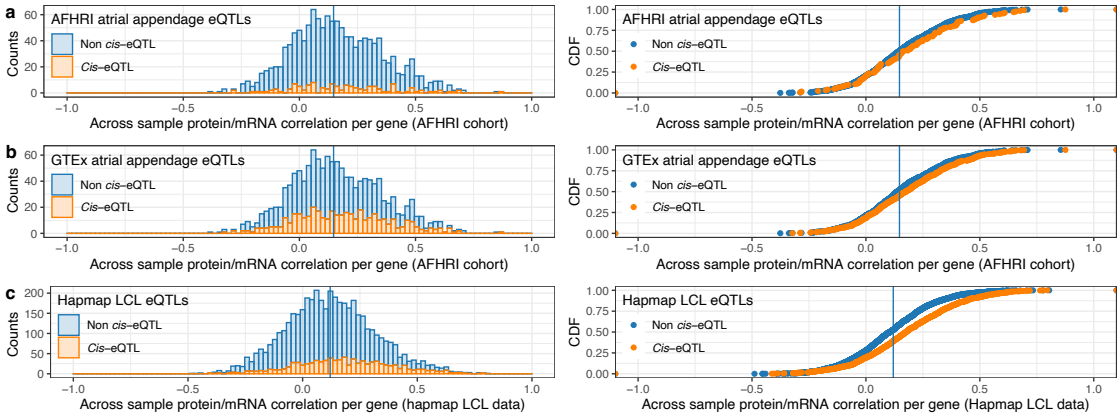


Figure 4.9: Pearson correlation between transcript and protein levels dependent on *cis* eQTL annotations in different datasets.
a: AFHRI cohort and *cis* eQTLs. Histogram and cumulative density function.
b: AFHRI cohort and GTEx *cis* eQTLs annotations. Histogram and cumulative density function.
c: Hapmap LCL data and Hapmap LCL *cis* eQTL annotations. Histogram and cumulative density function.
eQTL, expression quantitative trait loci; LCL, lymphoblastoid cell line;
Figure and legend taken from the supplementary material Assum et al. [2022b] [Assum et al., 2022a] `https://doi.org/10.1038/s41467-022-27953-1`.

4.1.5 Functional *cis* QTL categories

In general, high correlation between mRNA and protein indicates good data quality, however information about divergent effects in the two omics might be even more valuable. Specifically, genetic variation which does not progress to protein level might possibly be discarded while variants influencing protein abundance directly might be highly relevant.
Functionally characterizing *cis*-regulatory variants remains challenging due to the diverse mechanisms of potential transcriptional and post-transcriptional regulation. However, analyzing shared and independent effects of mRNA and protein helps narrowing down the number of possible regulations to consider. In this work, we explored three ways of grouping QTLs in three categories of shared effects across mRNA and protein, effects only present on mRNA level and effects exclusive to proteomics:

- **Lead SNP of a LD clump:**
  LD-clumping can be performed taking both eQTL and pQTL summary statistics per gene into account. For each clump, the eQTL and pQTL of the lead SNP can be evaluated as a proxy.
- **Colocalization analysis:**
  Approximate Bayes' Factor analysis evaluates shared or independent causal variants based on the paired summary statistics per gene.
- **Residual regression approach:**
  QTL computations on the residuals enable the evaluation of variance shared across the omics versus independent effects for each variant but does not include the LD structure.

If we consider all SNP-gene pairs with mRNA and protein measurements (4 198 168 pairs for 2 249 758 SNPs and 1 243 genes) and either a significant *cis* eQTL or *cis* pQTL (FDR < 0.05), then only 8.2 % overlapped, i.e. only 642 out of 4 603 eQTLs were also one of the 3 906 pQTLs. Similarly, when performing LD-clumping on both eQTL and pQTL summary statistics, 314 independent loci for 190 genes with an eQTL or pQTL were identified (200 for eQTLs and 133 for pQTLs), of which only 19 (approx. 6.1 %) overlapped, i.e. the lead SNP was a significant eQTL and pQTL (FDR < 0.05).

4.1.5.1 Residual regression approach

As described in the methods (see 3.4.1.3, 3.4.3.3), residual QTLs were computed by taking advantage of the matched individual transcriptomics and proteomics data. These results were combined to assign three types of functional *cis* QTL categories (results shown in Table 4.5).
First, we defined truly shared *cis* eQTLs/pQTLs as all SNP-gene pairs with a significant eQTL and pQTL, where additionally both residual eQTLs and pQTLs vanished, i.e. the variant-specific variation was removed by computing the residuals and therefore shared across both omics. In total, 430 shared *cis* eQTLs/pQTLs for eleven different genes or the lead SNP of 14 independent LD clumps for eight genes were defined based on that. These shared QTLs represent the classical case of transcriptional regulation which progresses to protein level.

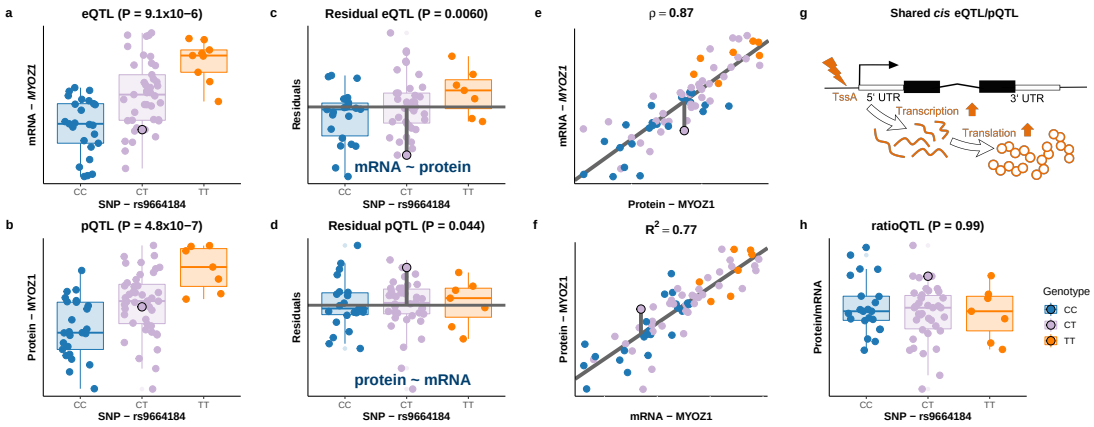


Figure 4.10: Definition of shared *cis* eQTLs/pQTLs.
Shared eQTLs/pQTLs represent QTLs where the effect of transcriptional regulation translates into mRNA and protein abundance exemplified by the significant SNP-gene pair rs9664184-MYOZ1. No corresponding ratioQTL can be observed as the genetic variation is shared across both omics levels.
eQTL, expression quantitative trait loci; pQTL, protein quantitative trait loci; ratioQTL, ratio quantitative trait loci; TssA, active transcription start site; UTR, untranslated region; TFBS, transcription factor binding site; RBP, RNA binding protein; SNP, single-nucleotide polymorphism; IQR, interquartile range;
Figure adapted from Assum et al. [2022a] `https://www.nature.com/articles/s41467-022-27953-1/figures/2` licensed under CC BY 4.0 `https://creativecommons.org/licenses/by/4.0/` including supplementary materials [Assum et al., 2022b].

Similarly, we defined independent *cis* eQTLs as all the SNP-gene pairs with a significant eQTL and residual eQTL but no pQTL or residual pQTL. Besides the strong eQTL,

we also required that the residual eQTL was significant in order to prevent false identification of an independent eQTL in case of only attenuated pQTL effect sizes. We found 1 593 such independent eQTLs for 37 genes which can be reduced to 62 lead SNPs of independent loci in 34 genes. Here, next to transcriptional regulation, there are additional factors preventing the mRNA changes to take effect on protein level. Although this phenomenon is often observed, the underlying cause often remains elusive. Possible mechanisms could be the adaptation of translational rates, protein degradation, protein complex formation and interfering long-noncoding RNAs [Liu et al., 2016, Eraslan et al., 2019].

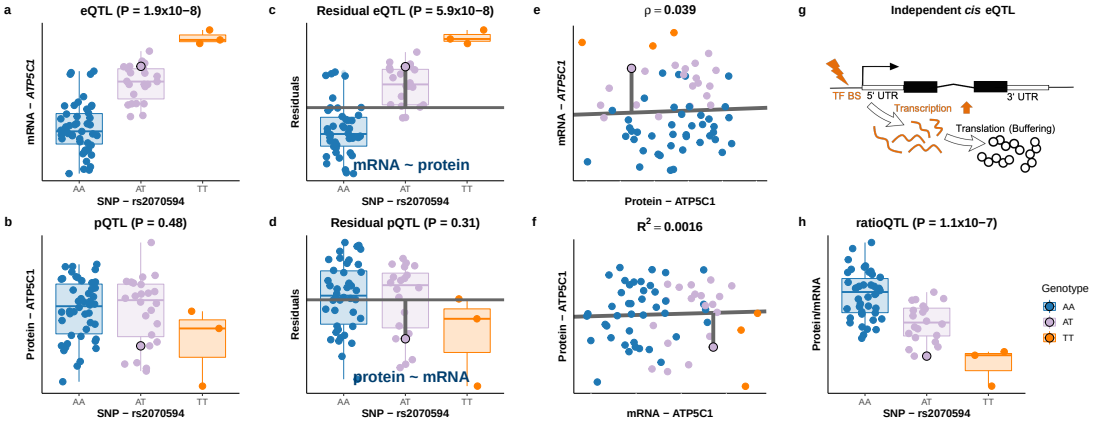


Figure 4.11: Definition of independent *cis* eQTLs.
Independent eQTLs depict variants with regulation on mRNA but not on protein level displayed by the significant SNP-transcript pair rs2070594-*ATP5C1*.
eQTL, expression quantitative trait loci; pQTL, protein quantitative trait loci; ratioQTL, ratio quantitative trait loci; TssA, active transcription start site; UTR, untranslated region; TFBS, transcription factor binding site; RBP, RNA binding protein; SNP, single-nucleotide polymorphism; IQR, interquartile range;
Figure adapted from Assum et al. [2022a] `https://www.nature.com/articles/s41467-022-27953-1/figures/2` licensed under CC BY 4.0 `https://creativecommons.org/licenses/by/4.0/` including supplementary materials [Assum et al., 2022b].

Lastly, independent *cis* pQTLs were defined accordingly as SNP-gene pairs with a significant pQTL and residual pQTL but no eQTL or residual eQTL. Such QTLs which are only observable on protein level were found for 1 083 SNP-gene pairs for 21 genes, also represented by 25 lead variants of LD clumps for 19 genes. As opposed to the two categories before, independent pQTLs are only regulated on post-transcriptional level. *Cis* QTL variants can possibly influence RNA binding proteins (RBPs) which influence mRNA translation [Robert and Pelletier, 2018] and lead to the changes on protein level only.

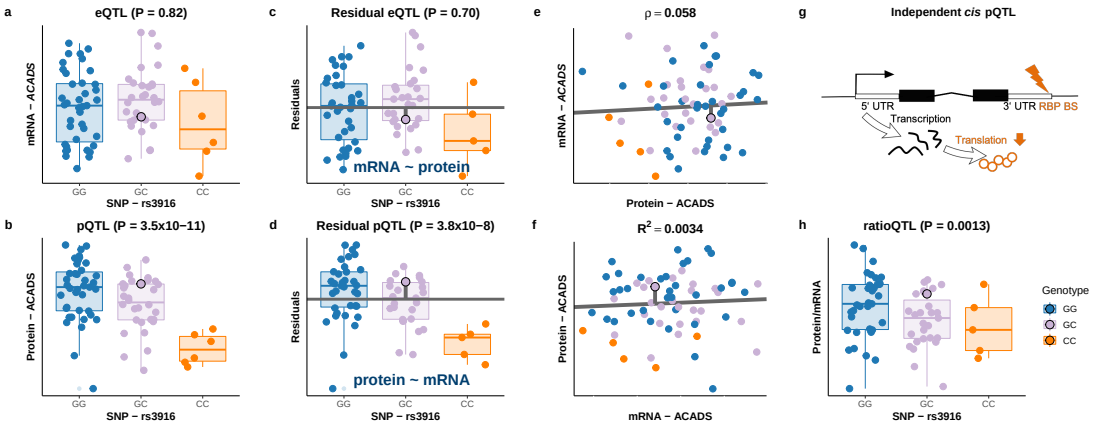


Figure 4.12: Definition of independent *cis* pQTLs.
Independent pQTLs represent variants which show regulation only on protein level as shown for the SNP-protein pair rs3916-ACADS. Genetic influence is not observable on transcript level.
eQTL, expression quantitative trait loci; pQTL, protein quantitative trait loci; ratioQTL, ratio quantitative trait loci; TssA, active transcription start site; UTR, untranslated region; TFBS, transcription factor binding site; RBP, RNA binding protein; SNP, single-nucleotide polymorphism; IQR, interquartile range;
Figure adapted from Assum et al. [2022a] `https://www.nature.com/articles/s41467-022-27953-1/figures/2` licensed under CC BY 4.0 `https://creativecommons.org/licenses/by/4.0/` including supplementary materials [Assum et al., 2022b].

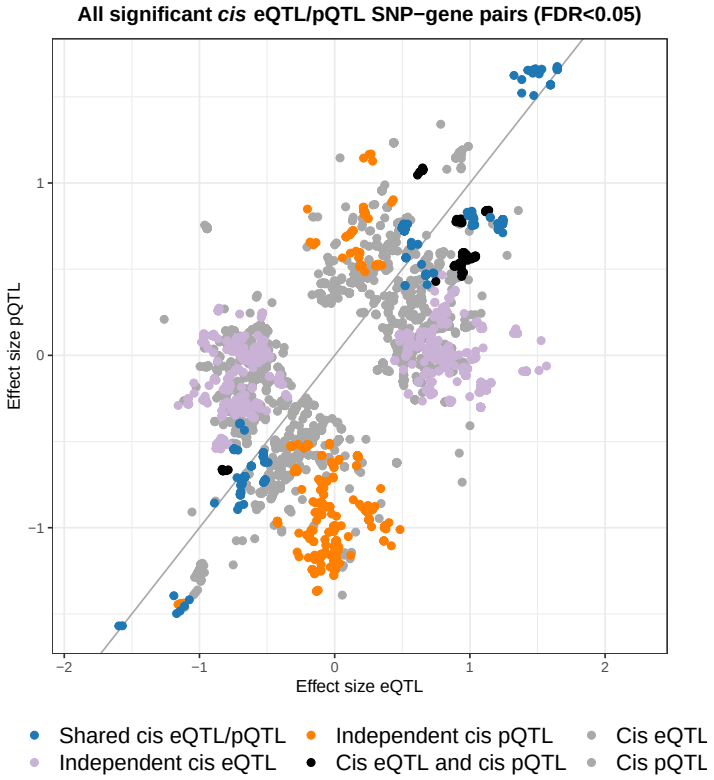


Figure 4.13: Comparison of *cis* eQTL and pQTLs effect sizes.
Comparison of effect sizes of eQTLs and pQTLs in the AFHRI cohort for all significant *cis* eQTL and *cis* pQTL SNP-gene pairs (FDR < 0.05, Benjamini-Hochberg procedure).
eQTL, expression quantitative trait loci; pQTL, protein quantitative trait loci; SNP, single-nucleotide polymorphism; FDR, false discovery rate;
Figure and legend adapted from the supplementary material Assum et al. [2022b] [Assum et al., 2022a] `https://doi.org/10.1038/s41467-022-27953-1`.

Table 4.5: Summary of tested data and discovered *cis* quantitative trait loci.
Significant *cis* QTLs (FDR < 0.05) in heart atrial appendage tissue for mRNA and protein measurements. Loci denote the number of independent loci derived by LD-clumping.
QTL, quantitative trait loci; FDR, false discovery rate; LD, linkage disequilibrium; eQTL, expression quantitative trait loci; pQTL, protein quantitative trait loci; res eQTL, residual expression quantitative trait loci; res pQTL, residual protein quantitative trait loci; N, sample size;
Table and legend adapted from the supplementary material Assum et al. [2022b] [Assum et al., 2022a] `https://doi.org/10.1038/s41467-022-27953-1`.

	Tested:			FDR < 0.05:			
	SNPs	**Pairs**	**Genes**	**Pairs**	**Genes**	**Loci (genes)**	**N**
eQTL	2 249 758	4 198 168	1 243	4 603	124	201 (124)	75
pQTL	2 249 758	4 198 168	1 243	3 906	87	133 (87)	75
res eQTL	2 249 758	4 198 168	1 243	2 261	63	99 (63)	66
res pQTL	2 249 758	4 198 168	1 243	1 316	34	45 (34)	66
Shared eQTL/pQTL	2 249 758	4 198 168	1 243	430	11	14 (8)	66
Independent eQTL	2 249 758	4 198 168	1 243	1 593	37	62 (34)	66
Independent pQTL	2 249 758	4 198 168	1 243	1 083	21	25 (19)	66

4.1.5.2 Colocalization analysis

Another common way to compare shared or independent effects between two modalities is colocalization analysis. Using Approximate Bayes' Factors, posterior probabilities for the five different hypothesis were derived for each of the LD clumps with either an eQTL or pQTL. For 18 out of 19 LD clumps, where the lead SNP was a significant eQTL and pQTL for the same gene, colocalization analysis suggested a shared causal variant, i.e. the posterior probability for H4 was larger than 0.5. A significant eQTL without a pQTL was found for 181 LD clumps. Colocalization analysis suggested an independent eQTL (H1 > 0.5) only for 64 of those. Vice versa, an independent pQTL (H2 > 0.5) was suggested for 33 out of 114 LD clumps.

Figure 4.14 summarizes the results of all three approaches. While there was strong agreement between the colocalization and residual approach, colocalization analysis tended to be more strict. This was to be expected by the additional individual information used for the regression analysis, as also illustrated by the example of the gene COQ5 which is visualized in Figure 3.5. On the other hand, there were clumps classified by each approach exclusively for all three categories.

In summary, our three functional *cis* QTL categories characterized extreme cases of regulation, which is also visualized in Figure 4.1.5.1. While the overall correlation of effect sizes of eQTLs and pQTLs is very high, clearly different regulatory mechanisms influence gene expression at different levels, manifesting in differences between transcriptomics and proteomics.
For a specific locus, all QTL types can be evaluated to get a better understanding of the strength of the genetic effects on transcript and protein expression and therefore reduce the number of possible mechanisms to investigate. This can be used to, first, pinpoint the specific mechanism, and second, to also prioritize between different candidate SNP

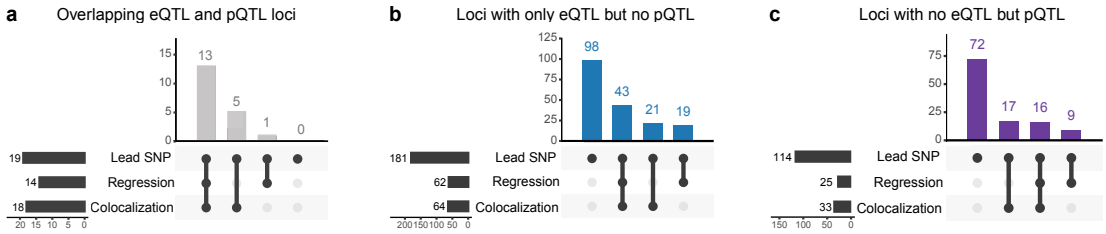


Figure 4.14: Characterization of significant *cis* eQTLs, *cis* pQTLs and their overlap.
a: Characterization of overlapping eQTL and pQTL loci. All 19 LD clumps (based on eQTL and pQTL summary statistics) where the lead SNP-gene-pair was a significant eQTL and pQTL were classified as a shared QTL by either our residual regression approach or colocalization analysis.
b: Characterization of eQTL loci without a corresponding pQTL. Only 83 out of 181 LD clumps (based on eQTL and pQTL summary statistics) which had a lead SNP-gene-pair with a significant eQTL but no pQTL were classified as an independent eQTL by either our residual regression approach or colocalization analysis.
c: Characterization of pQTL loci without a corresponding eQTL. Only 42 out of 114 LD clumps (based on eQTL and pQTL summary statistics) which had a lead SNP-gene-pair with a significant pQTL but no eQTL were classified as an independent pQTL by either our residual regression approach or colocalization analysis.
eQTL, expression quantitative trait loci; pQTL, protein quantitative trait loci; QTL, quantitative trait loci; FDR, false discovery rate; LD, linkage disequilibrium; SNP, single nucleotide polymorphism;
Figure and legend adapted from Assum et al. [2022a] `https://www.nature.com/articles/s41467-022-27953-1/figures/1`, licensed under CC BY 4.0 `https://creativecommons.org/licenses/by/4.0/`.

and gene sets. In order to get a clearer idea of possible mode of actions, we can also functionally annotate all SNP-gene pairs and query enrichments of specific regulatory elements for the different QTL types.

4.1.6 Enrichment *cis* QTLs for functional elements

Each SNP-gene pair was functionally annotated using public data such as chromatin states, gene annotations, transcription factor, mRNA and RBP binding sites as well as predictions for possible missense mutations or variants which introduce nonsense-mediated decay. Enrichments of functional elements were then evaluated for the single most significant QTL SNPs or the five most significant QTL SNPs per gene, ranked on the eQTL P value for eQTLs, pQTL P value for pQTLs, ratioQTL P value for ratioQTLs, pQTL P value for shared *cis* eQTLs/pQTLs, res eQTL P value for independent *cis* eQTLs and res pQTL P value for independent *cis* pQTLs. The results for different types of QTLs are summarized as heatmaps in Figure 4.15a-b and presented in more detail in Figure 4.16. In general, stronger enrichments were observed for the top five QTL SNPs, possibly due to the higher number of QTL SNPs used for the comparisons: 4 040 SNPs for 1 058 genes for eQTLs, 324 SNPs for 91 genes for pQTLs, 53 SNPs for 16 genes for ratioQTLs, 51 SNPs for eleven genes with a shared *cis* eQTL/pQTL, 151 SNPs for 37 genes with an independent *cis* eQTL and 73 SNPs for 21 genes with an independent *cis* pQTL.

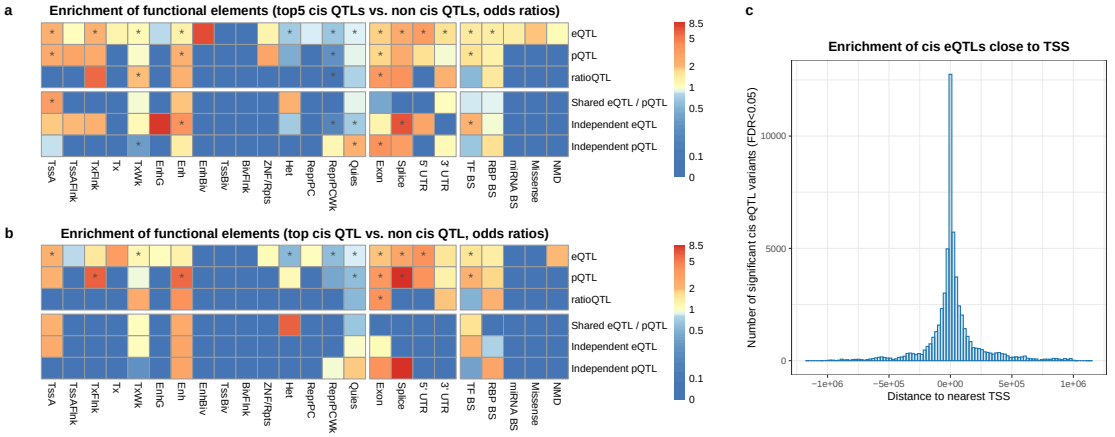


Figure 4.15: Enrichment of functional elements for different *cis* QTL categories.
a: Annotations of the top five *cis* QTL hits per gene were compared to a background distribution (100 background SNPs per QTL SNP) matched for MAF and distance to TSS. Displayed are odds ratios which represent enrichment or depletion, stars mark P values < 0.05 for the corresponding Fisher's exact test (two-sided).
b: Same as a, but for only the top *cis* QTL SNP per gene.
c: Enrichment of *cis* eQTL hits close to the nearest TSS.
QTL, quantitative trait loci; eQTL, expression quantitative trait loci; pQTL, protein quantitative trait loci; ratioQTL, ratio quantitative trait loci; UTR, untranslated region; TF, transcription factor; BS, binding site; NMD, nonsense-mediated decay; TSS, transcription start site; FDR, false discovery rate;
Figures a and c as well as the legend are adapted from the supplementary material Assum et al. [2022b] [Assum et al., 2022a] `https://doi.org/10.1038/s41467-022-27953-1`.

In line with previous studies [Lappalainen et al., 2013], eQTL SNPs were enriched in transcriptionally active or enhancer regions, such as chromatin states TssA (active transcription start site) or Enh (enhancer) but depleted in heterochromatin (Het), repressed (weak repressed polycomb ReprPCWk) and quiescent (Quies). When comparing the distance of an eQTL SNP to the nearest transcription start site (TSS), there is a strong enrichment of QTLs close to the TSS (see Figure 4.15c). Similar results - enrichments for TssA/Enh and depletion for ReprPCWk regions - were observed for pQTL SNPs. Furthermore, regulatory regions like the 3′ and 5′ UTR were significantly enriched for eQTL SNPs. Exons and splice sites were enriched for eQTLs (Figure 4.15a, Figure 4.16a) as well as pQTLs (Figure 4.15b, Figure 4.16b).

Considering the functional *cis* QTL categories (see Figure 4.15a-b and Figure 4.16b), variants with a shared *cis* eQTLs/pQTLs were enriched for regions of active TSS (chromatin state TssA). Independent *cis* eQTL variants were more often found in enhancer regions or overlapping with TF BS as well as within splicing sites. As already found by Battle et al. [2015], variants of protein specific QTLs which were comparable to our independent *cis* pQTL were significantly more often located in exons. Furthermore, even though not significant, independent pQTL variants appeared more often in RBP BS.

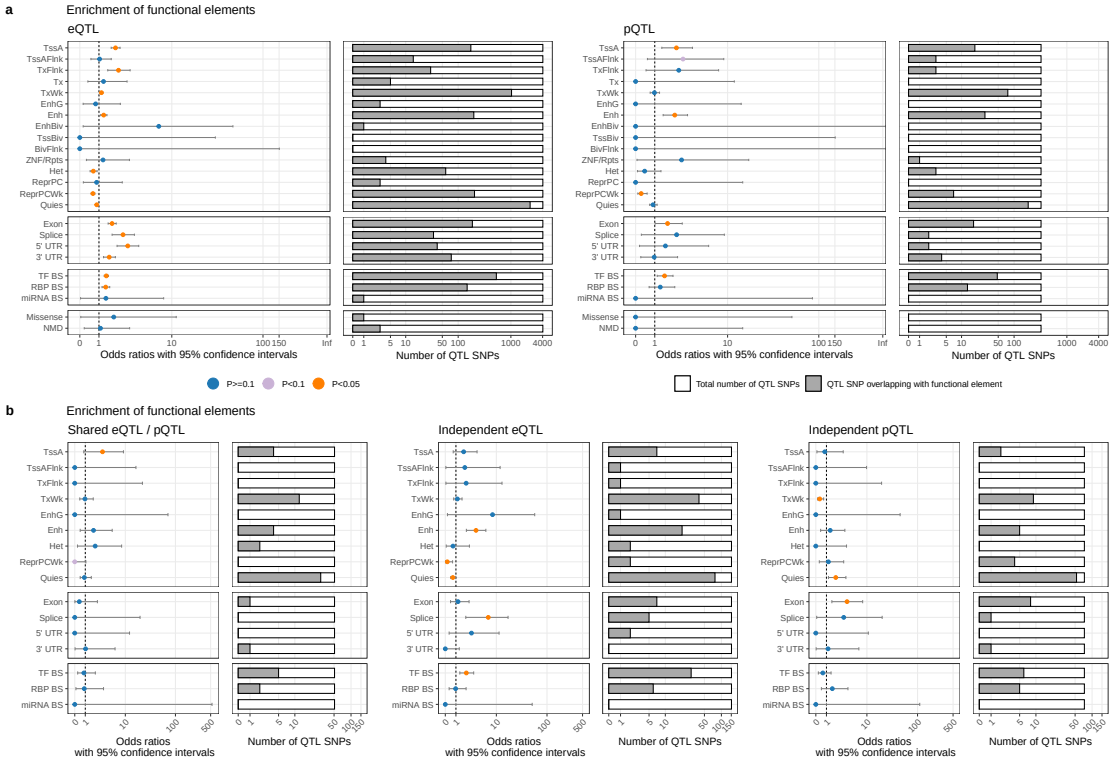


Figure 4.16: Enrichment of functional elements for different *cis* QTL categories.
a: Visualizations of the enrichment of functional elements for the top five *cis* eQTL and top five *cis* pQTL hits per gene. The filled dots represent the the estimated odds ratios with error bars for the 95 % confidence intervals (left panels). The right panels represent the absolute number of QTL SNPs evaluated (total bar length) and the fraction of those SNPs with the corresponding SNP-gene annotation (grey filling).
b: Similarly to a, enrichment results of functional elements for the top five SNPs per gene for the three functional *cis* QTL categories are shown. Left panels depict odds ratios with 95 % confidence intervals and right panels show absolute numbers of QTL SNPs.
QTL, quantitative trait loci; eQTL, expression quantitative trait loci; pQTL, protein quantitative trait loci; ratioQTL, ratio quantitative trait loci; UTR, untranslated region; TF, transcription factor; BS, binding site; NMD, nonsense-mediated decay; TSS, transcription start site; FDR, false discovery rate;

Additionally, nine out 1 083 SNPs (for four out of 21 genes), as shown in Table 4.6, were missense mutation which changed the amino acid sequence of the corresponding protein. Three of those (rs1801690, rs3858340 and rs1126501) were located in regions which were used for the protein identification and quantification, but none of those mutated peptides could be identified in any of the samples. This is in line with low or non-existent expression, specifically since the peptides in question were among the most abundant for the corresponding proteins and should - if expressed - be above the limit of detection.

In summary - especially in consideration of the small sample size - enrichments clearly reflect plausible links to assumed modes of action for the different QTL types.

Table 4.6: Missense mutations for independent *cis* pQTLs.
Missense mutations for all SNPs with an independent *cis* pQTLs with their corresponding allele and amino acid changes.
pQTL, protein quantitative trait loci;

SNP	Chromosome	Position	Gene	Allele change	Amino acid change
rs2186564	chr11	77 583 266	AAMDC	G>A	Val>Met
rs1799958	chr12	121 176 083	ACADS	G>A	Gly>Ser
rs1801690	chr17	64 208 285	APOH	C>G	Trp>Ser
rs8178847	chr17	64 216 815	APOH	C>T	Arg>His
rs52797880	chr17	64 216 854	APOH	A>G	Ile>Thr
rs3858340	chr10	121 436 286	BAG3	C>T	Pro>Leu
rs10761084	chr9	107 531 152	NIPSNAP3B	G>C	Ala>Pro
rs3739741	chr9	107 533 175	NIPSNAP3B	C>G	Ala>Gly
rs1126501	chr12	10 875 488	YBX3	T>C	Thr>Ala

4.1.7 Overlap with GWAS hits

Thousands of loci associated to cardiovascular diseases have been discovered in GWAS. For most of them, however, the functional mechanisms how genetics actually influence specific phenotypes remain unknown, especially due to differences in gene expression for different tissues.
QTL analyses are an important tool to investigate genotype-phenotype relationships by simultaneously analyzing gene expression changes and genetic associations for the same variants to pinpoint possible causal SNPs, genes and biological mechanisms.

First of all, any genetic variation which influences a trait needs to alter gene or protein expression. In this case, we were specifically interested in traits related to atrial tissue and therefore systematically evaluated any overlap of *cis* QTLs with SNPs annotated to any term in the categories cardiovascular disease (EFO_0000319) or cardiovascular measurements (EFO_0004298) in the GWAS catalog [Buniello et al., 2019]. The results are shown in Table 4.7 and visualized in Figure 4.17a.

Indeed, the most prevalent traits overlapping with *cis* QTL was (prevalent) atrial fibrillation with 17 eQTLs and four pQTLs. The individual GWAS loci and QTL SNP-gene pairs are listed in Table 4.8. While we only evaluated the best QTL per gene per GWAS hit, not all entries in the GWAS catalog were independent. Therefore, loci like the MYOZ1 10q22.2 appear more than once in this table. To remove that bias, we additionally evaluated the QTL-GWAS overlap on the combined *cis* eQTL/pQTL LD clumped results. Therefore, we first filtered our QTLs for all SNPs matching the Roselli et al. [2018] AF GWAS and then retained only the best eQTL/pQTL per LD clump per gene. Similarly to the enrichment of functional annotations, for each eQTL/pQTL hit we further sampled 100 background SNPs which matched for MAF and distance to the TSS, in order to then compare the number of SNPs with an AF GWAS P value $< 5 \times 10^{-8}$. As shown in Figure 4.18 and Table 4.9, we observed a strong enrichment of genome-wide significant SNPs for *cis* eQTL (P = 0.0016) and pQTL (P = 0.012) lead variants.

Table 4.7: *Cis* QTLs overlapping with GWAS hits.
Number of significant *cis* QTLs (FDR < 0.05) overlapping with variants annotated to cardiovascular traits in the GWAS catalog (or proxy with $R^2 > 0.8$).
QTL, quantitative trait loci; GWAS, genome-wide association study; COPD, Chronic obstructive pulmonary disease; eQTL, expression quantitative trait loci; pQTL, protein quantitative trait loci; ratioQTL, ratio quantitative trait loci; res pQTL, protein residual quantitative trait loci; res eQTL, expression residual quantitative trait loci;
Table and legend taken from the supplementary material Assum et al. [2022b] [Assum et al., 2022a] `https://doi.org/10.1038/s41467-022-27953-1`, licensed under CC BY 4.0 `https://creativecommons.org/licenses/by/4.0/`.

Trait	eQTL	pQTL	ratioQTL	res pQTL	res eQTL
Atrial fibrillation	15	3	0	0	0
Pulse pressure	12	1	0	0	0
Coronary artery disease	5	0	0	0	0
QT interval	5	0	0	0	0
Creatine kinase levels	1	1	1	1	0
COPD or resting heart rate (pleiotropy)	2	1	0	0	0
PR interval	3	0	0	0	0
Serum uric acid levels	2	0	0	0	1
Incident atrial fibrillation	1	1	0	0	0
Large artery stroke	2	0	0	0	0
Sudden cardiac arrest	2	0	0	0	0
Age-related disease endophenotypes	1	0	0	0	0
Birdshot chorioretinopathy	1	0	0	0	0
Carotid plaque burden (smoking interaction)	1	0	0	0	0
Circulating myeloperoxidase levels (serum)	1	0	0	0	0
Conotruncal heart defects (inherited effects)	1	0	0	0	0
Coronary artery disease (...)	1	0	0	0	0
Heart rate	1	0	0	0	0
Hematology traits	1	0	0	0	0
Homocysteine levels	1	0	0	0	0
Ischemic stroke	0	0	0	1	0
Left atrial antero-posterior diameter	1	0	0	0	0
Migraine	1	0	0	0	0
Peripheral arterial disease (...)	1	0	0	0	0
Prevalent atrial fibrillation	1	0	0	0	0
QRS complex (Sokolow-Lyon)	1	0	0	0	0
Resting heart rate	1	0	0	0	0
RR interval (heart rate)	1	0	0	0	0
Venous thromboembolism	1	0	0	0	0

When evaluating possible functional mechanisms for GWAS hits, analyses are often focused on transcriptomics information only. However, genetic variants can also just affect protein expression, such as the independent *cis* pQTL of the SNP rs1801690 for the gene APOH which was also a GWAS hit for creatine kinase levels (Figure 4.17b). Patients with the CG genotype show much lower levels of the APOH protein. Indeed, the C to G substitution changes the amino acid sequence, such that the Trypsin is switched to a Serin in the corresponding peptide. Slightly elevated transcript expression for patients with the CG genotype might be explained by a possible compensation approach.

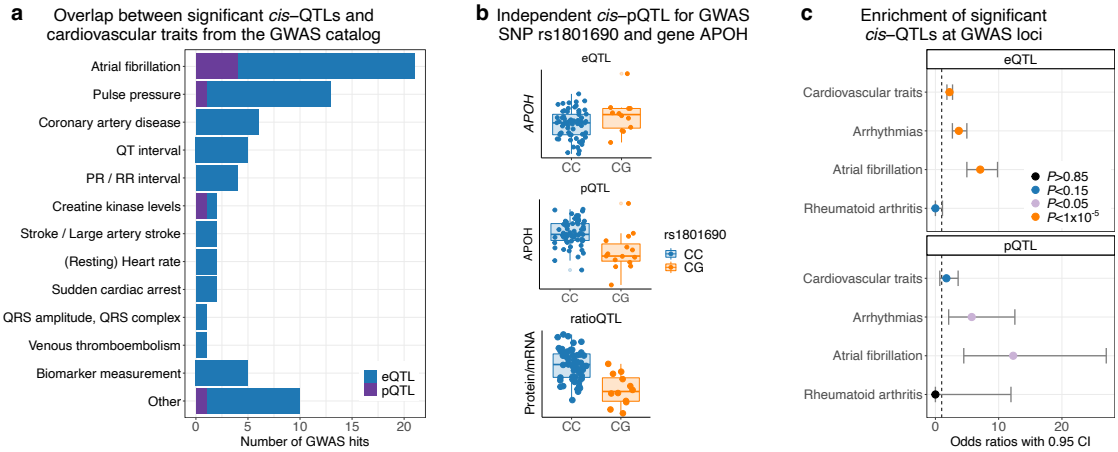


Figure 4.17: Overlap of *cis* QTL associations with GWAS hits annotated in the GWAS catalog.
a: Overview of significant *cis* eQTLs and pQTLs (FDR < 0.05) overlapping with GWAS hits for different disease traits.
b: Independent pQTL for GWAS hit creatine kinase levels. Shown are the non-significant *cis* eQTL as well as the significant *cis* pQTL and ratioQTL for the SNP rs1801690 and the gene APOH (FDR < 0.05). Statistics were derived based on two-sided T tests for $N = 75$ (eQTLs), $N = 75$ (pQTL) and $N = 66$ (ratioQTL) biologically independent samples. A FDR < 0.05 per omic based on the Benjamini-Hochberg procedure was applied to assess significance and to account for multiple comparisons. In the boxplots, the lower and upper hinges correspond to the first and third quartiles (the 25th and 75th percentiles). The median is denoted by the central line in the box. The upper/lower whisker extends from the hinge to the largest/smallest value no further than 1.5·IQR from the hinge.
c: For three different trait categories (cardiovascular traits, arrhythmias and atrial fibrillation) as well as rheumatoid arthritis as a negative control, the enrichment of GWAS hits at significant *cis* QTLs (FDR < 0.05) was evaluated. Enrichments were calculated based on Tables 4.10 using Fisher's exact test (two-sided). Odds ratios are presented with their 95% CI.
QTL, quantitative trait loci; GWAS, genome-wide association study; SNP, single-nucleotide polymorphism; eQTL, expression quantitative trait loci; pQTL, protein quantitative trait loci; ratioQTL, ratio quantitative trait loci; CI, confidence interval; IQR, interquartile range;
Figure and legend adapted from Assum et al. [2022a] `https://www.nature.com/articles/s41467-022-27953-1/figures/3`, licensed under CC BY 4.0 `https://creativecommons.org/licenses/by/4.0/`.

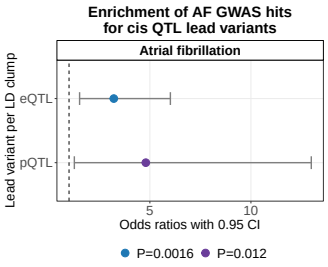


Figure 4.18: Enrichment of AF GWAS hits for *cis* QTLs (LD clumps).
AF GWAS annotations ($P < 5 \times 10^{-8}$) for *cis* eQTL/pQTL LD clump lead variants (FDR < 0.05) compared to a non-QTL background set. Enrichments were calculated using Fisher's exact test (two-sided). Odds ratios are presented with their 95% CI.
AF, atrial fibrillation; GWAS, genome-wide association study; QTL, quantitative trait loci; LD, linkage disequilibrium; eQTL, expression quantitative trait loci; pQTL, protein quantitative trait loci; CI, confidence interval;

We already evaluated that the lead variants of an *cis* eQTL or pQTL LD clump is more likely to be associated with AF based on the SNP-specific GWAS P value than a matched background SNP without a QTL. Similarly, we checked more general GWAS annotations starting with all SNPs and their high LD proxies ($R^2 > 0.8$) listed with a GWAS hit for any trait belonging to cardiovascular measurements and cardiovascular diseases. We

Table 4.8: Overlap of *cis* QTLs with GWAS loci for atrial fibrillation.
QTL hits that overlap with GWAS hits for atrial fibrillation in the GWAS catalog (or proxy with R^2<0.8). QTL, quantitative trait loci; GWAS, genome-wide association study; eQTL, expression quantitative trait loci; pQTL, protein quantitative trait loci;
Table and legend taken from the supplementary material Assum et al. [2022b] [Assum et al., 2022a] `https://doi.org/10.1038/s41467-022-27953-1`, licensed under CC BY 4.0 `https://creativecommons.org/licenses/by/4.0/`.

Region	QTL variant	Reported variant	Reported gene	QTL gene	QTL type	Pubmed-ID	First author
2p13.3	rs13028508	rs10165883	*SNRNP27*	*SNRNP27*	eQTL	29892015	Roselli C
10q22.2	rs10824026	rs10824026	*SYNPO2L*	*MYOZ1*	eQTL	29290336	Nielsen JB
3q22.3	rs6791611	rs1278493	*PPP2R3A*	*PCCB*	eQTL	30061737	Nielsen JB
5q31.2	rs9327807	rs2040862	*WNT8A*, *NPY6R*, M*YOT*, *FAM13B*	*FAM13B*	eQTL	30061737	Nielsen JB
2q33.1	rs159321	rs295114	*SPATS2L*	*SPATS2L*	eQTL	29892015	Roselli C
2q33.1	rs1347551	rs3820888	*SPATS2L*	*SPATS2L*	eQTL	30061737	Nielsen JB
3q26.33	rs2339798	rs4855074	*GNB4*	*GNB4*	eQTL	29892015	Roselli C
3q26.33	rs2339798	rs4855075	*GNB4*	*GNB4*	eQTL	29892015	Roselli C
1q32.1	rs951366	rs4951258	*NUCKS1*, *SLC41A1*	*NUCKS1*	eQTL	30061737	Nielsen JB
1q32.1	rs951366	rs4951261	*NUCKS1*	*NUCKS1*	eQTL	29892015	Roselli C
10q22.2	rs3740293	rs60212594	*SYNPO2L*	*MYOZ1*	eQTL	29892015	Roselli C
10q22.2	rs10824026	rs6480708	*SYNPO2L*	*MYOZ1*	eQTL	29892015	Roselli C
2p13.3	rs13028508	rs6546550	*ANXA4/GMCL1*	*SNRNP27*	eQTL	28416818	Christophersen IE
2p13.3	rs13028508	rs6546553	*GMCL1*	*SNRNP27*	eQTL	29892015	Roselli C
2p13.3	rs13028508	rs6747542	*GMCL1*, *ANXA4*	*SNRNP27*	eQTL	30061737	Nielsen JB
10q22.2	rs10824026	rs7394190	*SYNPO2L*	*MYOZ1*	eQTL	28416818	Christophersen IE
1q32.1	rs951366	rs951366	*NUCKS1*	*NUCKS1*	eQTL	29892015	Roselli C
10q22.2	rs10824026	rs10824026	*SYNPO2L*	*MYOZ1*	pQTL	28416818	Christophersen IE
10q22.2	rs3740293	rs60212594	*SYNPO2L*	*MYOZ1*	pQTL	29892015	Roselli C
10q22.2	rs12570126	rs6480708	*SYNPO2L*	*MYOZ1*	pQTL	29892015	Roselli C
10q22.2	rs12570126	rs7394190	*SYNPO2L*	*MYOZ1*	pQTL	28416818	Christophersen IE

Table 4.9: Enrichment of AF GWAS hits for *cis* QTLs LD clumps.
AF GWAS annotations ($P < 5 \times 10^{-8}$) for *cis* eQTL/pQTL LD clump lead variants (FDR < 0.05) compared to a non-QTL background set.
AF, atrial fibrillation; GWAS, genome-wide association study; QTL, quantitative trait loci; LD, linkage disequilibrium; eQTL, expression quantitative trait loci; pQTL, protein quantitative trait loci; CI, confidence interval;

***Cis* eQTL LD clump**	**Atrial fibrillation: GWAS $P < 5 \times 10^{-8}$**		***Cis* pQTL LD clump**	**Atrial fibrillation: GWAS $P < 5 \times 10^{-8}$**	
	False	**True**		**False**	**True**
Background SNP	109 187	313	**Background SNP**	16 015	85
Lead variant	1 085	10	**Lead variant**	157	4

summarized those as cardiovascular traits and then cross-referenced whether the SNP-gene pair had a significant *cis* QTL. Enrichments were then calculated using Fisher's exact test which showed a very strong enrichment of eQTLs ($P = 5.0 \times 10^{-13}$) and the trend of an enrichment for pQTLs (P = 0.13). Accordingly, we evaluated traits connected to arrhythmias, specifically atrial fibrillation, cardiac arrhythmia, sudden cardiac arrest, supraventricular ectopy, early cardiac repolarization measurement, heart rate, heart rate variability measurement, P wave duration, P wave terminal force measurement, PR interval, PR segment, QRS amplitude, QRS complex, QRS duration, QT interval, R wave amplitude, resting heart rate, RR interval, S wave amplitude and T wave amplitude. GWAS hits for those traits were strongly enriched for both *cis* eQTLs ($P = 5.7 \times 10^{-13}$)

and pQTLs (P = 7.6×10^{-4}), only exceeded by the enrichments of AF traits (Atrial fibrillation or QT interval) with P = 2.2×10^{-16} for eQTLs and P = 1.3×10^{-5} for pQTLs as visualized in Figure 4.17c. Furthermore, we evaluated GWAS hits for rheumatoid arthritis (RA), a disease non-specific to atrial tissue, as a negative control. No significant *cis* eQTLs or pQTLs overlapped with any of the RA loci. Cross tables for all the comparisons can be found in Table 4.10.

Table 4.10: Enrichment of GWAS hits at significant *cis* QTLs.
Cross tables show for each SNP tested in the *cis* eQTL or pQTL analyses, whether it had a significant eQTL/pQTL and whether it was annotated as a GWAS hit for the different trait categories.
QTL, quantitative trait loci; GWAS, genome-wide association study; eQTL, expression quantitative trait loci; pQTL, protein quantitative trait loci;

	Cardiovascular traits: GWAS hit		**Arrhythmias: GWAS hit**		**Atrial fibrillation: GWAS hit**		**Rheumatoid arthritis: GWAS hit**	
eQTL SNP	**False**	**True**	**False**	**True**	**False**	**True**	**False**	**True**
False	4 774 845	33 718	4 779 722	33 777	4 780 685	33 784	4 780 976	33 822
True	6 599	104	1 722	45	759	38	468	0

	Cardiovascular traits: GWAS hit		**Arrhythmias: GWAS hit**		**Atrial fibrillation: GWAS hit**		**Rheumatoid arthritis: GWAS hit**	
pQTL SNP	**False**	**True**	**False**	**True**	**False**	**True**	**False**	**True**
False	2 295 616	2 401	2 298 465	2 402	2 298 998	2 402	2 299 168	2 408
True	3 849	7	1 000	6	467	6	297	0

Due to the strong enrichment of AF-associated variants with *cis* QTL results, we further investigated specific AF loci. One of the strongest genetic associations was a locus near the SYPOL2L gene. As already reported before, the most likely causal gene is MYOZ1 which shows a strong *cis* eQTL and pQTL, where the QTL P values strongly correlate with the genetic association reported by Roselli et al. [2018]. Out of seven genes with overlaps of eQTLs and AF GWAS hits from the GWAS catalog, only MYOZ1 and PCCB were measured on protein level.

4.2 Discussion

4.2.1 Summary

We systematically mapped genome-wide *cis*-regulatory consequences of common variants on the expression of transcripts and proteins in human atrial tissue. Integrating genotype, mRNA and protein measurements enabled us to better understand potential functional mechanisms. Large differences were found between the effects on the different omic levels where the genetic variant would affect mRNA or protein exclusively. This results in severe restrictions, when considering the overlap of regulation on transcript and protein level.

4.2.2 Omic-specific regulation

A possible explanation for proteome-specific pQTLs are post-transcriptional regulatory elements which are affected by the corresponding variants. This was supported by the finding in our and other studies [Battle et al., 2015], showing that those SNPs were actually enriched in the coding sequence.
While our study was the first to integrate genomics, transcriptomics and proteomics data in human atrial tissue, similar missing co-regulation was already observed when comparing *cis*-pQTLs in human plasma to *cis*-eQTLs in GTEx tissues [Sun et al., 2018] as shown in Table 4.3. In addition, instead of using post-mortem tissues, our results were derived from tissue samples of living donors harvested while undergoing surgery. Post-mortem tissues potentially only show restricted or changed pathway regulation e.g. for metabolims.
Notably, we have observed slightly larger differences between omics, however, certain factors need to be considered when comparing to other studies, such as inherent differences between heterogeneous human tissue and cell-type specific lymphoblastoid cell lines [Battle et al., 2015], different stringency in significance cutoffs and applying multiple testing correction as well as different transcriptomics and proteomics measurement techniques.

4.2.3 Limitations

The benefits and the necessity of studying this highly specialized tissue were of course closely connected to limiting biological and technical factors.

Difficulties in accessibility resulted in challenges such as a restriction of the sample size. Furthermore, tissue samples are commonly very heterogeneous with respect to their cellular composition. Both factors majorly impact the statistical power of our QTL analysis and prohibited the application of common models for inferring causality of molecular regulation, such as Mendelian randomization [Lawlor et al., 2008, Smith and

Hemani, 2014].
To date, higher quality expression data can be acquired with RNA-sequencing technology instead of using micro-arrays.
Additionally to general tissue heterogeneity, muscle tissue of the human heart is characterized by very high contributions of mitochondrial and sarcomere proteins [Gramolini et al., 2007]. This results in a lack of detection of proteins which are only present in very low molecular numbers such as TFs and is therefore explained by a systematic failure to detect those proteins, not due to restrictions based on data quality.
Just like blood samples, also tissue specimens are a mixture of diverse cell-types. While estimating the different cell-type compositions of every tissue sample remains challenging, we did include a correction taking into account the amount of fibroblasts as one of the most important cell types next to cardiomyocytes. The utilized fibroblast-score was derived from gene expression values for a fibroblast-specific gene signature previously used and published by Heinig et al. [2017].

Finally, most of those restrictions also limited the amount of publicly available functional genomic annotations for our tissue type, including for example binding sites of TFs, miRNAs and RBPs.

4.2.4 Conclusion

In summary, compared to other studies [Battle et al., 2015, Sun et al., 2018, Hause et al., 2014] we investigated differences in *cis*-genetic regulation specific to human atrial tissue. The observation of large differences in transcript and protein expression as well as their regulatory mechanisms proves the need and advantages of multi-omic integration for a better understanding of molecular changes as consequences to common genetic variation.

5 Omnigenic effects and *trans*-regulatory networks in atrial fibrillation

In the following, we introduce the use of quantitative trait loci (QTL) and polygenic risk scores (PRS) to investigate consequences of genetic variation to transcript and protein expression and how this can establish a link between GWAS hits and transcription factor regulation in the human atria.

This chapter is based on and partly identical to the publication by Assum et al. [2022a] and also available as a preprint on *bioRxiv*[1,2]:
Tissue-specific multi-omics analysis of atrial fibrillation[3]
Ines Assum†, Julia Krause†, Markus O. Scheinhardt, Christian Müller, Elke Hammer, Christin S. Börschel, Uwe Völker, Lenard Conradi, Bastiaan Geelhoed, Tanja Zeller*, Renate B. Schnabel* and Matthias Heinig*, *Nature Communications* **13**, 441 (2022). Authors marked with † or * contributed equally to this work.
Code related to this project is available at `https://github.com/heiniglab/symatrial`[4] [Assum and Heinig, 2021].

We present a candidate selection approach which incorporates genotypes, transcript and protein expression in combination with public annotations, such as gene set and GWAS annotations, to prioritize SNPs and genes for *trans* QTL testing. By exploiting properties described for core genes, which are central, disease-associated genes in the context of the omnigenic model described in the introduction (section 1.2.3.1), we reduce the possible search space and therefore make *trans* analyses possible in settings where sample size is highly restricted.
The integration of new methods as well as our deeply phenotyped AFHRI-cohort enabled the analyses of genetic contributions of AF and elucidate the complex network of specific cases of transcription factor regulation.

[1] `https://www.biorxiv.org/content/10.1101/2020.04.06.021527v1`
[2] `https://www.biorxiv.org/content/10.1101/2020.04.06.021527v2`
[3] `https://doi.org/10.1038/s41467-022-27953-1`
[4] `https://doi.org/10.5281/zenodo.5094276`

5.1 Omnigenic effects and *trans*-regulatory networks in atrial fibrillation

GWAS studies and *cis* QTL analyses have been integrated to identify possible causal SNPs and genes underlying genetic causes of cardiovascular disease. While GWAS loci were enriched for significant *cis* eQTLs and pQTLs, for the majority of AF-associated loci, the molecular mechanisms remain unknown. A possible explanation are the more complex, polygenic and additive effects of genes not only acting on genes close to the variant of interest, but also on genes which are located at a greater distance.

5.1.1 Investigation of the omnigenic architecture of atrial fibrillation

Compared to the more than 100 identified AF loci [Roselli et al., 2018, Nielsen et al., 2018], Wang et al. [2019b] found 1 931 genes which were genetically influenced and associated with AF using a multi-omics approach leveraging GWAS, epigenome-wide association study (EWAS) and transcriptome-wide association study (TWAS) results. Furthermore, tissue-specific gene interaction plays a vital role in understanding disease relevant molecular networks.
Similarly, Choi et al. [2020] investigated the contribution of rare genetic variants by identifying genes with loss of function (LOF) mutations associated to AF and only identified one gene *TTN*. Only a small fraction (0.44 % of the corresponding cohort) were carriers and therefore, the TTN LOF mutations only explained 0.2 % of AF variability. Compared to that, considering the top 0.44 percentile of individuals with the highest polygenic risk score (PRS), the PRS explained 4.7 % of AF susceptibility [Choi et al., 2020].

Given the limited number of GWAS hits which could be explained by *cis* eQTLs and pQTLs 4.1.7, *trans* QTLs, where the variant is distant to the target gene, as well as more complex genetic or epigenetic interactions need to be investigated [van Ouwerkerk et al., 2019, Westra et al., 2013, Lemire et al., 2015]. This is in line with the estimates from the omnigenic model [Boyle et al., 2017], where *trans* associations contribute more than 70 % of disease heritability [Liu et al., 2019].
Even though the sum of all *trans* contributions is supposed to contribute to most of the heritability, effect sizes of each individual locus are often very small [Westra et al., 2013, Võsa et al., 2021]. However, important disease-relevant genes in the center of gene-regulatory networks can accumulate multiple *trans* effects by propagation within biological processes. Genes with extremely high impact on phenotypes often show less severe genetic variation due to evolutionary pressure but in turn, smaller changes in gene expression are often directly linked to disease [Boyle et al., 2017]. Therefore, core genes, i.e. central genes which accumulate *trans* genetic effects and are directly functionally connected to a phenotype, play a vital role in understanding gene regulation and disease mechanism.

A central hypothesis which we will use to identify putative AF core genes, is that in

general, genome-wide polygenic risk scores (PRSs) can act as proxies of cumulated *trans* associations [Võsa et al., 2021] and can be used for AF risk prediction [Kalsto et al., 2019, Khera et al., 2018]. Võsa et al. [2021] postulate that we can investigate *trans*-regulated genes by evaluating the correlation of PRSs with transcript or protein expression. Due to the already small effect sizes, the additive representation of genetic effects in the PRS as well as other factors influencing gene expression, correlations are expected to be rather small.

5.1.1.1 Genome-wide polygenic scores for atrial fibrillation

As a first step, we evaluated different approaches of computing the PRS and their association with AF. With risk scores being a continuous measure, the distribution of risk score values across healthy and diseased individuals is more important than the actual risk score values. Therefore, having a representative background distribution for non-diseased individuals is of great importance. Taking this into account, the PRS published by Khera et al. [2018] was computed not only on the AFHRI-B cohort, but it was calculated together with unrelated individuals from the 1000 Genomes project. The distribution of the risk score values was highly similar in the two cohorts (Figure 5.1a), with slightly higher values in the AFHRI-B cohort caused by the enrichment of disease cases in this cohort (Figure 5.1b). To include these shifts also when only considering AFHRI-B patients, we used percentiles of the risk score, calculated across all individuals for the final analysis as shown in the lower panel of (Figure 5.1b).

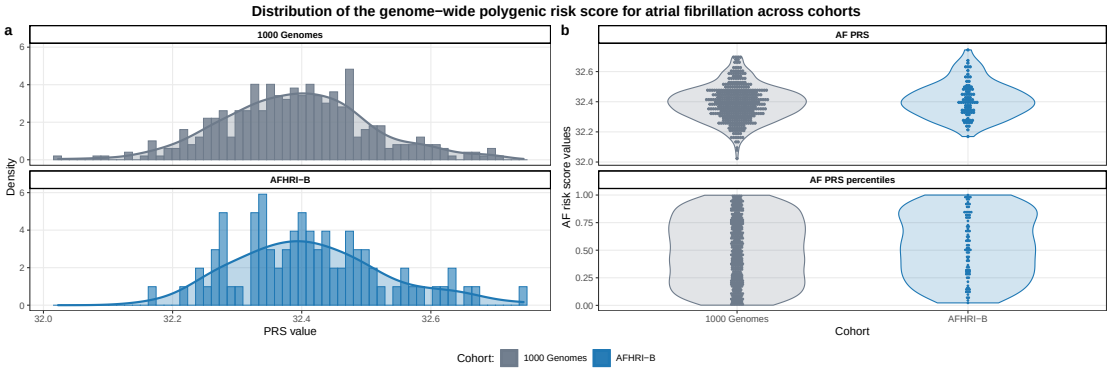


Figure 5.1: PRS distribution across different cohorts.
a: Similar PRS value distribution for the 1000 Genomes and AFHRI-B cohort.
b: Comparison of raw versus percentile-transformed PRS values across the 1000 Genomes and AFHRI cohort.
PRS, genome-wide polygenic risk score;

The PRS includes 6 730 540 SNPs, of which all but one were measured in the 1000 Genomes cohort with an overall genotyping rate of 99.9 %. The plink function score computes risk score values, by summing over the number of risk alleles multiplied by the SNP weight for each variant. Missing individual SNP contributions to the score are imputed by the expectation based on SNP weight and MAF inferred from all non-missing genotypes. As to be expected by the chip design and corresponding

imputation quality, the coverage was much lower for the AFHRI-B cohort, where 66.5 % of SNPs were measured in at least 80 % of the individuals.

Rare variants often have larger effect sizes in GWAS or, in this case, SNP weights in the risk score as they correlate with the original GWAS signal. However, rare variants often also show a lower genotyping rate due to the smaller allele frequency. Therefore, a higher number of 57.5 % SNPs with weights ranging in the top quartile of the score had a genotyping rate which was below the 80 % threshold in the AFHRI-B cohort. Due to their rare occurrence, rare variants contribute relatively little to the overall heritability compared to common genetic variation, as also stated by Weng et al. [2017], where 20.4 % out of a total SNP heritability of 22.1 % (MAF $\geq$ 0.01) were attributed to common variants (MAF $\geq$ 0.05).

Since almost all variants were measured for the 1000 Genomes cohort, we were able to evaluate any changes between the full PRS and a risk score restricted to a subset of SNPs.

When including only non-missing variants (genotyping rate per SNP > 80 %) in our smaller cohort, there were only subtle differences, resulting in a Pearson correlation of 0.96 (P = 9.7×10^{-226}) (Figure 5.2a).
Similarly, we can quantify the information contained in the missing SNPs by considering the proportion of variance for the total score, which is explained by the subset. While all higher weight SNPs which make up 25 % of variants account for 79 % of variance (R^2 in a linear regression model), all non-missing variants in the AFHRI-B cohort covered 92 % (R^2) of the total risk score variance, even though more than half of the higher weight SNPs were missing.

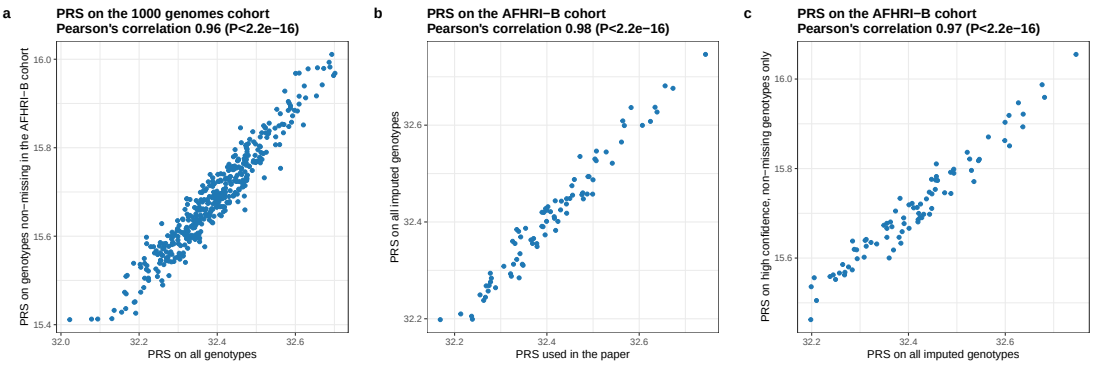


Figure 5.2: Comparison of polygenic risk score variation.
Scores computed on slightly varying SNP sets for lower confidence or missing genotypes show consistently high positive correlation.
a: For the 1000 Genomes cohort, a score with all SNPs was compared to a score restricted to SNPs with a genotyping rate > 80 % in the AFHRI-B cohort.
b: For the AFHRI-B cohort, a score with all SNPs which passed quality control was compared to a score computed on all available dosages independent of confidence in the imputation call.
c: For the AFHRI-B cohort, a score computed on all available SNP dosages was compared to a score evaluating only SNPs with high confidence imputation calls and a per-SNP genotyping rate > 80 %.
SNP, single-nucleotide polymorphism;

We further assessed the influence on the score, when using unfiltered, including low confidence dosages rather than missing genotype values, in order to derive the score for our AFHRI-B cohort. Here, an even higher Pearson correlation of 0.98 ($P = 3.2 \times 10^{-58}$) was observed (Figure 5.2b).

Finally, we evaluated the correlation between the dosage derived risk score values and the score computed only on the high confidence SNPs with also high genotyping rate (> 80 % per SNP) which showed again a Pearson correlation of 0.97 with a P value of 3.9×10^{-54}.

If we want to utilize the polygenic risk score as a proxy for accumulated *trans*-genetic disease susceptibility, we need to evaluate its association with AF. Figure 5.3a-c shows the discrimination between AF cases and controls for the three scores described above. While the restriction to non-missing variants shows a slight increase in disease discrimination, the use of low confidence imputation dosages decreased risk score performance. The best performing score, however, was achieved by taking the percentile-transformed values of the original score with a T value of -2.0 and P value of 0.026 (Figure 5.4a). Additionally, we observed a strong decrease of performance when only considering the top GWAS SNPs as incorporated in the 97 SNP score published by Kloosterman et al. [2020] which was recovered from the PGS catalog[5].

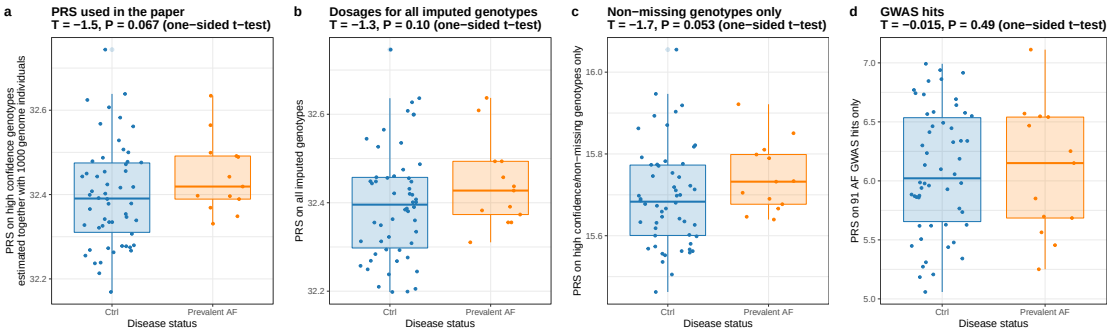


Figure 5.3: Comparison of polygenic risk score performance.
Performance of different variants of the polygenic risk score to discriminate between AF cases and controls.
a: The original genome-wide polygenic score, estimated together with unrelated individuals of the 1000 Genomes cohort and evaluated for prevalent AF in the AFHRI-B cohort.
b: Score performance with respect to AF when using all available SNP dosages independent of confidence in the imputation call for risk score derivation.
c: Better separation of prevalent AF cases and controls in the AFHRI-cohort, when restricting evaluated SNPs to those with a per-SNP genotyping rate > 80 %.
d: Limited differentiation between AF cases and controls for a score including only 97 top GWAS variants.
AF, atrial fibrillation; GWAS, genome-wide association study; SNP, single-nucleotide polymorphism;

[5] `http://www.pgscatalog.org/`

5.1.1.2 Genetic and non-genetic contributions to atrial fibrillation risk

We established the percentile-transformed PRS as most informative in our cohort (Figure 5.3, Figure 5.4a). To underline the importance of the genetic risk to AF susceptibility, we compared the contribution of the PRS to other classical risk factors such as age, sex, BMI, blood pressure and CRP or NTproBNP measurements. Results from the full logistic regression model are shown in Table 5.1 and visualized in Figure 5.4b. Of all considered factors, the PRS has the largest estimated β value and the second largest R^2 contribution (McFadden's pseudo R^2, age: 0.091, AF PRS percentiles: 0.034). Therefore, we established the genetic risk represented by our score as a vital predictor for disease.

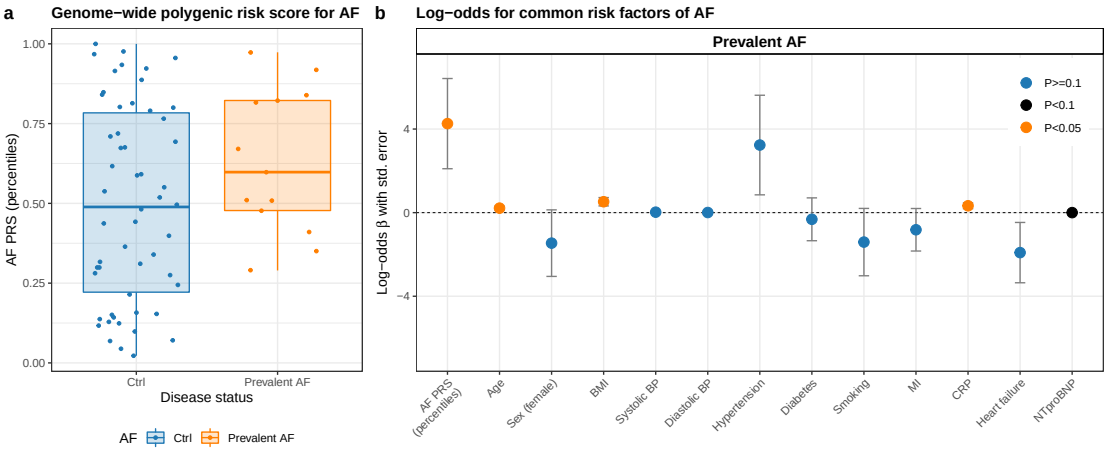


Figure 5.4: Genome-wide polygenic score adds relevant information in classifying atrial fibrillation disease status.
a: Percentiles of the atrial fibrillation polygenic risk score by disease status (T = -1.8, P = 0.043, one-sided t-test, N = 67).
b: Logistic regression results for common risk factors of AF. Significant and comparably strong effect for the PRS variable (β = 4.3, P = 0.048, two-sided z-test, N = 64, df = 51). Data are presented as log-odds ratio +/- standard error.
In the boxplots, the lower and upper hinges correspond to the first and third quartiles (the 25th and 75th percentiles). The median is denoted by the central line in the box. The upper/lower whisker extends from the hinge to the largest/smallest value no further than 1.5·IQR from the hinge.
AF, atrial fibrillation; PRS, genome-wide polygenic score; BMI, body mass index; BP, blood pressure; MI, myocardial infarction; CRP, C-reactive protein; NT-proBNP, N-terminal prohormone of brain natriuretic peptide; IQR, interquartile range;
Figure and legend adapted from the supplementary material Assum et al. [2022b] [Assum et al., 2022a] `https://doi.org/10.1038/s41467-022-27953-1`.

5.1.2 Investigation of AF core genes

In this part, we want to present a pathway enrichment approach to make *trans* QTL analyses feasible in clinical cohorts with a limited sample size. We efficiently integrate public annotations such as GWAS hits, PRS information and prior biological knowledge in order to narrow down the search space of *trans* QTLs.

Table 5.1: Logistic regression results for genetic and non-genetic contributions in AF.
Logistic regression to assess the contribution of different AF risk factors. Besides age, the polygenic risk score percentiles are the most important feature based on McFadden's pseudo R^2 with a large effect size. McFadden's pseudo R^2 for the whole model was 0.44.
AF, atrial fibrillation; PRS, genome-wide polygenic score; BMI, body mass index; BP, blood pressure; MI, myocardial infarction; CRP, C-reactive protein; NT-proBNP, N-terminal prohormone of brain natriuretic peptide;

	β	Std. error	Z value	P value	McFadden's pseudo R^2
Intercept	-40	14	-2.78	0.0052	
AF PRS (percentiles)	4.3	2.2	2.0	0.048	0.034
Age	0.22	0.095	2.3	0.023	0.091
Sex (female)	-1.5	1.6	-0.92	0.36	
BMI	0.52	0.20	2.6	0.0092	0.029
Diabetes	-0.32	1.0	-0.31	0.76	
Systolic BP	0.022	0.044	0.51	0.61	
Diastolic BP	0.0058	0.067	0.087	0.93	
Hypertension	3.2	2.4	1.4	0.17	
MI	-0.82	1.0	-0.81	0.42	
Smoking	-1.4	1.6	-0.87	0.38	
Heart failure	-1.9	1.4	-1.3	0.19	
CRP	0.33	0.14	2.4	0.016	0.023
NTproBNP	8.7×10^{-4}	4.5×10^{-4}	1.9	0.053	

One key hypothesis of the omnigenic model introduced by Liu et al. [2019] is the existence of core genes. Very important properties which we want to leverage in this context is their central role in biological networks, the genetic association as well as the direct link to the phenotype. Moreover, due to those properties the investigation of core genes is of great interest to better understand molecular disease mechanisms. Therefore, we applied the following steps to identify putative AF core genes visualized in Figure 5.5:

1. We were interested in genes which accumulate *trans*-genetic effects. Taking the PRS percentiles as a proxy for overall accumulated genetic effects, for each gene we calculated the association between risk score values and transcript (expression quantitative trait score, eQTS) as well as protein expression (protein quantitative trait score, pQTS) as previously proposed by Võsa et al. [2021]. To ensure rankings based on *trans* effects only, we additionally included the lead LD clump variant for each locus with a significant *cis* eQTL or pQTL for the corresponding gene in the linear regression model to compute the eQTSs/pQTSs.
2. Since core genes should be in the center of biological networks, we used GO biological process annotations as a simplified version to represent the biological networks propagating *trans* effects and identify groups of genes with shared biological functions. We therefore proceeded with applying gene set enrichment analysis (GSEA) [Subramanian et al., 2005, Korotkevich et al., 2019] on the eQTS and pQTS rankings to integrate the corresponding prior knowledge. Candidate genes for *trans* QTL testing were identified from the top genes which came up in the leading edge, i.e. driving the enrichment, of multiple terms.

3. Finally, we were interested in both possible *trans* regulation as well as potential core genes. Hence, we evaluated any candidate transcript or protein for association with the top independent GWAS SNPs and defined putative core genes as all genes with a significant *trans* eQTL or pQTL. Additionally, disease association was supported by analyzing AF differential protein abundance.

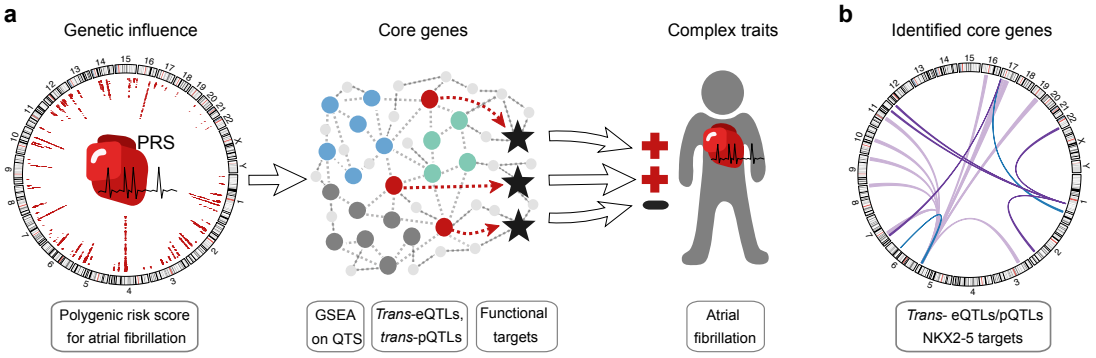


Figure 5.5: Graphical illustration of the strategy for *trans* QTL analysis to identify AF-relevant genes.
a: Overview: Based on patient-specific PRS values for AF correlated with transcript and protein expression, we performed GSEA to preselect genes for *trans* eQTL and pQTL analyses from the leading edge of enriched pathways. Core genes were identified as significant *trans* eQTLs or *trans* pQTLs. We further assessed their functional targets to investigate the genotype-phenotype relationship in the context of AF. b: Identified core genes as *trans* eQTLs (blue), *trans* pQTLs (purple) (FDR < 0.2) and functional NKX2-5 targets (light purple).
PRS, genome-wide polygenic risk score; GSEA, gene set enrichment analysis; QTS, quantitative trait score; eQTL, expression quantitative trait loci; pQTL, protein quantitative trait loci; FDR, false discovery rate; AF, atrial fibrillation; blue, green or gray dots = core gene candidates; red dots = core genes with *trans* eQTL/pQTL; stars = functional targets of core genes;
Circular plots were created with the R package circlize [Gu et al., 2014].
Figure and legend taken from Assum et al. [2022a] `https://www.nature.com/articles/s41467-022-27953-1/figures/4`, licensed under CC BY 4.0 `https://creativecommons.org/licenses/by/4.0/`.

5.1.2.1 Evaluation of the accumulation of genetic effects using polygenic risk scores

Each of the 26 376 transcripts and 1 469 proteins was associated with the PRS percentiles while taking into account any significant, independent *cis* QTL for that gene by the lead SNP of the corresponding LD clump based on the results from chapter 4. As to be expected due to the convolution, small effect sizes and limited number of samples (N = 74 for eQTS, N = 73 for pQTS), there were no significant associations when correcting for multiple testing. Genes were further ranked by their T value, with the 36 top eQTS genes (P < 0.002) and 24 top pQTS genes (P < 0.05) listed in Table 5.2 and Table 5.3.

Table 5.2: Top eQTS genes.
Linear regression results of transcript expression associating with the polygenic risk score for AF. Two-sided t-tests were derived from a linear model with covariates and *cis* SNPs. As T values were further used for ranking genes and not to assess statistical significance, no adjustments for multiple comparisons were applied.
eQTS, expression quantitative trait score; AF, atrial fibrillation; SNP, single-nucleotide polymorphism; N, sample size; df, degrees of freedom;

Gene	β	T value	P value	N	Df
LOC105377927	0.311	4.10	0.000116	74	65
SEMA3F	-0.254	-3.95	0.000196	74	65
LOC105371436	-0.306	-3.89	0.000240	74	65
STEAP3	-0.343	-3.80	0.000326	74	65
FBXL7	0.201	3.69	0.000465	74	65
VTN	-0.415	-3.68	0.000476	74	64
WDR82	0.159	3.66	0.000501	74	65
MIR544A	0.577	3.62	0.000570	74	65
NR1I2	0.266	3.59	0.000639	74	65
SNORD91A	0.384	3.58	0.000664	74	65
PLCL1	0.383	3.57	0.000668	74	65
LOC105374629	0.218	3.53	0.000758	74	65
LOC729080	0.282	3.53	0.000780	74	65
OIP5-AS1	0.225	3.52	0.000805	74	65
TXNDC12-AS1	-0.330	-3.50	0.000846	74	65
CRELD2	-0.189	-3.48	0.000896	74	65
CEP85L	0.226	3.45	0.00100	74	65
FKBP11	-0.330	-3.41	0.00112	74	65
SLC22A24	-0.268	-3.39	0.00119	74	65
VAT1L	-0.520	-3.39	0.00119	74	65
TFIP11	0.242	3.37	0.00128	74	65
LOC105377151	0.283	3.36	0.00131	74	65
TRIM43	0.239	3.36	0.00132	74	65
MIR376A2	0.481	3.35	0.00133	74	65
UPK1A-AS1	0.277	3.35	0.00134	74	65
WAC-AS1	0.226	3.33	0.00142	74	65
LOC101928257	0.212	3.33	0.00145	74	65
LOC105374674	0.230	3.31	0.00152	74	65
LINCR-0002	-0.196	-3.31	0.00152	74	65
FAM78A	0.219	3.31	0.00152	74	64
CFAP57	-0.232	-3.30	0.00157	74	65
TMPRSS9	-0.221	-3.29	0.00163	74	65
KLKB1	0.616	3.27	0.00171	74	65
RUNDC3A-AS1	0.255	3.27	0.00173	74	65
SGK2	-0.275	-3.24	0.00191	74	65
C1orf56	0.273	3.23	0.00195	74	65

Table 5.3: Top pQTS genes.
Linear regression results of protein expression associating with the polygenic risk score for AF. Two-sided t-tests were derived from a linear model with covariates and *cis* SNPs. As T values were further used for ranking genes and not to assess statistical significance, no adjustment for multiple comparisons was applied.
pQTS, protein quantitative trait score; AF, atrial fibrillation; SNP, single-nucleotide polymorphism; N, sample size; df, degrees of freedom;
Table and legend taken from the supplementary material Assum et al. [2022b] [Assum et al., 2022a] `https://doi.org/10.1038/s41467-022-27953-1`.

Gene	β	T value	P value	N	Df
NAMPT	-0.158	-3.07	0.00315	73	63
MARK1	-0.176	-2.58	0.0121	73	64
GYG1	0.0941	2.57	0.0125	73	64
AOC3	0.157	2.51	0.0146	73	64
PGM2	-0.184	-2.46	0.0167	73	64
CD36	0.0986	2.45	0.0171	73	64
CDIPT	0.126	2.40	0.0195	73	64
RPL37AP8	0.214	2.35	0.0219	73	64
RPL37L	0.214	2.35	0.0219	73	64
RAB2A	0.113	2.32	0.0235	73	64
RPS4X	0.102	2.27	0.0263	73	64
RPL39	-0.148	-2.23	0.0292	73	64
MYO1C	0.0818	2.20	0.0311	73	64
PCBP2	-0.0961	-2.20	0.0314	73	64
ERP44	-0.0954	-2.17	0.0337	73	64
BLVRA	-0.110	-2.15	0.0350	73	64
NIPSNAP3A	-0.0955	-2.13	0.0376	73	61
XRCC6	-0.0651	-2.06	0.0431	73	64
CLIC4	-0.0771	-2.04	0.0460	73	64
BCL11B	0.139	2.02	0.0479	73	64
TALDO1	-0.0884	-2.01	0.0488	73	64
MAP1LC3B	0.121	2.00	0.0495	73	64
MAP1LC3B2	0.121	2.00	0.0495	73	64
RPL13A	-0.170	-2.00	0.0498	73	64

5.1.2.2 PRS-based pathway enrichment approach for candidate gene selection

Gene set enrichment analysis was utilized to identify genes which share molecular function and, at the same time, accumulate *trans* effects. Due to the genetic association with AF, resulting enriched processes should also be connected to the phenotype AF. Indeed, when using the GO biological processes which are a very general representation of molecular networks and not linked to any specific diseases a priori, we identify processes highly relevant to the trait studied.

Using the eQTS ranking evaluating both positive and negative enrichments, 81 GO biological processes listed in Table 5.4 were enriched at a Benjamini-Hochberg FDR adjusted P value of 0.05. Many of them were connected to heart muscle or energy metabolism, such as *Generation of precursor metabolites and energy* (GO:0006091), *Regulation of cardiac muscle contraction* (GO:0055117), *Cardiac muscle tissue develop-*

ment (GO:0048738) and, especially relevant for an arrhythmia, *Regulation of heart rate* (GO:0002027) (see also Figure 5.6a-d).

Furthermore, arrhythmias are connected to pathways involving calcium homeostasis which we find also represented in the three significantly enriched terms *Regulation of release of sequestered calcium ion into cytosol by sarcoplasmic reticulum* (GO:0010880), *Regulation of cardiac muscle contraction by regulation of the release of sequestered calcium ion* (GO:0010881) and *Regulation of cardiac muscle contraction by calcium ion signaling* (GO:0010882) as shown in Figure 5.6e-g.

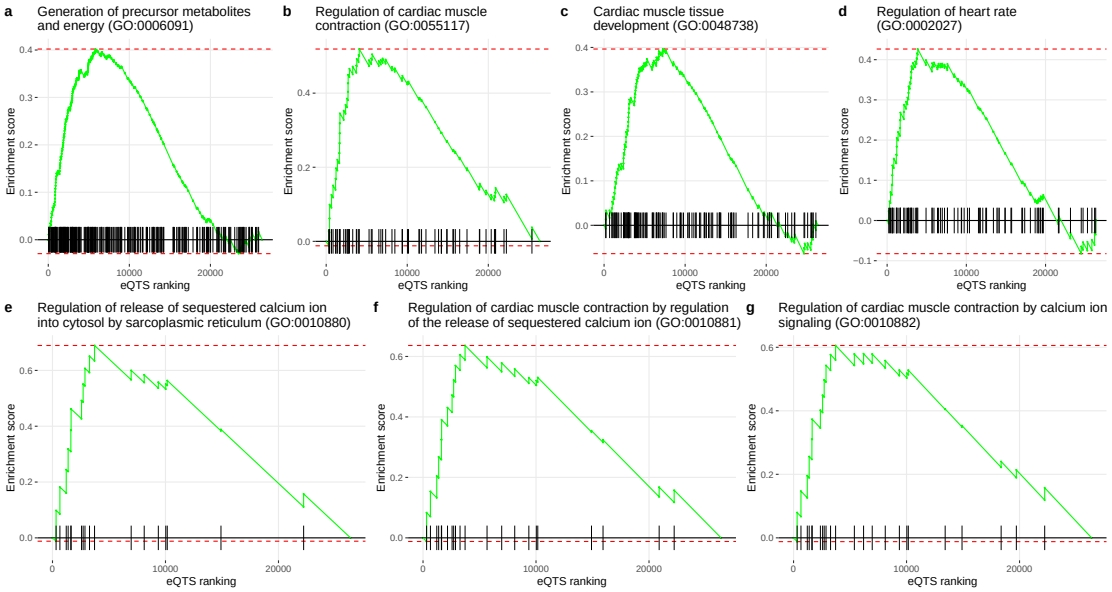


Figure 5.6: eQTS gene set enrichment results.
a-d: Visualization of the enrichment (FDR < 0.05) of selected GO biological processes connected to heart muscle and energy metabolism.
e-g: Enrichment of three gene sets (FDR < 0.05) in the context of calcium homeostasis.
eQTS, expression quantitative trait score; GO, gene ontology; FDR, false discovery rate;

On protein level, one GO biological process was significantly enriched (FDR < 0.05, Table 5.5). The corresponding term *Small molecule metabolic process* (GO:0044281) was again connected to metabolism.

The GO biological processes annotations are organized in a hierarchical fashion with more general, large gene sets consisting of smaller, more specialized child terms. Genes driving the enrichment of a specific term will also drive the enrichment of higher-level parent terms. Therefore, we selected genes for *trans* eQTL testing only if they were contained in the leading edge of multiple enriched gene sets.

Table 5.4: Enriched GO terms for the eQTS GSEA.
GSEA results on eQTS T value rankings. Shown are all GO terms enriched at FDR < 0.05 (Benjamini-Hochberg procedure to account for multiple comparisons of the two-sided, permutation-derived P values). Size refers to the number of genes in the gene set after removing those not evaluated for eQTS. Full summary statistics on all tables are available under `https://doi.org/10.5281/zenodo.5080229`.
GO, Gene Ontology; eQTS, expression quantitative trait score; GSEA, gene set enrichment analysis; NES, normalized enrichment score; FDR, false discovery rate;
Table and legend taken from the supplementary material Assum et al. [2022b] [Assum et al., 2022a] `https://doi.org/10.1038/s41467-022-27953-1`.

GO ID	GO term	NES	P value	FDR	Size
GO:0006091	Generation of precursor metabolites and energy	2.03	1.61×10^{-5}	0.00397	268
GO:0003012	Muscle system process	1.85	1.61×10^{-5}	0.00397	264
GO:0015980	Energy derivation by oxidation of organic compounds	2.21	1.65×10^{-5}	0.00397	197
GO:0045333	Cellular respiration	2.40	1.69×10^{-5}	0.00397	129
GO:0022900	Electron transport chain	2.17	1.74×10^{-5}	0.00397	86
GO:0003015	Heart process	2.17	1.75×10^{-5}	0.00397	80
GO:0006119	Oxidative phosphorylation	2.35	1.76×10^{-5}	0.00397	75
GO:0006942	Regulation of striated muscle contraction	2.13	1.76×10^{-5}	0.00397	74
GO:0055117	Regulation of cardiac muscle contraction	2.11	1.77×10^{-5}	0.00397	63
GO:0033108	Mitochondrial respiratory chain complex assembly	2.26	1.78×10^{-5}	0.00397	57
GO:0097031	Mitochondrial respiratory chain complex I biogenesis	2.38	1.80×10^{-5}	0.00397	46
GO:0009060	Aerobic respiration	2.32	1.80×10^{-5}	0.00397	48
GO:0002455	Humoral immune response mediated by circulating immunoglobulin	-2.22	2.25×10^{-5}	0.00397	46
GO:0006956	Complement activation	-2.27	2.27×10^{-5}	0.00397	52
GO:0019724	B cell mediated immunity	-2.02	2.32×10^{-5}	0.00397	76
GO:0072599	Establishment of protein localization to endoplasmic reticulum	-2.29	2.36×10^{-5}	0.00397	90
GO:0030216	Keratinocyte differentiation	-1.99	2.37×10^{-5}	0.00397	97
GO:0070972	Protein localization to endoplasmic reticulum	-2.07	2.40×10^{-5}	0.00397	109
GO:0006413	Translational initiation	-1.93	2.44×10^{-5}	0.00397	127
GO:0006612	Protein targeting to membrane	-2.03	2.46×10^{-5}	0.00397	141
GO:0006959	Humoral immune response	-1.77	2.48×10^{-5}	0.00397	153
GO:0008544	Epidermis development	-1.69	2.60×10^{-5}	0.00398	244
GO:1903034	Regulation of response to wounding	-1.57	2.80×10^{-5}	0.00410	386
GO:0090257	Regulation of muscle system process	1.79	3.31×10^{-5}	0.00465	184
GO:2000257	Regulation of protein activation cascade	-2.16	4.43×10^{-5}	0.00537	33
GO:0031424	Keratinization	-2.05	4.51×10^{-5}	0.00537	48
GO:0002920	Regulation of humoral immune response	-2.07	4.51×10^{-5}	0.00537	47
GO:0072376	Protein activation cascade	-1.95	4.63×10^{-5}	0.00537	74
GO:0072521	Purine-containing compound metabolic process	1.59	4.70×10^{-5}	0.00537	356
GO:0006936	Muscle contraction	1.79	4.91×10^{-5}	0.00537	217
GO:0009141	Nucleoside triphosphate metabolic process	1.79	4.95×10^{-5}	0.00537	201
GO:0072350	Tricarboxylic acid metabolic process	2.11	5.43×10^{-5}	0.00571	37
GO:0006941	Striated muscle contraction	1.93	6.95×10^{-5}	0.00706	90
GO:1903115	Regulation of actin filament-based movement	2.08	7.28×10^{-5}	0.00706	32
GO:0009913	Epidermal cell differentiation	-1.82	7.34×10^{-5}	0.00706	137
GO:0006937	Regulation of muscle contraction	1.84	8.45×10^{-5}	0.00790	139
GO:0048738	Cardiac muscle tissue development	1.82	1.02×10^{-4}	0.00924	129
GO:0055006	Cardiac cell development	1.99	1.08×10^{-4}	0.00955	47
GO:0010881	Regulation of cardiac muscle contraction by regulation of the release of sequestered calcium ion	2.08	1.11×10^{-4}	0.00962	18
GO:0000184	Nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	-1.80	1.19×10^{-4}	0.0100	103
GO:0050727	Regulation of inflammatory response	-1.59	1.33×10^{-4}	0.0109	274
GO:0086004	Regulation of cardiac muscle cell contraction	2.01	1.47×10^{-4}	0.0117	27

GO ID	GO term	NES	P value	FDR	Size
GO:0002443	Leukocyte mediated immunity	-1.69	1.49×10^{-4}	0.0117	160
GO:0090150	Establishment of protein localization to membrane	-1.60	1.56×10^{-4}	0.0119	241
GO:0000209	Protein polyubiquitination	1.65	1.63×10^{-4}	0.0122	229
GO:0055086	Nucleobase-containing small molecule metabolic process	1.47	1.83×10^{-4}	0.0134	476
GO:0042787	Protein ubiquitination involved in ubiquitin-dependent protein catabolic process	1.75	2.20×10^{-4}	0.0155	126
GO:0010882	Regulation of cardiac muscle contraction by calcium ion signaling	2.03	2.21×10^{-4}	0.0155	22
GO:0098901	Regulation of cardiac muscle cell action potential	2.01	2.41×10^{-4}	0.0165	19
GO:1901657	Glycosyl compound metabolic process	1.56	2.53×10^{-4}	0.0170	327
GO:0002673	Regulation of acute inflammatory response	-1.84	2.78×10^{-4}	0.0180	69
GO:0009123	Nucleoside monophosphate metabolic process	1.63	2.78×10^{-4}	0.0180	219
GO:0030855	Epithelial cell differentiation	-1.43	2.89×10^{-4}	0.0184	469
GO:0008016	Regulation of heart contraction	1.61	3.29×10^{-4}	0.0201	205
GO:0003013	Circulatory system process	1.53	3.31×10^{-4}	0.0201	343
GO:0070252	Actin-mediated cell contraction	1.86	3.35×10^{-4}	0.0201	70
GO:0030048	Actin filament-based movement	1.78	3.65×10^{-4}	0.0213	89
GO:0010880	Regulation of release of sequestered calcium ion into cytosol by sarcoplasmic reticulum	1.97	3.68×10^{-4}	0.0213	24
GO:0002027	Regulation of heart rate	1.82	4.37×10^{-4}	0.0249	82
GO:0060306	Regulation of membrane repolarization	1.94	4.75×10^{-4}	0.0267	28
GO:0040029	Regulation of gene expression, epigenetic	1.61	5.30×10^{-4}	0.0292	193
GO:0043588	Skin development	-1.57	6.11×10^{-4}	0.0332	202
GO:0002064	Epithelial cell development	-1.58	6.54×10^{-4}	0.0349	177
GO:0044033	Multi-organism metabolic process	-1.66	6.80×10^{-4}	0.0358	121
GO:0045165	Cell fate commitment	-1.54	6.95×10^{-4}	0.036	218
GO:0061448	Connective tissue development	-1.57	7.31×10^{-4}	0.0367	183
GO:0060968	Regulation of gene silencing	1.85	7.37×10^{-4}	0.0367	47
GO:0002063	Chondrocyte development	-1.99	7.42×10^{-4}	0.0367	21
GO:0033561	Regulation of water loss via skin	-2.00	7.80×10^{-4}	0.0381	18
GO:0002066	Columnar/cuboidal epithelial cell development	-1.86	8.11×10^{-4}	0.0385	45
GO:0009896	Positive regulation of catabolic process	1.46	8.12×10^{-4}	0.0385	372
GO:0060537	Muscle tissue development	1.52	8.42×10^{-4}	0.0388	252
GO:0061136	Regulation of proteasomal protein catabolic process	1.60	8.50×10^{-4}	0.0388	167
GO:0018149	Peptide cross-linking	-1.82	8.63×10^{-4}	0.0388	54
GO:0061035	Regulation of cartilage development	-1.82	8.65×10^{-4}	0.0388	57
GO:0099623	Regulation of cardiac muscle cell membrane repolarization cytosol by sarcoplasmic reticulum	1.92	9.64×10^{-4}	0.0427	19
GO:0061061	Muscle structure development	1.44	9.92×10^{-4}	0.0429	399
GO:1903522	Regulation of blood circulation	1.50	9.95×10^{-4}	0.0429	273
GO:0045669	Positive regulation of osteoblast differentiation	-1.80	1.02×10^{-3}	0.0437	56
GO:0050818	Regulation of coagulation	-1.70	1.05×10^{-3}	0.0439	85
GO:0014013	Regulation of gliogenesis	-1.71	1.06×10^{-3}	0.0439	87

Table 5.5: Enriched GO term for the pQTS GSEA.
GSEA results on pQTS T value rankings. Shown is the GO term enriched at FDR < 0.05 (Benjamini-Hochberg procedure to account for multiple comparisons of the two-sided, permutation-derived P values). Size refers to the number of genes in the gene set after removing those not evaluated for pQTS.
GO, Gene Ontology; pQTS, protein quantitative trait score; GSEA, gene set enrichment analysis; NES, normalized enrichment score; FDR, false discovery rate;
Table and legend taken from the supplementary material Assum et al. [2022b] [Assum et al., 2022a] `https://doi.org/10.1038/s41467-022-27953-1`.

GO ID	GO term	NES	P value	FDR	Size
GO:0044281	Small molecule metabolic process	-1.68	1.34×10^{-5}	0.0270	332

To assess the number of genes to be considered in the *trans* analysis together with 108 SNPs representing independent AF GWAS loci, Matthias Heinig performed a power analysis. Based on the sample size of 74 individuals, 108 SNPs to be tested, a Bonferroni-adjusted P value of 0.05 and the strongest *trans* eQTL in blood with an effect size of 21.8 % (R^2) [van der Wijst et al., 2020], we determined that no more than 23 genes could be tested with a power of 50 %.

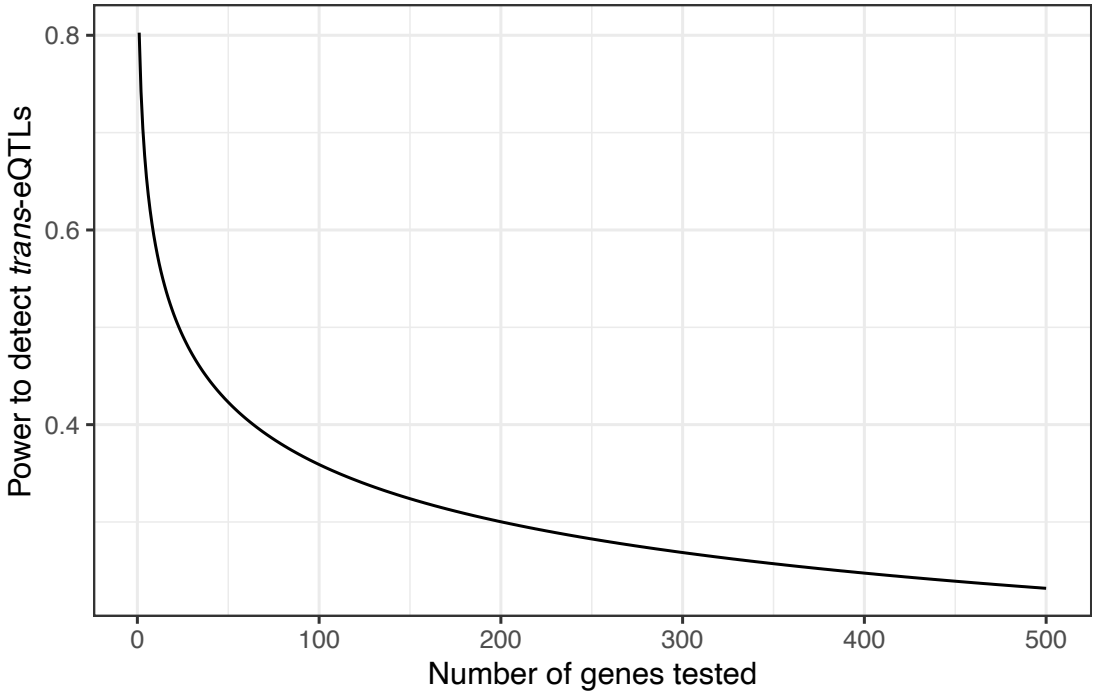


Figure 5.7: *Trans* eQTL power analysis.
Power analysis for the strongest *trans* eQTLs (effect size 21.8 %), considering 74 samples, 108 SNPs, $\alpha < 0.05$ and Bonferroni correction. 23 genes correspond to 50% power.
eQTL, expression quantitative trait loci; SNP, single-nucleotide polymorphism;
Figure and legend taken from the supplementary material Assum et al. [2022b] [Assum et al., 2022a] `https://doi.org/10.1038/s41467-022-27953-1`.

We therefore selected only the 23 genes appearing most often of all the 1 261 leading edge genes for *trans* eQTL testing, in this case all genes from the leading edges of at least 14 different significantly enriched gene sets which are listed in Table 5.6. For the 23 genes selected, only 14 were measured on protein level.

For proteins, we could not apply the same strategy, as there was only one significantly enriched gene set (FDR < 0.05). We therefore proceeded with all 152 leading edge proteins (Table 5.7) to the *trans* pQTL analysis.

Table 5.6: Transcriptomics core gene candidates extracted from eQTS GSEA leading edge.
Transcripts that appeared in 14 or more GSEA (on eQTS) leading edges for enriched GO terms (FDR < 0.05). eQTS, expression quantitative trait score; GSEA, gene set enrichment analysis; #, number of times that the gene appeared in the leading edge;
Table and legend taken from the supplementary material Assum et al. [2022b] [Assum et al., 2022a] `https://doi.org/10.1038/s41467-022-27953-1`.

Gene	#	Gene	#	Gene	#	Gene	#	Gene	#
ANK2	23	*MYH7*	21	*DMD*	18	*NKX2-5*	15	*NDUFS6*	14
RYR2	23	*CACNA1C*	20	*SCN5A*	18	*TNNT2*	15	*NDUFV2*	14
CAV3	22	*PKP2*	20	*KCNJ2*	17	*CALM1*	14	*TNNI3*	14
ATP1A2	21	*TAZ*	20	*PLN*	17	*CALM3*	14		
GJA5	21	*SLC8A1*	19	*MYL2*	16	*MYBPC3*	14		

Table 5.7: Proteomics core gene candidates extracted from pQTS GSEA leading edge.
Proteins that appeared in the leading edge of the significantly enriched pQTS gene set (FDR < 0.05). pQTS, protein quantitative trait score; GSEA, gene set enrichment analysis;
Table and legend taken from the supplementary material Assum et al. [2022b] [Assum et al., 2022a] `https://doi.org/10.1038/s41467-022-27953-1`.

Gene	Gene	Gene	Gene	Gene	Gene	Gene
ABHD10	APOC1	COX5A	GBAS	MPST	NME1	PRELP
ACAD10	AQP1	COX6C	GCDH	MSRA	NT5C	PRKAR2B
ACADVL	ATP1B1	CRAT	GLRX	MTAP	OGDH	PTGES2
ACAT1	ATP5A1	CS	GLUD1	MTHFD1	OGN	PTGR2
ACO1	ATP5D	CYB5R3	GOT1	NAMPT	OLA1	QDPR
ACO2	ATP5F1	CYC1	GPD1L	NDUFA10	OXCT1	SDHA
ACSBG2	ATP5J2	CYGB	GPI	NDUFA3	P4HB	SDHB
ADA	BCAT2	DBT	GSS	NDUFA5	PAFAH1B1	SDHD
ADI1	BDH1	DCXR	GSTO1	NDUFA7	PAM	SLC25A11
ADK	BGN	DDAH2	HIBADH	NDUFA9	PCCA	SLC25A12
AHCY	BRP44	DLAT	HPRT1	NDUFB10	PCCB	STOML2
AK4	C3	ECH1	HSD17B10	NDUFB11	PCYOX1	SUCLA2
AKR1A1	CA1	ECHS1	HSD17B4	NDUFB2	PDE5A	SUCLG1
AKR7A2	CA3	ECI2	IDH2	NDUFB3	PDHA1	TALDO1
ALDH1A1	CAT	ENO1	IDH3A	NDUFB8	PDXK	TPI1
ALDH1A2	CBR1	ENO3	IVD	NDUFB9	PEPD	UQCR10
ALDH2	CKM	ERLIN2	LHPP	NDUFC2	PGK1	UQCRC2
ALDH4A1	COQ3	ESD	MCCC1	NDUFS3	PGM1	UQCRFS1
ALDH5A1	COQ5	ETFA	MCCC2	NDUFS4	PGM2	VCAN
ALDH6A1	COQ7	ETFB	MCEE	NDUFS7	PGM5	WARS
ALDH7A1	COQ9	FH	ME2	NDUFS8	PPA1	
APOBEC2	COX4I1	FMOD	ME3	NDUFV1	PRDX4	

5.1.2.3 *Trans* QTL analyses for atrial fibrillation core gene candidates

QTL analyses were performed for 108 AF GWAS hits with 23 transcripts on 74 samples for *trans* eQTLs and 152 proteins on 73 samples for *trans* pQTLs.

We identified two significant *trans* eQTLs (FDR < 0.2): rs11658168-*TNNT2*, with the transcript encoding a cardiac structural protein and rs9481842-*NKX2-5*, where the regulated gene was a cardiac-specific transcription factor (TF). The TNNT2 protein which was also measured, was not associated to the same SNP (see also Table 5.8).

Using the same significance threshold (FDR < 0.2), we found five *trans* pQTLs for two SNPs (rs11588763-CYB5R3/NDUFB3/NDUFA9/DLAT, rs11658168-HIBADH) with all proteins connected to metabolism with HIBADH being a mitochondrial enzyme, as well as NDUFA9, NDUFB3 and DLAT being mitochondrial enzyme subunits. Only DLAT showed a nominal significant association of the *DLAT* transcript for the SNP rs11588763 (P = 0.0126), but with opposing effect sizes (Table 5.8).

Table 5.8: *Trans* QTL results.
Significant *trans* eQTLs and pQTLs for a FDR < 0.2 (Benjamini-Hochberg procedure to account for multiple comparisons per omic). Two-sided t-tests were performed on 23 transcripts with 74 samples and 152 proteins with 73 samples of human heart right atrial appendage tissue for 108 variants associated with atrial fibrillation from the GWAS catalog (or their proxy, if the GWAS SNP was not measured). Calculations were carried out using the SNP rs11658168 as a proxy for the GWAS SNP rs9675122 as well as rs11588763 instead of the GWAS SNP rs34292822.
QTL, quantitative trait loci; eQTL, expression quantitative trait loci; pQTL, protein quantitative trait loci; FDR, false discovery rate;
*Mutation known to affect cardiovascular phenotypes; **Mutation known to affect arrhythmias; $^{+}$Differential expression or functional impairment for cardiovascular phenotypes; $^{++}$Differential expression or functional impairment for arrhythmias;
Table and legend adapted from the supplementary material Assum et al. [2022b] [Assum et al., 2022a] `https://doi.org/10.1038/s41467-022-27953-1`.

	Variant		Gene			*Trans* eQTL				*Trans* pQTL		
SNP	**Chr**	**Position**	**Transcript**	**Chr**	β	***T* value**	**P value**	**FDR**	**fi**	***T* value**	**P value**	**FDR**
rs11658168	chr17	7 406 134	*TNNT2**,++	chr1	-0.517	-4.27	6.43×10^{-5}	**0.0812**	-0.213	-1.42	0.160	0.928
rs9481842	chr6	118 974 798	*NKX2-5***	chr5	-0.593	-4.27	6.54×10^{-5}	**0.0812**				
	Variant		**Gene**			***Trans* eQTL**				***Trans* pQTL**		
SNP	**Chr**	**Position**	**Protein**	**Chr**	β	**T value**	**P value**	**FDR**	β	**T value**	**P value**	**FDR**
rs11588763	chr1	154 813 584	CYB5R3	chr22	-0.119	-0.527	0.600	0.998	-0.786	-4.89	6.86×10^{-6}	**0.113**
rs11588763	chr1	154 813 584	NDUFB3^{+}	chr2	0.291	1.31	0.193	0.973	-0.916	-4.44	3.56×10^{-5}	**0.133**
rs11658168	chr17	7 406 134	HIBADH	chr7	-0.144	-0.861	0.393	0.997	-0.512	-4.43	3.66×10^{-5}	**0.133**
rs11588763	chr1	154 813 584	NDUFA9^{++}	chr12	0.257	1.16	0.249	0.985	-0.752	-4.42	3.85×10^{-5}	**0.133**
rs11588763	chr1	154 813 584	DLAT	chr11	0.470	2.56	0.0126	0.759	-0.716	-4.40	4.05×10^{-5}	**0.133**

5.1.3 NKX2-5 transcription factor network

5.1.3.1 NKX2-5 *trans* pQTL

In line with general low abundance of transcription factors compared to other proteins in heart tissue, the NKX2-5 protein was not detectable by mass spectrometry. Therefore, Julia Krause performed additional Western blot experiments to analyze transcription

factor levels in remaining tissue samples. Differential protein abundance for the SNP rs9481842 was clearly visible by two different housekeepers alpha-actinin and GAPDH as shown in Figure 5.8. Alpha-actinin was chosen for the final measurements and indeed, we were able to replicate the *trans* eQTL with the respective significant rs9481842-NKX2-5 *trans* pQTL in a linear regression (β = -0.45, T = -2.9, P = 0.049, N = 29, df = 21, two-sided t-test, Figure 5.9).

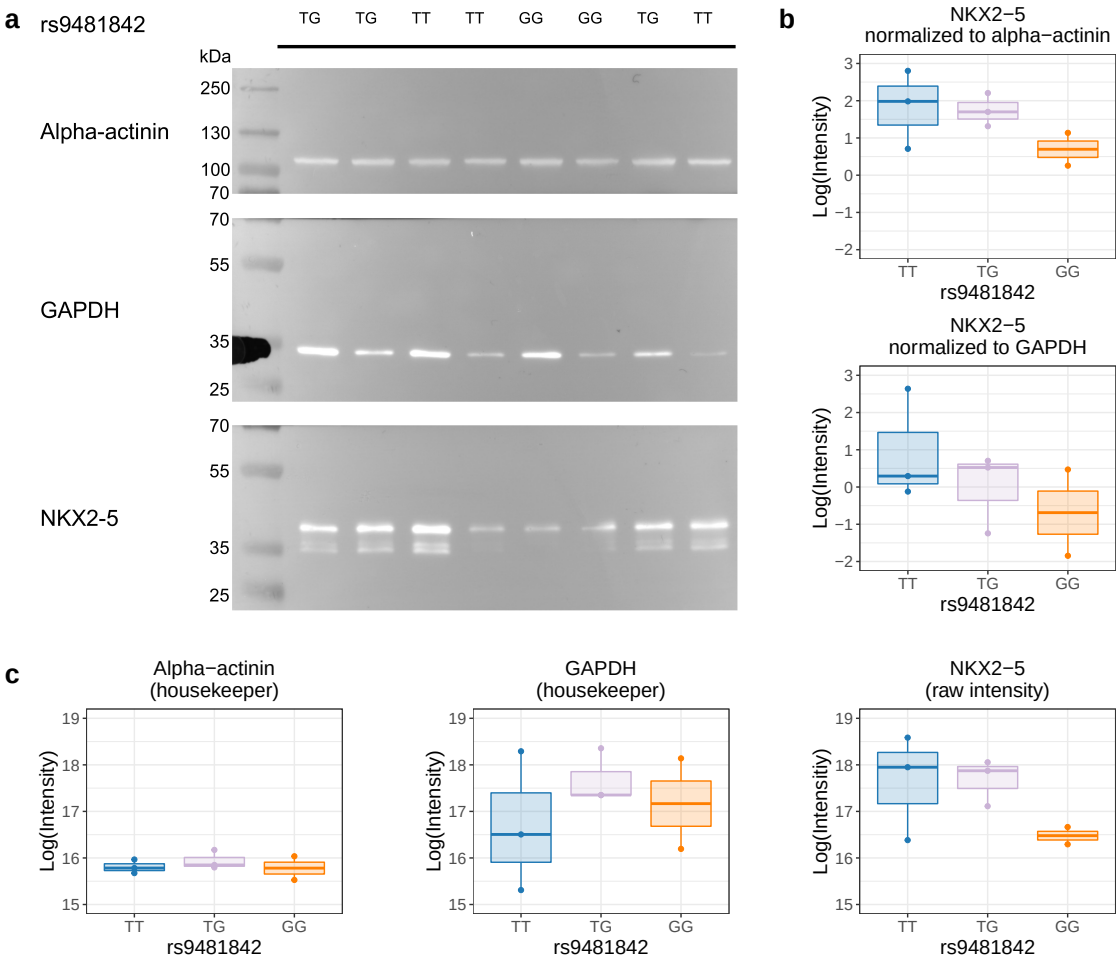


Figure 5.8: Western blot analysis for NKX2-5 quantification.
a: Exemplary additional Western blot image for alpha-actinin (100 kDa), GAPDH (37 kDa) and NKX2-5 (30-42 kDa) for N = 8 biologically independent samples with different genotypes. The membrane was cut in two parts to stain for alpha-actinin and NKX2-5 in parallel. The NKX2-5 membrane was reprobed with GAPDH antibody after incubation with stripping buffer. In total, five Western blots for NKX2-5 and alpha-actinin were performed to measure $N = 29$ biologically independent samples (presented in Figure 5.9).
b: Quantification of the NKX2-5 *trans* protein quantitative trait loci for housekeepers alpha-actinin and GAPDH (N = 8 independent biological samples).
c: Raw, absolute quantifications of alpha-actinin, GAPDH and NKX2-5 (N = 8 independent biological samples).
Figure and legend taken from the supplementary material Assum et al. [2022b] [Assum et al., 2022a] `https://doi.org/10.1038/s41467-022-27953-1`.

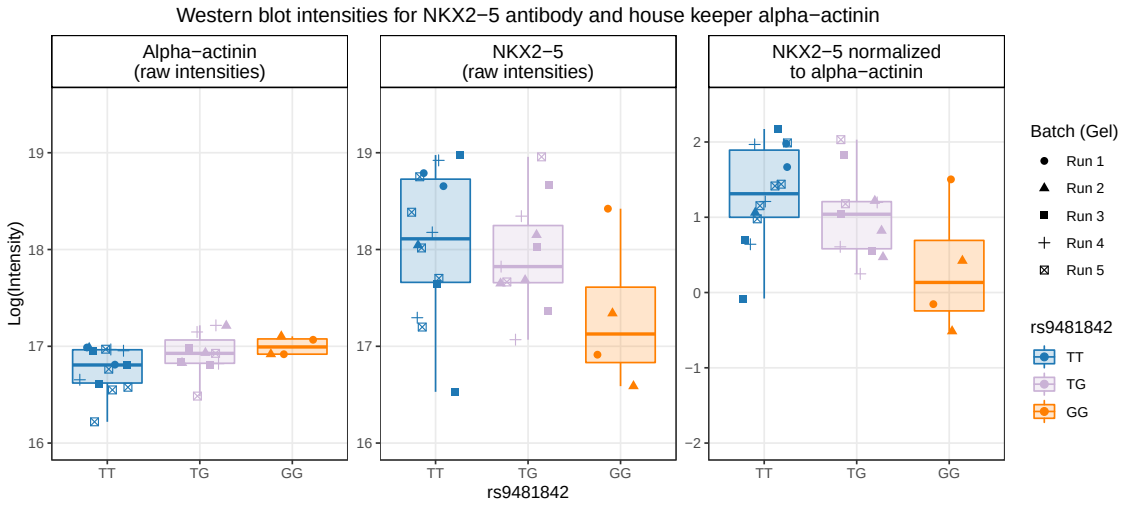


Figure 5.9: Replication of the NKX2-5 *trans* eQTL on proteomics level using Western blot analysis.
Association of rs9481842 genotypes with NKX2-5 protein expression and the housekeeper alpha-actinin (5 different Western blots, N = 29).
In the boxplots, the lower and upper hinges correspond to the first and third quartiles (the 25th and 75th percentiles). The median is denoted by the central line in the box. The upper/lower whisker extends from the hinge to the largest/smallest value no further than 1.5·IQR from the hinge.
eQTL, expression quantitative trail loci; IQR, interquartile range;
Figure and legend taken from the supplementary material Assum et al. [2022b] [Assum et al., 2022a] `https://doi.org/10.1038/s41467-022-27953-1`.

5.1.3.2 NKX2-5 transcription factor activity

We were further interested in the downstream consequences of the transcription factor variability. Based on genome-wide transcriptomics data, we estimated the NKX2-5 transcription factor activity (TFA) from gene expression and the number of most likely functional binding sites (BS). BS were derived from tissue-specific public annotations, starting with NKX2-5 ChIP-seq data from human iPSC-derived cardiomyocytes [Benaglio et al., 2019] which were fine-mapped for promoter and open chromatin regions. Promoter regions were first determined using Gencode genome annotations [Frankish et al., 2019] as well as regions linked to those by promoter-capture HiC in human iPSC-derived cardiomyocytes [Montefiori et al., 2018] and then overlapped with open chromatin regions defined by ChromHMM chromatin states for human atrial appendage tissue [Roadmap Epigenomics Consortium et al., 2015].

TFA was then computed by summing over the number of functional binding sites multiplied by Z-score transformed expression values for each of the 9 960 genes. While we observed a very strong Pearson correlation between the SNP rs9481842 and the NKX2-5 transcript (ρ = -0.43, P = 1.4×10^{-4}, two-sided Pearson's correlation, Figure 5.10a), we still saw a more subtle genome-wide expression change with respect to the GWAS SNP when considering the TFA (ρ = -0.13, P = 0.145, one-sided Pearson's correlation, Figure 5.10b). More importantly, the measured NKX2-5 transcript values and the estimated TFA activity, which are directly molecularly related, show a high positive

correlation of ρ = 0.36 (P = 1.3×10^{-4}, one-sided Pearson's correlation, Figure 5.10c).

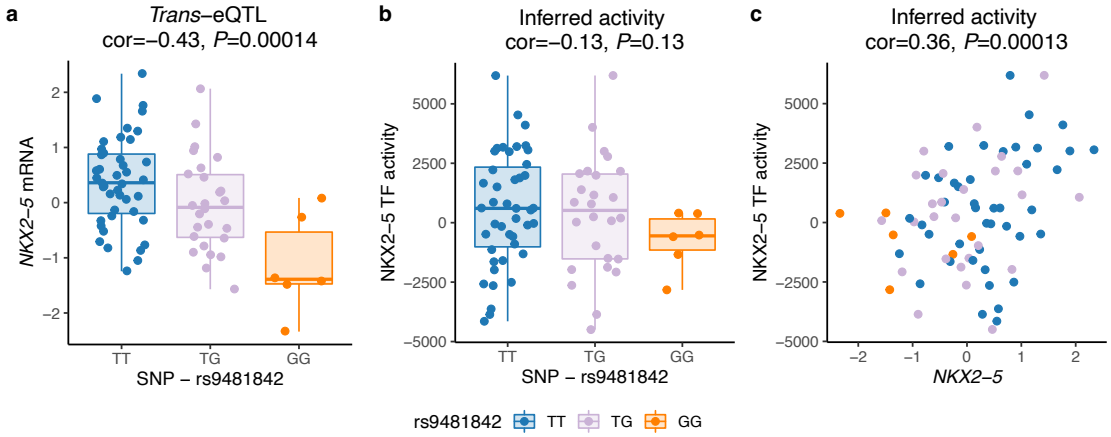


Figure 5.10: Causal modeling of NKX2-5 and TF activity.
a: *Trans* eQTL rs948182-*NKX2-5* (two-sided Pearson's correlation).
b: Dependence of TF activity on rs9481842 (one-sided Pearson's correlation).
c: Correlation between *NKX2-5* transcript and TF activity (one-sided Pearson's correlation).
In the boxplots, the lower and upper hinges correspond to the first and third quartiles (the 25th and 75th percentiles). The median is denoted by the central line in the box. The upper/lower whisker extends from the hinge to the largest/smallest value no further than 1.5·IQR from the hinge. N = 75 independent biological samples are displayed.
eQTL, expression quantitative trait loci; TF, transcription factor; IQR, interquartile range;
Figure and legend taken from the supplementary material Assum et al. [2022b] [Assum et al., 2022a] `https://doi.org/10.1038/s41467-022-27953-1`.

In order to investigate causality in linking the different measurements, we used partial correlation analysis to support our hypothesis of the SNP variability being propagated from transcript to TFA changes. Indeed, the SNP-TFA correlation decreased drastically when conditioning on transcript expression, while the SNP-transcript correlation remained relatively unchanged when considering the TFA. Additionally, we saw a small decrease in the transcript-TFA correlation when conditioning on the SNP genotype as summarized in Table 5.9.

Table 5.9: Partial correlation analysis of *NKX2-5* expression linking the SNP rs9481842 and TF activity.
Two-sided Pearson's correlation tests unless stated otherwise.
TF, transcription factor; *one-sided Pearson's correlation test;
Table and legend taken from the supplementary material Assum et al. [2022b] [Assum et al., 2022a] `https://doi.org/10.1038/s41467-022-27953-1`.

Measure	SNP and *NKX2-5* mRNA	SNP and TF activity	*NKX2-5* mRNA and TF activity
Correlation	-0.43 (P = 1.4×10^{-4})	-0.13 (P = 0.13)*	0.36 (P = 1.3×10^{-4})*
Partial correlation	-0.41 (P = 3×10^{-4})	0.007 (P = 0.95)	0.3 (P = 0.011)
Condition	TF activity	*NKX2-5* mRNA	SNP

To some extent, the estimated TFA can also be used as a correlate of the NKX2-5 protein levels. We therefore further investigated the relation between NKX2-5 transcript, protein and TFA. For all three entities, the dependence on the GWAS SNP rs9481842 is summarized in Figure 5.11a-c.

As described previously, we have observed significant differences between transcript and protein expression. Again, we found a rather modest correlation of only 0.14 (ρ = 0.14, P = 0.47, two-sided Spearman's rank correlation) between NKX2-5 transcript and protein abundance as shown in Figure 5.11d. On the contrary, the inferred TFA was a good representation of actual observed protein measurements with a correlation of 0.42 (ρ = 0.42, P = 0.026, two sided Spearman's rank correlation, Figure 5.11e).

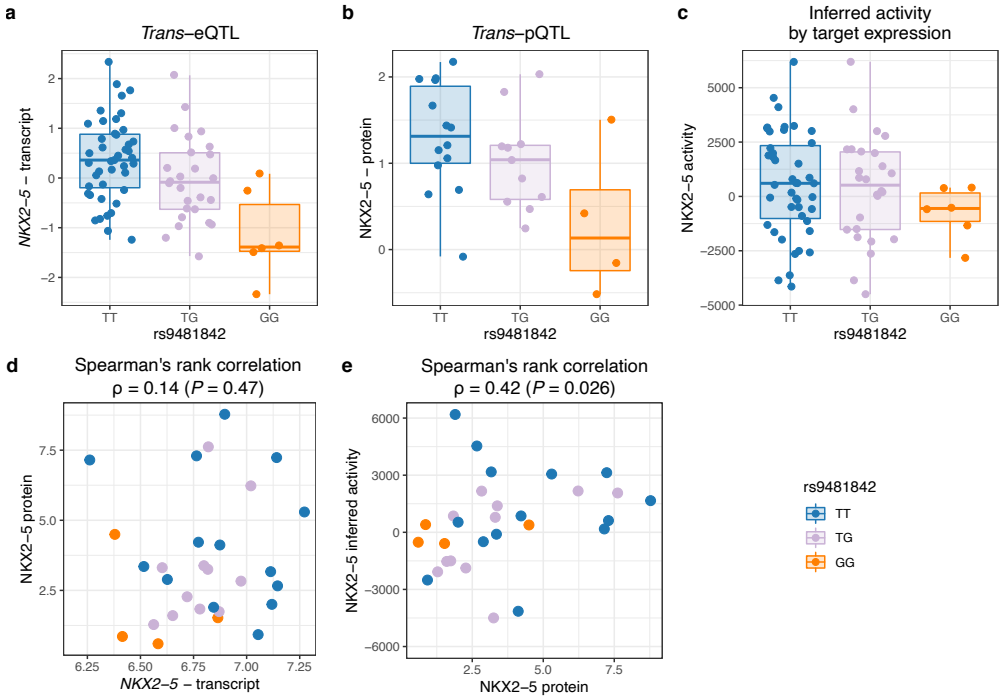


Figure 5.11: Inferred transcription factory activity strongly correlates with protein intensity.
a: *Trans* eQTL of the SNP rs9481842 and the transcription factor NKX2-5 discovered by our polygenic risk score based enrichment approach for *trans* QTL gene candidate selection using microarray mRNA quantifications in N = 75 biologically independent samples (two-sided Spearman's rank correlation, ρ = -0.38, P = 0.00086).
b: *Trans* pQTL validation of the rs9481842-NKX2-5 *trans* eQTL in remaining tissue samples using Western blot analysis in N = 29 biologically independent samples (two-sided Spearman's rank correlation, ρ = -0.41, P = 0.029).
c: Estimated transcription factor activity inferred by genome-wide transcriptomics data and independent tissue or cell type specific annotations in N = 75 biologically independent samples (one-sided Spearman's rank correlation, ρ = -0.13, P = 0.14).
d: Correlation between the NKX2-5 mRNA and protein in N = 29 biologically independent samples (two-sided Spearman's rank correlation).
e: High two-sided Spearman's rank correlation between the inferred NKX2-5 transcription factor activity and actual protein intensities in N = 29 biologically independent samples.
In the boxplots, the lower and upper hinges correspond to the first and third quartiles (the 25th and 75th percentiles). The median is denoted by the central line in the box. The upper/lower whisker extends from the hinge to the largest/smallest value no further than 1.5·IQR from the hinge.
eQTL, expression quantitative trait loci; SNP, single-nucleotide polymorphism; QTL, quantitative trait loci; pQTL, protein quantitative trait loci; IQR, interquartile range;
Figure and legend taken from the supplementary material Assum et al. [2022b] [Assum et al., 2022a] `https://doi.org/10.1038/s41467-022-27953-1`.

5.1.3.3 NKX2-5 targets

With the rs9481842 we discovered a new link between a GWAS variant and a very important TF in the human heart with a significant *trans* eQTL, a significant *trans* pQTL and more subtle consequences for general gene expression patterns (Figure 5.13b-d). However, NKX2-5 most likely acts as a key regulator of further disease-relevant genes which would be of great interest (Figure 5.13a).

We therefore analyzed genes which showed strong regulation of the SNP rs9481842 via the NKX2-5 transcript on mRNA and protein level and identified 13 specific targets which we also considered as putative core genes.

As visualized in Figure 5.12 and described in section 3.4.4.6, we first used the most likely functional TF BS from the estimation of the TFA (a) and selected all those genes which were measured on mRNA and protein level and had at least one BS.
We then evaluated whether the association of the SNP with the target transcript was most likely due to the NKX2-5 TF, i.e. if any significant association of the target transcript with the SNP (b) disappeared when including the NKX2-5 transcript (c) in the corresponding linear model. We further only selected the genes with the strongest correlation of the target protein with the NKX2-5 transcript (d) resulting in the 13 proteins PPIF, MYL4, CKM, MYL7, PGAM2, TNNC1, CYC1, ETFB, PRDX5, AK1, ALDOA, TCAP and TOM1L2.

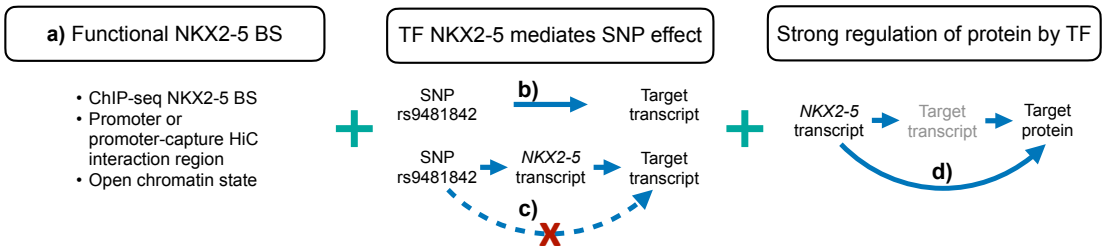


Figure 5.12: Definition of functional NKX2-5 targets.
Functional NKX2-5 targets were derived in a three step process: first, by the presence of a most likely functional TF BS, second by a likely regulation of target transcript by the SNP through the TF, and finally, we checked for strong regulation of the target protein by the TF.
TF, transcription factor; BS, binding site; SNP, single-nucleotide polymorphism;
Figure and legend taken from the supplementary material Assum et al. [2022b] [Assum et al., 2022a] `https://doi.org/10.1038/s41467-022-27953-1`.

Neither the rs9481842-target pQTL, nor the AF phenotype were used to derive the NKX2-5 targets, yet a very consistent down regulation of all target protein levels in prevalent AF cases was observed as shown in Figure 5.13e.
When quantifying the disease association, the 13 NKX2-5 targets were highly negatively enriched with a GSEA P value of 7.17×10^{-5} (ES = -0.781, NES = -2.18).

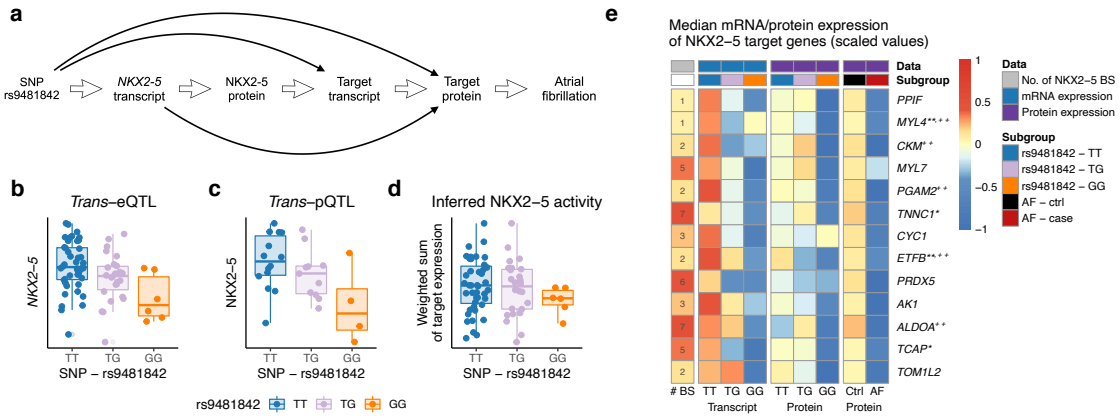


Figure 5.13: NKX2-5 activity controlled by AF GWAS variant rs9481842.
a: Graphical illustration of NKX2-5 TF target gene analysis in AF.
b: Strong *trans* eQTL of the SNP rs9481842 with the *NKX2-5* transcript for N = 75 independent biological samples.
c: Validation of the NKX2-5 *trans* eQTL on protein level (*trans* pQTL) using Western blot analysis in remaining tissue samples (N = 29 independent biological samples).
d: NKX2-5 activity estimation based on target mRNA expression stratified by the rs9481842 genotype for N = 75 independent biological samples.
e: Depicted are functional NKX2-5 targets with the number of TF binding sites (column 1), *trans* eQTL strength (columns 2-4), *trans* pQTL strength (columns 5-7) and protein level in AF (columns 8-9). The color scale represents median transcript or protein values per group (=columns). Residuals corrected for fibroblast-score and RIN-score/protein concentration with subsequent normal-quantile normalization per gene were used to calculate the medians per group. A quantitative description of the qualitative results presented in the heatmap can be found in Table 5.10 and Table 5.12.
In the boxplots, the lower and upper hinges correspond to the first and third quartiles (the 25th and 75th percentiles). The median is denoted by the central line in the box. The upper/lower whisker extends from the hinge to the largest/smallest value no further than 1.5·IQR from the hinge.
AF, atrial fibrillation; QTL, quantitative trait loci; BS, binding site; IQR, interquartile range;
*Mutation known to affect cardiovascular phenotypes; **Mutation known to affect arrhythmias; $^{+}$Differential expression or functional impairment for cardiovascular phenotypes; $^{++}$Differential expression or functional impairment for arrhythmias.
Figure and legend taken from Assum et al. [2022a] `https://www.nature.com/articles/s41467-022-27953-1/figures/5`, licensed under CC BY 4.0 `https://creativecommons.org/licenses/by/4.0/`.

Table 5.10: NKX2-5 target correlations with *trans* eQTL SNP rs9481842 and *NKX2-5* transcript as well as AF disease association.
All associations were computed by two-sided t-tests based on a linear model with covariates (fibroblast-score, RIN-score or protein concentration). Correlations of SNP rs9481842 with target transcript (*trans* eQTL) and target protein (*trans* pQTL) were evaluated as well as the correlation between *NKX2-5* transcript expression and target protein expression. FDR based on the Benjamini-Hochberg procedure was only assessed for a preselected subset of genes when evaluating the correlation between the TF and target protein expression (see methods and Figure 5.12). Proteomics differential expression results are the same as reported in Table 5.12, were derived by two-sided t-tests and additionally included common risk factors for AF as covariates (age, sex, BMI, diabetes, systolic blood pressure, hypertension medication, myocardial infarction and smoking) in the linear model.
SNP, single-nucleotide polymorphism; BS, number of binding sites; AF, atrial fibrillation; eQTL, expression quantitative trait loci; pQTL, protein quantitative trait loci; FDR, false discovery rate;
*Mutation known to affect cardiovascular phenotypes; **Mutation known to affect arrhythmias; $^{+}$Differential expression or functional impairment for cardiovascular phenotypes; $^{++}$Differential expression or functional impairment for arrhythmias;
Table and legend taken from the supplementary material Assum et al. [2022b] [Assum et al., 2022a] `https://doi.org/10.1038/s41467-022-27953-1`.

NKX2-5 target		**SNP rs9481842**				***NKX2-5* target protein association**			**Disease association protein and AF**	
		***trans* eQTL**		***trans* pQTL**						
Gene	**BS**	β	***P* value**	β	***P* value**	β	***P* value**	**FDR**	β	***P* value**
PPIF	1	-0.131	0.0081	-0.0388	0.0145	0.172	2.10×10^{-4}	0.00824	-0.0342	0.261
MYL4$^{**,++}$	1	-0.087	0.00921	-0.0359	0.0820	0.211	3.98×10^{-4}	0.00824	-0.0270	0.509
CKM^{++}	2	-0.115	0.0101	-0.0188	0.303	0.180	4.56×10^{-4}	0.00824	-0.0875	0.00705
MYL7	5	-0.125	0.0038	-0.0298	0.177	0.208	4.74×10^{-4}	0.00824	-0.0421	0.304
PGAM2^{++}	2	-0.214	0.00307	-0.0293	0.284	0.255	7.35×10^{-4}	0.0107	-0.175	0.000452
*TNNC1**	7	-0.0699	0.0359	-0.00809	0.623	0.147	1.69×10^{-3}	0.0211	-0.0557	0.0929
CYC1	3	-0.121	0.00278	-0.00651	0.749	0.175	3.16×10^{-3}	0.0307	-0.0946	0.0360
ETFB$^{**,++}$	2	-0.0924	0.0110	-0.0336	0.0563	0.152	3.39×10^{-3}	0.0307	-0.0553	0.105
PRDX5	6	-0.0710	0.00641	-0.0149	0.307	0.131	3.52×10^{-3}	0.0307	-0.0524	0.0789
AK1	3	-0.0654	0.0268	-0.0224	0.190	0.138	4.17×10^{-3}	0.0312	-0.0669	0.0341
ALDOA^{++}	7	-0.0555	0.0469	-0.00187	0.909	0.125	5.50×10^{-3}	0.0368	-0.0646	0.0341
*TCAP**	5	-0.114	0.00707	-0.0365	0.266	0.244	6.90×10^{-3}	0.0429	-0.0178	0.779
TOM1L2	2	-0.0886	0.0157	-0.0292	0.222	0.170	8.16×10^{-3}	0.0473	-0.0771	0.0849

5.1.4 Validation and replication

Due to the highly specific cohort, replication of the exact analyses in an independent dataset was not feasible. We still proceeded to validate or replicate as many parts of the analysis as possible.

5.1.4.1 Disease links in literature

A key property of AF core genes is the direct link to disease. As mutations affecting the function of core genes might have severe consequences, core genes are likely to be causal for rare, Mendelian disorders.

In this context, NKX2-5, MYL4 and ETFB mutations are known to be causal for AF or other arrhythmias [Jhaveri et al., 2018, Huang et al., 2013, Orr et al., 2016, Florian and Yilmaz, 2019] whereas TNNT2, TNNC1 and TCAP mutations are the genetic reason for different cardiomyopathies [Hershberger et al., 2009, Parvatiyar et al., 2012, Hayashi et al., 2004] (see Table 5.11).

Table 5.11: Disease annotations for putative core genes and functional targets in literature.
Findings relating putative core genes and functional targets to cardiovascular phenotypes.
AF, atrial fibrillation; DCM, Dilated cardiomyopathy; HCM, Hypertrophic cardiomyopathy;
*Mutation known to affect cardiovascular phenotypes; **Mutation known to affect arrhythmias; $^{+}$Differential expression or functional impairment for cardiovascular phenotypes; $^{++}$Differential expression or functional impairment for arrhythmias;
Table and legend adapted from the supplementary material Assum et al. [2022b] [Assum et al., 2022a] `https://doi.org/10.1038/s41467-022-27953-1`.

Gene	Finding	Clinical phenotypes	First Author
TNNT2$^{*,++}$	TNNT2 mutations	DCM	Hershberger et al. [2009]
	Lower TNNT2 protein expression (human atrial tissue)	AF	Doll et al. [2017]
*NKX2-5***	NKX2-5 mutations including loss of function mutation	Arrhythmia, AF	Jhaveri et al. [2018], Huang et al. [2013]
NDUFA9^{++}	Impaired complex I function (human atrial tissue)	AF, diabetes	Kanaan et al. [2019]
NDUFB3^{+}	NDUFB3 deficiency	Cardiomyopathy	El-Hattab and Scaglia [2016]
MYL4$^{**,++}$	Mutation in MYL4	Familial AF	Orr et al. [2016]
	Lower MYL4 protein expression (human atrial tissue)	AF	Doll et al. [2017]
CKM^{++}	Lower CKM protein expression (human atrial tissue)	AF	Tu et al. [2014], Doll et al. [2017]
	Lower CKM protein expression (human myocardial tissue)	HCM	Coats et al. [2018] (Coats Suppl. Table 3)
PGAM2^{++}	Lower PGAM2 protein expression (human atrial tissue)	AF	Tu et al. [2014]
	Lower PGAM2 protein expression (human myocardial tissue)	HCM	Coats et al. [2018] (Coats Suppl. Table 3)
*TNNC1**	Mutation in TNNC1	Cardiomyopathy	Parvatiyar et al. [2012]
ETFB$^{**,++}$	Lower ETFB protein expression (human atrial tissue)	AF	Tu et al. [2014]
	ETFB mutation	Arrhythmias, HCM, DCM, conduction defects	Florian and Yilmaz [2019]
ALDOA^{++}	Lower ALDOA protein expression (human atrial tissue)	AF	Tu et al. [2014]
	Lower ALDOA protein expression (human myocardial tissue)	HCM	Coats et al. [2018] (Coats Suppl. Table 3)
*TCAP**	TCAP gene mutations	HCM, DCM	Hayashi et al. [2004]
TOM1L2	Differentially expressed	AF together with neurocognitive decline	Dalal et al. [2015]

Additionally, many of our putative core genes were differentially expressed or had impaired function on protein level in human atrial or myocardial tissue for AF or other arrhythmias (TNNT2, NDUFA9, MYL4, CKM, PGAM2, ETFB, ALDOA and TOM1L2) [Doll et al., 2017, Coats et al., 2018, Tu et al., 2014, Dalal et al., 2015] or for

cardiomyopathies (NDUFB3) [El-Hattab and Scaglia, 2016].

A more detailed description of the literature annotations can be found in Table 5.11 and multiple genes as stated by Wang et al. [2020] were also either identified by their integrative omics approach or mentioned in the OMIM database [Amberger et al., 2015].

5.1.4.2 AF association

One of the most important characteristics of our putative core genes is the disease association. By better understanding the function of specific genes and their link to AF, we can learn more about complex molecular causes.

Table 5.12: Putative core genes and functional targets with disease association.
Proteomics differential abundance results in human atrial appendage tissue for prevalent AF. Two-sided t-tests were calculated as part of a multiple linear regression model including the AF-related covariates sex, age, BMI, diabetes, systolic blood pressure, hypertension medication, myocardial infarction, and smoking status (see methods differential protein analysis, $N = 78$, df $= 66$). The Benjamini-Hochberg procedure was used to asses FDR and account for multiple comparisons.
AF, atrial fibrillation; BMI, body mass index; QTL, quantitative trait loci; FDR, false discovery rate;
*Mutation known to affect cardiovascular phenotypes; **Mutation known to affect arrhythmias; $^{+}$Differential expression or functional impairment for cardiovascular phenotypes; $^{++}$Differential expression or functional impairment for arrhythmias;
Table and legend taken from Assum et al. [2022a] `https://doi.org/10.1038/s41467-022-27953-1`.

			Protein AF association			
Gene	**Chr**	**Type**	β	**T value**	**P value**	**FDR**
TNNT2*,$^{++}$	chr1	*trans* eQTL	-0.0609	-1.61	0.113	1.00
NKX2-5**	chr5	*trans* eQTL				
CYB5R3	chr22	*trans* pQTL	-0.0212	-0.662	0.511	1.00
NDUFB3^{+}	chr2	*trans* pQTL	-0.0631	-1.35	0.182	1.00
HIBADH	chr7	*trans* pQTL	-0.0454	-1.24	0.218	1.00
NDUFA9^{++}	chr12	*trans* pQTL	-0.0533	-1.20	0.235	1.00
DLAT	chr11	*trans* pQTL	-0.0231	-0.579	0.564	1.00
PPIF	chr10	NKX2-5 target	-0.0342	-1.13	0.261	1.00
MYL4**,$^{++}$	chr17	NKX2-5 target	-0.0270	-0.664	0.509	1.00
CKM^{++}	chr19	NKX2-5 target	-0.0875	-2.78	0.00705	0.120
MYL7	chr7	NKX2-5 target	-0.0421	-1.04	0.304	1.00
PGAM2^{++}	chr7	NKX2-5 target	-0.175	-3.70	0.000452	0.00813
TNNC1*	chr3	NKX2-5 target	-0.0557	-1.71	0.0929	1.00
CYC1	chr8	NKX2-5 target	-0.0946	-2.14	0.036	0.545
ETFB**,$^{++}$	chr19	NKX2-5 target	-0.0553	-1.65	0.105	1.00
PRDX5	chr11	NKX2-5 target	-0.0524	-1.79	0.0789	1.00
AK1	chr9	NKX2-5 target	-0.0669	-2.17	0.0341	0.545
ALDOA^{++}	chr16	NKX2-5 target	-0.0646	-2.17	0.0341	0.545
TCAP*	chr17	NKX2-5 target	-0.0178	-0.282	0.779	1.00
TOM1L2	chr17	NKX2-5 target	-0.0771	-1.75	0.0849	1.00

Our *trans* analyses expanded on the TF NKX2-5, whose function in the context of cardiac development [Anderson et al., 2018], congenital heart disease [Akazawa and

Komuro, 2005] and of course AF [Huang et al., 2013, Jhaveri et al., 2018, Benaglio et al., 2019] has been extensively investigated.
Additionally, most of the proteins were known to affect key mechanisms relevant for AF, such as MYL4, MYL7, TNNC1 and TCAP for contractile function and PPIF, CKM, AK1, PGAM2, CYC1, ETFB and ALDOA for metabolism.
We further analyzed the association of prevalent AF for all putative core genes while adjusting for common non-genetic risk factors of AF (Table 5.12). With CKM, PGAM2, CYC1, AK1 and ALDOA, five out of 13 NKX2-5 targets were significantly associated (nominal P value < 0.05).

We further reanalyzed existing transcriptomics and proteomics datasets in order to assess the AF disease association as well as the co-regulation of NKX2-5 targets by the TF.

5.1.4.3 GSE128188

Table 5.13: Putative core genes and functional targets differential expression in the GSE128188 dataset. RNA-seq differential expression results for AF in right and left atrial appendage tissue. Statistics were calculated using edgeR's exact test and were reported from the original authors of the study. FDR_{core} denotes Benjamini-Hochberg false discovery rate applied only on the 20 putative core genes listed below. AF, atrial fibrillation; logFC, logarithm of the fold change; QTL, quantitative trait loci; FDR, false discovery rate;
*Mutation known to affect cardiovascular phenotypes; **Mutation known to affect arrhythmias; $^{+}$Differential expression or functional impairment for cardiovascular phenotypes; $^{++}$Differential expression or functional impairment for arrhythmias;
Table and legend taken from the supplementary material Assum et al. [2022b] [Assum et al., 2022a] `https://doi.org/10.1038/s41467-022-27953-1`.

		mRNA AF association (right atrial appendage)				mRNA AF association (left atrial appendage)			
Gene	**Type**	**logFC**	**P value**	**FDR**	**FDR_{core}**	**logFC**	**P value**	**FDR**	**FDR_{core}**
TNNT2$^{*,++}$	*Trans* eQTL	-0.117	0.382	1.00	0.616	-0.191	0.152	0.983	0.522
*NKX2-5***	*Trans* eQTL	-0.121	0.431	1.00	0.616	-0.104	0.498	1.00	0.712
CYB5R3	*Trans* pQTL	-0.00112	0.993	1.00	0.993	0.0529	0.672	1.00	0.829
NDUFB3^{+}	*Trans* pQTL	-0.293	0.0443	0.554	0.272	-0.113	0.439	1.00	0.701
HIBADH	*Trans* pQTL	-0.236	0.0817	0.686	0.272	-0.147	0.277	1.00	0.522
NDUFA9^{++}	*Trans* pQTL	-0.138	0.285	0.971	0.571	-0.0964	0.456	1.00	0.701
DLAT	*Trans* pQTL	-0.0498	0.735	1.00	0.817	-0.0373	0.800	1.00	0.889
PPIF	NKX2-5 target	-0.0746	0.615	1.00	0.724	0.0145	0.922	1.00	0.922
MYL4$^{**,++}$	NKX2-5 target	-0.549	0.000681	0.0546	0.0136	-0.412	0.0107	0.338	0.214
CKM^{++}	NKX2-5 target	-0.208	0.239	0.941	0.571	-0.194	0.271	1.00	0.522
MYL7	NKX2-5 target	-0.352	0.0731	0.658	0.272	-0.271	0.167	1.00	0.522
PGAM2^{++}	NKX2-5 target	-0.521	0.00292	0.129	0.0292	-0.367	0.0359	0.596	0.359
*TNNC1**	NKX2-5 target	-0.0311	0.833	1.00	0.877	0.167	0.259	1.00	0.522
CYC1	NKX2-5 target	-0.237	0.11	0.77	0.315	-0.158	0.287	1.00	0.522
ETFB$^{**,++}$	NKX2-5 target	-0.0999	0.460	1.00	0.616	-0.0513	0.704	1.00	0.829
PRDX5	NKX2-5 target	-0.226	0.074	0.660	0.272	-0.230	0.0688	0.764	0.459
AK1	NKX2-5 target	-0.136	0.357	0.998	0.616	-0.180	0.225	1.00	0.522
ALDOA^{++}	NKX2-5 target	-0.0992	0.492	1.00	0.616	-0.0568	0.694	1.00	0.829
*TCAP**	NKX2-5 target	-0.164	0.277	0.967	0.571	-0.170	0.259	1.00	0.522
TOM1L2	NKX2-5 target	-0.098	0.493	1.00	0.616	0.0205	0.886	1.00	0.922

Thomas et al. [2019] evaluated RNA-seq data[6] from human right and left atrial tissue for AF cases and controls. We were able to replicate the down regulation of the 13 NKX2-5 targets in patients with AF compared to controls in sinus rythm in both right and left atrial samples (Figure 5.14 and Table 5.13).

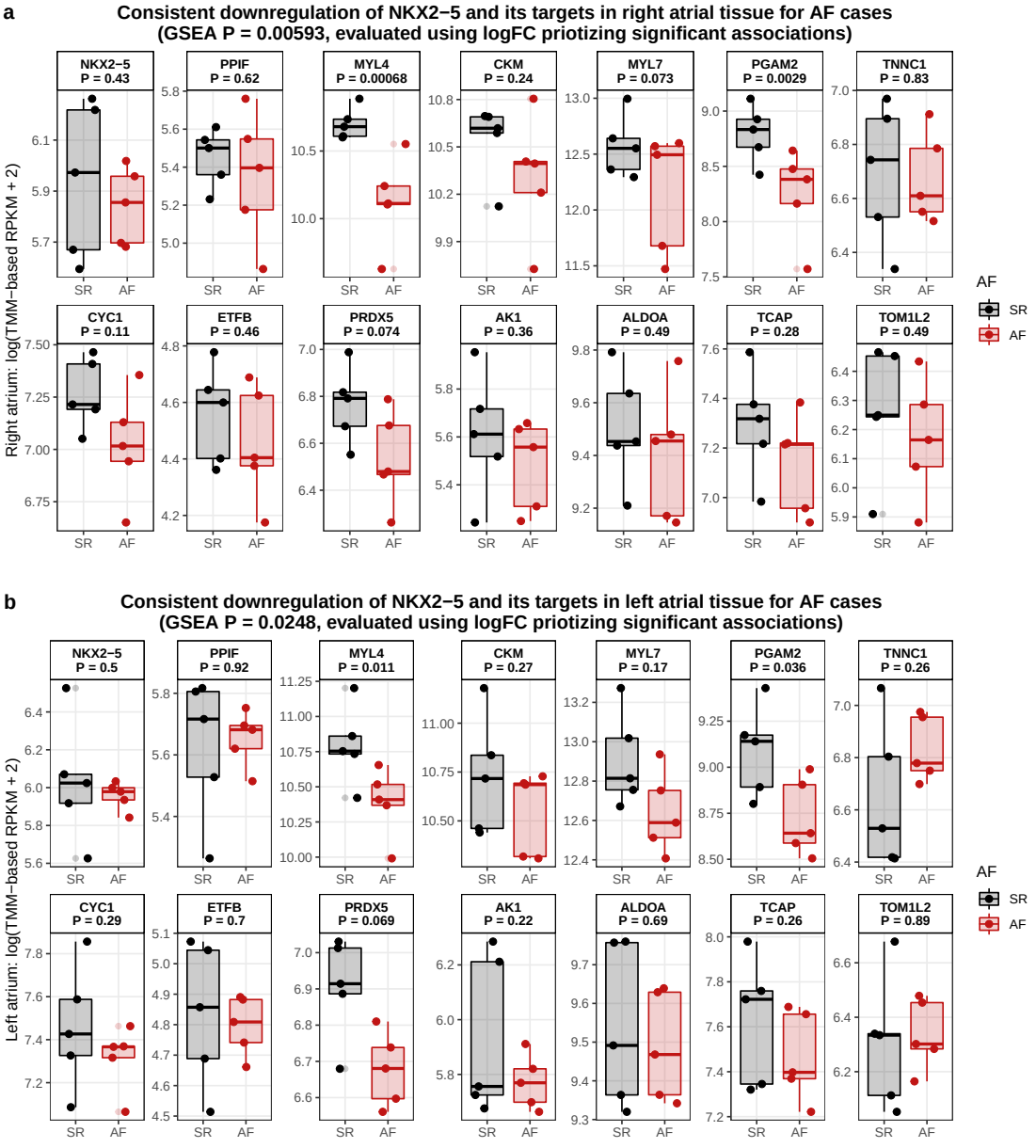


Figure 5.14: AF association of NKX2-5 and its functional targets in the GSE128188 dataset.
Log-transformed TMM-based RPKM RNA-seq expression counts for AF cases and controls in sinus rythm in right (a) and left (b) atrial tissue.
AF, atrial fibrillation; TMM, trimmed mean of M-values; RPKM, reads per kilobase of transcript per million mapped reads; SR, sinus rythm;

[6] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE128188

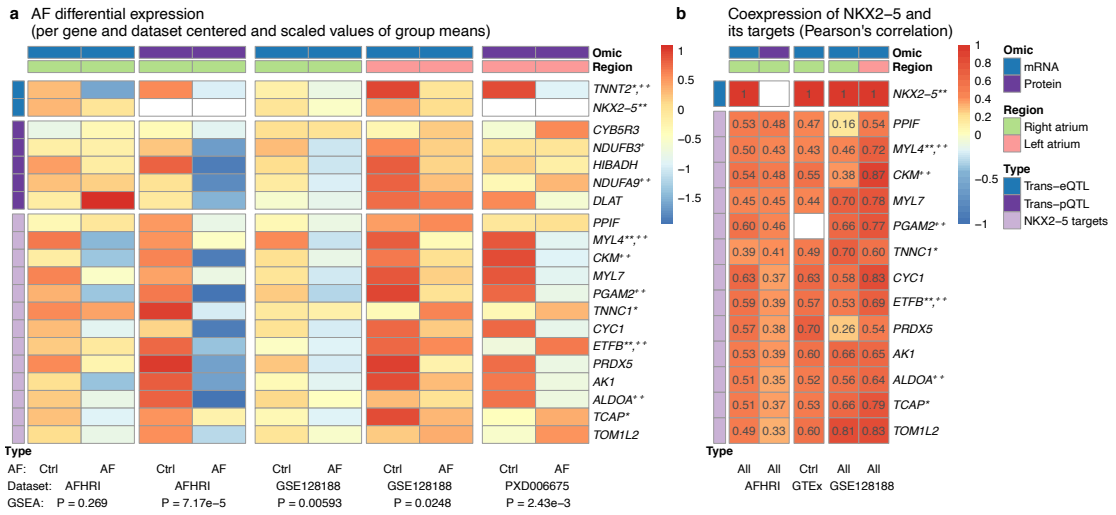


Figure 5.15: Replication of the core gene candidate AF association and NKX2-5 target coexpression in independent datasets.
Published proteomics data (PXD006675) as well as RNA-seq data (GSE128188, GTEx) generated from human atrial tissue samples were used for replication.
a: Centered and scaled values of the mean mRNA or protein expression in AF controls and cases, with stronger effects on protein level. GSEA P values quantify the negative association of NKX2-5 targets with respect to AF. Sample sizes per column: 69 controls, 14 prevalent AF cases, 69 controls, 14 prevalent AF cases (AFHRI, all right atrial appendage); five controls, five AF cases (GSE128188, both right atrial appendage); five controls, five AF cases (GSE128188, both left atrial appendage); three controls, three AF cases (PXD006675, both left atrium). A quantitative description of the qualitative results presented in the heatmap can be found in Table 5.12-5.14.
b: Coexpression of *NKX2-5* with the 13 identified NKX2-5 transcription factor targets (Pearson's correlation). Quantified is the correlation between *NKX2-5* and its targets on mRNA level for mRNA datasets and the correlation between the *NKX2-5* transcript expression with the target protein concentrations for the AFHRI proteomics (NKX2-5 not quantified in proteomics). Sample sizes used for the computation of correlations: 102 AFHRI mRNA, 96 AFHRI protein, 372 GTEx, ten GSE128188 right as well as ten left atrial appendage samples.
AF, atrial fibrillation; Ctrl, control i.e. individuals in sinus rhythm; GSEA, gene set enrichment analysis
*Mutation known to affect cardiovascular phenotypes; **Mutation known to affect arrhythmias; $^{+}$Differential expression or functional impairment for cardiovascular phenotypes; $^{++}$Differential expression or functional impairment for arrhythmias.
Figure and legend adapted from Assum et al. [2022a] `https://www.nature.com/articles/s41467-022-27953-1/figures/6`, licensed under CC BY 4.0 `https://creativecommons.org/licenses/by/4.0/`.

Similarly, we were also able to observe the collective negative disease association for the GSEA analysis (right atrium: P = 0.00593 and left atrium: P = 0.0248).
The differential expression results provided by the original authors are also listed in Table 5.13 and relative gene expression abundance compared to several other datasets is visualized in Figure 5.15a.

We were additionally interested in the co-expression of NKX2-5 and its identified targets. In right atrial tissue, we observed an overall high correlation of more than 0.5 for nine out of the 13 targets, with the most correlated being the *TOM1L2* gene with ρ = 0.81 and the lower ones being *PPIF* (ρ = 0.16), *PRDX5* (ρ = 0.26), *CKM* (ρ = 0.38) and *MYL4* (ρ = 0.46). Even stronger co-expression was found for the left atrial samples

with the lowest being again *PPIF* and *PRDX5* with a very strong correlation of $\rho = 0.54$ and *CKM* even going up to a $\rho = 0.87$. All results can be found in Figure 5.15b.

5.1.4.4 PXD006675

For the same purpose of replication, we investigated proteomics data[7] from Doll et al. [2017] which provided protein measurements of different anatomical regions of the human heart and additionally compared left atrial protein in three AF cases as well as in three controls.

Similar to our dataset, also in this case the NKX2-5 protein was not measured, but we were able to assess the differential protein abundance for our putative core genes as listed in Table 5.14 based on the findings of the original authors.

Table 5.14: Putative core genes and functional targets AF disease association for the proteomics dataset PXD006675.
Proteomics differential abundance results in human left atrial appendage tissue for AF. Two-sided t-tests were calculated and are reported from the original authors of the study.
AF, atrial fibrillation; QTL, quantitative trait loci; FDR, false discovery rate;
*Mutation known to affect cardiovascular phenotypes; **Mutation known to affect arrhythmias; $^{+}$Differential expression or functional impairment for cardiovascular phenotypes; $^{++}$Differential expression or functional impairment for arrhythmias;
Table and legend taken from the supplementary material Assum et al. [2022b] [Assum et al., 2022a] `https://doi.org/10.1038/s41467-022-27953-1`.

		Protein AF association			
Gene	**Type**	**Difference**	**P value**	**Significant**	**FDR**
TNNT2*,++	*Trans* eQTL	-1.300	0.000973	yes	0.0185
NKX2-5**	*Trans* eQTL				
CYB5R3	*Trans* pQTL	0.260	0.113		0.195
NDUFB3^{+}	*Trans* pQTL	-0.195	0.611		0.726
HIBADH	*Trans* pQTL	-0.197	0.504		0.638
NDUFA9^{++}	*Trans* pQTL	-0.075	0.727		0.727
DLAT	*Trans* pQTL	-0.409	0.0284		0.0674
PPIF	NKX2-5 target	-0.124	0.711		0.727
MYL4**,++	NKX2-5 target	-1.480	0.00594	yes	0.0376
CKM^{++}	NKX2-5 target	-0.955	0.00499	yes	0.0376
MYL7	NKX2-5 target	-0.771	0.0188		0.0562
PGAM2^{++}	NKX2-5 target	-0.776	0.0207		0.0562
TNNC1*	NKX2-5 target	0.293	0.692		0.727
CYC1	NKX2-5 target	-0.723	0.0187		0.0562
ETFB**,++	NKX2-5 target	0.562	0.127		0.201
PRDX5	NKX2-5 target	-0.647	0.0142		0.0562
AK1	NKX2-5 target	-0.609	0.0561		0.107
ALDOA^{++}	NKX2-5 target	-0.644	0.0391		0.0825
TCAP*	NKX2-5 target	0.531	0.360		0.489
TOM1L2	NKX2-5 target	0.288	0.183		0.268

TNNT2, MYL4 and CKM were differentially expressed after proteome-wide FDR

[7] `https://www.ebi.ac.uk/pride/archive/projects/PXD006675`

correction (FDR < 0.05). Moreover, nine out of 13 functional NKX2-5 targets were negatively associated with AF as visualized in Figure 5.15a and we were also able to replicate the collective down regulation in AF with a GSEA P value of P = 2.43×10^{-3}. As NKX2-5 was not measured, we were not able to assess co-expression with the TF targets.

5.1.4.5 GTEx project

Finally, although the GTEx project [Gamazon et al., 2018] does not have any disease phenotypes, it offers mRNA as well as protein measurements [Jiang et al., 2020] across different tissues and our putative core genes showed high tissue-specific expression (Figure 5.16).

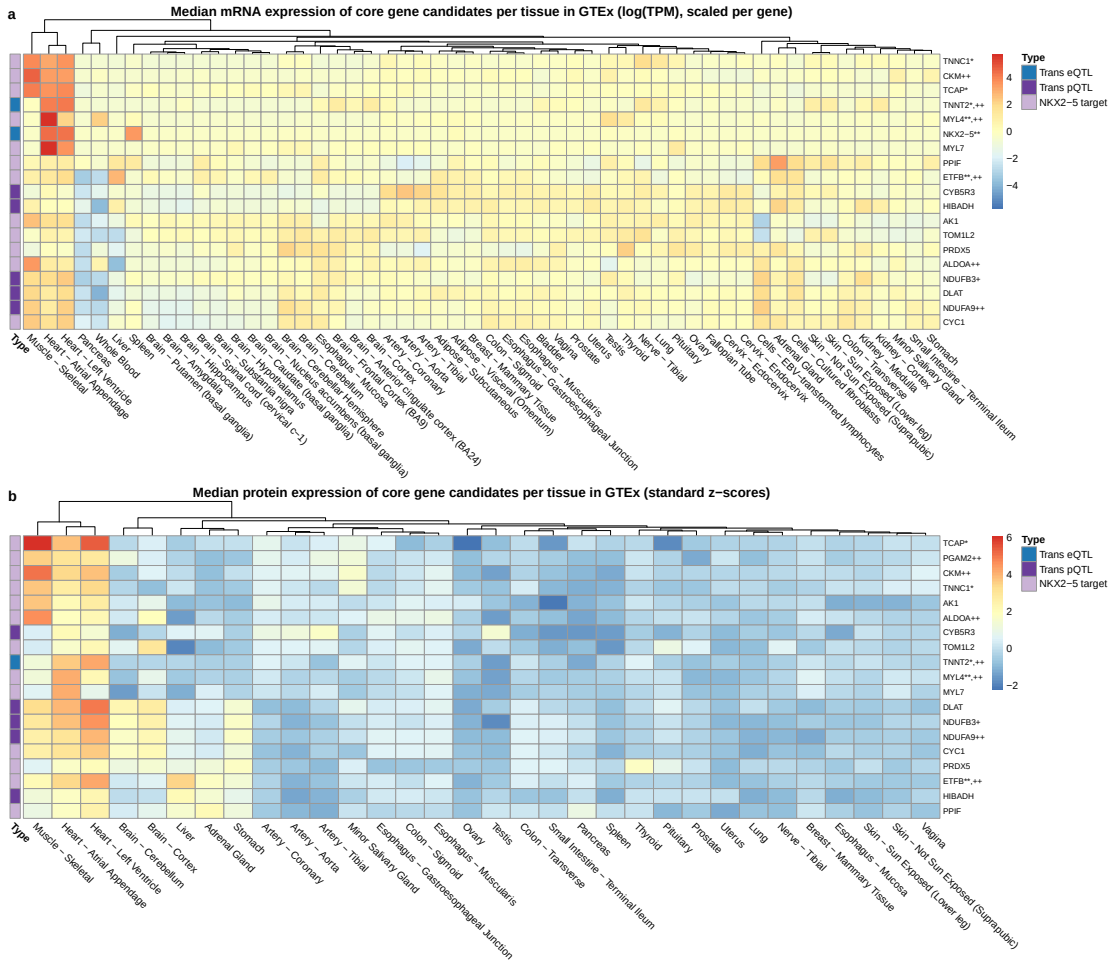


Figure 5.16: Tissue specific mRNA and protein expression profiles of putative core genes in GTEx.
Median mRNA (a) and protein (b) expression of putative core genes across all GTEx tissue types.
GTEx, genotype tissue expression;
*Mutation known to affect cardiovascular phenotypes; **Mutation known to affect arrhythmias; $^{+}$Differential expression or functional impairment for cardiovascular phenotypes; $^{++}$Differential expression or functional impairment for arrhythmias;

On mRNA level, transcript expression patterns of the heart tissues atrial appendage and left ventricle were almost identical and both were similar to skeletal muscle (Figure 5.16a) while those of whole blood and pancreas were most distinct. Rather low expression was observed for the rest of the tissues, especially coronary, aorta and tibial arteries.

The highest protein abundance of the core genes was observed in heart and muscle tissues followed by brain cerebellum/cortex, liver, adrenal gland and stomach tissues (Figure 5.16b). Other tissues showed mostly very low abundance of putative core genes.

The similarity of the heart and muscle tissues could be due to genes involved in shared biological functions such as TNNC1 and TCAP for contractile processes. Likewise, CKM is up-regulated on mRNA and protein level for the same tissues.

Finally, we were once again able to assess the co-expression of NKX2-5 and its targets using the RNA-seq data. In this case, independent of any disease context, we still saw a very strong correlation of all 13 functional targets with the NKX2-5 transcript ranging from $\rho = 0.43$ to $\rho = 0.70$ as shown in Figure 5.15.

5.2 Discussion

5.2.1 Summary

In this chapter, we established genome-wide polygenic risk scores as a proxy for accumulated genetic risk for AF which strongly represents *trans* effects and introduced the concept of core genes in the context of AF.

We carried out eQTS and pQTS analyses including a correction for *cis* effects in order to evaluate genetic contributions to gene variability. The following GSEA did not only recognize biological processes impacted by genetics, but also identified processes highly relevant for AF. Based on the leading edge genes from the GSEA results, we were able to prioritize a small subset of genes for *trans* QTL testing. By evaluating only independent GWAS SNPs together with those few transcripts and proteins, we narrowed down our search space such that *trans* analyses were feasible on our rather small clinical cohort which led us to discover two *trans* eQTLs as well as five *trans* pQTLs. *Trans* QTL genes were mostly involved in cardiac development, contractile function and energy metabolism.

We further investigated the *trans* eQTL between the GWAS SNP rs9481842 with unknown function and the TF NKX2-5 which is a key regulator of the heart development and strongly associated with AF and other arrhythmias. By estimating TFA from genome-wide transcriptomics data, we were able to show subtle but global expression changes with respect to the GWAS SNP. Using Western Blot analysis, we created independent protein measurements of NKX2-5 in remaining tissue samples and replicated the *trans* eQTL on protein level. We further identified 13 most likely functional NKX2-5 targets whose protein expression showed a collectively strong association with AF. We validated the disease link as well as the co-expression with NKX2-5 of those 13 targets in two independent cohorts each.

In summary, we were able to identify 21 putative AF core genes consisting of two *trans* eQTLs, five *trans* pQTLs and 13 NKX2-5 targets. We accumulated various sources of replicating the disease link using molecular data and literature annotations. Finally, our results can help to improve our understanding of the molecular disease pathology underlying AF and can be used for hypothesis generation and target prioritization. Additionally, our approach can also be used for other disease applications.

5.2.2 Targeted *trans*-QTL approach

Our PRS-based candidate selection approach leveraged specific characteristics of AF core genes [Boyle et al., 2017]. Accumulated *trans* effects were represented by the PRS and taken into account using the eQTS/pQTS analyses [Võsa et al., 2021], while correcting for the most important, often stronger *cis* associations per gene. The central role of genes at the core of complex biological networks was assured by the GSEA

analysis and by taking only leading edge genes into account. By choosing the GO biological processes gene set annotations instead of e.g. KEGG, we avoided introducing bias towards known human disease pathways, as the focus of the gene ontology is more to categorize groups of genes interacting in the same processes than deriving signature pathways for specific diseases. Finally, testing only independent AF GWAS loci for *trans* QTL associations with the selected candidates came with four desired properties: First, considering only really independent loci drastically reduces the amount of variants to test. Second, core genes are indeed expected to be enriched for *trans* QTLs. Third, even though by rather small effect sizes, *trans*-associated variants of core genes should also be enriched for GWAS hits. Finally, by choosing only AF-associated loci, we select for *trans* QTLs with a specific disease link.

Due to the tissue-specific omics data, we were able to identify *trans*-genetically influenced pathways which were highly specific for cardiac processes and comparable to results obtained from Wang et al. [2020], who investigated AF-associated genes and pathways from diverse sources of multi-omics data.
Enriched pathways on transcriptome level mostly identified processes involved in cardiac muscle contraction, signal transduction and metabolism, with the last also being relevant for the proteome enrichments. As to be expected, the same was true for the predominant function of the two identified *trans* eQTLs and the five *trans* pQTLs, which strongly associated with mitochondrial processes. All of these mechanisms have been described to be relevant for molecular changes in AF in previous studies [Opacic et al., 2016, Iwasaki et al., 2011, Ghezelbash et al., 2015].

5.2.3 Differences in transcriptomics and proteomics enrichment and *trans* QTLs

Similarly to the previous chapter analyzing *cis* QTLs, also *trans* QTLs showed significant differences on transcript and protein level as shown in Table 5.8. For *trans* eQTLs, *NKX2-5* was not originally measured but replicated on protein level in independent measurements using Western blot analysis. For *TNNT2*, the corresponding protein was associated with a P value of 0.16 showing at least a trend in the same direction of effect. None of the putative core genes had a significant *trans* association when testing a corresponding set of transcripts.
Additionally, we decided to consider different candidate genes on transcript and protein level. As an alternative, we could have prioritized genes based exclusively on the eQTS results, but then only 14 out of 23 genes tested for *trans* eQTLs would have been available for *trans* pQTL analysis as they were also measured on protein level.

Also in literature, a detailed description of the overlap between *trans* eQTLs and pQTLs is currently not available, especially not in heart tissue due to missing comparable datasets. Other studies in plasma have focused on *cis* eQTL/pQTL analysis [Sun et al., 2018] or considered both the overlap of *cis* and *trans* eQTLs/pQTLs but found no overlap for *trans* QTLs [Yao et al., 2018]. Finally, Suhre et al. [2017] replicated plasma

trans pQTLs in other proteomics datasets, but did not consider matched eQTL data.

Next to sample size restrictions influencing power, missing overlap could be due to other genetic factors confounding the associations, general effects of post-transcriptional regulation and also epigenetic or environmental influences.

5.2.4 NKX2-5 transcription factor network and transcription factor activity

The TF NKX2-5 plays a vital role in human heart development [Anderson et al., 2018] and is specifically known for its relation with cardiac arrhythmias, such as susceptibility to familial AF [Huang et al., 2013, Jhaveri et al., 2018]. While linking genetic variation in NKX2-5 to AF, those studies are lacking relevant mechanistic insights. Previous work describes altered *cis*-regulatory elements in NKX2-5 binding sites to overlap with GWAS hit for AF [Benaglio et al., 2019], which we now extend to *trans*-regulatory considerations of *NKX2-5* to an AF-associated GWAS variant.

In order to investigate downstream effects of the NKX2-5 TF, we assessed genome-wide consequences in the form of TFA as well as influences on specific target genes.
The TFA was inferred using promoter annotations [Frankish et al., 2019], epigenetic information on atrial appendage specific chromatin states [Roadmap Epigenomics Consortium et al., 2015], cardio myocyte specific HiC-promoter capture interactions [Montefiori et al., 2018], cardio myocyte specific NKX2-5 binding sites [Benaglio et al., 2019] and genome-wide transcriptomics data from our AFHRI-B cohort. The strong changes in TFA derived from functional NKX2-5 target gene expression and its strong correlation with the NKX2-5 transcript strongly suggest significant effect of the TF on target gene expression.
The TFA can be additionally viewed as an effective proxy for NKX2-5 protein expression, validated by the higher correlation of TFA with actual protein measurements derived by Western blot compared to correlation between NKX2-5 transcript and protein expression. Still, the strong correlation between NKX2-5 mRNA and estimated TFA confirmed consequences of genetic variation on NKX2-5 and its resulting consequences on NKX2-5 targets.
Due to its more direct link, a stronger rs9481842-NKX2-5 *trans* eQTL as well as pQTL, compared to the association of rs9481842 with the TFA, actually fit the assumed underlying causal mechanism.

The TFA was derived by a weighted sum of target gene expression assuming NKX2-5 to act as a transcriptional activator. Positive correlation between the *NKX2-5* transcript and the TFA as well as the NKX2-5 protein validated this hypothesis, which is also in line with previous studies [Anderson et al., 2018, Benaglio et al., 2019], even though NKX2-5 can potentially also act as a transcriptional repressor of *ISL1* as reported by Anderson et al. [2018] and Dorn et al. [2015].

Extending the general concept of TFA, we identified 13 specific NKX2-5 targets which were most likely strongly influenced by NKX2-5 and therefore also by the AF GWAS

SNP rs9481842. Indeed, most of the target genes are known to be involved in contractile function and metabolism which are highly relevant for AF pathophysiology. As they were derived exploiting regulatory links, public annotations and molecular data without taking the clinical phenotype into account, the collective down-regulation of NKX2-5 target protein abundance with respect to AF serves as an independent validation of the disease link.

5.2.5 Validation and replication

All analyses were performed on a to this day unique cohort integrating genotypes, transcript and protein measurements in human atrial tissue in a case control cohort of atrial fibrillation. Therefore, replication of all results in an independent cohort was not possible. Instead, we validated and replicated different parts of our findings whenever feasible.

Regulated pathways identified by our eQTS and pQTS gene set enrichment analyses were very similar to results obtained by Wang et al. [2020].
All *trans* eQTL and most of the *trans* pQTL genes were already mentioned to be either differentially expressed for or have mutations associated with arrhythmias or cardiovascular disease [Hershberger et al., 2009, Doll et al., 2017, Jhaveri et al., 2018, Huang et al., 2013, Kanaan et al., 2019, El-Hattab and Scaglia, 2016, Orr et al., 2016, Tu et al., 2014, Coats et al., 2018, Parvatiyar et al., 2012, Florian and Yilmaz, 2019, Hayashi et al., 2004, Dalal et al., 2015] as we described in detail in Table 5.11 and Table 5.13-5.14.
Specifically, we replicated the down-regulation of the 13 identified NKX2-5 targets in an independent RNA-seq dataset of left and right atrial appendage samples GSE128188 [Thomas et al., 2019] and in a cohort with left atrial appendage protein measurements PXD006675 [Doll et al., 2017] where AF case-control data was available.

Additionally, co-expression of *NKX2-5* and its targets was replicated in GTEx atrial appendage tissue and the GSE128188 dataset (Figure 5.15b).

5.2.6 Limitations

Similarly to the *cis* QTL analysis, the same challenges with respect to measurement techniques, tissue heterogeneity and cell type composition apply.
Specifically, the missing coverage of transcription factors on protein level might lead to a lack of discovering corresponding *trans* pQTLs, as was the case with NKX2-5 which was additionally measured via Western blot analysis.
Additionally, the sample and effect size considerations are especially important in the context of *trans* QTL analyses and strongly limit the power. Hence, it was only our targeted *trans* QTL approach which made the analysis feasible with this small sample size.

Focusing more deeply on the disease aspect, the relevance of right atrial appendage

tissue has to be questioned, as AF is known to originate from the pulmonary vein ostia. However, several studies have proved the relevance of atrial appendage tissue to assess AF GWAS hits [Roselli et al., 2018] or derived disease-mechanisms and candidate genes [Mayr et al., 2008, Martin et al., 2015]. Specifically keeping in mind that those tissue samples are not post-mortem specimens with potentially impaired metabolism, they serve as an accurate proxy for analyzing atrial impairment.

5.2.7 Conclusion

We derived a PRS-based candidate selection approach which uses gene set enrichment analysis in order to prioritize genes for *trans* eQTL and pQTL testing. By integrating multiple data sources and integrating them with molecular cohort data, we were able to reduce the search space enough to make such kind of analyses feasible in comparably small clinical cohorts. Of course, similar approaches can be generalized and applied to other traits as well.

Most importantly, being the first study to investigate consequences of genetic variation in human heart tissue on protein level, we provided new insights into *trans* regulation of proteins in atrial tissue by detecting two *trans* eQTLs and five *trans* pQTLs.
We further proposed mechanistic insights on the link between the GWAS SNP rs9481842 and the TF NKX2-5 as well as evaluated the disease association of AF for 13 specific NKX2-5 targets and replicated their connection to AF in two independent datasets.

In conclusion, we identified putative core genes of AF inspired by the omnigenic model. We investigated AF specific genetic variation and its consequences on transcript and protein levels to better understand molecular changes in AF. Our analyses suggest strong candidates for future follow-up analyses including molecular characterizations and experimental gene prioritization to enable clinical translation.

6 Multi-omics gene set enrichment for atrial fibrillation

Atrial fibrillation (AF) represents the most common atrial arrhythmia in the general population [Benjamin et al., 2019]. However, the disease is highly complex and despite similar cardiovascular risk factors, the underlying molecular pathology differs severely. Possible mechanisms that are involved and may also contribute differently to different AF subtypes range from ion channel modulation, inflammation, atrial fibrosis to cardiac developmental pathways [Staerk et al., 2017].

Additionally, proteomics analyses in human atrial tissue identified de-regulated genes involved in metabolism, specifically connected to mitochondria [Tu et al., 2014]. Since especially the tissue-specific molecular mechanisms underlying AF remain difficult to investigate, more research is needed to gain a better understanding and derive potential therapeutic targets.
With atrial tissue transcript and protein measurements available in the AFHRI-B cohort, we aim to better understand those mechanisms, which molecular phenotypes are affected and if certain effects manifest on different omics layers. For details on the cohort, the available molecular data, its preprocessing and the disease phenotypes, please refer to section 2.1.

The work of this chapter is currently being prepared for publication: Ines Assum[†], Julia Krause[†], Renate Schnabel, Tanja Zeller and Matthias Heinig. **Multi-omics pathway analysis in atrial fibrillation.** (*Manuscript in preparation*).

6.1 Multi-omics differential expression analyses

We first evaluated transcriptomics and proteomics measurements for differential expression of prevalent and incident AF in right atrial appendage tissue samples from our AFHRI-B cohort as visualized in Figure 6.1. Using Benjamini-Hochberg FDR correction per omic and AF subtype, none of the genes were differentially expressed on either mRNA or protein level at any common FDR significance level like 0.05, 0.1 or even 0.2 (Table 6.1 and Table 6.2).

Due to the very small effect sizes, even when applying less stringent significance cutoffs, we could not identify a clear group of genes that could be followed up on. Alternatively, pathway enrichment analysis can be more suitable to infer regulated biological processes

while leveraging our multi-omics data. For this, different methods were available for evaluation, including some unpublished extensions to existing methods that we will introduce in the following sections.

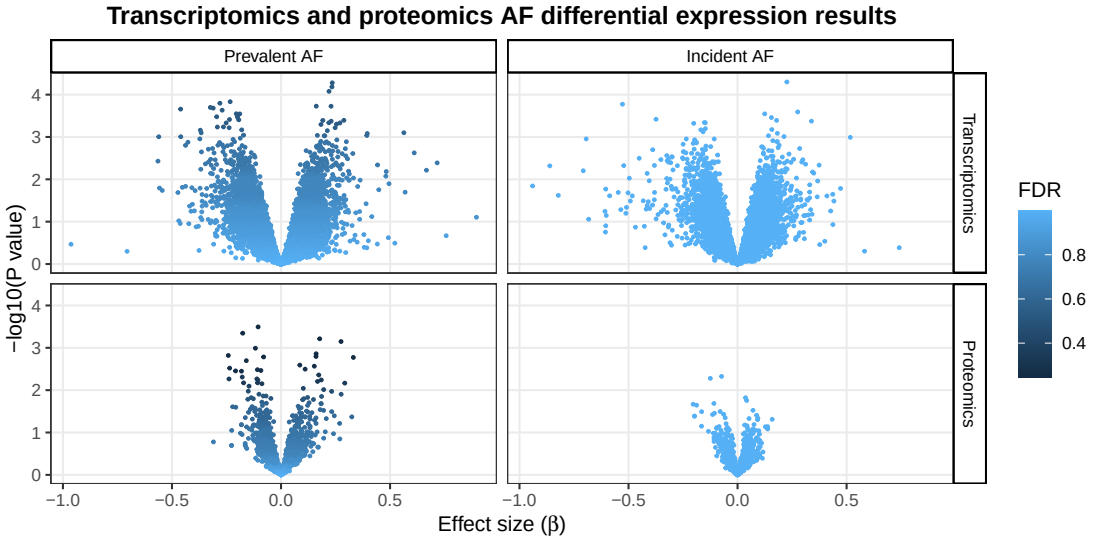


Figure 6.1: Transcriptomics and proteomics differential expression results.
Transcriptomics and proteomics differential expression results for prevalent and incident AF in the AFHRI-B right atrial appendage tissues. Two-sided t-tests were applied as part of a multiple linear regression model including covariates age, sex, BMI, diabetes, sysBP, hypertension medication, myocardial infarction, smoking status, RIN-score/protein concentration and fibroblast-score to correct for confounding factors and multiple testing correction was carried out by inferring the false discovery rate based on the Benjamini-Hochberg procedure. AF, atrial fibrillation; BMI, body mass index; sysBP, systolic blood pressure; RIN, RNA integrity number;

Table 6.1: Transcriptomics AF differential expression results.
Transcriptomics differential expression results for prevalent and incident AF in the AFHRI-B right atrial appendage tissues. Two-sided t-tests were applied as part of a multiple linear regression model including covariates age, sex, BMI, diabetes, sysBP, hypertension medication, myocardial infarction, smoking status, RIN-score and fibroblast-score to correct for confounding factors and multiple testing correction was carried out by inferring the false discovery rate based on the Benjamini-Hochberg procedure.
AF, atrial fibrillation; BMI, body mass index; sysBP, systolic blood pressure; RIN, RNA integrity number; FDR, false discovery rate;

Transcript	Prevalent AF association				Transcript	Incident AF association			
	β	T value	P value	FDR		β	T value	P value	FDR
NBEAL2	0.235	4.31	5.26×10^{-5}	0.534	*TMEM116*	0.226	4.30	5.02×10^{-5}	0.997
KLF8	0.234	4.25	6.57×10^{-5}	0.534	*CAPN6*	-0.528	-3.96	1.69×10^{-4}	0.997
OR51M1	0.220	4.18	8.36×10^{-5}	0.534	*IL18R1*	0.276	3.84	2.56×10^{-4}	0.997
ST8SIA6-AS1	-0.233	-4.02	1.47×10^{-4}	0.534	*KCNJ1*	0.125	3.81	2.85×10^{-4}	0.997
UCMA	-0.280	-3.99	1.60×10^{-4}	0.534	*LOC105374887*	0.155	3.75	3.47×10^{-4}	0.997
PIGS	0.162	3.94	1.87×10^{-4}	0.534	*HTR2A*	-0.374	-3.72	3.82×10^{-4}	0.997
ZNF337	0.230	3.94	1.88×10^{-4}	0.534	*LOC105378431*	0.179	3.70	4.06×10^{-4}	0.997
GNAL	-0.324	-3.92	2.01×10^{-4}	0.534	*LOC105370728*	0.339	3.69	4.24×10^{-4}	0.997
HTR1F	-0.311	-3.91	2.08×10^{-4}	0.534	*CHST10*	-0.152	-3.67	4.56×10^{-4}	0.997
ETNPPL	-0.460	-3.90	2.20×10^{-4}	0.534	*ZCCHC3*	-0.148	-3.66	4.61×10^{-4}	0.997

Table 6.2: Proteomics AF differential abundance results.
Proteomics differential abundance results for prevalent and incident AF in the AFHRI-B right atrial appendage tissues. Two-sided t-tests were applied as part of a multiple linear regression model including covariates age, sex, BMI, diabetes, sysBP, hypertension medication, myocardial infarction, smoking status, protein concentration and fibroblast-score to correct for confounding factors and multiple testing correction was carried out by inferring the false discovery rate based on the Benjamini-Hochberg procedure.
AF, atrial fibrillation; BMI, body mass index; sysBP, systolic blood pressure; FDR, false discovery rate;

Protein	Prevalent AF association				Protein	Incident AF association			
	β	T value	P value	FDR		β	T value	P value	FDR
PRDX1	-0.105	-3.80	3.21×10^{-4}	0.248	HEBP1	-0.0733	-2.92	4.74×10^{-3}	0.995
PGAM2	-0.175	-3.70	4.52×10^{-4}	0.248	GAA	-0.125	-2.88	5.25×10^{-3}	0.995
GGT5	0.177	3.60	6.12×10^{-4}	0.248	GAPDH	0.0366	2.49	1.51×10^{-2}	0.995
LBP	0.275	3.56	7.11×10^{-4}	0.248	PSMA1	0.0418	2.43	1.75×10^{-2}	0.995
ISOC1	-0.117	-3.44	1.02×10^{-3}	0.248	FBLN1	-0.108	-2.37	2.06×10^{-2}	0.995
EMILIN1	0.161	3.34	1.38×10^{-3}	0.248	OSBPL7	-0.204	-2.35	2.15×10^{-2}	0.995
MYBPHL	-0.242	-3.31	1.52×10^{-3}	0.248	KRT19	-0.188	-2.33	2.27×10^{-2}	0.995
VCAN	0.161	3.30	1.60×10^{-3}	0.248	CYB5R1	0.0710	2.22	3.00×10^{-2}	0.995
PRDX6	-0.0797	-3.29	1.64×10^{-3}	0.248	SELENBP1	-0.0796	-2.19	3.20×10^{-2}	0.995
ENTPD5	0.332	3.28	1.69×10^{-3}	0.248	KRT8	-0.165	-2.19	3.22×10^{-2}	0.995

6.2 Multi-omics pathways enrichment analysis

Differential expression analysis has become one of the most commonly used tools to interpret changes in molecular data. However, if very few or too many genes reach significance, interpretation of the results becomes extremely challenging. Instead of looking at each single gene individually, we are interested in a more general up- or down-regulation of multiple interacting genes, that e.g. are influencing the same molecular mechanism. In this case, smaller but consistent regulation of multiple genes often gives a better understanding of possible underlying biological mechanisms.

6.2.1 Inferring regulated biological pathways from differential expression results

Prior knowledge has been accumulated in numerous databases with pathway annotations such as Gene Ontology [Carbon et al., 2019, Ashburner et al., 2000], KEGG, STRING [Szklarczyk et al., 2021] and can be used for overrepresentation analysis (ORA) to infer regulated processes.

Several challenges have been addressed by different methods. With their GSEA approach, Subramanian et al. [2005] quantified the enrichment of specific gene set at the top or bottom of a ranked list. By taking into account the ranking instead of dividing genes into significant and non-significant genes, they circumvented the problem of picking arbitrary significance thresholds.
Bayesian models, such as MGSA proposed by Bauer et al. [2010] and MONA by Sass et al. [2013], greatly improved on better estimating the importance of hierarchical or

strongly overlapping gene sets. Notably, MONA was the first approach integrating multiple omics.

As we introduced in the previous chapters [Assum et al., 2022a], regulatory mechanisms can take effect at different steps of gene expression or even be influenced by environmental factors. Furthermore, the observation of specific types of regulation may be hindered by technical restrictions of specific measurement techniques, creating the strong need for multi-omics pathway considerations

However, while the Bayesian models outperform GSEA in general, the impact of choosing different significance cutoffs is unclear, also contrary to GSEA, both Bayesian approaches do not include the direction of effect. Also, little is known about how different coverages, or missing values in general, change the performance of different methods. Furthermore, individual model performance may strongly depend on types of regulation in the multi-omic data and currently, no specific recommendations for the use of different approaches exist.

To evaluate those questions and make specific recommendations, we propose a multi-omics simulation study focused on integrating transcript and protein measurements to compare the different methods across different simulation scenarios, significance cutoffs, protein coverages and correlation between omics.

First of all, we introduce different extensions to the existing MONA models to achieve best performance in recovering regulated processes in multi-omics data.

6.2.2 Extensions of the existing MONA models

The unique strength of MONA is the direct integration of multiple omics layers, which we described in the methods section 3.3.4. However, with the original design taking only P value cutoffs to evaluate the active/not-active observation of a gene, direction of effect was lost. This is a major restriction, as we will show later on.
Moreover, the original implementation was restricted to two modalities with the additional constraint that only nodes that were available for the first omic layer could also be evaluated for the second, but not vice versa.

6.2.2.1 Adding direction of effect to the pathway enrichment for MONA

Any kind of gene set annotation can be realized by the Bayesian network by manipulation the edges between the term and the hidden layer. In this case, instead of having one term per pathway which is connected to one representation of a gene, we can split each term T_j, $j = 1, \ldots, n$ into an up-regulated T_i^+ and down-regulated T_i^- version. Accordingly, we also need two hidden nodes H_i^+ and H_i^- for each original node H_i, $i = 1, \ldots, m$, requiring also up-regulated and down-regulated observations $O_i^{mRNA,+}$ and $H_i^{mRNA,-}$ for the first, as well as $O_i^{prot,+}$ and $H_i^{prot,-}$ for the second omic layer as

visualized in Figure 6.2. All observations of one omic layer, both up-regulated and down-regulated still share the same error rates α and β.

Instead of just taking into account if a regulation was observed or not for the original cooperative model, we now supply additional information about the direction of effect. As an example of t-test results with a significance threshold of 0.05, this would read then:

$$O_i^{mRNA,+} = \begin{cases} 1 & \text{if} \quad P < 0.05 \wedge T > 0 \text{ for mRNA}_i \\ 0 & \text{otherwise} \end{cases} \tag{6.1}$$

$$O_i^{mRNA,-} = \begin{cases} 1 & \text{if} \quad P < 0.05 \wedge T < 0 \text{ for mRNA}_i \\ 0 & \text{otherwise} \end{cases} \tag{6.2}$$

$$O_i^{prot,+} = \begin{cases} 1 & \text{if} \quad P < 0.05 \wedge T > 0 \text{ for protein}_i \\ 0 & \text{otherwise} \end{cases} \tag{6.3}$$

$$O_i^{prot,-} = \begin{cases} 1 & \text{if} \quad P < 0.05 \wedge T < 0 \text{ for protein}_i \\ 0 & \text{otherwise} \end{cases} \tag{6.4}$$

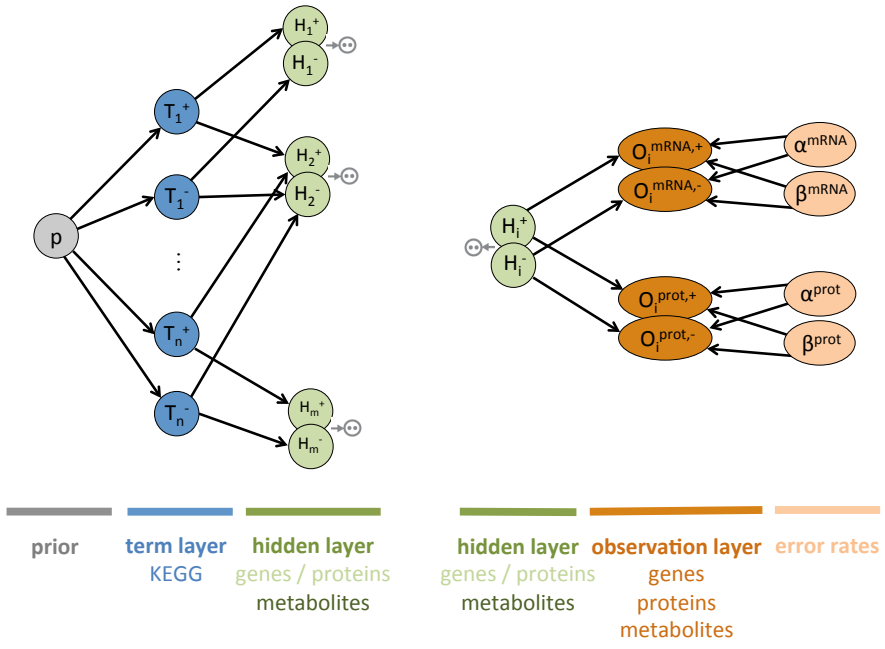


Figure 6.2: MONA cooperative model direction of effect extension.
By doubling all term, hidden and observation nodes into an up-regulated and a down-regulated node, we can include the direction of effect information from the differential analysis into the observation layer of the Bayesian network. Therefore, separate posteriors for up- and down-regulation of the different terms can be estimated. Error rates α and β are shared across up- and down-regulated observation nodes. MONA, multi-level ontology analysis; α, false positive rate, β, false negative rate;

6.2.2.2 Extending the cooperative MONA model to a more flexible, three omics version

The original cooperative model implementation only allowed for hidden nodes of the second omic layer to be evaluated, if their corresponding node in the first omic layer was observed. By extending the hidden and observation layer, the model can be extended to three molecular modalities. Additionally, we added missingness information for the first omic layer. That way, also omics that don't map to common hidden gene nodes, such as metabolites can be analyzed together with e.g. transcript and protein data as visualized in Figure 6.3.

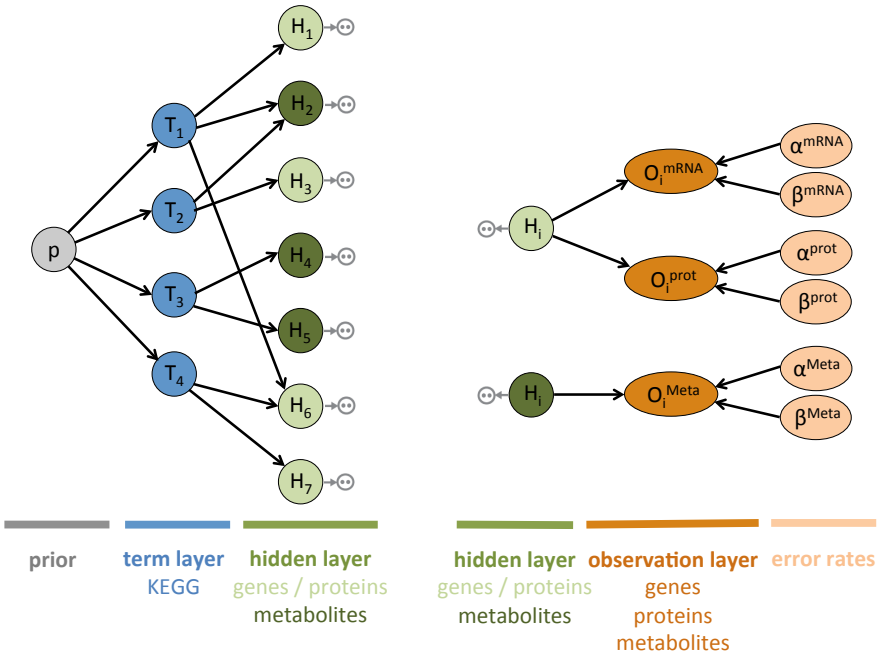


Figure 6.3: MONA cooperative model extension to three omics modalities.
The cooperative MONA model was extended to include three modalities. An addition of missingness information in all considered omic levels also allows the integration of different omics with identifiers that do not map to the same hidden nodes, such as genes and metabolites.
MONA, multi-level ontology analysis; α, false positive rate, β, false negative rate;

6.3 Multi-omics simulation study

With our extensions of the MONA models that can now take direction of effect into account, we have addressed a major drawback of this method compared to GSEA. While the cooperative MONA methods shows a clear benefit of analyzing multiple omics at the same time, little is known about the comparable performance of ad-hoc integration of single-omic methods on pathway level.
Furthermore, the effects of coverage or the choice of a significance threshold for MGSA and MONA have not been considered in detail. Similarly, different characteristics of pathway activations might strongly influence the performance of the various approaches.

In order to give sound recommendations on which method performs best under different assumptions, we performed a simulation study that evaluated and compared all models on the same simulated data. Next to different underlying pathway activations, we included simulation scenarios incorporating predefined false positive and false negative error rates.
Based on the idea of power analysis, we simulated summary statistics that can be used to represent correlated multi-omics differential expression results with predefined error rates. The false positive rate α should represent the false positive hits in the analysis, as well as the false negative rate β should represent the false negative, i.e. not discovered, hits.

6.3.1 Sampling procedure

Key assumption of our simulation study was to take KEGG gene set annotations as a ground truth of how genes are interacting with respect to important biological processes. Based on those annotations, we can simulate data to represent summary statistics from a differential expression analysis. By analyzing those data, we can evaluate the performance of different pathway enrichment methods. The complete procedure is visualized in Figure 6.4.
In this context, we also often refer to gene sets as pathways. While this is technically not correct, we do use the gene sets as a proxy for real world pathways representing e.g. disease-specific processes or signaling pathways.

Starting with a subset of regulated pathways, we can simulate all genes not involved in any of those pathways with Z scores of a background distribution of non-significant hits. Gene from regulated pathways can then be sampled from a shifted distribution.

The top right of Figure 6.4 further shows the exact derivation of Z score rankings.
Here, we chose the background distribution Q_{bg} being normally distributed with mean 0, i.e. $Q_{bg} \sim \mathcal{N}(0,\ \sigma^2_{bg})$. If the signal distribution is also a normal distribution with variance σ^2_{sig}, then we can shift the signal distribution by a mean μ_{sig} to satisfy the error rates α and β. In our case, we want to simulate the two sided background distribution,

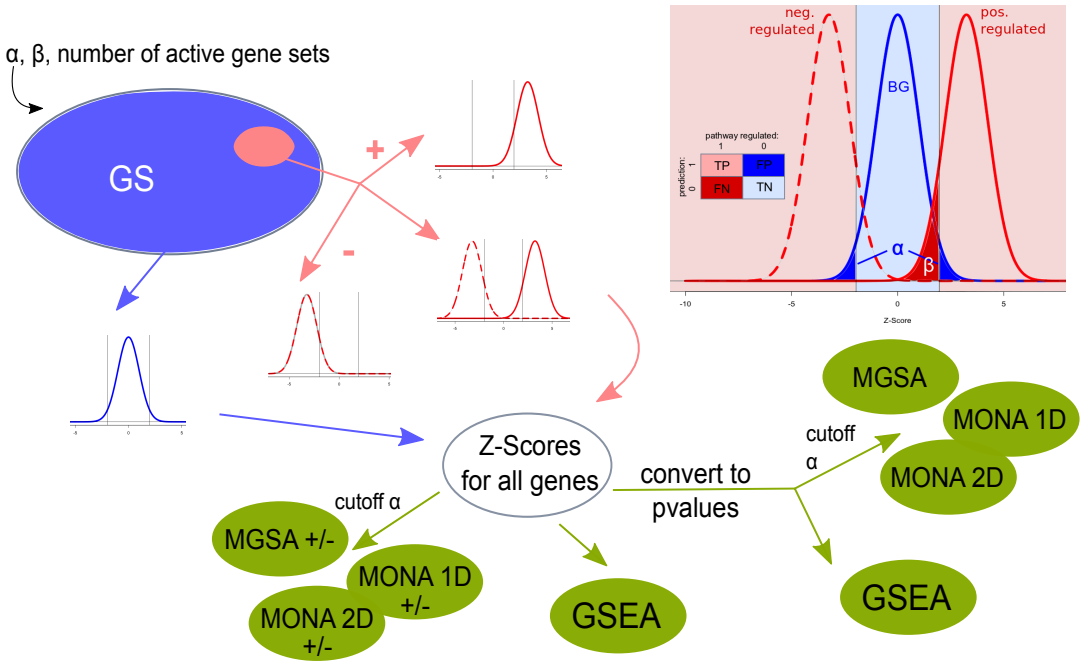


Figure 6.4: Graphical abstract of the simulation study including sampling procedure.
Summary statistics for unregulated and regulated genes are modeled as overlapping normal distributions characterized by parameters dependent on pre-defined error rates.
Based on a ground truth of active gene sets with specific signs, Z scores representing statistical testing from the background or signal distribution are assigned to every gene.
These simulated summary statistic allow the evaluation of different models by either taking Z score cutoffs including the sign or the absolute Z scores to mimic P values. MONA 1D denotes the single-omic, MONA2D the multi-omics cooperative model implementation of MONA. MGSA +/-, MONA 1D +/- and MONA 2D +/- indicate the extensions which take into account direction of effect.
GS, gene set; BG, background; TP, true positives; FP, false positives; FN, false negatives; TN, true negatives. MONA, multi-level ontology analysis; MGSA, model-based gene set analysis; GSEA, gene set enrichment analysis;

match it with only a positive signal distribution and assign the sign of the signal distribution later. Therefore, for μ_{sig}, we need to take into account that all but half of the α tail of the background distribution needs to overlap with the β tail of the signal distribution.
Hence, we choose μ_{sig} based on the cutoff γ_α based on the quantiles of the normal distribution and the corresponding β error rate, i.e.

$$\gamma_\alpha = -q(\frac{\alpha}{2},\ \mu = 0,\ \sigma^2 = \sigma^2_{bg}) \tag{6.5}$$

$$\mu_{sig} = \gamma_\alpha - q(\beta,\ \mu = 0,\ \sigma^2 = \sigma^2_{sig}). \tag{6.6}$$

Furthermore, if we consider two correlated omics T and P with an overall correlation of ρ, we can simulate the summary statistics of both omics together based on the sum of multiple bivariate normal distributions. For each gene, we consider if it belongs to the background or signal distribution, while allowing genes to be active in both or only single omics, i.e.

$$Q(T,\ P) = Q_{bg}(T,\ P) + Q_{T_{sig}P_{bg}}(T,\ P) + Q_{T_{bg}P_{sig}}(T,\ P) + Q_{T_{sig}P_{sig}}(T,\ P) \tag{6.7}$$

$$Q_{bg}(T,\ P) = \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_{bg}^2 & \rho\cdot\sigma_{bg}^2\\ \rho\sigma_{bg}^2 & \sigma_{bg}^2\end{pmatrix}\right) \tag{6.8}$$

$$Q_{T_{sig}P_{bg}}(T,\ P) = \mathcal{N}\left(\begin{pmatrix}\mu_{sig}\\0\end{pmatrix}, \begin{pmatrix}\sigma_{sig}^2 & \rho\cdot\sigma_{sig}\cdot\sigma_{bg}\\ \rho\cdot\sigma_{bg}\cdot\sigma_{sig} & \sigma_{bg}^2\end{pmatrix}\right) \tag{6.9}$$

$$Q_{T_{bg}P_{sig}}(T,\ P) = \mathcal{N}\left(\begin{pmatrix}0\\\mu_{sig}\end{pmatrix}, \begin{pmatrix}\sigma_{bg}^2 & \rho\cdot\sigma_{bg}\cdot\sigma_{sig}\\ \rho\cdot\sigma_{sig}\cdot\sigma_{bg} & \sigma_{sig}^2\end{pmatrix}\right) \tag{6.10}$$

$$Q_{T_{sig}P_{sig}}(T,\ P) = \mathcal{N}\left(\begin{pmatrix}\mu_{sig}\\\mu_{sig}\end{pmatrix}, \begin{pmatrix}\sigma_{sig}^2 & \rho\cdot\sigma_{sig}^2\\ \rho\cdot\sigma_{sig}^2 & \sigma_{sig}^2\end{pmatrix}\right). \tag{6.11}$$

By additionally adjusting the choice of the variance for the background and signal distribution, we can ensure proper sampling for a diverse combination of error rates.

6.3.1.1 Implementation of the summary statistics simulation

Based on the theoretical distributions shown in equations 6.7-6.11, we can now simulate summary statistics while always checking how much the actual sampled representations match the original, predefined error rates.

For each scenario we use the following procedure:

1. **Ground truth:**
   For each replicate, we define the truly regulated pathways incl. the sign of regulation per omic based on the current simulation scenario. Accordingly, active genes inherit the sign of regulation based on the sign of the corresponding pathway(s).
2. **Simulate summary statistics:**
   The background multi-omics summary statistics are sampled according to $Q_{bg} \sim \mathcal{N}(0,\ \sigma_{bg}^2)$ with specific protein coverage and correlation parameter ρ and then the scores of regulated genes are overwritten based on equations 6.9, 6.10 and 6.11.
   Summary statistics for the different combinations of pre-defined error rates for one replicate are all sampled according to the same ground truth.
3. **Diagnostics:**
   Due to the sampling, small deviations between the preset and actual error set can occur. In case of larger discrepancies, sampling is repeated and detailed diagnostics are recorded for all datasets.

6.3.2 Simulation scenarios

At the core of our simulation study, we were interested in evaluating different type of simulation scenarios. We always include scenarios, which are as close to reality as possible to make decisions on how to treat existing data. Additionally, further simulations also allow us to investigate the influence of different factors and the corresponding impact on the results.

In summary, we simulated data for different main factors, including:

- **Regulated pathways:**
  Pathways can be either regulated on both omics (shared effects) or different omics can also have different pathways that are regulated. Here, we evaluated an independent scenario, where three pathways were regulated in each omic independently and a shared scenario, with six pathways regulated each in both omics.
- **Coverage (of e.g. proteomics):**
  Usually, one omics is more challenging to measure. However, lack of identification of effects might be to the restriction in observation and might majorly influence the need and gain of multi-omics integration. We therefore evaluate simulation settings based on actual measured proteins, 30 % of genes measured on proteomics level and also if all genes are available on transcript and protein level.
- **Correlation (between omics):**
  Our simulation procedure includes a parameter for the strength of correlation between the different omics. In this case, we simulate different datasets for weak (ρ = 0.2) correlation between omics as well as a correlation of ρ = 0.3 and high correlation of ρ = 0.8.

An example what the simulated Z scores for both omics look like is shown in Figure 6.5.

6.3.3 Evaluated methods and performance measurements

This simulation study focused on methods applied to summary statistics to evaluate the enrichment or overrepresentation of differentially expressed genes in specific categories of gene sets. While the MONA cooperative model is able to directly integrate multiple omics, GSEA, MGSA and the single-omic MONA results need to be integrated on gene set level after running it on each omic layer separately. We propose two versions of ad-hoc integration, considering additive effects, i.e. checking for gene sets found active in either omic (OR combination) as well as overlapping effects, i.e. counting only gene sets found active in both omics (AND combination).
Additionally, we point out, that for GSEA we use the running-sum approach to derive the enrichment score, which is then evaluated based on a weighted Kolmogorov-Smirnov-like statistic. Empirical P values were derived by permuting gene lists as

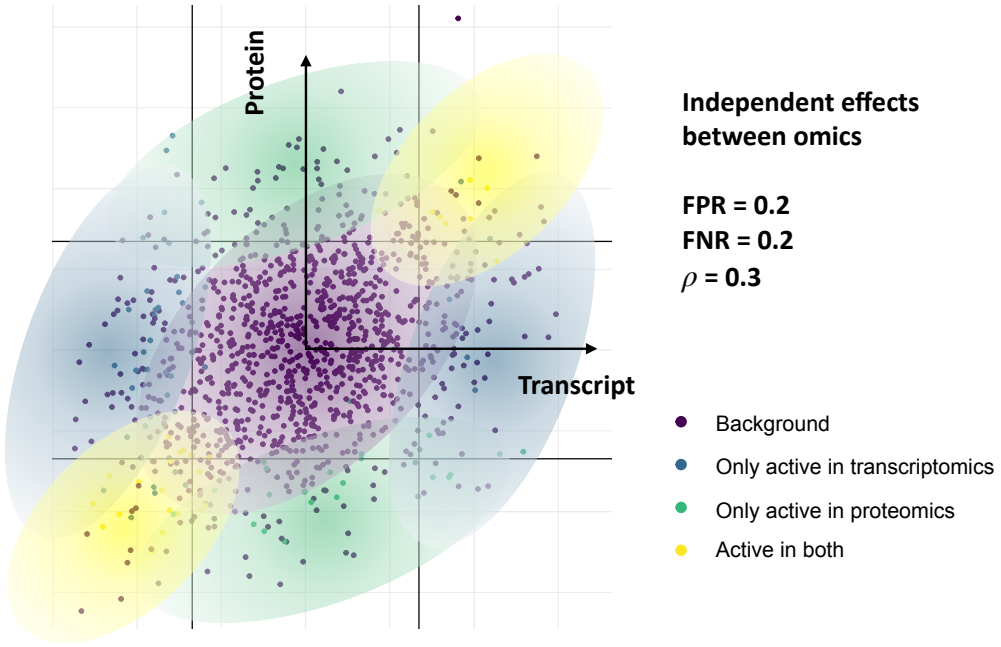


Figure 6.5: Example of the simulated Z score summary statistics.
Simulation of correlated transcript and protein test statistics based on the combination of four bivariate normal distributions describing genes that are inactive in both omics, active only in transcriptomics, active only in proteomics or active in both omics.
FPR, false positive rate; FNR, false negative rate;

implemented in the fgsea package[1] [Korotkevich et al., 2019] rather than permuting class labels of the individual level expression data which was the focus of the original publication [Subramanian et al., 2005].

For GSEA results, significant gene sets were defined based on a FDR-corrected (Benjamini-Hochberg procedure) P value with a significance cutoff of either 0.05 or 0.01. MGSA and the different MONA models estimate posterior probabilities for a gene set to be active. Posterior cutoffs of 0.5 and 0.6 were used to determine active gene sets. All the evaluated methods and their characteristics are also summarized in Table 6.3.

To evaluate the performance of the different methods, we focused on the capabilities to recover the gene sets that were simulated to be active. We therefore used classical metrics for classification methods such as accuracy, the F-score, sensitivity/recall, specificity, precision and false positive/negative as well as the false discovery rate.

Since we evaluated many different simulation scenarios using different parameter settings, different combinations of preset error rates and many replicates, we considered average rates calculated across all replicates for each simulation scenario.
Additionally, we considered the fraction of replicates, where the exact simulated ground truth was recovered.

[1] `https://bioconductor.org/packages/release/bioc/html/fgsea.html`

Table 6.3: Overview of the different methods evaluated.
GSEA, gene set enrichment analysis, MGSA, model-based gene set analysis, MONA, multi-level ontology analysis.

Method	Data (type of ranking)	Omics	Integration level	Input (data type)	Direction of effect
GSEA	P values	Single-omic	Gene sets (OR / AND)	Continuous	No
Signed GSEA	Signed P value or test statistic	Single-omic	Gene sets (OR / AND)	Continuous	Yes
MGSA	P values	Single-omic	Gene sets (OR / AND)	Binary	No
Signed MGSA	Signed P value or test statistic	Single-omic	Gene sets (OR / AND)	Binary	Yes
Single MONA	P values	Single-omic	Gene sets (OR / AND)	Binary	No
Single MONA signed	Signed P value or test statistic	Single-omic	Gene sets (OR / AND)	Binary	Yes
MONA	P values	Multi-omics	Genes	Binary	No
Signed MONA	Signed P value or test statistic	Multi-omics	Genes	Binary	Yes

6.3.4 Benchmarking results

To evaluate the performance of the different methods, we first compared the results for each individual method across the different simulation scenarios. Next, we assessed differences and similarities between the different methods to conclude the best approach in a specific context.
Results in this main text have been restricted to the most important comparisons. More detailed visualizations are provided in the Appendix A.

6.3.4.1 GSEA performance

GSEA applied to P value rankings We first evaluated GSEA performance when applied to rankings without direction of effect in a scenario where all gene sets were active in both omics (Supplementary Figure A.1A). Overall accuracy was higher for the combination using overlapping effects, i.e. the AND integration as shown in Supplementary Figure A.1Aa and A.1Ae. Both evaluation approaches have in general a very high sensitivity in picking up activated pathways until very high false positive rates (Supplementary Figure A.1Ab and A.1Af) in the simulated scenarios, at the same time, as previously discussed [Subramanian et al., 2005, Bauer et al., 2010], GSEA struggles with many false positive findings. Interestingly, this improves with higher preset false positive and false negative error rates in the simulation scenarios (Supplementary Figure A.1Ac). Compared to looking for additive effects using the OR combination of single-omic results, evaluating overlapping active pathways with the AND combination reduces the number of false positives (Supplementary Figure A.1Ag). Less

false positives outweigh the small disadvantage in sensitivity of the OR combination compared to the AND combination, however, when considering exact matches, GSEA is almost never able to recover the simulated ground truth as representatively shown for a coverage of 30 % and a correlation of 0.3 in Figure 6.6Aa.

Overall, the GSEA sensitivity is more attenuated by high false positive rates of the simulation setting than by high false negative rates.

Similar results were found for different pathway active in different omics as shown in Supplementary Figure A.1B. Due to the different activations in the different omics, sensitivity is much lower when evaluating the AND combination counting only overlapping pathways found in both omics (Supplementary Figure A.1Bb and A.1Bf). On the other hand, specificity is greatly increased as visualized in Supplementary Figure A.1Bg compared to Supplementary Figure A.1Bc.

When comparing the shared and independent effects in different omics scenarios, accuracy is improved for the independent effects by gaining power in specificity at the cost of much lower ability to identify actual activated pathways.

GSEA applied to rankings including direction of effect Instead of classical P value rankings, GSEA can also be applied to rankings based on either signed P values or any other suitable test statistic with direction of effect. While this improves the already high sensitivity for simulation scenarios with high preset false positive and false negative error rates (Supplementary Figure A.2Ab, A.2Af and A.2Bb), performance using the AND combination approach remains unchanged with comparable low sensitivity for the scenarios with independent pathways activated in different omics (Supplementary Figure A.2Bf).

Conversely, specificity remains unchanged (Supplementary Figure A.2Bg) and is slightly decreased in all but the aforementioned scenario (Supplementary Figure A.2Ac, A.2Ag and A.2Bc), which in summary leads to a very slight decrease in accuracy (Supplementary Figure A.2Aa, A.2Ae and A.2Be).

Dependence on GSEA enrichment score FDR cutoff Finally, we also evaluated the dependence of GSEA performance on different FDR cutoffs (Figure 6.6). In general, more stringent cutoffs showed slightly better results. For active gene sets shared across omics selecting only the top six gene sets improved performance, especially when using the information of direction of effects (Figure 6.6Ac). For independent gene sets active across omics, GSEA performance was overall very poor but best including direction of effect and combining information across omics by adding up effects (OR combination) at a FDR < 0.01.

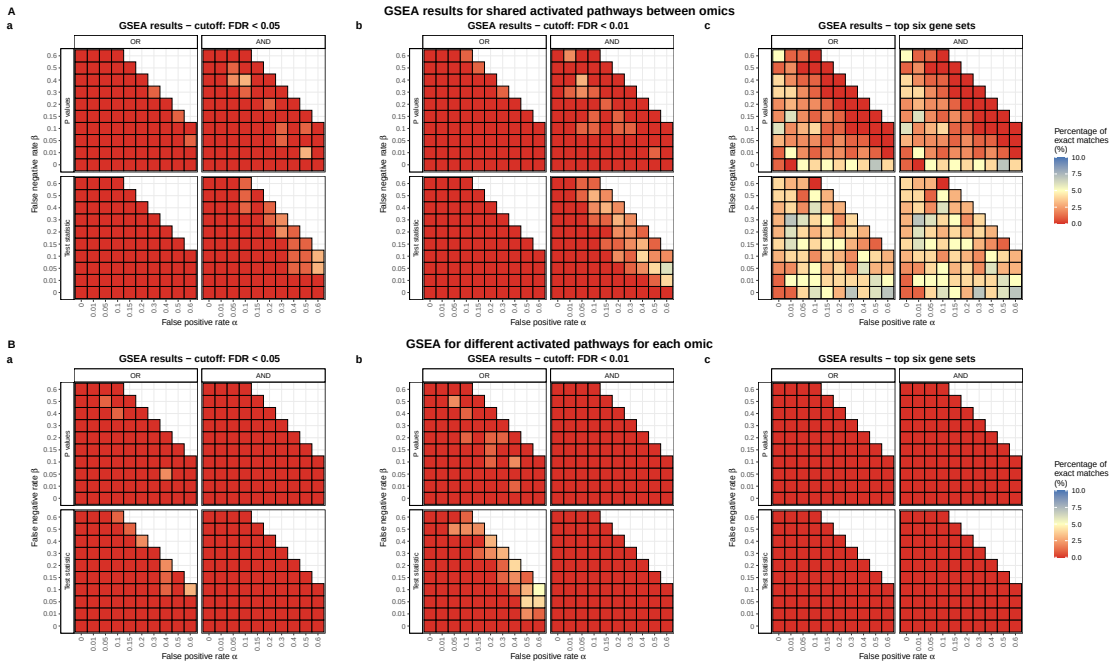


Figure 6.6: Summary of GSEA performance dependence on significance cutoffs.
To simplify visualizations, a subset of scenarios showing only data for a coverage of 30 % and a correlation of 0.3 between transcripts and proteins was chosen. We evaluated the percentage of cases, where the exact ground truth was recovered when evaluating enrichment results using a FDR < 0.05, FDR < 0.01 and the top six ranked gene sets.
Contrary to all other visualizations, the color code was restricted to a range of 0 % to 10 % instead 0 % to 100 % for better visibility.
GSEA, gene set enrichment analysis; FDR, false discovery rate;

6.3.4.2 MGSA performance

MGSA applied to absolute Z scores We first evaluated MGSA performance when applied to simulated rankings without direction of effect in a scenario where all gene sets were active in both omics (Supplementary Figure A.3A). Overall accuracy was close to one with the exception of very high false positive rates α of the simulated data. It was higher for the combination using overlapping effects, i.e. the AND integration as shown in Supplementary Figure A.3Ae compared to the additive combination (OR) in Supplementary Figure A.3Aa. Both evaluation approaches have a very high sensitivity in picking up activated pathways across all evaluated false positive and false negative rates α/β (Supplementary Figure A.3Ab, A.3Af and A.3Bb) in both simulated scenarios, with the exception of the overlapping combination (AND) for independent pathways across omics (Supplementary Figure A.3Bf). For this specific case, sensitivity was close to zero except for very high preset false positive rates of the simulation scenario, where it was still smaller than 0.5.

When evaluating how good the exact simulated ground truth was recovered, MGSA showed very high percentage of successes for the OR-combination except for very high

false positive rates ≥ 0.4 (Supplementary Figure A.3Ad and A.3Bd) for all simulation scenarios. Best performance was achieved using overlapping results across omics with the AND-combination for shared pathway activations across omics (Supplementary Figure A.3Ah) while almost never reconstructing activated pathways for independent effects between omics.

MGSA applied to Z scores When including direction of effect, i.e. evaluating Z scores representing either signed P values or a signed test statistic, almost perfect reconstruction of the simulated ground truth was achieved with both MGSA combinations OR and AND (Supplementary Figure A.4A). While in the case of the analysis excluding direction of effect was sensitive to high false positive rates α in the simulated dataset, only very high false negative rates impede the recovery of all simulated datasets when the sign of the test statistic was taken into account as shown in Supplementary Figure A.4Ad and A.4Ah.

The very same applies to MGSA performance when analyzing scenarios with different pathways being active in both omics and evaluating additive effects based on the OR-combination (Supplementary Figure A.4Ba-A.4Bd). However, in the case of combining multi-omics results by assessing overlapping effects (AND combination), active gene sets can no longer be picked up resulting in close to zero sensitivity and consequently, no reconstruction of the simulated ground truth (Supplementary Figure A.4Be-A.4Bh).

Dependence on the MGSA posterior cutoff Similar to GSEA, also for MGSA we evaluated the dependence on the choice of posterior cutoffs to consider a gene set as active. No significant differences were observed for using a posterior cutoff of 0.5 or more stringently 0.6 which is shown by comparing Figure 6.7Aa to Figure 6.7Ab and Figure 6.7Ba to Figure 6.7Bb. While those are again arbitrary choices of thresholds, the U-shaped distribution of posterior probability clearly supports this finding. Additionally, Figure 6.7 clearly demonstrates the improved performance when including direction of effect (panels: Test statistic versus P value).

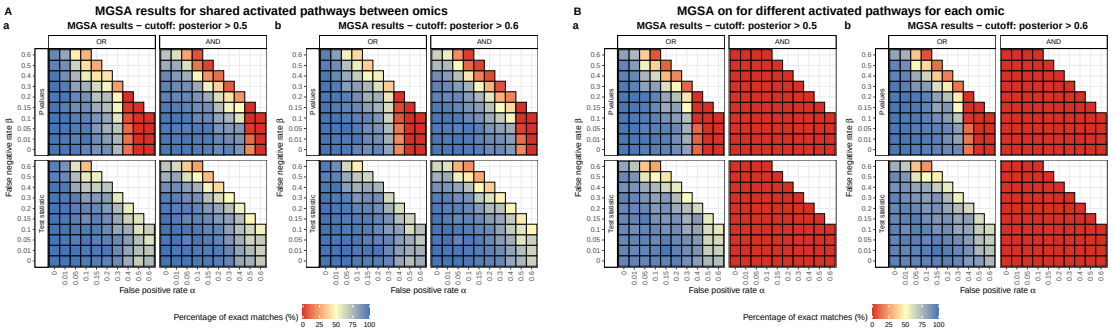


Figure 6.7: MGSA performance dependence on significance cutoffs.
To simplify visualizations, a subset of scenarios showing only data for a coverage of 30 % and a correlation of 0.3 between transcripts and proteins was chosen. We evaluated the percentage of cases, where the exact ground truth was recovered when evaluating enrichment results using a posterior > 0.5 and posterior > 0.6. MGSA, model-based gene set analysis;

Overall MGSA performance is also summarized in Figure 6.8.

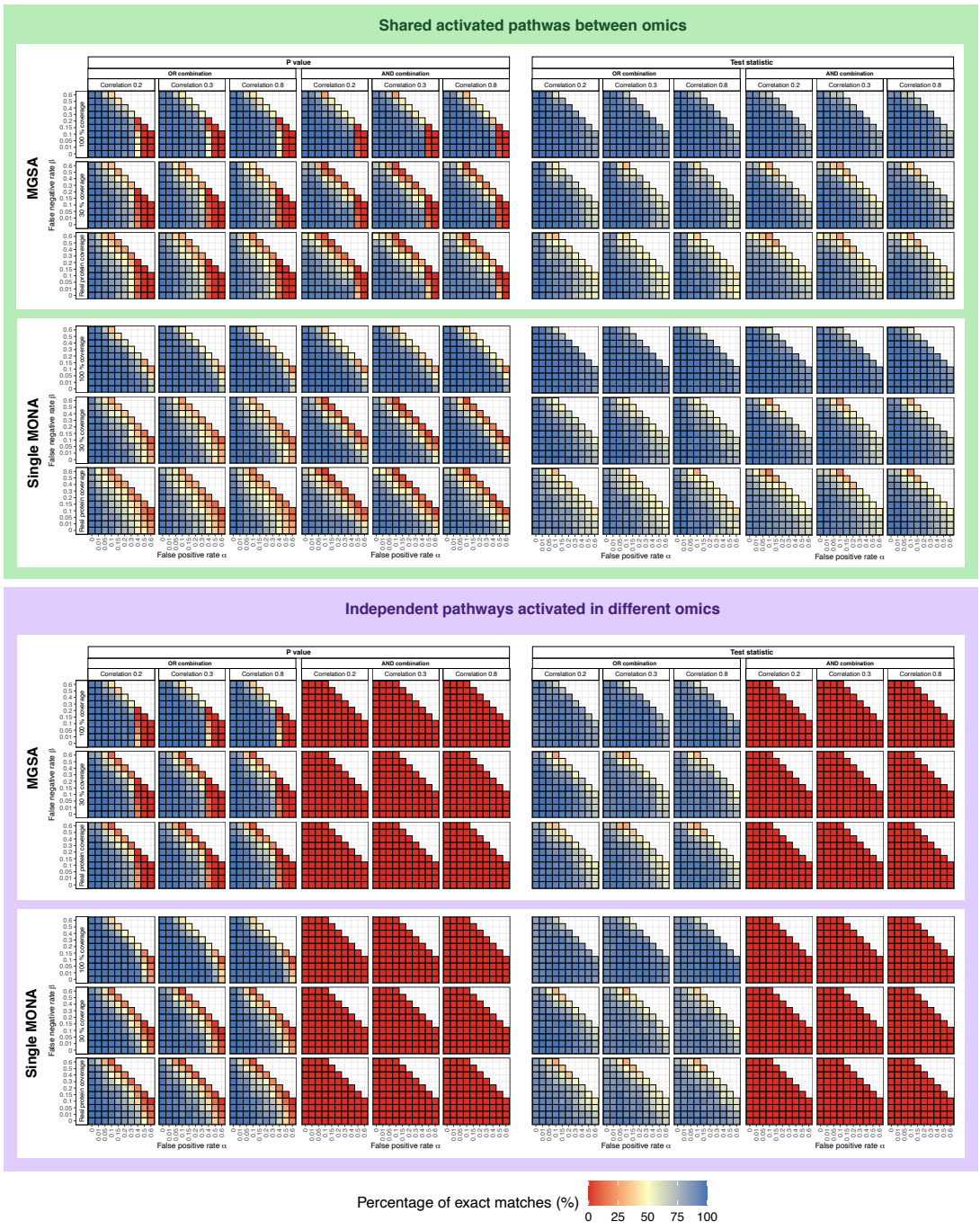


Figure 6.8: Performance overview of multi-omics combination of MGSA and MONA single-omic methods

Multi-omics analysis of simulated data using single-omic OR and AND combinations for MGSA and MONA. Percentage of cases, where the exact simulated pathway activations were recovered from either absolute Z scores or original Z scores across different simulation scenarios including shared and independent effects of pathway activations between omics, various combinations of coverage, correlation and error rates.

MGSA, model-based gene set analysis; MONA, multi-level ontology analysis;

6.3.4.3 MONA performance

Single-omic MONA results Capabilities of the single-omic MONA model were very similar to those of the MGSA. Almost perfect accuracy and specificity was observed across all evaluation types and simulation scenarios (Supplementary Figure A.5 and A.6). Sensitivity was only close to zero when evaluating overlapping effects (AND combination) in simulation scenarios with independent pathway activations across omics (Supplementary Figure A.5Bf and A.6Bf) resulting also in poor outcomes when considering the exact recovery of gene sets simulated as active.

Again, taking into account direction of effect greatly improved the performance by increasing the exact reconstruction of simulated pathways also for very high false positive rates α (Supplementary Figure A.5Ad, A.5Ah and A.5Bd compared to Supplementary Figure A.6Ad, A.6Ah and A.6Bd). Almost perfect performance was achieved by using the additive combination of single omic results (OR) across all scenarios with only a very slight decrease for very high false negative rates β (Supplementary Figure A.6Ad and A.6Bd).

Results for the single-omic MONA performance are summarized in Figure 6.8.

Similarly to MGSA, choosing a significance cutoff for the single MONA model posterior of 0.5 or 0.6 did not significantly impact performance as shown in Figure 6.9 which was also confirmed by the shape of the posterior distribution.

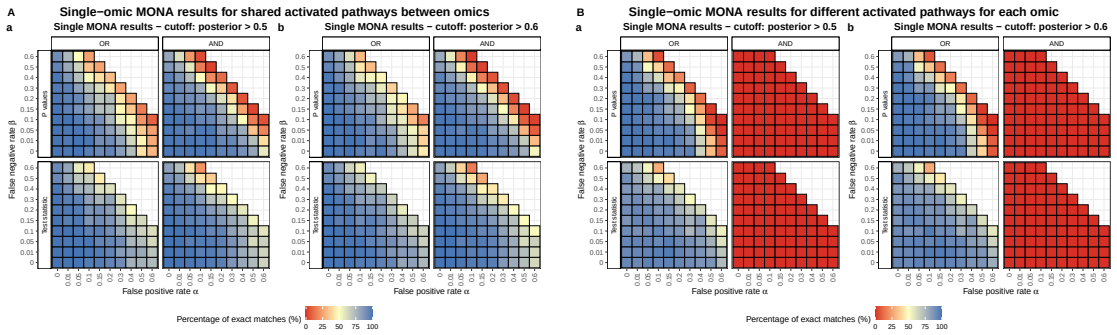


Figure 6.9: Single-omic MONA performance dependence on significance cutoffs.
To simplify visualizations, a subset of scenarios showing only data for a coverage of 30 % and a correlation of 0.3 between transcripts and proteins was chosen. We evaluated the percentage of cases, where the exact ground truth was recovered when evaluating enrichment results using a posterior > 0.5 and posterior > 0.6. MONA, multi-level ontology analysis;

MONA multi-omics integration results When applying the MONA two-omics cooperative model to our simulated transcriptomics and proteomics differential expression results for shared activated pathways across omics, we observe better performance than by any other method evaluated so far when using absolute Z scores (Supplementary Figure A.7Aa-A.7Ad). By additionally taking into account the direction of effect, i.e. evaluating the original Z scores, we reach almost perfect accuracy (Supplementary Figure A.7Ae), sensitivity (Supplementary Figure A.7Af) and specificity (Supplementary Figure A.7Ag) with a percentage of exact matches recovering the simulated ground truth in over 94 % of all runs in more than 80 % of all simulation scenarios (Supplementary Figure A.7Ah).

Conversely, sensitivity drops significantly when evaluating simulation scenarios with independent pathways active in different omics (Figure 6.10Ba). Approximately only half of the gene sets that were simulated to be active are recovered in most cases (Supplementary Figure A.7Bb), which is further investigated in the following section including Figure 6.11B. Direction of effect only slightly increases the sensitivity (Supplementary Figure A.7Bf) with the largest improvement shown for simulation scenarios with a 100 % coverage of proteomics and a higher correlation in the simulated dataset. While overall accuracy (Supplementary Figure A.7Ba and A.7Be) and specificity (Supplementary Figure A.7Bc and A.7Bg) remain high due to the imbalance in active and inactive pathways, we accordingly observe a clear drop to almost zero for the percentage of exact matches to the ground truth (Supplementary Figure A.7Bd and A.7Bh). The only exception are consequently the scenarios with very high coverage and high correlation between omics but even those settings are highly sensitive to very small false positive rates α in the simulated data.
Additionally, for the scenarios with a 100 % coverage (see also Figure 6.10Bb), the best overall performance is achieved for the highest false negative rates combined with very low false positive rates. This is contrary to what was observed for the other methods, where performance decreased for higher false positive and false negative error rates.

When considering different posterior cutoffs of 0.5 and 0.6 in Figure 6.10, no significant difference can be observed, which was also confirmed by the shape of the posterior distribution.

Comparing single-omic MONA combination to direct MONA integration Finally, we want to compare the single MONA combination approaches to the direct MONA multi-omics integration.

In Figure 6.11A (first column), we summarize the results for shared pathways across omics. When including direction of effect, the direct multi-omics integration (Figure 6.11Aa) outperforms all other methods, followed by the OR-combination (Figure 6.11Ab) and the AND-combination (Figure 6.11Ac).

Considering independent pathways activated in different omics, even when taking direction of effect into account, the direct multi-omics integration (Figure 6.11Ba)

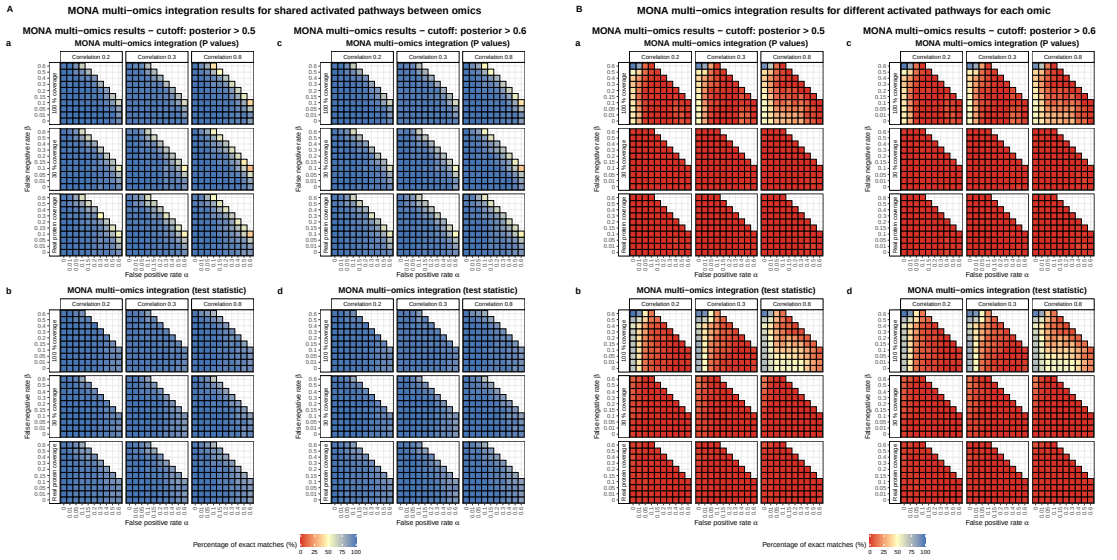


Figure 6.10: MONA direct multi-omics integration performance dependence on significance cutoffs. For the MONA cooperative two-omics model we evaluated the percentage of cases, where the exact ground truth was recovered when evaluating enrichment results using a posterior > 0.5 and posterior > 0.6. MONA, multi-level ontology analysis;

performs mostly very poorly and only the AND-combination (Figure 6.11Bc) shows even worse results. In contrast, the OR-combination (Figure 6.11Bb) shows a very robust strong performance comparable to that in the shared-effects setting. The performance of the MONA multi-omics integration is due to a strong bias of identifying gene sets active in the first omic layer exclusively from the multi-omics observations (Figure 6.11Bd), while missing those of the second omic. A similar behavior of only detecting the gene sets active in the second omic (Figure 6.11Be) - even for full coverage of the second omic - is not observed. As the single-omic MONA OR-combination does not show such restrictions, the ability of identifying active gene sets in the first omic from the first omic summary statistics (Figure 6.11Bf) and recovering active gene sets in the second omic type from the second omic summary statistics (Figure 6.11Bg) is only dependent on the general coverage and correlation parameters of the simulation scenario, i.e. higher coverages for any omic facilitate the reconstruction of the correct pathways.

Finally, when evaluating the capability of inferring all pathways from both omics using only the summary statistics for one omic, this is not possible for any of the simulated scenarios (Figure 6.11Bh and Figure 6.11Bi).

6.3.5 Conclusion and comparison of all evaluated methods

In summary, our simulation study clearly showed the importance of direction of effect, which strongly improved the performance of all methods.

Compared to all other approaches, GSEA showed much poorer results as shown in Figure 6.12, which is mostly connected to false positive discoveries. Due to their

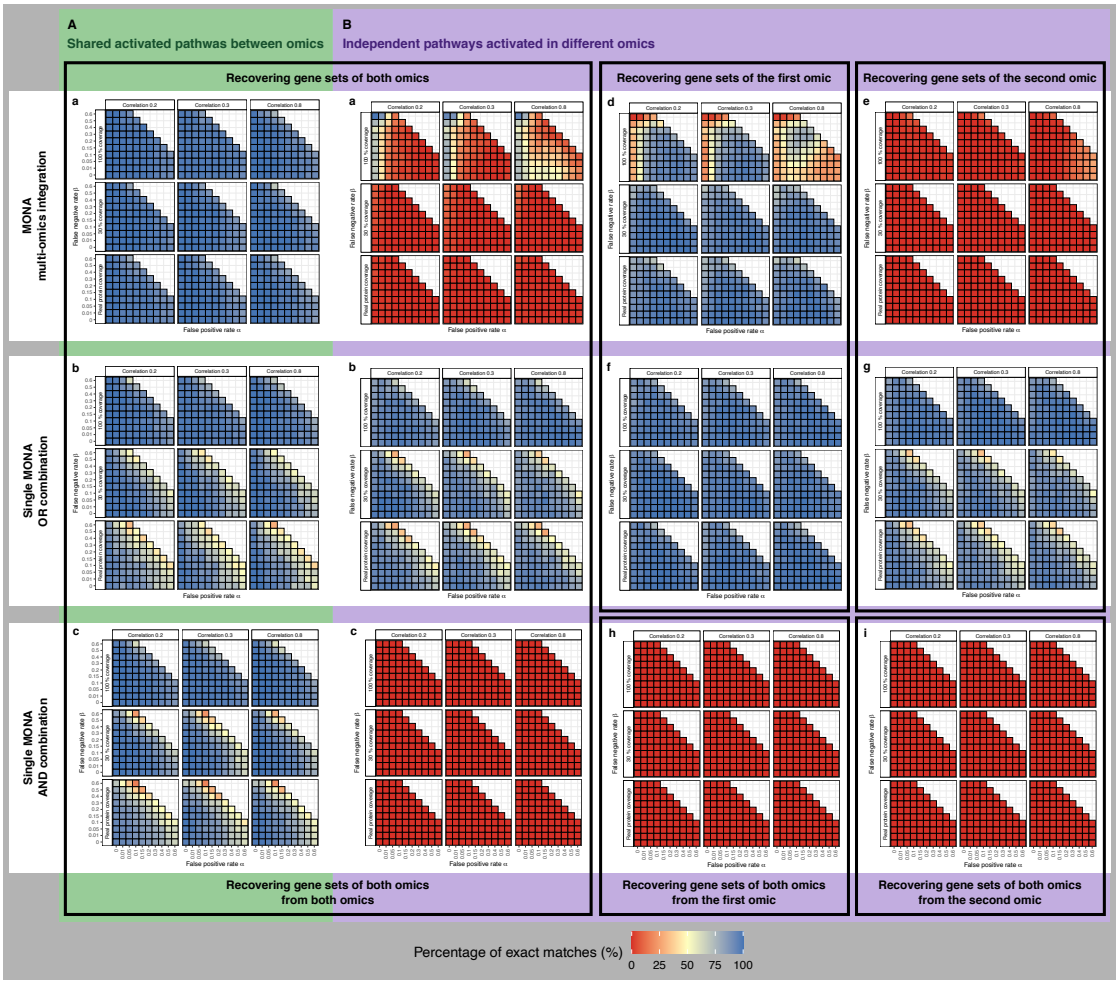


Figure 6.11: Comparison of single-omic MONA combinations and direct multi-omics MONA integration.
MONA multi-omics results for activated pathways that were shared across omics (panel A), and different pathways being active in different omics (panel B). The color code represents the percentage of exact matches of recovered pathways compared to the simulated ground truth.
A: Shared pathway activations across omics:
Aa: MONA multi-omics integration; Ab: Single-omic MONA OR-combination; Ac: Single-omic MONA AND-combination.
B: Independent pathway activations in different omics:
Ba: MONA multi-omics integration; Bb: Single-omic MONA OR-combination; Bc: Single-omic MONA AND-combination.
Bd: MONA multi-omics integration, recovering pathways active in the first omic (using information from both omics); Be: MONA multi-omics integration, recovering pathways active in second omic (using information from both omics).
Bf: Single-omic MONA for the first omic, recovering pathways active in the first omic; Bg: Single-omic MONA for the second omic, recovering pathways active in the second omic.
Bh: Single-omic MONA for the first omic, recovering all active pathways from both omics (using information from the first omic only); Bi: Single-omic MONA for the second omic, recovering all active pathways from both omics (using information from the second omic only).
MONA, multi-level ontology analysis;

common underlying model, MGSA and the single MONA were comparable, with MONA performing slightly better in the case of not taking direction of effect into account (Figure 6.8).

The MONA multi-omics cooperative model was only superior to the single-omic methods for the shared effect scenarios as shown in Figure 6.12A, specifically Figure 6.12Ad. In the case of independent pathways activated in the different omics, mostly the activations from the first omic layer were recovered leading to basically no exact cases of full reconstruction of the simulated activations (Figure 6.12Bd).

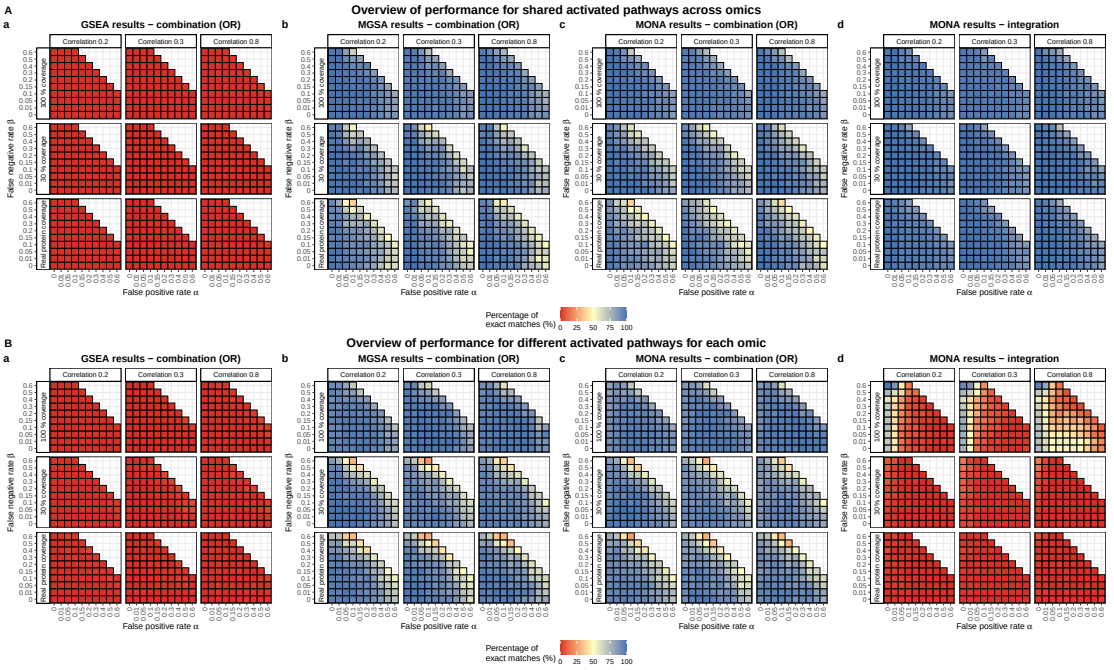


Figure 6.12: Summary: comparison of model performance across the different simulation scenarios.
Each method evaluated with direction of effect and including the combination of two omics by evaluating additive effects (OR integration) for GSEA, MGSA and MONA as well as the direct integration with the MONA cooperative model.
Aa-Ad: Percentage of exact reconstruction of the simulated ground truth for shared pathways across the different omics for GSEA, MGSA, single-omic MONA and MONA multi-omics cooperative model.
Ba-Bd: Percentage of exact reconstruction of the simulated ground truth for independent pathways in different omics for GSEA, MGSA, single-omic MONA and MONA multi-omics cooperative model.
GSEA, gene set enrichment analysis; MGSA, model-based gene set analysis; MONA, multi-level ontology analysis;

The MONA OR-integration (Figure 6.12Ac and Figure 6.12Bc) showed extremely robust and almost always best performing results.
Specifically, the high performance also holds true for very high false positive α and false negative β rates in the simulated data (Figure 6.12) and was not sensitive to the specific posterior cutoff to determine activated pathways (Figure 6.9).

Next, we use the knowledge gained from our simulation study to apply it to the pathway enrichment analysis of our atrial fibrillation cohort AFHRI-B.

6.4 Multi-omics pathway enrichment analysis of atrial fibrillation in the AFHRI-B cohort

In section 6.1, we evaluated differentially expressed genes in human atrial tissue transcriptomics and proteomics for the two AF subtypes prevalent and incident AF. Overall effect sizes were very low, which together with the relatively small sample size resulted in no differentially expressed genes even with very relaxed significance thresholds.
However, as introduced by Subramanian et al. [2005], consistent up- or down-regulation of multiple genes in the same pathway may actually be more informative, than single differentially expressed genes.

To proceed with the pathway enrichment analysis of this dataset, we took into account the results of the previous section 6.3 evaluating different gene set enrichment approaches. Our simulations showed that the evaluation of additive effects by a single-omic MONA model gives the most robust results, and that shared effects across omics are best identified using the multi-omics MONA cooperative model (Figure 6.12). We therefore chose to apply both approaches while including direction of effect for the most accurate outcome.

Analogously to the good performance even for high false positive rates, i.e. less stringent P value cutoffs, we chose a threshold of $P < 0.25$ for all analyses. A KEGG gene set was considered up- or down-regulated, if its posterior exceeded the threshold of 0.5. To better visualize the results, we used a signed posterior, i.e. adding a minus sign to the posterior in case of down-regulation. We further evaluated prevalent AF and post-operative or incident AF separately.

All pathway enrichment results are summarized in Figure 6.13. Since the coverage of an omic is strongly influencing enrichment results, we also added a heat map showing the pathway size and the number of genes measured on each omic level as well as for both omics.

6.4.1 Prevalent atrial fibrillation pathway enrichment results

We first considered the AF subtype of prevalent AF. Applying the single-omic MONA model on transcript level, three pathways were significantly up-regulated (Figure 6.13): ubiquitin mediated proteolysis, the insulin signaling pathway and RIG-I-like receptor signaling pathways. Even though the activations were rather small, these three pathways belong to recurrent groups of pathways involved in metabolism, cellular processes and immune response that were most seen across the different omics and AF subtypes and represent major factors of known AF disease pathology [Staerk et al., 2017, Hindricks et al., 2021].

A much stronger signal was observed on protein level, where overall 25 pathways

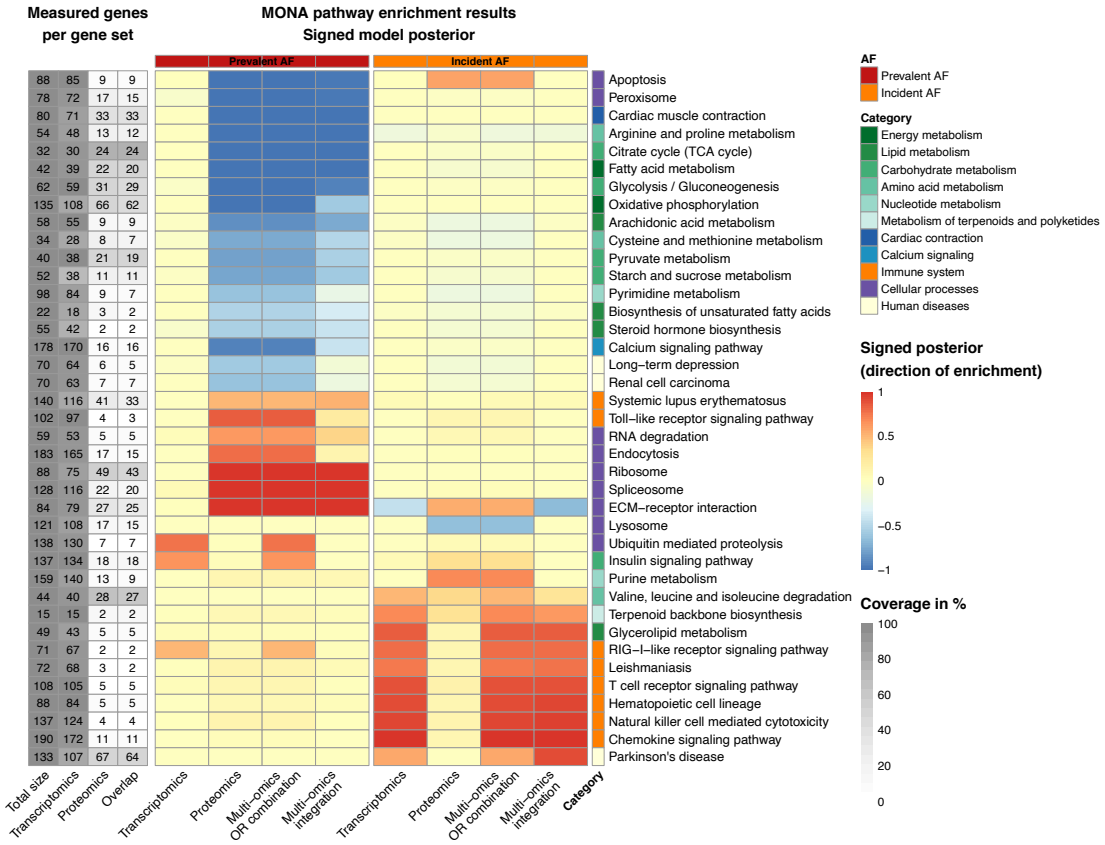


Figure 6.13: Pathway enrichment results for AF.
AF multi-omics pathway enrichment results for the AFHRI-B cohort. The heat map in grey-scales depicts the total size of each gene set, as well as the number of genes measured in transcriptomics, proteomics and in both. Darker shades represent a higher coverage. The single-omic MONA model was first applied to transcriptomics and proteomics separately. From those results, the multi-omics OR combination was derived. Both omics were subsequently analyzed together with the MONA cooperative model (multi-omics integration). Two AF subtypes were considered, the first four columns of the colored heat map show the results for prevalent AF, the fifth to eight column the results for incident AF.
AF, atrial fibrillation; MONA, multi-level ontology analysis;

were estimated to be significantly changed. Down-regulated metabolic pathways involved in energy, lipid and carbohydrate metabolism were the most prominent. Changed metabolism affecting pathways such as the fatty-acid metabolism or more specific, oxidative phosphorylation and the biosynthesis of unsaturated fatty acids with matching down-regulation of the associated proteins has been described before [Tu et al., 2014]. Similarly, proteins of pathways associated with carbohydrate metabolism like TCA cycle, glycolysis/gluconeogenesis, pyruvate and starch/sucrose metabolism were also collectively down-regulated just as described before by Tu et al. [2014].
Furthermore, we find two more pathways that are highly relevant in this disease context, a down-regulation of cardiac muscle contraction and the calcium signaling pathway.
In contrast to the transcriptome analysis no inflammatory pathways came up for proteomics, probably due to the low amount of proteins measured for those specific

pathways.

The MONA cooperative model multi-omics integration showed an overall profile that was very similar to the single-omic MONA proteomics results but with attenuated sizes of the posteriors.

6.4.2 Incident atrial fibrillation pathway enrichment results

The inflammatory response is in general very important for the pathophysiology of AF [Hu et al., 2015], but Watt et al. [2021] acknowledged its specific role in post-operative AF. Indeed, this was confirmed by the single-omic MONA model, as the majority of pathways that were up-regulated in transcriptomics were connected to pro-inflammatory pathways such as T cell receptor signaling pathway, chemokine signaling pathway and natural killer cell mediated cytotoxicity.
In addition, multiple pathways involved in amino-acid and nucleotide metabolism were up-regulated, while the ECM receptor interaction was the only one down-regulated.

A much smaller signal was found on protein level with a down-regulation of the lysosome and up-regulation of ECM-receptor interaction, purine metabolism and apoptosis.

When evaluating the multi-omics cooperative MONA model, the pathway activations were almost identical to those of the transcript data with a similar strong focus on inflammatory pathways.

6.4.3 Summary

In summary, we found various groups of pathways activated for each of the AF subtypes that were in line with known mechanism in literature. Strong differences between the different single-omic analyses were observed with most of the activations replicated in the multi-omics approach. Different omics showed stronger signals for different AF subtypes with prevalent AF results being dominated by metabolic pathways on protein level and incident AF showing various regulated immune pathways on transcript level.

6.5 EnrichmentNodes - a KNIME plugin to perform multi-omics enrichment analyses

When interpreting differential expression results, most tools do not offer more sophisticated analyses than simple overrepresentation analysis, such as AmiGO[2] [Carbon et al., 2009] or Panther[3] [Mi et al., 2019]. Additionally, they mainly only consider one omic level.

To make multi-omics pathway enrichment analyses more accessible, we created the EnrichmentNodes[4] plugin for the KNIME framework.
The Java-based Konstanz Information Miner (KNIME) KNIME Analytics Platform[5] [Berthold et al., 2009] is an open source project for interactive data analysis using visual workflows to build pipelines in a graphical user interface.
Topic specific nodes add functionality by encoding single analysis steps and can be incorporated into analysis pipelines via drag-and-drop.

6.5.1 Generic KNIME nodes

To further facilitate the development of new nodes, the Generic KNIME nodes (GKN)[6] [Fillbrunn et al., 2017] node generator can create custom nodes by mapping command line tools into a graphical user interface. This opens up a lot of flexibility with respect to the tools and dependencies, while it retains the easy accessibility of visual interaction. Additionally, tools can be run using a specified docker container instead of the host system, making it possible to provide operating system independent solutions.

An example of how to build a Hello-World node is shown in Figure 6.14.

Examples for easy graphical interfaces to more complex programming tasks using generic workflow nodes are the OpenMS package for mass spectrometry analysis[7,8] [Pfeuffer et al., 2017], the SeqAn library for efficient sequence analysis[9,10] [Döring et al., 2008] and ImmunoNodes[11] an immunoinformatics toolbox [Schubert et al., 2017].

[2] http://amigo.geneontology.org/amigo
[3] http://www.pantherdb.org/
[4] https://github.com/InesAssum/EnrichmentNodes
[5] https://www.knime.com/knime-analytics-platform
[6] https://github.com/genericworkflownodes/GenericKnimeNodes
[7] https://github.com/genericworkflownodes/de.openms.knime
[8] https://openms.de
[9] https://github.com/genericworkflownodes/de.openms.knime
[10] https://www.seqan.de/
[11] https://github.com/FRED-2/ImmunoNodes

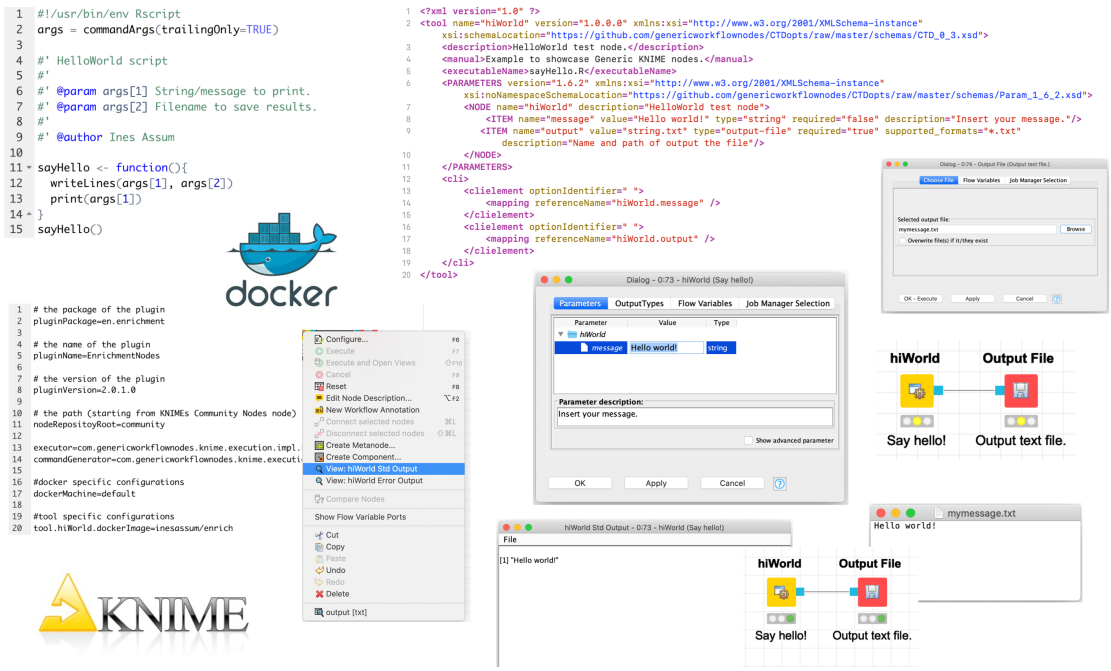


Figure 6.14: Summary of R-based Generic KNIME node development.
Process of creating generic workflow nodes for KNIME. Here, starting from a R script, the graphical user interface options are designed. Using the parameters configured in a KNIME workflow and selecting where output files should be saved, inputs for the corresponding R script are mapped to shell commands and run by a specified docker. After executing a node, output or error messages are directly accessible in KNIME and the result has been saved as a text file.
KNIME, Konstanz Information Miner;

6.5.2 EnrichmentNodes

Goal of the EnrichmentNodes plugin was to provide access to multi-omics pathway analyses without any programming. The graphical interface makes the setup of parameters easy and intuitive. Instead of complicated file handling, data is provided by supplying the path to the corresponding tables, e.g. excel files.

6.5.2.1 Enabling multi-omics enrichment analyses

Leveraging docker containers for platform independent and reproducible research
A docker container with all necessary tools including various packages and R version 4.1.1 has been created to run all the analyses. Its source files are available on GitHub and the docker image named **enrich** is also accessible via Docker Hub[12].
On top, being based on a RStudio rocker[13], i.e. a docker from the rocker project including a RStudio server installation[14], larger computational projects can be easily

[12] https://hub.docker.com/repository/docker/inesassum/enrich
[13] https://www.rocker-project.org/
[14] https://www.rstudio.com/

carried out in a reliable and reproducible fashion.

Installation instructions Installing EnrichmentNodes requires only a few steps.

1. Visit `https://www.knime.com/downloads`.
   Download and install the KNIME Analytics Platform (Version 4.3.4 or higher).
2. Visit `https://www.docker.com/products/docker-desktop`.
   Download and install Docker Desktop (including creating a Docker id).
3. Download the EnrichmentNodes plugin
   `https://github.com/InesAssum/EnrichmentNodes/blob/master/export_plugin/en.enrichment_2.0.1.2.jar`
4. Place it in the folder **C:/Programme/KNIME x.x.x/plugins** (on Windows) and **/Applications/KNIME x.x.x.app/Contents/Eclipse/plugins** (on Mac OS).

The docker image for running the analyses will be automatically pulled from the Docker Hub when needed.

6.5.2.2 EnrichmentNodes functionalities

Currently, our EnrichmentNodes offer the gene set enrichment methods described in this thesis including the simulation of correlated multi-omics summary statistics.

For all the pathway enrichment methods, i.e. GSEA, the single-omic MONA model, MGSA as well as the MONA two-omics and three-omics cooperative model, example datasets are included in the docker, so that all nodes can be tested without providing individual input data. Additionally, we provide a function to acquire commonly used pathway annotations for multiple biological species (i.e. human, mouse, rat and zebrafish), convert and save them in .gmt format, which is both readable as a tab-separated text file and convenient to use with our enrichment framework.

An overview of the KNIME Analytics Platform and our EnrichmentNodes can be found in Figure 6.15.

6.5.2.3 EnrichmentNodes data structures

The gene set enrichment analysis tools implemented in our EnrichmentNodes work on mostly rankings or sets of significant genes derived from summary statistics.
To be able to access all necessary information, we expect specifically named columns in the input data tables. Depending on the analysis, either rankings, a ranking including a cutoff or a set of significant genes are needed:

- Identifier information, that must match the names used in the annotations must be in a column named **id**.

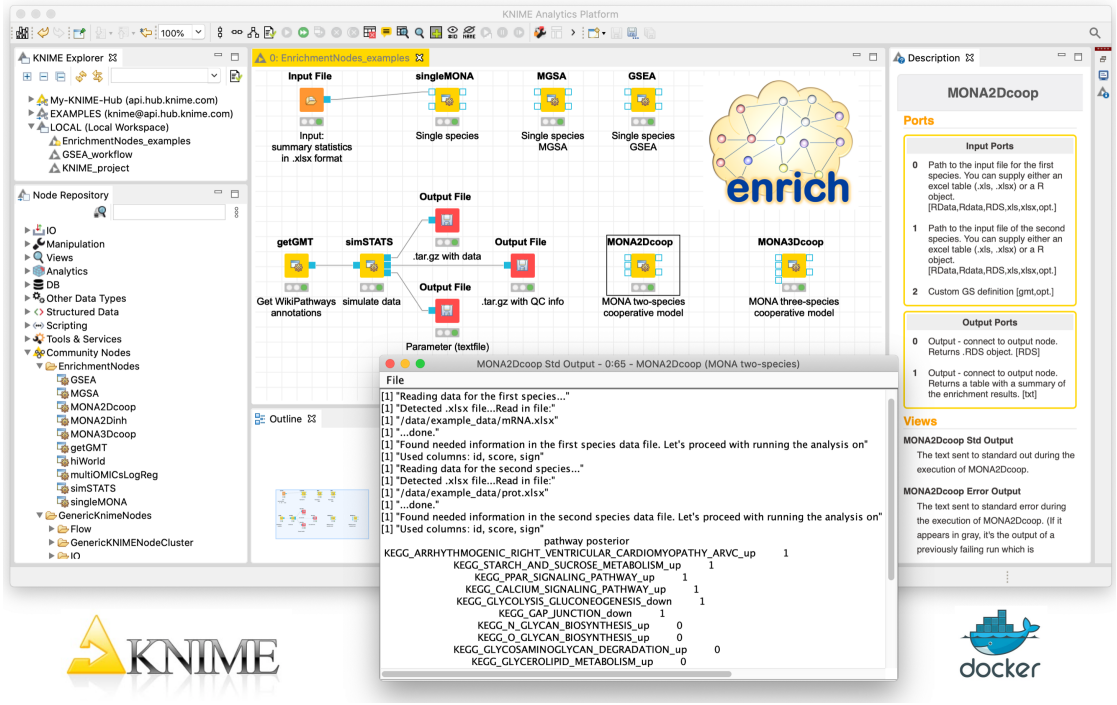


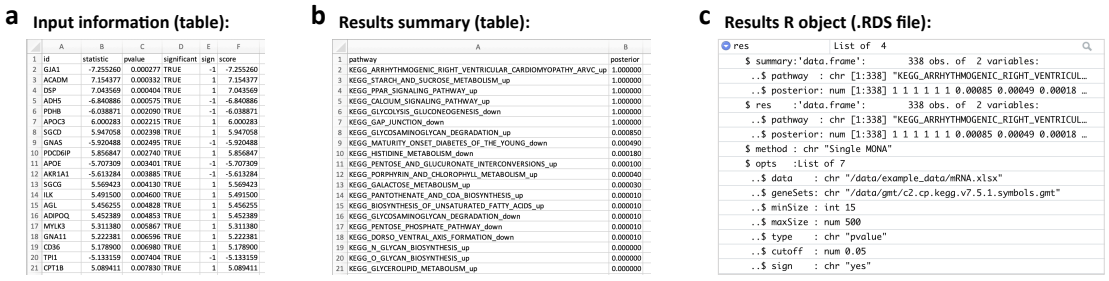
Figure 6.15: Overview and example workflow for the EnrichmentNodes Generic KNIME node plugin. Example for using EnrichmentNodes with the KNIME Analytics Platform. Nodes from the EnrichmentNodes plugin can be selected via the folder **EnrichmentNodes** under **Community Nodes** in the Node repository. Using the KNIME Explorer, workflows can be managed. Pipelines can be built and executed in the graphical workspace and information about the different nodes is displayed on the right.
In the left lower part of the workspace, we show an example for combining the *getGMT* node to get WikiPathways annotations and use them to simulate multi-omics summary statistics. The resulting files containing the simulated data (.tar.gz file), quality control information about the sampling (.tar.gz file) and a text file with the used parameters are then saved locally.
Every gene set enrichment analysis node can be run without additional input using supplied example data. Progress information and the top ten pathways are displayed in the Std Output of every node.
KNIME, Konstanz Information Miner;

- **P values:** *type == "pvalue"*
  Analyses will be performed on P value information contained in the column **pvalue**, which should contain positive values. Smaller values are considered more regulated. If the direction of effect should be included, the parameter *sign* must be set to *"yes"* and an additional column **sign** must be included.
  For MGSA and MONA, a cutoff to determine significance must be provided as *cutoff* parameter.
- **Scores, rankings, fold change or test statistic:** *type == "score"*
  Analyses will be performed on the ranking contained in the column **score**. Bigger absolute values are considered more regulated. If the direction of effect should be included, the parameter *sign* must be set to *"yes"*. Up- or down-regulation is inferred from the sign of the score unless an additional column **sign** is available. If present, information from that column is used.
  For MGSA and MONA, a cutoff to determine significance must be provided as

cutoff parameter.

- **Set of significant observations:** *type == "significant"*
  This only applies to MGSA and MONA. Analyses will be performed on binary 0/1 or False/True information in the column **significant**. If the direction of effect should be included, the parameter *sign* must be set to *"yes"* and an additional column **sign** must be supplied to determine up- or down-regulation.

Enrichment results will be summarized as a table and the top ten regulated gene sets will also be printed as node Std output. Additionally, the summary, the original result R object as well as information on the method and parameters used are saved as a .RDS file.
Examples for input information (Figure 6.16a), the summary table (Figure 6.16b) and the R object (Figure 6.16) are visualized in Figure 6.16. For MONA, the enrichment results are a simple table with the term names and inferred posterior. However, for GSEA a similar table contains lists with leading edge information that could impair readability and is therefore removed for the summary and for MGSA, it is an object of class *MgsaMcmcResults*.



	A	B	C	D	E	F
1	id	statistic	pvalue	significant	sign	score
2	GJA1	-7.255260	0.000277	TRUE	-1	-7.255260
3	ACADM	7.154377	0.000332	TRUE	1	7.154377
4	DSP	7.043569	0.000404	TRUE	1	7.043569
5	ADH5	-6.840886	0.000575	TRUE	-1	-6.840886
6	PDHB	-6.038871	0.002090	TRUE	-1	-6.038871
7	APOC3	6.000283	0.002215	TRUE	1	6.000283
8	SGCD	5.947058	0.002398	TRUE	1	5.947058
9	GNAS	-5.920488	0.002495	TRUE	-1	-5.920488
10	PDCD6IP	5.856847	0.002740	TRUE	1	5.856847
11	APOE	-5.707309	0.003401	TRUE	-1	-5.707309
12	ARL1A1	-5.613284	0.003885	TRUE	-1	-5.613284
13	SGCG	5.569423	0.004130	TRUE	1	5.569423
14	ILK	5.491500	0.004600	TRUE	1	5.491500
15	AGL	5.456255	0.004828	TRUE	1	5.456255
16	ADIPOQ	5.452389	0.004853	TRUE	1	5.452389
17	MYLK3	5.311380	0.005867	TRUE	1	5.311380
18	GNA11	5.222381	0.006596	TRUE	1	5.222381
19	CD36	5.178900	0.006980	TRUE	1	5.178900
20	TPI1	-5.133159	0.007404	TRUE	-1	-5.133159
21	CPT1B	5.089411	0.007830	TRUE	1	5.089411

	A	B
1	pathway	posterior
2	KEGG_ARRHYTHMOGENIC_RIGHT_VENTRICULAR_CARDIOMYOPATHY_ARVC_up	1.000000
3	KEGG_STARCH_AND_SUCROSE_METABOLISM_up	1.000000
4	KEGG_PPAR_SIGNALING_PATHWAY_up	1.000000
5	KEGG_CALCIUM_SIGNALING_PATHWAY_up	1.000000
6	KEGG_GLYCOLYSIS_GLUCONEOGENESIS_down	1.000000
7	KEGG_GAP_JUNCTION_down	1.000000
8	KEGG_GLYCOSAMINOGLYCAN_DEGRADATION_up	0.000850
9	KEGG_MATURITY_ONSET_DIABETES_OF_THE_YOUNG_down	0.000490
10	KEGG_HISTIDINE_METABOLISM_down	0.000180
11	KEGG_PENTOSE_AND_GLUCURONATE_INTERCONVERSIONS_up	0.000100
12	KEGG_PORPHYRIN_AND_CHLOROPHYLL_METABOLISM_up	0.000040
13	KEGG_GALACTOSE_METABOLISM_up	0.000030
14	KEGG_PANTOTHENATE_AND_COA_BIOSYNTHESIS_up	0.000010
15	KEGG_BIOSYNTHESIS_OF_UNSATURATED_FATTY_ACIDS_up	0.000010
16	KEGG_GLYCOSAMINOGLYCAN_DEGRADATION_down	0.000010
17	KEGG_PENTOSE_PHOSPHATE_PATHWAY_down	0.000010
18	KEGG_DORSO_VENTRAL_AXIS_FORMATION_down	0.000010
19	KEGG_N_GLYCAN_BIOSYNTHESIS_up	0.000000
20	KEGG_O_GLYCAN_BIOSYNTHESIS_up	0.000000
21	KEGG_GLYCEROLIPID_METABOLISM_up	0.000000

```
res                List of  4
 $ summary:'data.frame':        338 obs. of  2 variables:
  ..$ pathway  : chr [1:338] "KEGG_ARRHYTHMOGENIC_RIGHT_VENTRICUL…
  ..$ posterior: num [1:338] 1 1 1 1 1 1 0.00085 0.00049 0.00018 …
 $ res     :'data.frame':        338 obs. of  2 variables:
  ..$ pathway  : chr [1:338] "KEGG_ARRHYTHMOGENIC_RIGHT_VENTRICUL…
  ..$ posterior: num [1:338] 1 1 1 1 1 1 0.00085 0.00049 0.00018 …
 $ method : chr "Single MONA"
 $ opts   :List of 7
  ..$ data     : chr "/data/example_data/mRNA.xlsx"
  ..$ geneSets: chr "/data/gmt/c2.cp.kegg.v7.5.1.symbols.gmt"
  ..$ minSize : int 15
  ..$ maxSize : num 500
  ..$ type     : chr "pvalue"
  ..$ cutoff   : num 0.05
  ..$ sign     : chr "yes"
```

Figure 6.16: Data structures used for running EnrichmentNodes.
Example for input data (a), results summary table (b) and detailed results R object for the single MONA analysis (c).

6.6 Discussion

In this chapter, we focused on analyzing the differences between AF patients and controls in sinus rhythm using atrial tissue transcriptomics and proteomics data. The small effect sizes together with a very heterogeneous phenotype resulted in no significantly differentially expressed transcripts or proteins after multiple testing correction.

We further evaluated different methods and custom extensions to apply pathway enrichment analysis on a multi-omics dataset. We conducted a simulation study to make informed decisions on diverse parameters and evaluate performance of the different methods and across various simulation scenarios, including realistic scenarios but also exploring hypothetical settings taking into account pathway activations in different omics, second-omic layer coverage of genes measured and correlation between omics. The ad-hoc combination of single-omic MONA approaches proved to be the most robust and performed extremely well across all simulated scenarios, while the direct integration of multi-omics in the MONA cooperative model only showed rather negligible benefits in case of shared pathways across omics.

Therefore, we applied both the single-omic MONA approach with the multi-omic OR combination as well as the direct multi-omics MONA integration cooperative model to our AF cohort with taking into account direction of effect. We recovered specific differences between omics and AF subtypes with respect to metabolic changes and immune responses. Most of the identified processes were directly related to the disease phenotype, such as cardiac muscle contraction or in line with current literature on the topic describing e.g. inflammation and changes in the energy metabolism [Tu et al., 2014, Hu et al., 2015, Staerk et al., 2017].

In the following, we would like to address specific challenges at the different steps that were just summarized.

6.6.1 Differential expression analysis

The lack of detection of differentially expressed genes or proteins can have multiple reasons. As mentioned before, new measurement technologies could reduce the noise level and increase sensitivity. Additionally, we are dealing a with small sample size and with rather small effect sizes which restricts the statistical power. Tissue samples in general display varying cell type composition. Moreover, donor specific effects including risk factors and comorbidities further contribute to substrate heterogeneity. Heterogeneity is additionally observed for the AF phenotype. We considered the two subtypes of prevalent and incident AF, but even more forms of AF exist and further research on the characterization of those patients and possible subgroups are needed. Finally, tissue specimens were derived from living donors, that still retained most of their cardiac function. Therefore, physiological changes are limited compared to those observable in deceased donors or from animal models.

6.6.2 Simulation study

Pathway enrichment analysis was a possible solution to overcome the limitations of individual gene differential expression analysis, additionally we wanted to make use of our multi-omics data available for the AFHRI-B cohort. However, currently there is a very limited amount of multi-omics gene set enrichment methods available. Each method comes with its own challenges including what cutoffs to use to define significance in the differential analysis as well as to determine significantly enriched or activated pathways.
As part of this thesis, we also introduced two extensions to the MONA model, making it possible to take into account direction of effect and properly combining gene-centered data modalities with e.g. metabolite data.

The significance threshold for the differential analyses directly influences the probability of including false positive and false negative findings. Therefore, we wanted to evaluate how different false positive and false negative rates, the coverage and the correlation between omics and different simulation scenarios for the activated pathways impacted the different methods.
Our simulation approach inspired by classical power analysis allowed us to model simulated multi-omics summary statistics with pre-defined false positive and false negative rates in the form of Z scores mimicking a test statistic. By using one set of input, we were able to apply and evaluate all the different methods utilizing these scores.

When broadly comparing the different methods, all of them profited from the consideration of direction of effect. Higher coverage of the second omic and higher correlation between omics improved performance with the strongest effects being observed for the cases with full coverage of the second omic.
Overall, GSEA performed much worse than MGSA and MONA. For the ad-hoc integration of multi-omics on pathway level, we determined that combining additive effects between omics was more accurate than evaluating the overlap, as MGSA and the single-omic MONA were not prone to false positive activations even with very high false positive rates in the simulations. This is especially important when considering the choice of the significance cutoff for the original differential expression results.

The second central aim was to uncover how to gain most from multi-omics data. As reported before, the MONA cooperative multi-omics model [Sass et al., 2013] outperformed the single MONA and MGSA single-omic approaches for shared pathways across omics. This was also true, when applying the ad-hoc multi-omics OR and AND combinations. However, with our simulation study, we also specifically addressed the scenario of pathways being exclusively activated in one omic only and how multi-omics integration approaches recover those. Here, the MONA cooperative model showed a much poorer performance than the single-omic OR-combination. Additionally, running the single-omic methods separately first also gives more information about what effects being how strongly observed on the different levels. This might give important insights and have consequences with respect to follow-up experiments or validation of biological

hypotheses.
The reduced performance of the multi-omics MONA cooperative model was due to a surprising bias of detecting activations from the first omic layer only, which was an artifact of the current implementation and not based on the underlying Bayesian network model. Similarly, this is also the most likely explanation for the counterintuitive increase in performance for higher false negative rates for scenarios with a 100 % coverage of the second omic. Additionally, interference due to the artificial nature of the simulated data cannot be ruled out either, even though no similar problems occurred for any of the other methods. These uncertainties need to be taken into account when interpreting corresponding results where independent pathway activations between omics are possible.

Finally, it has to be mentioned, that while GSEA was performing worse in comparison to the other methods, it is the only one giving concrete information about what genes might be driving the enrichment by the leading edge returned for each pathway. Unfortunately, GSEA completely looses all information about any significance threshold, which is still incorporated in all Bayesian approaches.

In conclusion, the best method to use for multi-omics data highly depends on the research question at hand. Specifically, false positive results have to be taken into account for GSEA, at the same time the additional leading edge genes can provide valuable information on which genes drive the enrichment. Due to this fact, GSEA was chosen for our targeted *trans* QTL approach [Assum et al., 2022a].
MGSA and the single-omic MONA method are very comparable, however, the MGSA implementation as a R package[15,16] is much more convenient than the current implementation of MONA, which on unix systems is run as a Windows application using the mono framework[17].
If the research question is to accumulate evidence across different omics to uncover the same kind of activation, the MONA cooperative model is best suited. Here, we focused on transcriptomics and proteomics data integration, where the OR combination of multiple data types proved most reliable across all simulation scenarios. As a next step, this could be extended to also incorporate other omics, such as metabolomics.

6.6.3 AFHRI-B AF pathway enrichment

Using the insights gained from the simulation study, we applied the single- and multi-omics MONA approaches to the AFHRI-B differential expression results of atrial tissue transcriptomics and proteomics for prevalent and incident AF.

As described in detail in the previous chapters, we again observe large differences in the different omics.
Next to actual biological effects, also technical factors, i.e. the measurement of relevant

[15] `https://www.bioconductor.org/packages/release/bioc/html/mgsa.html`
[16] `https://github.com/sba1/mgsa-bioc`
[17] `https://www.mono-project.com/`

genes can be the reason. Therefore, we included the gene set sizes of each gene set and coverage across the different omics as a heat map in Figure 6.13. In general, around 90 % of the genes annotated in the KEGG pathways were measured on transcript level while only around 10 % were available in proteomics.
Diverging activations, such as the disagreement in regulation of the ECM receptor interaction, which is up-regulated in one omic and down-regulated in the other can also be caused due to different groups of individual genes being observed and activated in the different omics.

We first analyzed the prevalent AF subtype. Multiple metabolic pathways were identified as being regulated, including multiple gene sets associated to mitochondrial metabolism. The involvement of such pathways has been discussed in depth by Tu et al. [2014]. Similar groups of genes were also identified in our *trans* pQTL analysis [Assum et al., 2022a], with a focus of the genetic variation and molecular consequences on protein level to AF. Additionally, the two very specific pathways cardiac muscle contractions and calcium signaling pathway were highly relevant with respect to the arrhythmogenic disease that we consider.
While we saw the strongest effects on protein level, one single metabolic pathway was discovered in transcriptomics by the up-regulation of insulin signaling pathway. This finding is specifically interesting in the context of diabetes being a major risk factor [Staerk et al., 2017] of AF and more follow-up analyses due to the complex nature of possible interactions are necessary.
Most of the pathways not regulated on transcriptome but on protein level had sufficient coverage of the respective mRNA. Here, technical limitations of microarray measurements compared to RNA-sequencing as already discussed in the previous chapters might be a possible explanation. More importantly though, a general lack of correlation between transcript and protein abundance, especially when considering cross sample variation for the same gene, has been observed consistently [Schwanhäusser et al., 2011, Liu et al., 2016, Edfors et al., 2016, Wang et al., 2019a, Eraslan et al., 2019, Jiang et al., 2020] and variation on protein level is significantly driven by post-transcriptional regulation. Such mechanisms have been underreported due to the focus on transcriptomics experiments.

For post-operative AF, the majority of identified pathways were involved in inflammation. This is again in accordance with literature, where inflammation does not only play a crucial role in general AF pathogenesis [Hu et al., 2015], but Watt et al. [2021] specifically identified the up-regulation of various inflammatory genes in left atrial tissue samples of patient developing post-operative AF.
Although the dis-regulated ECM receptor interaction may indicate changes in tissue structure which are commonly observed in AF patients, it is difficult to draw specific conclusion especially due to discrepancy of transcript and protein results.
The same holds true for the interpretations of activations for the pathways leishmaniasis, RIG-I-like receptor signaling pathway and Parkinson's disease. Follow-up analyses focusing on the specific genes and pathways annotations are therefore needed.
Furthermore, multiple pathways are connected to cellular processes that are relevant in many general biological processes. At this point, no clear interpretation of the exact

involvement of those mechanisms in the context of AF can be derived.

Given the insights by our simulation study about how pathway activations in the different omics are recovered and the strong bias towards the first omic layer, it is very interesting to observe that the majority of the regulated gene sets on protein level were also replicated by the multi-omics approach. Since we observe most of the relevant genes in this pathway, missing coverage should not to be the reason. Hence, most likely activations were not strong enough to be picked up in the transcriptomics data alone, but some signal might still be contained even on this level, leading to the higher posterior in the multi-omics model.
On the same note, it is not surprising that the activated pathways identified on transcript level for incident AF were replicated on multi-omics level rather than those on protein level, specifically since the coverage of proteomics for the relevant pathways was extremely low.

In summary, we see the strongest single-omics activations for each of the AF subtypes are also represented by the multi-omics integration. The prevalent AF phenotype shows most effects on protein level, while incident AF is more pronounced in transcriptomics. As prevalent AF may include more long-term changes compared to incident AF expedited by newly introduced stress through cardiac surgery, differences in the regulation of transcription and translation might also be explained by different timescales [Schwanhäusser et al., 2011, Liu et al., 2016].

6.6.4 EnrichmentNodes

In the last part of this chapter, we present the EnrichmentNodes plugin enabling the use of all the different pathway enrichment methods as easily and convenient as possible. We specifically addressed researchers from an experimental rather than computational background.

Therefore, we implemented basic R functions performing diverse pathway enrichment techniques as generic KNIME nodes in the KNIME framework. This enables the creation of complex workflows by selecting and connecting the corresponding analysis steps as single nodes in a graphical user interface.

All that is required is the installation of KNIME, Docker and the Generic KNIME nodes. From there, our EnrichmentNodes plugin can be installed by adding a single file in a corresponding folder.
While this is very convenient on the user side, KNIME plugin creation and maintenance comes with some challenges for the developers. Documentation was in parts incomplete, outdated or missing. This was specifically true for automatic deployment using the buckminster tool[18], which is why those options were not pursued further.

Our data structures are kept very general by providing a basic summary table, a list

[18] `https://projects.eclipse.org/projects/tools.buckminster`

of used parameters and also a highly flexible slot for any kind of method-specific object, which makes it easily extendable to other methods. Additionally, separate docker images can be specified for each node. Therefore, independent development for different analysis nodes is possible in case of conflicting dependencies.

In the case of computational support or high performance computing infrastructure, the setup of other graphical workflow managers, such as Galaxy [Afgan et al., 2018] might be better suited. Also, for more computationally versed end user, we highly recommend making use of our docker available at `https://hub.docker.com/r/inesassum/enrich`. This docker offers a Rstudio session with all pre-installed dependencies and R functions used for the enrichment analyses described in this thesis. Additionally, even more enrichment and simulation methods are available in this setting, such as random forest-guided pathway analysis[19] [Seifert et al., 2020], and get extended continuously. Furthermore, it is much more flexible to use and adapt. Hence, to build custom pipelines and workflows, Nextflow [DI Tommaso et al., 2017] or Snakemake [Köster et al., 2021] might be better suited than the KNIME implementation.

6.6.5 Current limitations and outlook

Altogether, we showed the importance of integrating multiple omics and evaluating grouped activations rather than single gene associations. However, some limitations need to be taken into account.

First, in this specific case we used data from atrial tissue. We have discussed the limitations connected to biological and technical factors in detail before. At this point, we want to specifically extend this to the definition and heterogeneity of the clinical AF phenotypes, while acknowledging that other tissues may be different and require other methodological approaches.
Second, for the setting of pathway enrichment analysis, we highly depend on the accuracy and quality of the used gene set annotations. We explicitly want to point out that this was not evaluated at all in our simulation study. The choice of gene sets has large implications with respect to the representation of biological processes, how fine-grained, context specific or generalizable they are in an experimental setting.
In this context, it is very important to also consider that one pathway can have multiple components that can also have opposite effect with respect to up- or down-regulation.
Third, the simulation study was designed to evaluate and highlight specific areas of interest. Even though we tried to include scenarios as close as possible to our actual data or measurements that potentially might be available in the future, the simulated data remains a simplification of complex biological systems in order to make larger scale analyses feasible.

A new and more flexible implementation of Bayesian models underlying MONA might drastically improve its capabilities. Additionally, it might be interesting to pair the

[19] `https://github.com/szymczak-lab/PathwayGuidedRF`

enrichment results from MONA with leading edge information from GSEA. Also, pathway approaches relying on interaction networks such as e.g. PathwAX[20] [Ogris et al., 2016] have not been evaluated in this thesis.
Furthermore, detailed follow-up analyses of every pathway are needed. Pathways often integrate multiple subunits, including different up- and down-regulated genes. Also, the investigation of rate-limiting enzymes might add important information.

In this work, we only considered multi-omics measurements from the same species, e.g. human atrial transcriptomics and proteomics. However, similar work could easily be extended to multiple species such as comparing mouse and human data.

[20] http://pathwax.sbc.su.se/

7 Discussion and Outlook

In this thesis, we present a comprehensive multi-omics analysis of gene regulation in human atrial tissue and how genetics as well as intermediate molecular phenotypes are related to atrial fibrillation (AF).

7.1 Discussion

Let us come back to the original motivation and challenges:

1. Gene regulation is a complex process, which is very specific to the tissue of question. However, downstream consequences of genetic variation, which would be vital to understand the underlying mechanisms, are mostly unknown.

   By integrating genomics, transcriptomics and proteomics data, we were able to derive a comprehensive map of genome-wide *cis*-regulatory mechanisms highlighting differences in regulation on transcript and protein level.

As the first study to integrate proteomics data in human atrial tissue, we provide valuable information about gene regulation going past transcriptional regulation which could be derived from existing eQTL results [Gamazon et al., 2018]. The general lack of overlap of eQTLs and pQTLs and specifically our functional *cis* QTL categories further illustrate the vast differences in consequences of genetic variation which can affect transcriptional as well as post-transcriptional processes independently and is only observable with matched transcriptomics and proteomics data. Our integrative approach enables us to narrow down possible underlying mechanisms by specifically considering transcriptional or post-transcriptional regulatory elements affecting processes such as transcription, splicing, translation or degradation. Different mechanisms are also defined by the position of *cis*-regulatory elements in the gene sequence, e.g. regulation of translation by variants located in exons. Using public annotations, we assessed overrepresented regulatory elements to infer possible mechanisms underlying the genetic regulation, such as enhancer regions or TSSs for transcriptional regulation and exonic regions for translational regulation. Similar enrichments were already described by Lappalainen et al. [2013] and Battle et al. [2015]. Moreover, tissue specificity of regulation is witnessed by generally low overlap of quantified proteins as well as corresponding identified pQTLs compared to existing pQTL results in plasma [Sun

et al., 2018].

However, more accurate transcript quantifications using RNA-sequencing, and a better coverage of proteins, specifically for lowly expressed regulatory proteins such as transcription factors (TF) or inflammatory proteins could further improve the discovery of eQTLs and pQTLs.

2. More than one hundred genetic loci have been associated with AF. However, their function and affected genes often remain elusive.

   The integration of *cis* eQTLs and pQTLs in combination with a multitude of annotations makes it possible to evaluate the context of each GWAS hit individually. Our functional *cis* QTL categories aid in shortlisting possible underlying mechanisms or causal factors. Specifically, also non-significant associations contribute valuable information.

Cis eQTL and pQTL analyses can help elucidate some of those relations by proposing candidates for causal SNP-gene relations. We confirmed the AF disease relevance of the discovered QTLs by an overrepresentation of *cis* eQTLs and pQTLs for GWAS loci associated with cardiovascular measurements or disease phenotypes. Proteins are suggested as a more direct determinant for phenotypic consequence [Battle et al., 2015]. In this context, independent pQTLs are especially important as they cannot be detected with transcriptome based studies. Therefore, the link between the genetic variant and corresponding gene can only be made when including proteomics. Additionally, due to arbitrary significance cutoffs, also information about hits which did not reach significance but do show regulation can be used to select relevant candidate genes as well as discarding those with a lack of association. Full summary statistics, including additional annotations such as miRNA, RBP and TF binding sites of all SNP-gene pairs, are publicly available[1] [Assum et al., 2021] for future investigations.

We established numerous links for GWAS hits with unknown function. While some of them might have already been known and the initial links are very important, further analyses including experimental evaluation are needed to uncover concrete mechanisms.

3. *Trans*-genetic effects have been shown to majorly contribute to common disease. However, current analyses are largely impaired due to the vast search space in combination with limited sample sizes.

   We propose a PRS-based candidate selection approach to make targeted *trans* QTL analyses possible in a relatively small clinical cohort.

Based on the idea of the omnigenic model [Boyle et al., 2017, Liu et al., 2019, Võsa et al., 2021], our QTS analyses integrated molecular data from our relatively small clinical cohort with public data derived from large cohorts - such as the polygenic risk score and AF GWAS summary statistics. By correcting for often times stronger *cis* effects,

[1] `https://doi.org/10.1101/2020.04.06.021527`

our enrichment-based approach pre-selected transcripts and proteins based on their correlation with the PRS as a proxy for accumulated *trans* effects. Additionally, the GO biological processes were used to ensure shared molecular function. Restricting the variants to be tested to AF GWAS hits secured the disease link and further reduced the number of tests.

Of course, due to the strong selection criteria, many relevant associations remain unknown, until larger datasets become available.

4. The omnigenic model proposes the existence of core genes. Due to their direct link to the phenotype, these core genes could be a key component of understanding more complex disease mechanisms. However, so far it has been challenging to identify core genes due to complex interactions and small effect sizes.

   The resulting two *trans* eQTLs and five *trans* pQTLs, as well as the analysis of the NKX2-5 TF network with 13 identified targets result in overall 20 putative core genes for AF.

The omnigenic model also proposes the existence of core genes [Boyle et al., 2017], which are highly relevant genes at the center of molecular networks that accumulate *trans*-genetic variation and are directly linked to a phenotype. However, they have not yet been explored in the context of AF.
Trans QTLs genes were characterized by driving the enrichment of biological processes correlating with the accumulation of *trans*-genetic effects. Further exploring the role of NKX2-5, we followed up on the NKX2-5 TF network. By exploiting the molecular link of the *trans* QTL SNP, the TF transcript, transcription factor activity as well as transcript and protein data of potential target genes, we derived 13 NKX2-5 TF targets, which also showed co-expression with the *NKX2-5* transcript in two independent RNA-seq datasets. Importantly, we observed two key properties that were predicted for core genes [Boyle et al., 2017]. First, various putative core genes had mutations which were known to be causal for AF, arrhythmias or also other cardiomyopathies, and second, core genes are expected to be directly linked to the phenotype [Boyle et al., 2017]. The 13 NKX2-5 targets showed a very strong collective down-regulation with respect to AF on protein level, even though the original cohort AF phenotypes were not used in any form until this point. The disease association of the NKX2-5 TF targets was further replicated by an independent transcriptomics as well as proteomics dataset and supported by findings in previous literature (see Table 5.11).
Neither the link of the GWAS SNP rs9481842 to NKX2-5 nor the relation between the TF and its targets were known in the context of AF. Also, when considering transcriptomics and proteomics, stronger disease associations have been observed on protein level for the NKX2-5 TF targets in both, our AFHRI-B cohort as well as the two independent replication datasets.

Still, due to the unique nature of this cohort including the AF case-control context as well as the disease- and tissue-specific analyses, replication of the *trans* QTLs or the

link between rs9481842 and the NKX2-5 TF targets was not feasible.

5. Current measurement technologies enable the evaluation of multiple molecular omics. However, it is difficult to leverage the full potential of multi-omics data to derive de-regulated biological processes in AF.

   We have extended existing methods for multi-omics gene set enrichment analysis and made them accessible to a broader audience in the graphical workflow framework KNIME. Our simulation study warranted insights on the performance of different methods under varying underlying pathway activations. By leveraging the gained knowledge, we were able to identify common molecular mechanisms underlying AF even in the case of extremely small effect sizes.

With the exception of MONA [Sass et al., 2013], most pathway enrichment methods like GSEA [Subramanian et al., 2005, Korotkevich et al., 2019] or MGSA [Bauer et al., 2010] focus on one omic only. Therefore, we conducted a simulation study comparing different multi-omics pathway enrichment approaches including extensions of existing methods as well as ad-hoc combinations of single data modality analyses. Using multi-omics data in general and including direction of effect greatly improved model performance. The ad-hoc OR combination of single-omic MGSA and MONA approaches showed consistently good performance across all simulation scenarios, including shared and independent pathway activations per omic, different correlation between the two data types and various coverages of the second omic, even for very high false positive and false negative rates in the simulated data.
The direct multi-omic MONA integration only outperformed the ad-hoc OR combination in the case of shared pathway activations across omics but failed to detect the corresponding activations if they were independent for each modality due to a tendency to exclusively detect the pathways active in the first omic.
Given the insight, that the MONA approaches can cope even with very high false positive rates, we were able to apply both the multi-omics MONA ad-hoc OR combination as well as the direct integration to our AFHRI-B transcriptomics and proteomics AF differential expression results with relaxed significance thresholds. We observed a broad range of mechanisms, including cardiac, metabolic and immune pathways which were in line with previous literature [Tu et al., 2014, Staerk et al., 2017]. The two AF subtypes, prevalent and incident AF, showed distinct regulation. For both subtypes and both omics, we found processes which were regulated in only transcriptomics or proteomics exclusively. As to be expected from the results of our simulation study, the OR combination and direct integration matched very closely.
To make similar multi-omics analyses more accessible without programming skills, we developed the EnrichmentNodes plugin for the KNIME Analytics Platform for integrating the presented analysis approaches in an interactive graphical workflow development framework.

Many of the AF core genes, as well as the enriched processes which were identified,

were related to metabolism. Therefore, as those measurements are also available for the AFHRI-B cohort, tissue as well as serum metabolomics analyses provide promising opportunities for validating the underlying mechanisms.
In this regard serum measurements in other cohort studies revealed that higher serum concentrations of long-chain acylcarnitines, which are essential for mitochondrial energy metabolism, associated with prevalent and incident AF. The corresponding paper *An arrhythmogenic metabolite in atrial fibrillation*, by J. Krause, ..., I. Assum, M. Heinig, ..., J. Stenzig* and T. Zeller* (*authors contributed equally) is currently under revision.

Integrating metabolites in gene-regulatory networks is still challenging. Manual harmonization and curation of shared pathway maps is tedious and also highly dependent on the metabolomics measurement techniques. Also, general annotations will not take tissue-specific effects into account. Instead, we propose linking the metabolomics, genomics, transcriptomics and proteomics data of this unique cohort via mQTL associations and correlation networks.

In general, further combination of even more omics types for complex diseases such as AF could be the next step. Important risk factors include a large variety of genetic but also clinical and molecular features. To derive more personalized risk predictions, we propose a risk score incorporating classical risk factors, the PRS which was also used in the QTL analyses as well as molecular markers such as transcripts, proteins or metabolites. Moreover, besides omics measured in tissue samples we want to apply the same strategy to serum metabolomics, as such biomarkers are more feasible to be applied in a larger clinical setting. First results in the AFHRI-B cohort look promising and the KORA cohort [Holle et al., 2005] can be used as replication.

As part of the e:Med networking fonds Networks of heart disease: Systems medicine approach to improve heart health (coNfirm)[2] pathway enrichment methods were also generalized across different species and studied for a broader range of cardiovascular diseases, such as heart failure, myocardial infarct and atrial fibrillation. By combining data and knowledge from different stages, including animal models, human-induced pluripotent stem cells to human cohort data, we want to identify and refine our understanding of disease mechanisms. Specifically, we want to leverage stronger effect sizes in experimental model organisms compared to human data with the goal of developing better treatment options.

[2] https://www.sys-med.de/en/networking/spalte-2/networking-fonds/confirm/

7.2 Conclusion

In this thesis, we have showcased how multi-omics integration improves our understanding of complex traits such as AF and helps to overcome challenges such as data heterogeneity, small effects and restricted sample sizes.
The methods used in this thesis have been tailored specifically to AF, but the underlying concepts can be applied in a much more general context. In conclusion, we can emphasize not only the benefit but also the necessity of evaluating multiple omic types to investigate molecular mechanisms.

A Supplementary Figures

A.1 Simulation study - extended benchmarking results

A.1.1 Extended GSEA performance results

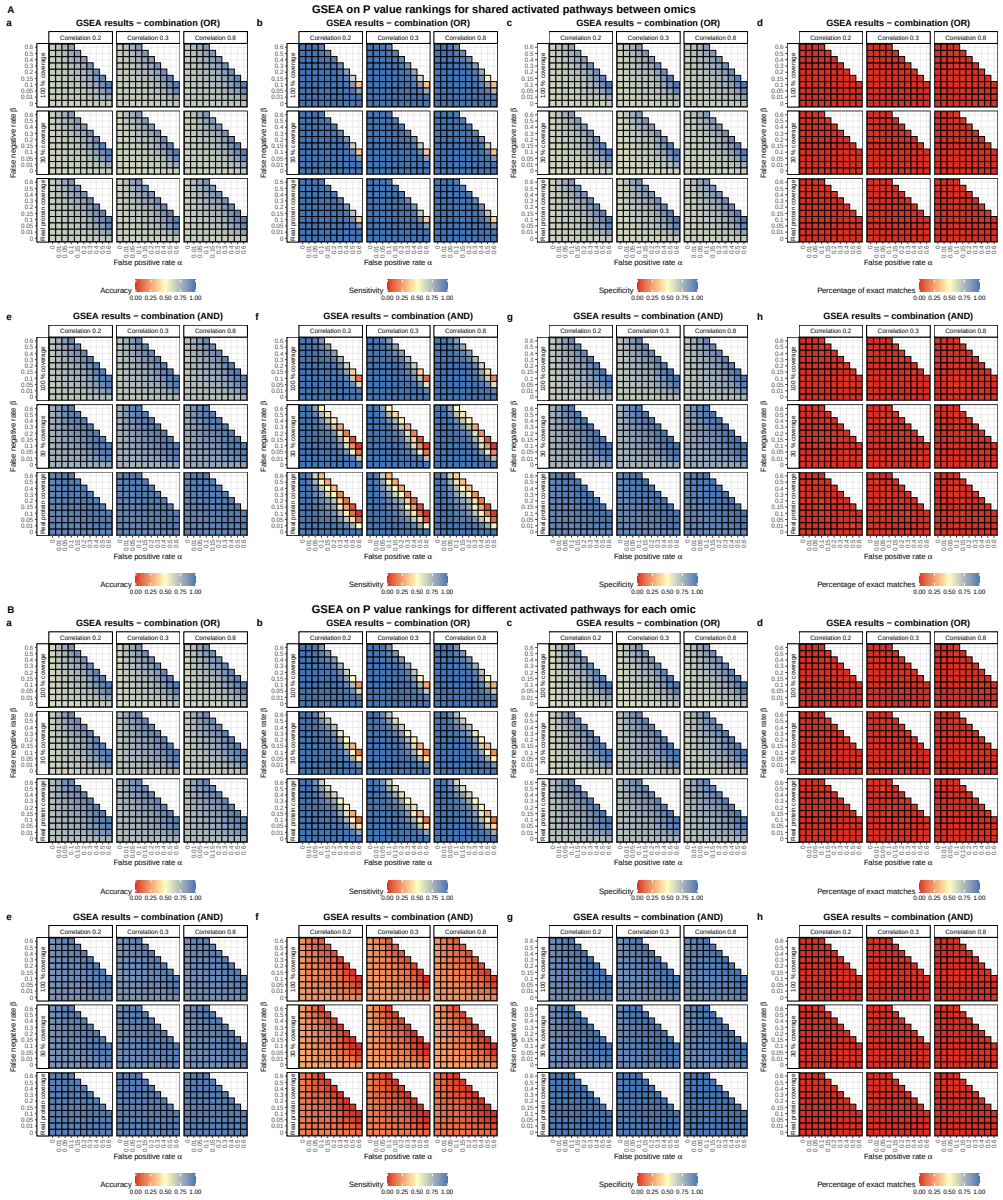


Figure A.1: GSEA performance using P value/absolute Z score rankings.
GSEA applied to absolute Z scores of the simulated summary statistics representing P value rankings. Panel A presents the results for the scenarios, where the activated pathways were shared across omics, for the results in panel B different pathways were active in different omics. The average accuracy, sensitivity and specificity across 100 simulation replicatess is visualized across the different combinations of predefined false positive and false negative rates and the last panel of each row represents the percentage of cases, where the exact match of simulated gene sets were recovered. To evaluate the multi-omics performance, GSEA results on each omic were combined using additive effects, i.e. a gene set is considered to be found active by the method, if it came up significant in either omic (OR, panel a-d) as well as overlapping effects, i.e. a gene set is considered active if it was significant in both omics (AND, panel e-h). GSEA, gene set enrichment analysis;

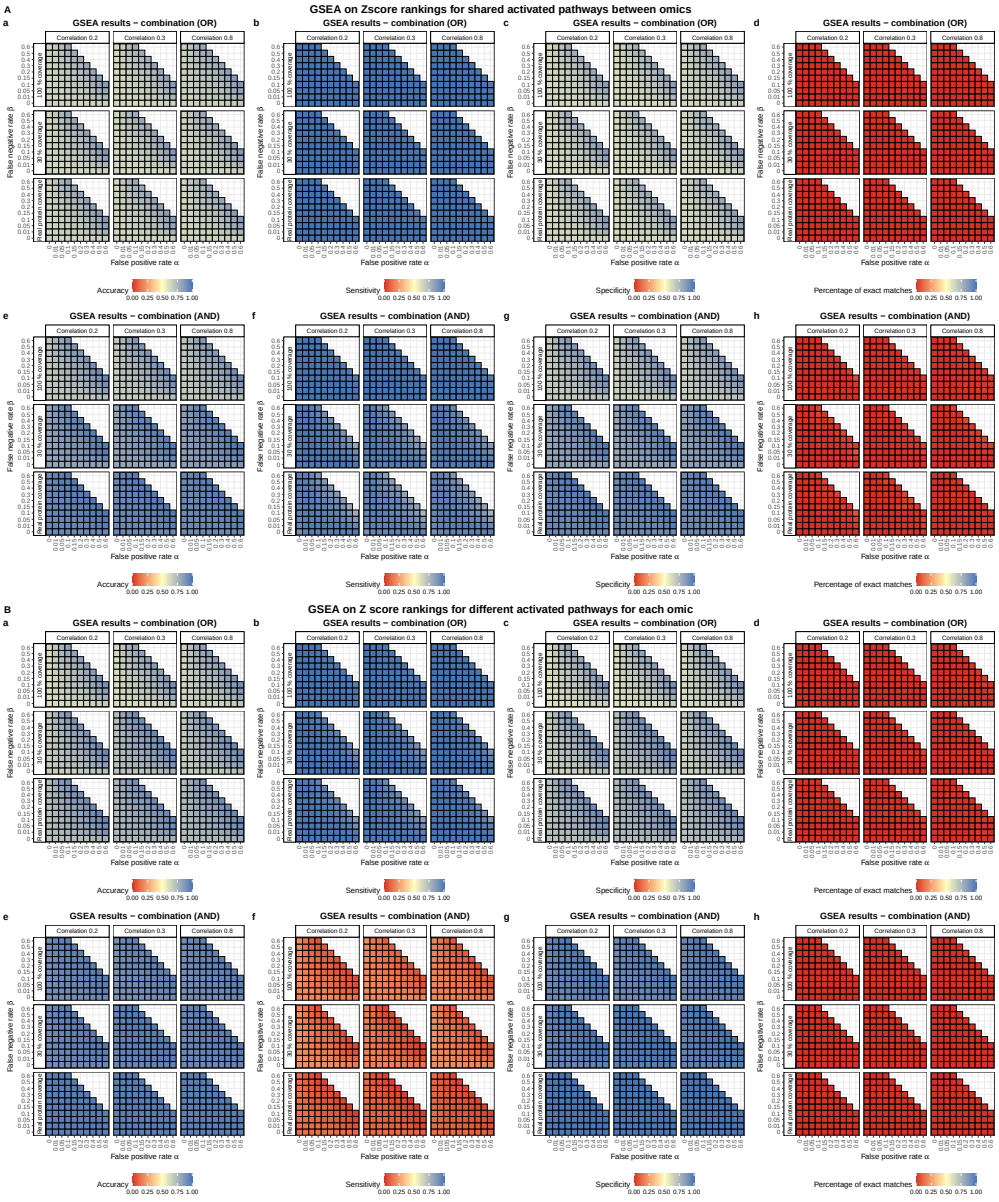


Figure A.2: GSEA performance using signed test statistic/Z score rankings.
GSEA applied to Z scores of the simulated summary statistics representing rankings based on a test statistic or signed P values. Panel A presents the results for the scenarios, where the activated pathways were shared across omics, for the results in panel B different pathways were active in different omics. The average accuracy, sensitivity and specificity across 100 simulation replicates is visualized across the different combinations of predefined false positive and false negative rates and the last panel of each row represents the percentage of cases, where the exact match of simulated gene sets were recovered. To evaluate the multi-omics performance, GSEA results on each omic were combined using additive effects, i.e. a gene set is considered to be found active by the method, if it came up significant in either omic (OR, panel a) as well as overlapping effects, i.e. a gene set is considered active if it was significant in both omics (AND, panel b).
GSEA, gene set enrichment analysis;

A.1.2 Extended MGSA performance results

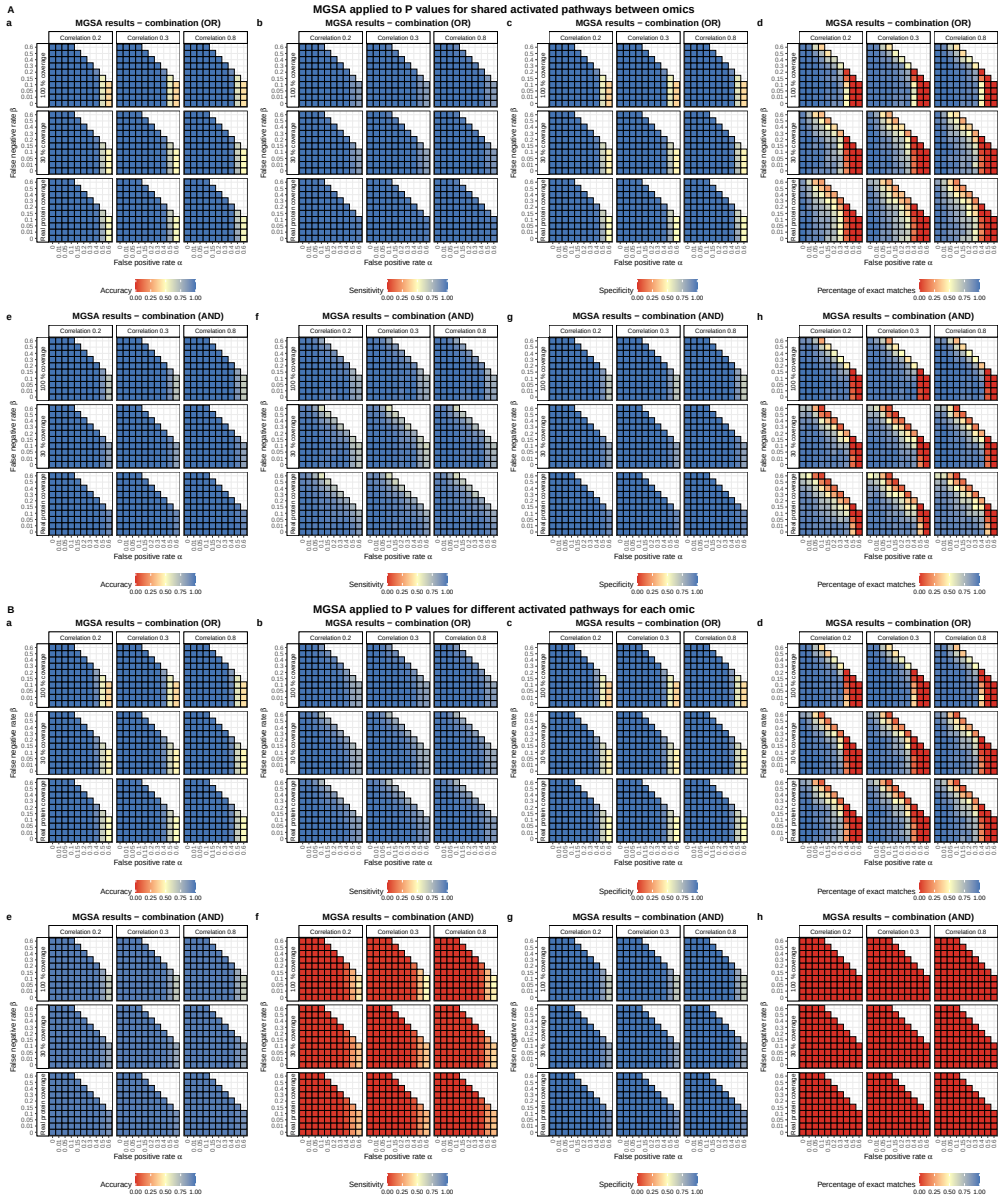


Figure A.3: MGSA performance using P values/absolute Z scores.
MGSA applied to absolute Z scores of the simulated summary statistics representing P values without taking into account direction of effect. Panel A presents the results for the scenarios, where the activated pathways were shared across omics, for the results in panel B different pathways were active in different omics. The average accuracy, sensitivity and specificity across 100 simulation replicates is visualized across the different combinations of predefined false positive and false negative rates and the last panel of each row represents the percentage of cases, where the exact match of simulated gene sets were recovered. To evaluate the multi-omics performance, MGSA results on each omic were combined using additive effects, i.e. a gene set is considered to be found active by the method, if it came up significant in either omic (OR, panel a-d) as well as overlapping effects, i.e. a gene set is considered active if it was significant in both omics (AND, panel e-h).
MGSA, model-based gene set analysis;

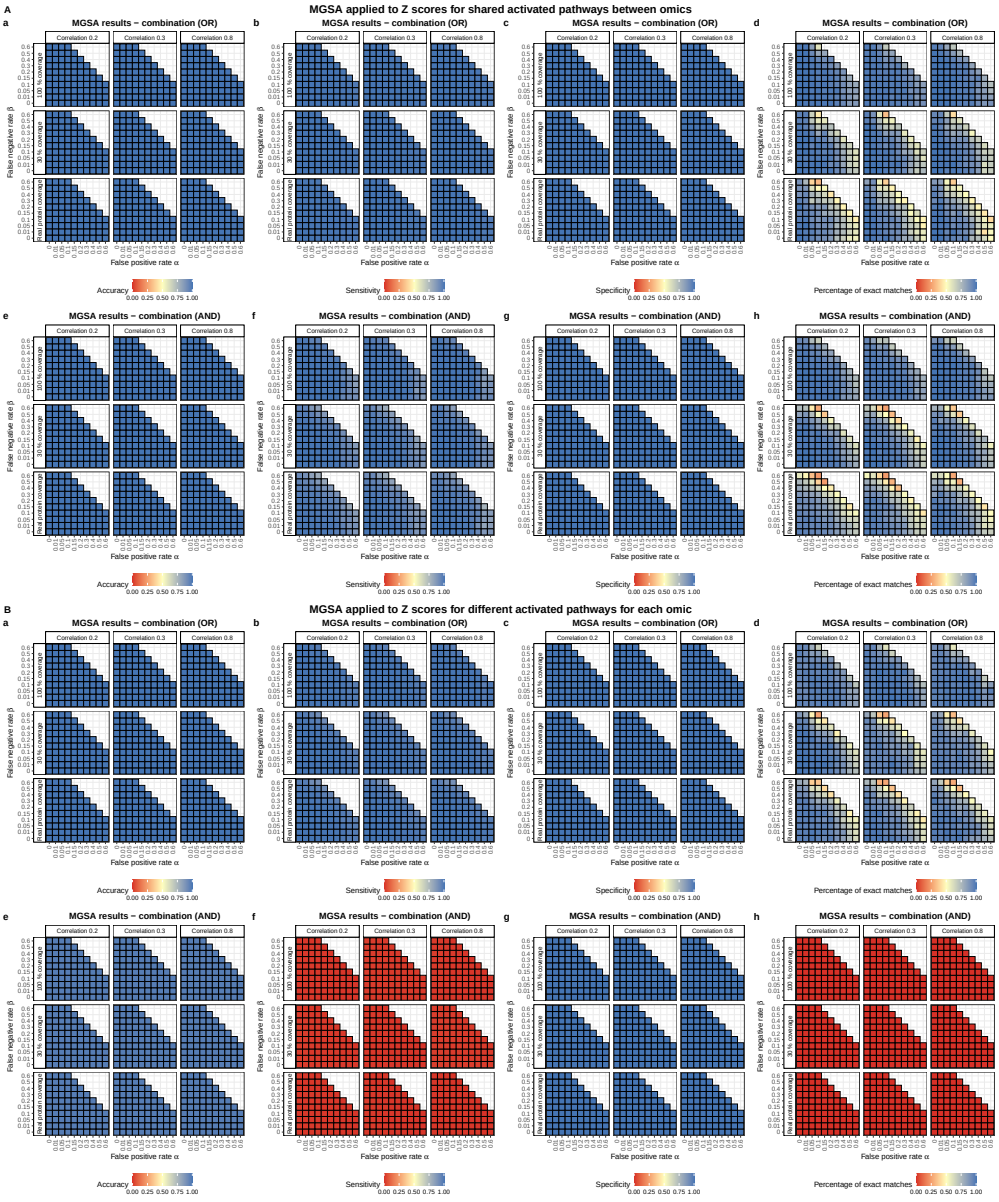


Figure A.4: MGSA performance using a test statistic/Z scores.
MGSA applied to absolute Z scores of the simulated summary statistics representing a test statistic or signed P values taking into account direction of effect. Panel A presents the results for the scenarios, where the activated pathways were shared across omics, for the results in panel B different pathways were active in different omics. The average accuracy, sensitivity and specificity across 100 simulation replicates is visualized across the different combinations of predefined false positive and false negative rates and the last panel of each row represents the percentage of cases, where the exact match of simulated gene sets were recovered. To evaluate the multi-omics performance, MGSA results on each omic were combined using additive effects, i.e. a gene set is considered to be found active by the method, if it came up significant in either omic (OR, panel a-d) as well as overlapping effects, i.e. a gene set is considered active if it was significant in both omics (AND, panel e-h).
MGSA, model-based gene set analysis;

A.1.3 Extended single-omic MONA performance results

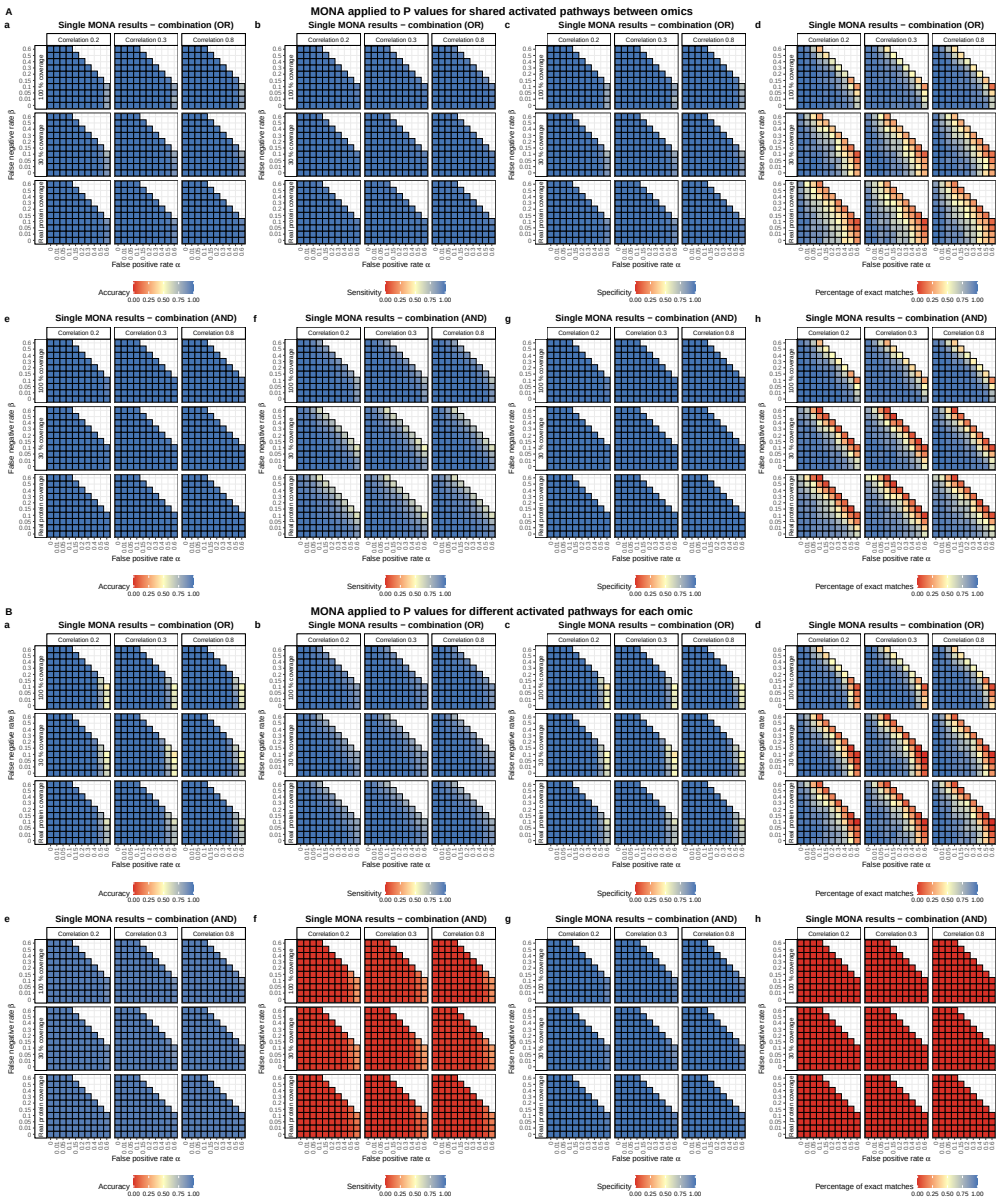


Figure A.5: Single-omic MONA performance using P values/absolute Z scores.
Single-omic MONA applied to absolute Z scores of the simulated summary statistics representing P values without taking into account direction of effect. Panel A presents the results for the scenarios, where the activated pathways were shared across omics, for the results in panel B different pathways were active in different omics. The average accuracy, sensitivity and specificity across 100 simulation replicates is visualized across the different combinations of predefined false positive and false negative rates and the last panel of each row represents the percentage of cases, where the exact match of simulated gene sets were recovered. To evaluate the multi-omics performance, MONA results on each omic were combined using additive effects, i.e. a gene set is considered to be found active by the method, if it came up significant in either omic (OR, panel a-d) as well as overlapping effects, i.e. a gene set is considered active if it was significant in both omics (AND, panel e-h).
MONA, multi-level ontology analysis;

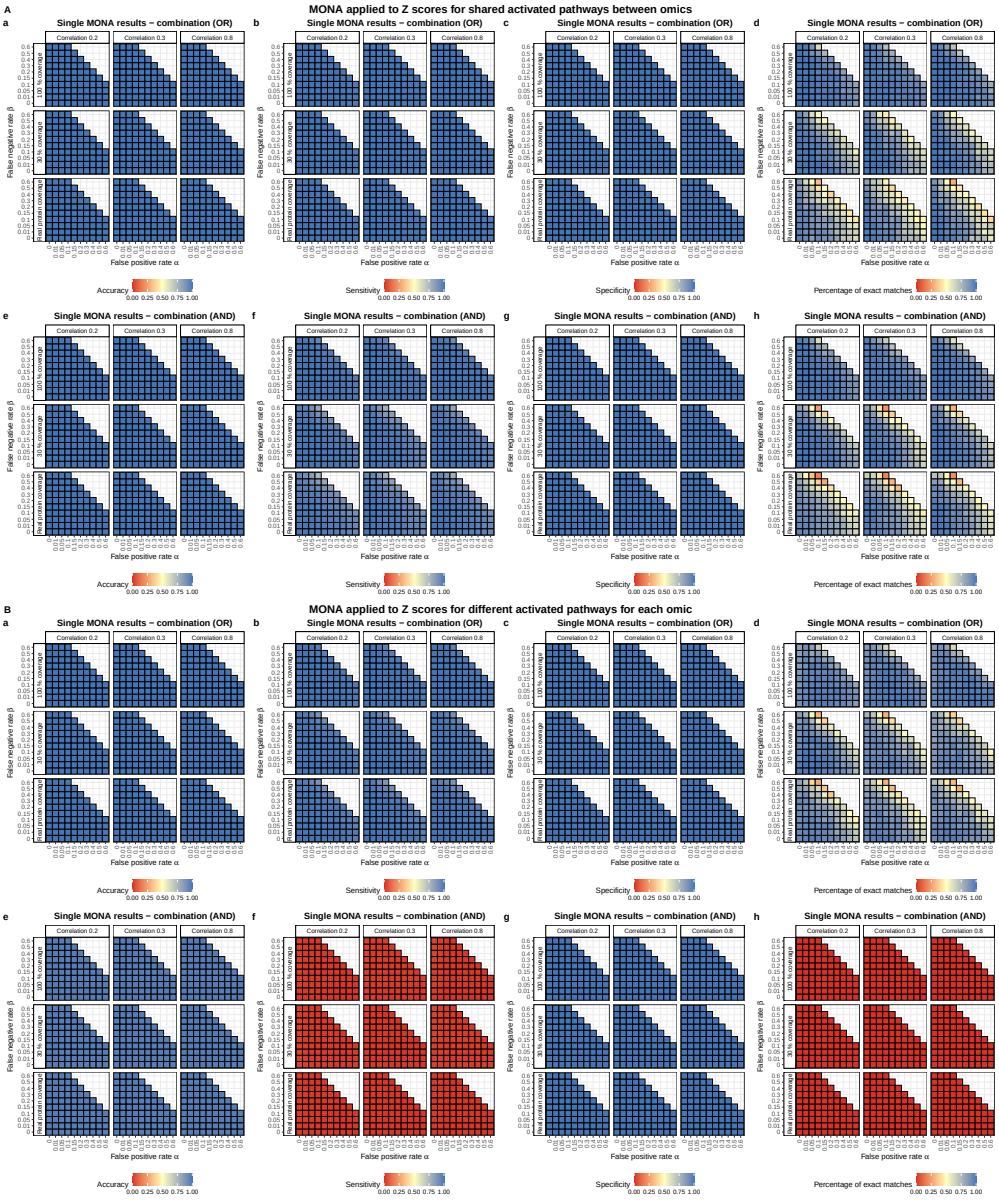


Figure A.6: Single-omics MONA performance using a test statistic/Z scores.
Single-omics MONA applied to absolute Z scores of the simulated summary statistics representing a test statistic or signed P values taking into account direction of effect. Panel A presents the results for the scenarios, where the activated pathways were shared across omics, for the results in panel B different pathways were active in different omics. The average accuracy, sensitivity and specificity across 100 simulation replicates is visualized across the different combinations of predefined false positive and false negative rates and the last panel of each row represents the percentage of cases, where the exact match of simulated gene sets were recovered. To evaluate the multi-omics performance, MONA results on each omic were combined using additive effects, i.e. a gene set is considered to be found active by the method, if it came up significant in either omic (OR, panel a-d) as well as overlapping effects, i.e. a gene set is considered active if it was significant in both omics (AND, panel e-h).
MONA, multi-level ontology analysis;

A.1.4 Extended MONA multi-omics integration performance results

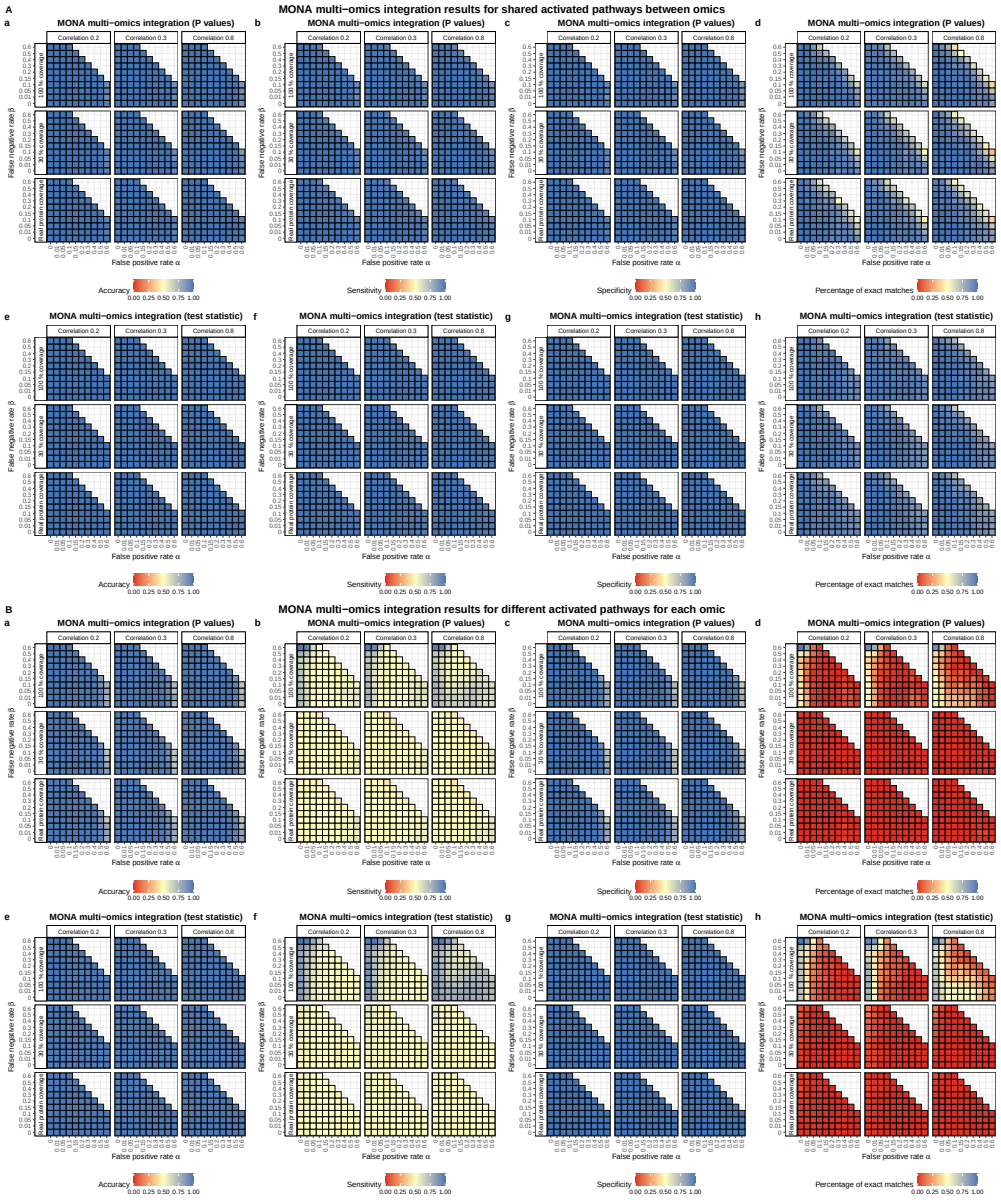


Figure A.7: MONA direct multi-omics integration performance.
MONA direct multi-omics integration results for activated pathways that were shared across omics (panel A), and independent pathways being active in different omics (panel B). Absolute Z scores of the simulated summary statistics representing P values without taking into account direction of effect were evaluated in panel A a-d and panel B a-d, and signed P values/test statistics including direction of effects in panel A e-h and panel B e-h. The average accuracy, sensitivity and specificity across 100 simulation replicates is visualized across the different combinations of predefined false positive and false negative rates and the last panel of each row represents the percentage of cases, where the exact match of simulated gene sets were recovered.
MONA, multi-level ontology analysis;

B List of Tables

C List of Figures

Bibliography

E. Afgan, D. Baker, B. Batut, M. Van Den Beek, D. Bouvier, M. Ech, J. Chilton, D. Clements, N. Coraor, B. A. Grüning, A. Guerler, J. Hillman-Jackson, S. Hiltemann, V. Jalili, H. Rasche, N. Soranzo, J. Goecks, J. Taylor, A. Nekrutenko, and D. Blankenberg. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research*, 46(W1):W537–W544, 2018. ISSN 13624962. doi: 10.1093/nar/gky379.

V. Agarwal, G. W. Bell, J. W. Nam, and D. P. Bartel. Predicting effective microRNA target sites in mammalian mRNAs. *eLife*, 4:e05005, 2015. ISSN 2050084X. doi: 10.7554/eLife.05005. URL `https://doi.org/10.7554/eLife.05005`.

H. Akazawa and I. Komuro. Cardiac transcription factor Csx/Nkx2-5: Its role in cardiac development and diseases. *Pharmacology and Therapeutics*, 107(2):252–268, 2005. ISSN 01637258. doi: 10.1016/j.pharmthera.2005.03.005.

J. S. Amberger, C. A. Bocchini, F. Schiettecatte, A. F. Scott, and A. Hamosh. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an Online catalog of human genes and genetic disorders. *Nucleic Acids Research*, 43(D1):D789–D798, 2015. ISSN 13624962. doi: 10.1093/nar/gku1205. URL `https://doi.org/10.1093/nar/gku1205`.

C. A. Anderson, F. H. Pettersson, G. M. Clarke, L. R. Cardon, A. P. Morris, and K. T. Zondervan. Data quality control in genetic case-control association studies. *Nature Protocols*, 5(9):1564–1573, 2010. ISSN 17502799. doi: 10.1038/nprot.2010.116. URL `https://doi.org/10.1038/nprot.2010.116`.

D. J. Anderson, D. I. Kaplan, K. M. Bell, K. Koutsis, J. M. Haynes, R. J. Mills, D. G. Phelan, E. L. Qian, A. R. Leitoguinho, D. Arasaratnam, T. Labonne, E. S. Ng, R. P. Davis, S. Casini, R. Passier, J. E. Hudson, E. R. Porrello, M. W. Costa, A. Rafii, C. L. Curl, L. M. Delbridge, R. P. Harvey, A. Oshlack, M. M. Cheung, C. L. Mummery, S. Petrou, A. G. Elefanty, E. G. Stanley, and D. A. Elliott. NKX2-5 regulates human cardiomyogenesis via a HEY2 dependent transcriptional network. *Nature Communications*, 9(1):1–13, 2018. ISSN 20411723. doi: 10.1038/s41467-018-03714-x.

C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan. An Introduction to MCMC for Machine Learning. *Machine Learning*, 50:5–43, 2003. doi: 10.1007/978-0-387-21811-3_1.

M. Arnold, J. Raffler, A. Pfeufer, K. Suhre, and G. Kastenmüller. SNiPA: An interactive, genetic variant-centered annotation browser. *Bioinformatics*, 31(8):1334–1336, 2015. ISSN 14602059. doi: 10.1093/bioinformatics/btu779.

M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, and G. Rubin, Gerald M. Sherlock. Gene Ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000. doi: 10.1038/75556.Gene.

I. Assum and M. Heinig. heiniglab/symatrial: nat_comm_1.1, jul 2021. URL `https://doi.org/10.5281/zenodo.5094276`.

I. Assum, J. Krause, T. Zeller, R. B. Schnabel, and M. Heinig. Extended data: Tissue-specific multi-omics analysis of atrial fibrillation, jul 2021. URL `https://doi.org/10.5281/zenodo.5080229`.

I. Assum, J. Krause, M. O. Scheinhardt, C. Müller, E. Hammer, C. S. Börschel, U. Völker, L. Conradi, B. Geelhoed, T. Zeller, R. B. Schnabel, and M. Heinig. Tissue-specific multi-omics analysis of atrial fibrillation. *Nature Communications*, 13:441, jan 2022a. doi: 10.1038/s41467-022-27953-1. URL `https://doi.org/10.1038/s41467-022-27953-1`.

I. Assum, J. Krause, M. O. Scheinhardt, C. Müller, E. Hammer, C. S. Börschel, U. Völker, L. Conradi, B. Geelhoed, T. Zeller, R. B. Schnabel, and M. Heinig. Supplementary material: Tissue-specific multi-omics analysis of atrial fibrillation, jan 2022b. URL `https://doi.org/10.1038/s41467-022-27953-1`.

A. Auton, G. R. Abecasis, D. M. Altshuler, R. M. Durbin, D. R. Bentley, A. Chakravarti, A. G. Clark, P. Donnelly, E. E. Eichler, P. Flicek, S. B. Gabriel, R. A. Gibbs, E. D. Green, M. E. Hurles, B. M. Knoppers, J. O. Korbel, E. S. Lander, C. Lee, H. Lehrach, E. R. Mardis, G. T. Marth, G. A. McVean, D. A. Nickerson, J. P. Schmidt, S. T. Sherry, J. Wang, R. K. Wilson, E. Boerwinkle, H. Doddapaneni, Y. Han, V. Korchina, C. Kovar, S. Lee, D. Muzny, J. G. Reid, Y. Zhu, Y. Chang, Q. Feng, X. Fang, X. Guo, M. Jian, H. Jiang, X. Jin, T. Lan, G. Li, J. Li, Y. Li, S. Liu, X. Liu, Y. Lu, X. Ma, M. Tang, B. Wang, G. Wang, H. Wu, R. Wu, X. Xu, Y. Yin, D. Zhang, W. Zhang, J. Zhao, M. Zhao, X. Zheng, N. Gupta, N. Gharani, L. H. Toji, N. P. Gerry, A. M. Resch, J. Barker, L. Clarke, L. Gil, S. E. Hunt, G. Kelman, E. Kulesha, R. Leinonen, W. M. McLaren, R. Radhakrishnan, A. Roa, D. Smirnov, R. E. Smith, I. Streeter, A. Thormann, I. Toneva, B. Vaughan, X. Zheng-Bradley, R. Grocock, S. Humphray, T. James, Z. Kingsbury, R. Sudbrak, M. W. Albrecht, V. S. Amstislavskiy, T. A. Borodina, M. Lienhard, F. Mertes, M. Sultan, B. Timmermann, M. L. Yaspo, L. Fulton, V. Ananiev, Z. Belaia, D. Beloslyudtsev, N. Bouk, C. Chen, D. Church, R. Cohen, C. Cook, J. Garner, T. Hefferon, M. Kimelman, C. Liu, J. Lopez, P. Meric, C. O'Sullivan, Y. Ostapchuk, L. Phan, S. Ponomarov, V. Schneider, E. Shekhtman, K. Sirotkin, D. Slotta, H. Zhang, S. Balasubramaniam, J. Burton, P. Danecek, T. M. Keane, A. Kolb-Kokocinski, S. McCarthy, J. Stalker, M. Quail, C. J. Davies, J. Gollub, T. Webster, B. Wong, Y. Zhan, C. L. Campbell, Y. Kong, A. Marcketta, F. Yu, L. Antunes, M. Bainbridge, A. Sabo, Z. Huang, L. J. Coin, L. Fang, Q. Li, Z. Li, H. Lin, B. Liu, R. Luo, H. Shao, Y. Xie, C. Ye, C. Yu, F. Zhang, H. Zheng, H. Zhu, C. Alkan, E. Dal, F. Kahveci, E. P. Garrison, D. Kural, W. P. Lee, W. F. Leong, M. Stromberg, A. N. Ward, J. Wu, M. Zhang, M. J.

Daly, M. A. DePristo, R. E. Handsaker, E. Banks, G. Bhatia, G. Del Angel, G. Genovese, H. Li, S. Kashin, S. A. McCarroll, J. C. Nemesh, R. E. Poplin, S. C. Yoon, J. Lihm, V. Makarov, S. Gottipati, A. Keinan, J. L. Rodriguez-Flores, T. Rausch, M. H. Fritz, A. M. Stütz, K. Beal, A. Datta, J. Herrero, G. R. Ritchie, D. Zerbino, P. C. Sabeti, I. Shlyakhter, S. F. Schaffner, J. Vitti, D. N. Cooper, E. V. Ball, P. D. Stenson, B. Barnes, M. Bauer, R. K. Cheetham, A. Cox, M. Eberle, S. Kahn, L. Murray, J. Peden, R. Shaw, E. E. Kenny, M. A. Batzer, M. K. Konkel, J. A. Walker, D. G. MacArthur, M. Lek, R. Herwig, L. Ding, D. C. Koboldt, D. Larson, K. Ye, S. Gravel, A. Swaroop, E. Chew, T. Lappalainen, Y. Erlich, M. Gymrek, T. F. Willems, J. T. Simpson, M. D. Shriver, J. A. Rosenfeld, C. D. Bustamante, S. B. Montgomery, F. M. De La Vega, J. K. Byrnes, A. W. Carroll, M. K. DeGorter, P. Lacroute, B. K. Maples, A. R. Martin, A. Moreno-Estrada, S. S. Shringarpure, F. Zakharia, E. Halperin, Y. Baran, E. Cerveira, J. Hwang, A. Malhotra, D. Plewczynski, K. Radew, M. Romanovitch, C. Zhang, F. C. Hyland, D. W. Craig, A. Christoforides, N. Homer, T. Izatt, A. A. Kurdoglu, S. A. Sinari, K. Squire, C. Xiao, J. Sebat, D. Antaki, M. Gujral, A. Noor, K. Ye, E. G. Burchard, R. D. Hernandez, C. R. Gignoux, D. Haussler, S. J. Katzman, W. J. Kent, B. Howie, A. Ruiz-Linares, E. T. Dermitzakis, S. E. Devine, H. M. Kang, J. M. Kidd, T. Blackwell, S. Caron, W. Chen, S. Emery, L. Fritsche, C. Fuchsberger, G. Jun, B. Li, R. Lyons, C. Scheller, C. Sidore, S. Song, E. Sliwerska, D. Taliun, A. Tan, R. Welch, M. K. Wing, X. Zhan, P. Awadalla, A. Hodgkinson, Y. Li, X. Shi, A. Quitadamo, G. Lunter, J. L. Marchini, S. Myers, C. Churchhouse, O. Delaneau, A. Gupta-Hinch, W. Kretzschmar, Z. Iqbal, I. Mathieson, A. Menelaou, A. Rimmer, D. K. Xifara, T. K. Oleksyk, Y. Fu, X. Liu, M. Xiong, L. Jorde, D. Witherspoon, J. Xing, B. L. Browning, S. R. Browning, F. Hormozdiari, P. H. Sudmant, E. Khurana, C. Tyler-Smith, C. A. Albers, Q. Ayub, Y. Chen, V. Colonna, L. Jostins, K. Walter, Y. Xue, M. B. Gerstein, A. Abyzov, S. Balasubramanian, J. Chen, D. Clarke, Y. Fu, A. O. Harmanci, M. Jin, D. Lee, J. Liu, X. J. Mu, J. Zhang, Y. Zhang, C. Hartl, K. Shakir, J. Degenhardt, S. Meiers, B. Raeder, F. P. Casale, O. Stegle, E. W. Lameijer, I. Hall, V. Bafna, J. Michaelson, E. J. Gardner, R. E. Mills, G. Dayama, K. Chen, X. Fan, Z. Chong, T. Chen, M. J. Chaisson, J. Huddleston, M. Malig, B. J. Nelson, N. F. Parrish, B. Blackburne, S. J. Lindsay, Z. Ning, Y. Zhang, H. Lam, C. Sisu, D. Challis, U. S. Evani, J. Lu, U. Nagaswamy, J. Yu, W. Li, L. Habegger, H. Yu, F. Cunningham, I. Dunham, K. Lage, J. B. Jespersen, H. Horn, D. Kim, R. Desalle, A. Narechania, M. A. Sayres, F. L. Mendez, G. D. Poznik, P. A. Underhill, D. Mittelman, R. Banerjee, M. Cerezo, T. W. Fitzgerald, S. Louzada, A. Massaia, F. Yang, D. Kalra, W. Hale, X. Dan, K. C. Barnes, C. Beiswanger, H. Cai, H. Cao, B. Henn, D. Jones, J. S. Kaye, A. Kent, A. Kerasidou, R. Mathias, P. N. Ossorio, M. Parker, C. N. Rotimi, C. D. Royal, K. Sandoval, Y. Su, Z. Tian, S. Tishkoff, M. Via, Y. Wang, H. Yang, L. Yang, J. Zhu, W. Bodmer, G. Bedoya, Z. Cai, Y. Gao, J. Chu, L. Peltonen, A. Garcia-Montero, A. Orfao, J. Dutil, J. C. Martinez-Cruzado, R. A. Mathias, A. Hennis, H. Watson, C. McKenzie, F. Qadri, R. LaRocque, X. Deng, D. Asogun, O. Folarin, C. Happi, O. Omoniwa, M. Stremlau, R. Tariyal, M. Jallow, F. S. Joof, T. Corrah, K. Rockett, D. Kwiatkowski, J. Kooner, T. T. Hien, S. J. Dunstan, N. ThuyHang, R. Fonnie, R. Garry, L. Kanneh, L. Moses, J. Schieffelin, D. S. Grant, C. Gallo, G. Poletti, D. Saleheen, A. Rasheed, L. D. Brooks, A. L. Felsenfeld, J. E. McEwen, Y. Vaydylevich, A. Duncanson, M. Dunn, and J. A. Schloss. A global

reference for human genetic variation. *Nature*, 526(7571):68–74, 2015. ISSN 14764687. doi: 10.1038/nature15393. URL https://doi.org/10.1038/nature15393.

S. N. Baharum and K. A. Azizan. *Metabolomics in Systems Biology*, pages 51–68. Springer International Publishing, Cham, 2018. ISBN 978-3-319-98758-3. doi: 10.1007/978-3-319-98758-3_4. URL https://doi.org/10.1007/978-3-319-98758-3_4.

A. Battle, Z. Khan, S. H. Wang, A. Mitrano, M. J. Ford, J. K. Pritchard, and Y. Gilad. Impact of regulatory variation from RNA to protein. *Science*, 347(6222):664 LP – 667, feb 2015. doi: 10.1126/science.1260793. URL http://science.sciencemag.org/content/347/6222/664.abstract.

S. Bauer, J. Gagneur, and P. N. Robinson. Going Bayesian: Model-based gene set analysis of genome-scale data. *Nucleic Acids Research*, 38(11):3523–3532, 2010. ISSN 03051048. doi: 10.1093/nar/gkq045.

P. Benaglio, A. D'Antonio-Chronowska, W. Ma, F. Yang, W. W. Young Greenwald, M. K. Donovan, C. DeBoever, H. Li, F. Drees, S. Singhal, H. Matsui, J. van Setten, N. Sotoodehnia, K. J. Gaulton, E. N. Smith, M. D'Antonio, M. G. Rosenfeld, and K. A. Frazer. Allele-specific NKX2-5 binding underlies multiple genetic associations with human electrocardiographic traits. *Nature Genetics*, 51(10):1506–1517, 2019. ISSN 15461718. doi: 10.1038/s41588-019-0499-3.

E. J. Benjamin, P. Muntner, A. Alonso, M. S. Bittencourt, C. W. Callaway, A. P. Carson, A. M. Chamberlain, A. R. Chang, S. Cheng, S. R. Das, F. N. Delling, L. Djousse, M. S. Elkind, J. F. Ferguson, M. Fornage, L. C. Jordan, S. S. Khan, B. M. Kissela, K. L. Knutson, T. W. Kwan, D. T. Lackland, T. T. Lewis, J. H. Lichtman, C. T. Longenecker, M. S. Loop, P. L. Lutsey, S. S. Martin, K. Matsushita, A. E. Moran, M. E. Mussolino, M. O'Flaherty, A. Pandey, A. M. Perak, W. D. Rosamond, G. A. Roth, U. K. Sampson, G. M. Satou, E. B. Schroeder, S. H. Shah, N. L. Spartano, A. Stokes, D. L. Tirschwell, C. W. Tsao, M. P. Turakhia, L. B. VanWagner, J. T. Wilkins, S. S. Wong, and S. S. Virani. Heart Disease and Stroke Statistics-2019 Update: A Report From the American Heart Association. *Circulation*, 139(10):e56–e528, 2019. ISSN 15244539. doi: 10.1161/CIR.0000000000000659.

M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, K. Thiel, and B. Wiswedel. KNIME-the Konstanz information miner: version 2.0 and beyond. *AcM SIGKDD explorations Newsletter*, 11(1):26–31, 2009. ISSN 19310145. URL http://portal.acm.org/citation.cfm?doid=1656274.1656280.

E. A. Boyle, Y. I. Li, and J. K. Pritchard. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell*, 169(7):1177–1186, jun 2017. ISSN 10974172. doi: 10.1016/j.cell.2017.05.038.

R. B. Brem, G. Yvert, R. Clinton, and L. Kruglyak. Genetic dissection of transcriptional regulation in budding yeast. *Science*, 296(5568):752–755, apr 2002. ISSN 00368075. doi: 10.1126/science.1069516.

B. J. Brundel, R. H. Henning, H. H. Kampinga, I. C. Van Gelder, and H. J. Crijns. Molecular mechanisms of remodeling in human atrial fibrillation. *Cardiovascular Research*, 54(2):315–324, 2002. ISSN 00086363. doi: 10.1016/S0008-6363(02)00222-5.

C. Buccitelli and M. Selbach. mRNAs, proteins and the emerging principles of gene expression control. *Nature Reviews Genetics*, 21(10):630–644, 2020. ISSN 14710064. doi: 10.1038/s41576-020-0258-4. URL `http://dx.doi.org/10.1038/s41576-020-0258-4`.

A. Buniello, J. A. Macarthur, M. Cerezo, L. W. Harris, J. Hayhurst, C. Malangone, A. McMahon, J. Morales, E. Mountjoy, E. Sollis, D. Suveges, O. Vrousgou, P. L. Whetzel, R. Amode, J. A. Guillen, H. S. Riat, S. J. Trevanion, P. Hall, H. Junkins, P. Flicek, T. Burdett, L. A. Hindorff, F. Cunningham, and H. Parkinson. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, 47(D1):D1005–D1012, 2019. ISSN 13624962. doi: 10.1093/nar/gky1120. URL `https://doi.org/10.1093/nar/gky1120`.

S. Carbon, A. Ireland, C. J. Mungall, S. Shu, B. Marshall, S. Lewis, J. Lomax, C. Mungall, B. Hitz, R. Balakrishnan, M. Dolan, V. Wood, E. Hong, and P. Gaudet. AmiGO: Online access to ontology and annotation data. *Bioinformatics*, 25(2):288–289, 2009. ISSN 13674803. doi: 10.1093/bioinformatics/btn615.

S. Carbon, E. Douglass, N. Dunn, B. Good, N. L. Harris, S. E. Lewis, C. J. Mungall, S. Basu, R. L. Chisholm, R. J. Dodson, E. Hartline, P. Fey, P. D. Thomas, L. P. Albou, D. Ebert, M. J. Kesling, H. Mi, A. Muruganujan, X. Huang, S. Poudel, T. Mushayahama, J. C. Hu, S. A. LaBonte, D. A. Siegele, G. Antonazzo, H. Attrill, N. H. Brown, S. Fexova, P. Garapati, T. E. Jones, S. J. Marygold, G. H. Millburn, A. J. Rey, V. Trovisco, G. Dos Santos, D. B. Emmert, K. Falls, P. Zhou, J. L. Goodman, V. B. Strelets, J. Thurmond, M. Courtot, D. S. Osumi, H. Parkinson, P. Roncaglia, M. L. Acencio, M. Kuiper, A. Lreid, C. Logie, R. C. Lovering, R. P. Huntley, P. Denny, N. H. Campbell, B. Kramarz, V. Acquaah, S. H. Ahmad, H. Chen, J. H. Rawson, M. C. Chibucos, M. Giglio, S. Nadendla, R. Tauber, M. J. Duesbury, N. T. Del, B. H. Meldal, L. Perfetto, P. Porras, S. Orchard, A. Shrivastava, Z. Xie, H. Y. Chang, R. D. Finn, A. L. Mitchell, N. D. Rawlings, L. Richardson, A. Sangrador-Vegas, J. A. Blake, K. R. Christie, M. E. Dolan, H. J. Drabkin, D. P. Hill, L. Ni, D. Sitnikov, M. A. Harris, S. G. Oliver, K. Rutherford, V. Wood, J. Hayles, J. Bahler, A. Lock, E. R. Bolton, J. De Pons, M. Dwinell, G. T. Hayman, S. J. Laulederkind, M. Shimoyama, M. Tutaj, S. J. Wang, P. D'Eustachio, L. Matthews, J. P. Balhoff, S. A. Aleksander, G. Binkley, B. L. Dunn, J. M. Cherry, S. R. Engel, F. Gondwe, K. Karra, K. A. MacPherson, S. R. Miyasato, R. S. Nash, P. C. Ng, T. K. Sheppard, A. Shrivatsav Vp, M. Simison, M. S. Skrzypek, S. Weng, E. D. Wong, M. Feuermann, P. Gaudet, E. Bakker, T. Z. Berardini, L. Reiser, S. Subramaniam, E. Huala, C. Arighi, A. Auchincloss, K. Axelsen, G. P. Argoud, A. Bateman, B. Bely, M. C. Blatter, E. Boutet, L. Breuza, A. Bridge, R. Britto, H. Bye-A-Jee, C. Casals-Casas, E. Coudert, A. Estreicher, L. Famiglietti, P. Garmiri, G. Georghiou, A. Gos, N. Gruaz-Gumowski, E. Hatton-Ellis, U. Hinz, C. Hulo, A. Ignatchenko, F. Jungo, G. Keller, K. Laiho, P. Lemercier, D. Lieberherr, Y. Lussi, A. Mac-Dougall, M. Magrane, M. J. Martin, P. Masson, D. A. Natale, N. N. Hyka, I. Pedruzzi, K. Pichler, S. Poux,

C. Rivoire, M. Rodriguez-Lopez, T. Sawford, E. Speretta, A. Shypitsyna, A. Stutz, S. Sundaram, M. Tognolli, N. Tyagi, K. Warner, R. Zaru, C. Wu, J. Chan, J. Cho, S. Gao, C. Grove, M. C. Harrison, K. Howe, R. Lee, J. Mendel, H. M. Muller, D. Raciti, K. Van Auken, M. Berriman, L. Stein, P. W. Sternberg, D. Howe, S. Toro, and M. Westerfield. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, 47(D1):D330–D338, 2019. ISSN 13624962. doi: 10.1093/nar/gky1055.

B. S. Carvalho and R. A. Irizarry. A framework for oligonucleotide microarray preprocessing. *Bioinformatics*, 26(19):2363–2367, 2010. ISSN 13674803. doi: 10.1093/bioinformatics/btq431.

J. Chèneby, M. Gheorghe, M. Artufel, A. Mathelier, and B. Ballester. ReMap 2018: An updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. *Nucleic Acids Research*, 46(D1):D267–D275, 2018. ISSN 13624962. doi: 10.1093/nar/gkx1092. URL `https://doi.org/10.1093/nar/gkx1092`.

S. H. Choi, S. J. Jurgens, L.-c. Weng, J. P. Pirruccello, C. Roselli, M. Chaffin, C. J. Lee, A. W. Hall, A. V. Khera, K. L. Lunetta, S. A. Lubitz, and P. T. Ellinor. Monogenic and Polygenic Contributions to Atrial. *Circulation Research*, 126:200–209, 2020. doi: 10.1161/CIRCRESAHA.119.315686.

S. S. Chugh, R. Havmoeller, K. Narayanan, D. Singh, M. Rienstra, E. J. Benjamin, R. F. Gillum, Y. H. Kim, J. H. McAnulty, Z. J. Zheng, M. H. Forouzanfar, M. Naghavi, G. A. Mensah, M. Ezzati, and C. J. Murray. Worldwide epidemiology of atrial fibrillation: A global burden of disease 2010 study. *Circulation*, 129(8):837–847, 2014. ISSN 00097322. doi: 10.1161/CIRCULATIONAHA.113.005119.

M. Civelek and A. J. Lusis. Systems genetics approaches to understand complex traits, 2014. ISSN 14710056. URL `www.nature.com/reviews/genetics`.

C. J. Coats, W. E. Heywood, A. Virasami, N. Ashrafi, P. Syrris, C. Dos Remedios, T. A. Treibel, J. C. Moon, L. R. Lopes, C. G. McGregor, M. Ashworth, N. J. Sebire, W. J. McKenna, K. Mills, and P. M. Elliott. Proteomic Analysis of the Myocardium in Hypertrophic Obstructive Cardiomyopathy. *Circulation. Genomic and precision medicine*, 11(12):e001974, 2018. ISSN 25748300. doi: 10.1161/CIRCGEN.117.001974.

D. Corradi, S. Callegari, R. Maestri, S. Benussi, and O. Alfieri. Structural remodeling in atrial fibrillation. *Nature Clinical Practice Cardiovascular Medicine*, 5(12):782–796, 2008. ISSN 17434297. doi: 10.1038/ncpcardio1370.

R. S. Dalal, A. A. Sabe, N. Y. Elmadhun, B. Ramlawi, and F. W. Sellke. Atrial fibrillation, neurocognitive decline and gene expression after cardiopulmonary bypass. *Brazilian Journal of Cardiovascular Surgery*, 30(5):520–532, 2015. ISSN 16789741. doi: 10.5935/1678-9741.20150070. URL `https://doi.org/10.5935/1678-9741.20150070`.

J. O. Davies, A. M. Oudelaar, D. R. Higgs, and J. R. Hughes. How best to identify chromosomal interactions: A comparison of approaches. *Nature Methods*, 14(2):125–134, 2017. ISSN 15487105. doi: 10.1038/nmeth.4146.

C. A. Davis, B. C. Hitz, C. A. Sloan, E. T. Chan, J. M. Davidson, I. Gabdank, J. A. Hilton, K. Jain, U. K. Baymuradov, A. K. Narayanan, K. C. Onate, K. Graham, S. R. Miyasato, T. R. Dreszer, J. S. Strattan, O. Jolanki, F. Y. Tanaka, and J. M. Cherry. The Encyclopedia of DNA elements (ENCODE): Data portal update. *Nucleic Acids Research*, 46(D1):D794–D801, 2018. ISSN 13624962. doi: 10.1093/nar/gkx1081.

P. DI Tommaso, M. Chatzou, E. W. Floden, P. P. Barja, E. Palumbo, and C. Notredame. Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35(4): 316–319, 2017. ISSN 15461696. doi: 10.1038/nbt.3820.

P. Diaconis. The markov chain monte carlo revolution. *Bulletin of the American Mathematical Society*, 46(2):179–205, 2009. ISSN 02730979. doi: 10.1090/S0273-0979-08-01238-X.

P. Diaconis and L. Saloff-Coste. What Do We Know about the Metropolis Algorithm? *Journal of Computer and System Sciences*, 57:20–36, 1998. ISSN 00220000. doi: 10.1006/jcss.1998.1576.

D. Dobrev and S. Nattel. New insights into the molecular basis of atrial fibrillation: Mechanistic and therapeutic implications. *Cardiovascular Research*, 89(4):689–691, 2011. ISSN 17553245. doi: 10.1093/cvr/cvr021.

D. Dobrev, M. Aguilar, J. Heijman, J. B. Guichard, and S. Nattel. Postoperative atrial fibrillation: mechanisms, manifestations and management. *Nature Reviews Cardiology*, 16(7):417–436, 2019. ISSN 17595010. doi: 10.1038/s41569-019-0166-5. URL `http://dx.doi.org/10.1038/s41569-019-0166-5`.

S. Doll, M. Dreßen, P. E. Geyer, D. N. Itzhak, C. Braun, S. A. Doppler, F. Meier, M. A. Deutsch, H. Lahm, R. Lange, M. Krane, and M. Mann. Region and cell-type resolved quantitative proteomic map of the human heart. *Nature Communications*, 8(1):1–13, 2017. ISSN 20411723. doi: 10.1038/s41467-017-01747-2.

A. Döring, D. Weese, T. Rausch, and K. Reinert. SeqAn an efficient, generic C++ library for sequence analysis. *BMC Bioinformatics*, 9:11, 2008. ISSN 14712105. doi: 10.1186/1471-2105-9-11.

T. Dorn, A. Goedel, J. T. Lam, J. Haas, Q. Tian, F. Herrmann, K. Bundschu, G. Dobreva, M. Schiemann, R. Dirschinger, Y. Guo, S. J. Kühl, D. Sinnecker, P. Lipp, K. L. Laugwitz, M. Kühl, and A. Moretti. Direct Nkx2-5 transcriptional repression of isl1 controls cardiomyocyte subtype identity. *Stem Cells*, 33(4):1113–1129, 2015. ISSN 15494918. doi: 10.1002/stem.1923.

I. Dunham, A. Kundaje, S. F. Aldred, P. J. Collins, C. A. Davis, F. Doyle, C. B. Epstein, S. Frietze, J. Harrow, R. Kaul, J. Khatun, B. R. Lajoie, S. G. Landt, B. K. Lee, F. Pauli, K. R. Rosenbloom, P. Sabo, A. Safi, A. Sanyal, N. Shoresh, J. M. Simon, L. Song, N. D. Trinklein, R. C. Altshuler, E. Birney, J. B. Brown, C. Cheng, S. Djebali, X. Dong, J. Ernst, T. S. Furey, M. Gerstein, B. Giardine, M. Greven, R. C. Hardison, R. S. Harris, J. Herrero, M. M. Hoffman, S. Iyer, M. Kellis, P. Kheradpour, T. Lassmann, Q. Li, X. Lin, G. K. Marinov, A. Merkel, A. Mortazavi, S. C. Parker, T. E. Reddy,

J. Rozowsky, F. Schlesinger, R. E. Thurman, J. Wang, L. D. Ward, T. W. Whitfield, S. P. Wilder, W. Wu, H. S. Xi, K. Y. Yip, J. Zhuang, B. E. Bernstein, E. D. Green, C. Gunter, M. Snyder, M. J. Pazin, R. F. Lowdon, L. A. Dillon, L. B. Adams, C. J. Kelly, J. Zhang, J. R. Wexler, P. J. Good, E. A. Feingold, G. E. Crawford, J. Dekker, L. Elnitski, P. J. Farnham, M. C. Giddings, T. R. Gingeras, R. Guigó, T. J. Hubbard, W. J. Kent, J. D. Lieb, E. H. Margulies, R. M. Myers, J. A. Stamatoyannopoulos, S. A. Tenenbaum, Z. Weng, K. P. White, B. Wold, Y. Yu, J. Wrobel, B. A. Risk, H. P. Gunawardena, H. C. Kuiper, C. W. Maier, L. Xie, X. Chen, T. S. Mikkelsen, S. Gillespie, A. Goren, O. Ram, X. Zhang, L. Wang, R. Issner, M. J. Coyne, T. Durham, M. Ku, T. Truong, M. L. Eaton, A. Dobin, A. Tanzer, J. Lagarde, W. Lin, C. Xue, B. A. Williams, C. Zaleski, M. Röder, F. Kokocinski, R. F. Abdelhamid, T. Alioto, I. Antoshechkin, M. T. Baer, P. Batut, I. Bell, K. Bell, S. Chakrabortty, J. Chrast, J. Curado, T. Derrien, J. Drenkow, E. Dumais, J. Dumais, R. Duttagupta, M. Fastuca, K. Fejes-Toth, P. Ferreira, S. Foissac, M. J. Fullwood, H. Gao, D. Gonzalez, A. Gordon, C. Howald, S. Jha, R. Johnson, P. Kapranov, B. King, C. Kingswood, G. Li, O. J. Luo, E. Park, J. B. Preall, K. Presaud, P. Ribeca, D. Robyr, X. Ruan, M. Sammeth, K. S. Sandhu, L. Schaeffer, L. H. See, A. Shahab, J. Skancke, A. M. Suzuki, H. Takahashi, H. Tilgner, D. Trout, N. Walters, H. Wang, Y. Hayashizaki, A. Reymond, S. E. Antonarakis, G. J. Hannon, Y. Ruan, P. Carninci, C. A. Sloan, K. Learned, V. S. Malladi, M. C. Wong, G. P. Barber, M. S. Cline, T. R. Dreszer, S. G. Heitner, D. Karolchik, V. M. Kirkup, L. R. Meyer, J. C. Long, M. Maddren, B. J. Raney, L. L. Grasfeder, P. G. Giresi, A. Battenhouse, N. C. Sheffield, K. A. Showers, D. London, A. A. Bhinge, C. Shestak, M. R. Schaner, S. K. Kim, Z. Z. Zhang, P. A. Mieczkowski, J. O. Mieczkowska, Z. Liu, R. M. McDaniell, Y. Ni, N. U. Rashid, M. J. Kim, S. Adar, Z. Zhang, T. Wang, D. Winter, D. Keefe, V. R. Iyer, M. Zheng, P. Wang, J. Gertz, J. Vielmetter, E. C. Partridge, K. E. Varley, C. Gasper, A. Bansal, S. Pepke, P. Jain, H. Amrhein, K. M. Bowling, M. Anaya, M. K. Cross, M. A. Muratet, K. M. Newberry, K. McCue, A. S. Nesmith, K. I. Fisher-Aylor, B. Pusey, G. DeSalvo, S. L. Parker, S. Balasubramanian, N. S. Davis, S. K. Meadows, T. Eggleston, J. S. Newberry, S. E. Levy, D. M. Absher, W. H. Wong, M. J. Blow, A. Visel, L. A. Pennachio, H. M. Petrykowska, A. Abyzov, B. Aken, D. Barrell, G. Barson, A. Berry, A. Bignell, V. Boychenko, G. Bussotti, C. Davidson, G. Despacio-Reyes, M. Diekhans, I. Ezkurdia, A. Frankish, J. Gilbert, J. M. Gonzalez, E. Griffiths, R. Harte, D. A. Hendrix, T. Hunt, I. Jungreis, M. Kay, E. Khurana, J. Leng, M. F. Lin, J. Loveland, Z. Lu, D. Manthravadi, M. Mariotti, J. Mudge, G. Mukherjee, C. Notredame, B. Pei, J. M. Rodriguez, G. Saunders, A. Sboner, S. Searle, C. Sisu, C. Snow, C. Steward, E. Tapanari, M. L. Tress, M. J. Van Baren, S. Washietl, L. Wilming, A. Zadissa, Z. Zhang, M. Brent, D. Haussler, A. Valencia, N. Addleman, R. P. Alexander, R. K. Auerbach, S. Balasubramanian, K. Bettinger, N. Bhardwaj, A. P. Boyle, A. R. Cao, P. Cayting, A. Charos, Y. Cheng, C. Eastman, G. Euskirchen, J. D. Fleming, F. Grubert, L. Habegger, M. Hariharan, A. Harmanci, S. Iyengar, V. X. Jin, K. J. Karczewski, M. Kasowski, P. Lacroute, H. Lam, N. Lamarre-Vincent, J. Lian, M. Lindahl-Allen, R. Min, B. Miotto, H. Monahan, Z. Moqtaderi, X. J. Mu, H. O'Geen, Z. Ouyang, D. Patacsil, D. Raha, L. Ramirez, B. Reed, M. Shi, T. Slifer, H. Witt, L. Wu, X. Xu, K. K. Yan, X. Yang, K. Struhl, S. M. Weissman, L. O. Penalva, S. Karmakar, R. R. Bhanvadia, A. Choudhury, M. Domanus, L. Ma, J. Moran, A. Victorsen, T. Auer,

L. Centanin, M. Eichenlaub, F. Gruhl, S. Heermann, B. Hoeckendorf, D. Inoue, T. Kellner, S. Kirchmaier, C. Mueller, R. Reinhardt, L. Schertel, S. Schneider, R. Sinn, B. Wittbrodt, J. Wittbrodt, G. Jain, G. Balasundaram, D. L. Bates, R. Byron, T. K. Canfield, M. J. Diegel, D. Dunn, A. K. Ebersol, T. Frum, K. Garg, E. Gist, R. S. Hansen, L. Boatman, E. Haugen, R. Humbert, A. K. Johnson, E. M. Johnson, T. V. Kutyavin, K. Lee, D. Lotakis, M. T. Maurano, S. J. Neph, F. V. Neri, E. D. Nguyen, H. Qu, A. P. Reynolds, V. Roach, E. Rynes, M. E. Sanchez, R. S. Sandstrom, A. O. Shafer, A. B. Stergachis, S. Thomas, B. Vernot, J. Vierstra, S. Vong, H. Wang, M. A. Weaver, Y. Yan, M. Zhang, J. M. Akey, M. Bender, M. O. Dorschner, M. Groudine, M. J. MacCoss, P. Navas, G. Stamatoyannopoulos, K. Beal, A. Brazma, P. Flicek, N. Johnson, M. Lukk, N. M. Luscombe, D. Sobral, J. M. Vaquerizas, S. Batzoglou, A. Sidow, N. Hussami, S. Kyriazopoulou-Panagiotopoulou, M. W. Libbrecht, M. A. Schaub, W. Miller, P. J. Bickel, B. Banfai, N. P. Boley, H. Huang, J. J. Li, W. S. Noble, J. A. Bilmes, O. J. Buske, A. D. Sahu, P. V. Kharchenko, P. J. Park, D. Baker, J. Taylor, and L. Lochovsky. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414): 57–74, 2012. ISSN 14764687. doi: 10.1038/nature11247.

F. Edfors, F. Danielsson, B. M. Hallström, L. Käll, E. Lundberg, F. Pontén, B. Forsström, and M. Uhlén. Gene-specific correlation of RNA and protein levels in human cells and tissues. *Molecular Systems Biology*, 12(10):883, 2016. ISSN 1744-4292. doi: 10.15252/msb.20167144.

A. W. El-Hattab and F. Scaglia. Mitochondrial Cardiomyopathies. *Frontiers in Cardiovascular Medicine*, 3:25, 2016. ISSN 2297055X. doi: 10.3389/fcvm.2016.00025. URL `https://doi.org/10.3389/fcvm.2016.00025`.

B. Eraslan, D. Wang, M. Gusic, H. Prokisch, B. M. Hallström, M. Uhlén, A. Asplund, F. Pontén, T. Wieland, T. Hopf, H. Hahne, B. Kuster, and J. Gagneur. Quantification and discovery of sequence determinants of protein-per-mRNA amount in 29 human tissues. *Molecular Systems Biology*, 15(2):1–25, 2019. ISSN 1744-4292. doi: 10.15252/msb.20188513.

L. Fahrmeir, T. Kneib, and S. Lang. *Regression*. Springer, Berlin, Heidelberg, 1 edition, 2013. ISBN 9783642343339. doi: 10.1007/978-3-642-34333-9.

A. Farcomeni. A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Statistical Methods in Medical Research*, 17(4):347–388, 2008. ISSN 09622802. doi: 10.1177/0962280206079046.

E. Ferkingstad, P. Sulem, B. A. Atlason, G. Sveinbjornsson, M. I. Magnusson, E. L. Styrmisdottir, K. Gunnarsdottir, A. Helgason, A. Oddsson, B. V. Halldorsson, B. O. Jensson, F. Zink, G. H. Halldorsson, G. Masson, G. A. Arnadottir, H. Katrinardottir, K. Juliusson, M. K. Magnusson, O. T. Magnusson, R. Fridriksdottir, S. Saevarsdottir, S. A. Gudjonsson, S. N. Stacey, S. Rognvaldsson, T. Eiriksdottir, T. A. Olafsdottir, V. Steinthorsdottir, V. Tragante, M. O. Ulfarsson, H. Stefansson, I. Jonsdottir, H. Holm, T. Rafnar, P. Melsted, J. Saemundsdottir, G. L. Norddahl, S. H. Lund, D. F. Gudbjartsson, U. Thorsteinsdottir, and K. Stefansson. Large-scale integration of the plasma

proteome with genetics and disease. *Nature Genetics*, 53(12):1712–1721, 2021. ISSN 15461718. doi: 10.1038/s41588-021-00978-w.

A. Fillbrunn, C. Dietz, J. Pfeuffer, R. Rahn, G. A. Landrum, and M. R. Berthold. KNIME for reproducible cross-domain analysis of life science data. *Journal of Biotechnology*, 261(February):149–156, 2017. ISSN 18734863. doi: 10.1016/j.jbiotec.2017.07.028.

A. R. Florian and A. Yilmaz. Mitochondrial Heart Involvement. In M. Mancuso and T. Klopstock, editors, *Diagnosis and Management of Mitochondrial Disorders*, chapter Mitchondri, pages 257–280. Springer, Basel, 2019.

A. Frankish, M. Diekhans, A. M. Ferreira, R. Johnson, I. Jungreis, J. Loveland, J. M. Mudge, C. Sisu, J. Wright, J. Armstrong, I. Barnes, A. Berry, A. Bignell, S. Carbonell Sala, J. Chrast, F. Cunningham, T. Di Domenico, S. Donaldson, I. T. Fiddes, C. García Girón, J. M. Gonzalez, T. Grego, M. Hardy, T. Hourlier, T. Hunt, O. G. Izuogu, J. Lagarde, F. J. Martin, L. Martínez, S. Mohanan, P. Muir, F. C. Navarro, A. Parker, B. Pei, F. Pozo, M. Ruffier, B. M. Schmitt, E. Stapleton, M. M. Suner, I. Sycheva, B. Uszczynska-Ratajczak, J. Xu, A. Yates, D. Zerbino, Y. Zhang, B. Aken, J. S. Choudhary, M. Gerstein, R. Guigó, T. J. Hubbard, M. Kellis, B. Paten, A. Reymond, M. L. Tress, and P. Flicek. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research*, 47(D1):D766–D773, 2019. ISSN 13624962. doi: 10.1093/nar/gky955.

E. R. Gamazon, A. V. Segrè, M. Van De Bunt, X. Wen, H. S. Xi, F. Hormozdiari, H. Ongen, A. Konkashbaev, E. M. Derks, F. Aguet, J. Quan, D. L. Nicolae, E. Eskin, M. Kellis, G. Getz, M. I. McCarthy, E. T. Dermitzakis, N. J. Cox, and K. G. Ardlie. Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nature Genetics*, 50(7):956–967, 2018. ISSN 15461718. doi: 10.1038/s41588-018-0154-4. URL `http://dx.doi.org/10.1038/s41588-018-0154-4`.

S. Ghezelbash, C. E. Molina, and D. Dobrev. Altered atrial metabolism: An underappreciated contributor to the initiation and progression of atrial fibrillation. *Journal of the American Heart Association*, 4(3):1–3, 2015. ISSN 20479980. doi: 10.1161/JAHA.115.001808.

C. Giambartolomei, D. Vukcevic, E. E. Schadt, L. Franke, A. D. Hingorani, C. Wallace, and V. Plagnol. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genetics*, 10(5), 2014. ISSN 15537404. doi: 10.1371/journal.pgen.1004383.

M. Gillespie, B. Jassal, R. Stephan, M. Milacic, K. Rothfels, A. Senff-Ribeiro, J. Griss, C. Sevilla, L. Matthews, C. Gong, C. Deng, T. Varusai, E. Ragueneau, Y. Haider, B. May, V. Shamovsky, J. Weiser, T. Brunson, N. Sanati, L. Beckman, X. Shao, A. Fabregat, K. Sidiropoulos, J. Murillo, G. Viteri, J. Cook, S. Shorser, G. Bader, E. Demir, C. Sander, R. Haw, G. Wu, L. Stein, H. Hermjakob, and P. D'Eustachio. The reactome pathway knowledgebase 2022. *Nucleic Acids Research*, 50(D1):D687–D692, 2022. ISSN 13624962. doi: 10.1093/nar/gkab1028.

A. O. Gramolini, T. Kislinger, P. Liu, D. H. MacLennan, and A. Emili. *Analyzing the Cardiac Muscle Proteome by Liquid Chromatography-Mass Spectrometry-Based Expression Proteomics*, pages 15–31. Humana Press, Totowa, NJ, 2007. ISBN 978-1-59745-214-4. doi: 10.1385/1-59745-214-9:15. URL https://doi.org/10.1385/1-59745-214-9:15.

Z. Gu, L. Gu, R. Eils, M. Schlesner, and B. Brors. Circlize implements and enhances circular visualization in R. *Bioinformatics*, 30(19):2811–2812, 2014. ISSN 14602059. doi: 10.1093/bioinformatics/btu393.

R. J. Hause, A. L. Stark, N. N. Antao, L. K. Gorsic, S. H. Chung, C. D. Brown, S. S. Wong, D. F. Gill, J. L. Myers, L. A. To, K. P. White, M. E. Dolan, and R. B. Jones. Identification and validation of genetic variants that influence transcription factor and cell signaling protein levels. *American Journal of Human Genetics*, 95(2):194–208, 2014. ISSN 15376605. doi: 10.1016/j.ajhg.2014.07.005. URL http://dx.doi.org/10.1016/j.ajhg.2014.07.005.

T. Hayashi, T. Arimura, M. Itoh-Satoh, K. Ueda, S. Hohda, N. Inagaki, M. Takahashi, H. Hori, M. Yasunami, H. Nishi, Y. Koga, H. Nakamura, M. Matsuzaki, B. Y. Choi, S. W. Bae, C. W. You, K. H. Han, J. E. Park, R. Knöll, M. Hoshijima, K. R. Chien, and A. Kimura. Tcap gene mutations in hypertrophic cardiomyopathy and dilated cardiomyopathy. *Journal of the American College of Cardiology*, 44(11):2192–2201, 2004. ISSN 07351097. doi: 10.1016/j.jacc.2004.08.058. URL http://dx.doi.org/10.1016/j.jacc.2004.08.058.

M. Heinig, M. E. Adriaens, S. Schafer, H. W. van Deutekom, E. M. Lodder, J. S. Ware, V. Schneider, L. E. Felkin, E. E. Creemers, B. Meder, H. A. Katus, F. Rühle, M. Stoll, F. Cambien, E. Villard, P. Charron, A. Varro, N. H. Bishopric, A. L. George, C. dos Remedios, A. Moreno-Moral, F. Pesce, A. Bauerfeind, F. Rüschendorf, C. Rintisch, E. Petretto, P. J. Barton, S. A. Cook, Y. M. Pinto, C. R. Bezzina, and N. Hubner. Natural genetic variation of the cardiac transcriptome in non-diseased donors and patients with dilated cardiomyopathy. *Genome Biology*, 18(1):1–21, 2017. ISSN 1474760X. doi: 10.1186/s13059-017-1286-z.

R. E. Hershberger, J. R. Pinto, S. B. Parks, J. D. Kushner, D. Li, S. Ludwigsen, J. Cowan, A. Morales, M. S. Parvatiyar, and J. D. Potter. Clinical and functional Characterization of TNNT2 mutations identified in patients with dilated cardiomyopathy. *Circulation: Cardiovascular Genetics*, 2(4):306–313, 2009. ISSN 1942325X. doi: 10.1161/CIRCGENETICS.108.846733. URL https://doi.org/10.1161/CIRCGENETICS.108.846733.

G. Hindricks, T. Potpara, N. Dagres, J. J. Bax, G. Boriani, G. A. Dan, L. Fauchier, J. M. Kalman, D. A. Lane, M. Lettino, F. J. Pinto, G. N. Thomas, M. Valgimigli, B. P. Van Putte, P. Kirchhof, M. Kühne, V. Aboyans, A. Ahlsson, P. Balsam, J. Bauersachs, S. Benussi, A. Brandes, F. Braunschweig, A. J. Camm, D. Capodanno, B. Casadei, D. Conen, H. J. Crijns, V. Delgado, D. Dobrev, H. Drexel, L. Eckardt, D. Fitzsimons, T. Folliguet, C. P. Gale, B. Gorenek, K. G. Haeusler, H. Heidbuchel, B. Iung, H. A. Katus, D. Kotecha, U. Landmesser, C. Leclercq, B. S. Lewis, J. Mascherbauer, J. L.

Merino, B. Merkely, L. Mont, C. Mueller, K. V. Nagy, J. Oldgren, N. Pavlović, R. F. Pedretti, S. E. Petersen, J. P. Piccini, B. A. Popescu, H. Pürerfellner, D. J. Richter, M. Roffi, A. Rubboli, D. Scherr, R. B. Schnabel, I. A. Simpson, E. Shlyakhto, M. F. Sinner, J. Steffel, M. Sousa-Uva, P. Suwalski, M. Svetlosak, R. M. Touyz, E. Arbelo, C. Blomström-Lundqvist, M. Castella, P. E. Dilaveris, G. Filippatos, M. La Meir, J. P. Lebeau, G. Y. Lip, G. Neil Thomas, I. C. Van Gelder, and C. L. Watkins. 2020 ESC Guidelines for the diagnosis and management of atrial fibrillation developed in collaboration with the European Association for Cardio-Thoracic Surgery (EACTS). *European Heart Journal*, 42(5):373–498, 2021. ISSN 15229645. doi: 10.1093/eurheartj/ehaa612.

M. Hollander, D. A. Wolfe, and E. Chicken. *The Two-Sample Dispersion Problem and Other Two-Sample Problems*, chapter 5, pages 151–201. John Wiley & Sons, Ltd, 2015. ISBN 9781119196037. doi: https://doi.org/10.1002/9781119196037.ch5. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119196037.ch5`.

R. Holle, M. Happich, H. Löwel, H. E. Wichmann, and f. t. M. S. Group. KORA - A Research Platform for Population Based Health Research TT - KORA - Eine Forschungsplattform für bevölkerungsbezogene Gesundheitsforschung. *Gesundheitswesen*, 67(S 01):19–25, 2005. ISSN 0941-3790. doi: 10.1055/s-2005-858235.

B. N. Howie, P. Donnelly, and J. Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, 5 (6):e1000529, 2009. ISSN 15537390. doi: 10.1371/journal.pgen.1000529. URL `https://doi.org/10.1371/journal.pgen.1000529`.

Y. F. Hu, Y. J. Chen, Y. J. Lin, and S. A. Chen. Inflammation and the pathogenesis of atrial fibrillation. *Nature Reviews Cardiology*, 12(4):230–243, 2015. ISSN 17595010. doi: 10.1038/nrcardio.2015.2. URL `http://dx.doi.org/10.1038/nrcardio.2015.2`.

R. T. Huang, S. Xue, Y. J. Xu, M. Zhou, and Y. Q. Yang. A novel NKX2.5 loss-of-function mutation responsible for familial atrial fibrillation. *International Journal of Molecular Medicine*, 31(5):1119–1126, 2013. ISSN 11073756. doi: 10.3892/ijmm.2013.1316.

Y. K. Iwasaki, K. Nishida, T. Kato, and S. Nattel. Atrial fibrillation pathophysiology: Implications for management. *Circulation*, 124(20):2264–2274, 2011. ISSN 00097322. doi: 10.1161/CIRCULATIONAHA.111.019893.

S. Jhaveri, P. F. Aziz, and E. Saarel. Expanding the electrical phenotype of NKX2-5 mutations: Ventricular tachycardia, atrial fibrillation, and complete heart block within one family. *HeartRhythm Case Reports*, 4(11):530–533, 2018. ISSN 22140271. doi: 10.1016/j.hrcr.2018.08.001. URL `https://doi.org/10.1016/j.hrcr.2018.08.001`.

L. Jiang, M. Wang, S. Lin, R. Jian, X. Li, J. Chan, G. Dong, H. Fang, A. E. Robinson, F. Aguet, S. Anand, K. G. Ardlie, S. Gabriel, G. Getz, A. Graubert, K. Hadley, R. E. Handsaker, K. H. Huang, S. Kashin, D. G. MacArthur, S. R. Meier, J. L. Nedzel, D. Y. Nguyen, A. V. Segrè, E. Todres, B. Balliu, A. N. Barbeira, A. Battle, R. Bonazzola,

A. Brown, C. D. Brown, S. E. Castel, D. Conrad, D. J. Cotter, N. Cox, S. Das, O. M. de Goede, E. T. Dermitzakis, B. E. Engelhardt, E. Eskin, T. Y. Eulalio, N. M. Ferraro, E. Flynn, L. Fresard, E. R. Gamazon, D. Garrido-Martín, N. R. Gay, R. Guigó, A. R. Hamel, Y. He, P. J. Hoffman, F. Hormozdiari, L. Hou, H. K. Im, B. Jo, S. Kasela, M. Kellis, S. Kim-Hellmuth, A. Kwong, T. Lappalainen, X. Li, Y. Liang, S. Mangul, P. Mohammadi, S. B. Montgomery, M. Muñoz-Aguirre, D. C. Nachun, A. B. Nobel, M. Oliva, Y. Park, Y. Park, P. Parsana, F. Reverter, J. M. Rouhana, C. Sabatti, A. Saha, A. D. Skol, M. Stephens, B. E. Stranger, B. J. Strober, N. A. Teran, A. Viñuela, G. Wang, X. Wen, F. Wright, V. Wucher, Y. Zou, P. G. Ferreira, G. Li, M. Melé, E. Yeger-Lotem, M. E. Barcus, D. Bradbury, T. Krubit, J. A. McLean, L. Qi, K. Robinson, N. V. Roche, A. M. Smith, L. Sobin, D. E. Tabor, A. Undale, J. Bridge, L. E. Brigham, B. A. Foster, B. M. Gillard, R. Hasz, M. Hunter, C. Johns, M. Johnson, E. Karasik, G. Kopen, W. F. Leinweber, A. McDonald, M. T. Moser, K. Myer, K. D. Ramsey, B. Roe, S. Shad, J. A. Thomas, G. Walters, M. Washington, J. Wheeler, S. D. Jewell, D. C. Rohrer, D. R. Valley, D. A. Davis, D. C. Mash, P. A. Branton, L. K. Barker, H. M. Gardiner, M. Mosavel, L. A. Siminoff, P. Flicek, M. Haeussler, T. Juettemann, W. J. Kent, C. M. Lee, C. C. Powell, K. R. Rosenbloom, M. Ruffier, D. Sheppard, K. Taylor, S. J. Trevanion, D. R. Zerbino, N. S. Abell, J. Akey, L. Chen, K. Demanelis, J. A. Doherty, A. P. Feinberg, K. D. Hansen, P. F. Hickey, F. Jasmine, R. Kaul, M. G. Kibriya, J. B. Li, Q. Li, S. E. Linder, B. L. Pierce, L. F. Rizzardi, K. S. Smith, J. Stamatoyannopoulos, H. Tang, L. J. Carithers, P. Guan, S. E. Koester, A. R. Little, H. M. Moore, C. R. Nierras, A. K. Rao, J. B. Vaught, S. Volpi, and M. P. Snyder. A quantitative proteome map of the human body. *Cell*, 183(1):269–283, 2020. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2020.08.036. URL `https://www.sciencedirect.com/science/article/pii/S0092867420310783`.

R. Joehanes, X. Zhang, T. Huan, C. Yao, S. xia Ying, Q. T. Nguyen, C. Y. Demirkale, M. L. Feolo, N. R. Sharopova, A. Sturcke, A. A. Schäffer, N. Heard-Costa, H. Chen, P. ching Liu, R. Wang, K. A. Woodhouse, K. Tanriverdi, J. E. Freedman, N. Raghavachari, J. Dupuis, A. D. Johnson, C. J. O'Donnell, D. Levy, and P. J. Munson. Integrated genome-wide analysis of expression quantitative trait loci aids interpretation of genomic association studies. *Genome Biology*, 18(1), 2017. ISSN 1474760X. doi: 10.1186/s13059-016-1142-6.

S. M. Kalsto, J. E. Siland, M. Rienstra, and I. E. Christophersen. Atrial Fibrillation Genetics Update: Toward Clinical Implementation. *Frontiers in Cardiovascular Medicine*, 6(September):1–16, 2019. ISSN 2297055X. doi: 10.3389/fcvm.2019.00127.

G. N. Kanaan, D. A. Patten, C. J. Redpath, and M. E. Harper. Atrial Fibrillation Is Associated With Impaired Atrial Mitochondrial Energetics and Supercomplex Formation in Adults With Type 2 Diabetes. *Canadian Journal of Diabetes*, 43(1):67–75, 2019. ISSN 23523840. doi: 10.1016/j.jcjd.2018.05.007. URL `https://doi.org/10.1016/j.jcjd.2018.05.007`.

A. V. Khera, M. Chaffin, K. G. Aragam, M. E. Haas, C. Roselli, S. H. Choi, P. Natarajan, E. S. Lander, S. A. Lubitz, P. T. Ellinor, and S. Kathiresan. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic

mutations. *Nature Genetics*, 50(9):1219–1224, 2018. ISSN 15461718. doi: 10.1038/s41588-018-0183-z. URL http://dx.doi.org/10.1038/s41588-018-0183-z.

R. J. Kinsella, A. Ka, G. Spudich, J. Almeida-king, D. Staines, P. Derwent, A. Kerhornou, P. Kersey, and P. Flicek. Original article Ensembl BioMarts : a hub for data retrieval across taxonomic space. *Database*, 2011(bar030):1–9, 2011. doi: 10.1093/database/bar030.

M. Kloosterman, B. T. Santema, C. Roselli, C. P. Nelson, A. Koekemoer, S. P. Romaine, I. C. Van Gelder, C. S. Lam, V. A. Artola, C. C. Lang, L. L. Ng, M. Metra, S. Anker, G. Filippatos, K. Dickstein, P. Ponikowski, P. van der Harst, P. van der Meer, D. J. van Veldhuisen, E. J. Benjamin, A. A. Voors, N. J. Samani, and M. Rienstra. Genetic risk and atrial fibrillation in patients with heart failure. *European Journal of Heart Failure*, 22(3):519–527, mar 2020. ISSN 18790844. doi: 10.1002/ejhf.1735. URL https://onlinelibrary.wiley.com/doi/full/10.1002/ejhf.1735https://onlinelibrary.wiley.com/doi/abs/10.1002/ejhf.1735https://onlinelibrary.wiley.com/doi/10.1002/ejhf.1735.

G. Korotkevich, V. Sukhov, and A. Sergushichev. Fast gene set enrichment analysis. pages 1–29, 2019. doi: 10.1101/060012. URL https://www.biorxiv.org/content/10.1101/060012v2.full.

J. Köster, F. Mölder, K. P. Jablonski, B. Letcher, M. B. Hall, C. H. Tomkins-Tinch, V. Sochat, J. Forster, S. Lee, S. O. Twardziok, A. Kanitz, A. Wilm, M. Holtgrewe, S. Rahmann, and S. Nahnsen. Sustainable data analysis with Snakemake. *F1000Research*, 10, 2021. ISSN 1759796X. doi: 10.12688/f1000research.29032.2.

T. Lappalainen, M. Sammeth, M. R. Friedländer, P. A. 'T Hoen, J. Monlong, M. A. Rivas, M. Gonzàlez-Porta, N. Kurbatova, T. Griebel, P. G. Ferreira, M. Barann, T. Wieland, L. Greger, M. Van Iterson, J. Almlöf, P. Ribeca, I. Pulyakhina, D. Esser, T. Giger, A. Tikhonov, M. Sultan, G. Bertier, D. G. Macarthur, M. Lek, E. Lizano, H. P. Buermans, I. Padioleau, T. Schwarzmayr, O. Karlberg, H. Ongen, H. Kilpinen, S. Beltran, M. Gut, K. Kahlem, V. Amstislavskiy, O. Stegle, M. Pirinen, S. B. Montgomery, P. Donnelly, M. I. McCarthy, P. Flicek, T. M. Strom, H. Lehrach, S. Schreiber, R. Sudbrak, Á. Carracedo, S. E. Antonarakis, R. Häsler, A. C. Syvänen, G. J. Van Ommen, A. Brazma, T. Meitinger, P. Rosenstiel, R. Guigó, I. G. Gut, X. Estivill, and E. T. Dermitzakis. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511, 2013. ISSN 00280836. doi: 10.1038/nature12531. URL https://doi.org/10.1038/nature12531.

D. A. Lawlor, R. M. Harbord, J. A. Sterne, N. Timpson, and G. D. Smith. Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*, 27(April):1133–1163, 2008. doi: 10.1002/sim.

M. Lemire, S. H. Zaidi, M. Ban, B. Ge, D. Aïssi, M. Germain, I. Kassam, M. Wang, B. W. Zanke, F. Gagnon, P. E. Morange, D. A. Trégouët, P. S. Wells, S. Sawcer, S. Gallinger, T. Pastinen, and T. J. Hudson. Long-range epigenetic regulation is conferred by

genetic variation located at thousands of independent loci. *Nature Communications*, 6, 2015. ISSN 20411723. doi: 10.1038/ncomms7326.

H. Lin, E. V. Dolmatova, M. P. Morley, K. L. Lunetta, D. D. McManus, J. W. Magnani, K. B. Margulies, H. Hakonarson, F. Del Monte, E. J. Benjamin, T. P. Cappola, and P. T. Ellinor. Gene expression and genetic variation in human atria. *Heart Rhythm*, 11(2): 266–271, 2014. ISSN 15475271. doi: 10.1016/j.hrthm.2013.10.051.

M. Litviňuková, C. Talavera-López, H. Maatz, D. Reichart, C. L. Worth, E. L. Lindberg, M. Kanda, K. Polanski, M. Heinig, M. Lee, E. R. Nadelmann, K. Roberts, L. Tuck, E. S. Fasouli, D. M. DeLaughter, B. McDonough, H. Wakimoto, J. M. Gorham, S. Samari, K. T. Mahbubani, K. Saeb-Parsy, G. Patone, J. J. Boyle, H. Zhang, H. Zhang, A. Viveiros, G. Y. Oudit, O. A. Bayraktar, J. G. Seidman, C. E. Seidman, M. Noseda, N. Hubner, and S. A. Teichmann. Cells of the adult human heart. *Nature*, 588 (7838):466–472, 2020. ISSN 14764687. doi: 10.1038/s41586-020-2797-4. URL `https://doi.org/10.1038/s41586-020-2797-4`.

X. Liu, Y. I. Li, and J. K. Pritchard. Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. *Cell*, 177(4):1022–1034.e6, 2019. ISSN 10974172. doi: 10.1016/j.cell.2019.04.014. URL `https://doi.org/10.1016/j.cell.2019.04.014`.

Y. Liu, A. Beyer, and R. Aebersold. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell*, 165(3):535–550, 2016. ISSN 10974172. doi: 10.1016/j.cell.2016.03.014. URL `http://dx.doi.org/10.1016/j.cell.2016.03.014`.

C. Magnussen, T. J. Niiranen, F. M. Ojeda, F. Gianfagna, S. Blankenberg, I. Njølstad, E. Vartiainen, S. Sans, G. Pasterkamp, M. Hughes, S. Costanzo, M. B. Donati, P. Jousilahti, A. Linneberg, T. Palosaari, G. De Gaetano, M. Bobak, H. M. Den Ruijter, E. Mathiesen, T. Jørgensen, S. Söderberg, K. Kuulasmaa, T. Zeller, L. Iacoviello, V. Salomaa, and R. B. Schnabel. Sex differences and similarities in atrial fibrillation epidemiology, risk factors, and mortality in community cohorts: Results from the biomarcare consortium (Biomarker for cardiovascular risk assessment in Europe). *Circulation*, 136 (17):1588–1597, 2017. ISSN 15244539. doi: 10.1161/CIRCULATIONAHA.117.028981.

T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. MacKay, S. A. McCarroll, and P. M. Visscher. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, oct 2009. ISSN 00280836. doi: 10.1038/nature08494. URL `https://www.nature.com/articles/nature08494`.

R. I. Martin, M. S. Babaei, M. K. Choy, W. A. Owens, T. J. Chico, D. Keenan, N. Yonan, M. S. Koref, and B. D. Keavney. Genetic variants associated with risk of atrial fibrillation regulate expression of PITX2, CAV1, MYOZ1, C9orf3 and FANCC. *Journal of Molecular and Cellular Cardiology*, 85:207–214, aug 2015. ISSN 10958584. doi: 10.1016/j.yjmcc.2015.06.005.

M. Mayr, S. Yusuf, G. Weir, Y. L. Chung, U. Mayr, X. Yin, C. Ladroue, B. Madhu, N. Roberts, A. De Souza, S. Fredericks, M. Stubbs, J. R. Griffiths, M. Jahangiri, Q. Xu, and A. J. Camm. Combined Metabolomic and Proteomic Analysis of Human Atrial Fibrillation. *Journal of the American College of Cardiology*, 51(5):585–594, feb 2008. ISSN 07351097. doi: 10.1016/j.jacc.2007.09.055.

W. McLaren, L. Gil, S. E. Hunt, H. S. Riat, G. R. Ritchie, A. Thormann, P. Flicek, and F. Cunningham. The Ensembl Variant Effect Predictor. *Genome Biology*, 17(1):1–14, 2016. ISSN 1474760X. doi: 10.1186/s13059-016-0974-4. URL `http://dx.doi.org/10.1186/s13059-016-0974-4`.

H. Mi, A. Muruganujan, D. Ebert, X. Huang, and P. D. Thomas. PANTHER version 14: More genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Research*, 47(D1):D419–D426, 2019. ISSN 13624962. doi: 10.1093/nar/gky1038.

L. E. Montefiori, D. R. Sobreira, N. J. Sakabe, I. Aneas, A. C. Joslin, G. T. Hansen, G. Bozek, I. P. Moskowitz, E. M. McNally, and M. A. Nóbrega. A promoter interaction map for cardiovascular disease genetics. *eLife*, 7:e35788, 2018. ISSN 2050084X. doi: 10.7554/eLife.35788. URL `https://doi.org/10.7554/eLife.35788`.

V. K. Mootha, C. M. Lindgren, K.-F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstråle, E. Laurila, N. Houstis, M. J. Daly, N. Patterson, J. P. Mesirov, T. R. Golub, P. Tamayo, B. Spiegelman, E. S. Lander, J. N. Hirschhorn, D. Altshuler, and L. C. Groop. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, 34(3):267–273, 2003.

S. Nattel, J. Heijman, L. Zhou, and D. Dobrev. Molecular Basis of Atrial Fibrillation Pathophysiology and Therapy: A Translational Perspective. *Circulation Research*, pages 51–72, 2020. ISSN 15244571. doi: 10.1161/CIRCRESAHA.120.316363.

J. B. Nielsen, R. B. Thorolfsdottir, L. G. Fritsche, W. Zhou, M. W. Skov, S. E. Graham, T. J. Herron, S. McCarthy, E. M. Schmidt, G. Sveinbjornsson, I. Surakka, M. R. Mathis, M. Yamazaki, R. D. Crawford, M. E. Gabrielsen, A. H. Skogholt, O. L. Holmen, M. Lin, B. N. Wolford, R. Dey, H. Dalen, P. Sulem, J. H. Chung, J. D. Backman, D. O. Arnar, U. Thorsteinsdottir, A. Baras, C. O'Dushlaine, A. G. Holst, X. Wen, W. Hornsby, F. E. Dewey, M. Boehnke, S. Kheterpal, B. Mukherjee, S. Lee, H. M. Kang, H. Holm, J. Kitzman, J. A. Shavit, J. Jalife, C. M. Brummett, T. M. Teslovich, D. J. Carey, D. F. Gudbjartsson, K. Stefansson, G. R. Abecasis, K. Hveem, and C. J. Willer. Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. *Nature Genetics*, 50(9):1234–1239, 2018. ISSN 15461718. doi: 10.1038/s41588-018-0171-3. URL `http://dx.doi.org/10.1038/s41588-018-0171-3`.

C. Ogris, T. Helleday, and E. L. Sonnhammer. PathwAX: a web server for network crosstalk based pathway annotation. *Nucleic acids research*, 44(W1):W105–W109, 2016. ISSN 13624962. doi: 10.1093/nar/gkw356.

D. Opacic, K. A. Van Bragt, H. M. Nasrallah, U. Schotten, and S. Verheule. Atrial metabolism and tissue perfusion as determinants of electrical and structural remodelling in atrial fibrillation. *Cardiovascular Research*, 109(4):527–541, 2016. ISSN 17553245. doi: 10.1093/cvr/cvw007.

N. Orr, R. Arnaout, L. J. Gula, D. A. Spears, P. Leong-Sit, Q. Li, W. Tarhuni, S. Reischauer, V. S. Chauhan, M. Borkovich, S. Uppal, A. Adler, S. R. Coughlin, D. Y. Stainier, and M. H. Gollob. A mutation in the atrial-specific myosin light chain gene (MYL4) causes familial atrial fibrillation. *Nature Communications*, 7:11303, 2016. ISSN 20411723. doi: 10.1038/ncomms11303. URL `http://dx.doi.org/10.1038/ncomms11303`.

P. J. Park. ChIP-seq: Advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10:669–680, 2009. ISSN 14710056. doi: 10.1038/nrg2641.

M. S. Parvatiyar, A. P. Landstrom, C. Figueiredo-Freitas, J. D. Potter, M. J. Ackerman, and J. R. Pinto. A mutation in TNNC1-encoded cardiac troponin C, TNNC1-A31S, predisposes to hypertrophic cardiomyopathy and ventricular fibrillation. *Journal of Biological Chemistry*, 287(38):31845–31855, 2012. ISSN 00219258. doi: 10.1074/jbc.M112.377713.

K. F. Pearson. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901. doi: 10.1080/14786440109462720. URL `https://doi.org/10.1080/14786440109462720`.

J. Pfeuffer, T. Sachsenberg, O. Alka, M. Walzer, A. Fillbrunn, L. Nilse, O. Schilling, K. Reinert, and O. Kohlbacher. OpenMS – A platform for reproducible analysis of mass spectrometry data. *Journal of Biotechnology*, 261(May):142–148, 2017. ISSN 18734863. doi: 10.1016/j.jbiotec.2017.05.016.

Roadmap Epigenomics Consortium, A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, V. Amin, J. W. Whitaker, M. D. Schultz, L. D. Ward, A. Sarkar, G. Quon, R. S. Sandstrom, M. L. Eaton, Y. C. Wu, A. R. Pfenning, X. Wang, M. Claussnitzer, Y. Liu, C. Coarfa, R. A. Harris, N. Shoresh, C. B. Epstein, E. Gjoneska, D. Leung, W. Xie, R. D. Hawkins, R. Lister, C. Hong, P. Gascard, A. J. Mungall, R. Moore, E. Chuah, A. Tam, T. K. Canfield, R. S. Hansen, R. Kaul, P. J. Sabo, M. S. Bansal, A. Carles, J. R. Dixon, K. H. Farh, S. Feizi, R. Karlic, A. R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T. R. Mercer, S. J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R. C. Sallari, K. T. Siebenthall, N. A. Sinnott-Armstrong, M. Stevens, R. E. Thurman, J. Wu, B. Zhang, X. Zhou, A. E. Beaudet, L. A. Boyer, P. L. De Jager, P. J. Farnham, S. J. Fisher, D. Haussler, S. J. Jones, W. Li, M. A. Marra, M. T. McManus, S. Sunyaev, J. A. Thomson, T. D. Tlsty, L. H. Tsai, W. Wang, R. A. Waterland, M. Q. Zhang, L. H. Chadwick, B. E. Bernstein, J. F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. A. Stamatoyannopoulos, T. Wang, and M. Kellis. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–329, 2015. ISSN 14764687. doi: 10.1038/nature14248. URL `https://doi.org/10.1038/nature14248`.

F. Robert and J. Pelletier. Exploring the Impact of Single-Nucleotide Polymorphisms on Translation. *Frontiers in Genetics*, 9(October):1–11, 2018. ISSN 1664-8021. doi: 10.3389/fgene.2018.00507.

M. D. Robinson, D. J. Mccarthy, and G. K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *BIOINFORMATICS APPLICATIONS NOTE*, 26(1):139–140, 2010. doi: 10.1093/bioinformatics/btp616. URL `http://bioconductor.org`.

C. Roselli, M. D. Chaffin, L. C. Weng, S. Aeschbacher, G. Ahlberg, C. M. Albert, P. Almgren, A. Alonso, C. D. Anderson, K. G. Aragam, D. E. Arking, J. Barnard, T. M. Bartz, E. J. Benjamin, N. A. Bihlmeyer, J. C. Bis, H. L. Bloom, E. Boerwinkle, E. B. Bottinger, J. A. Brody, H. Calkins, A. Campbell, T. P. Cappola, J. Carlquist, D. I. Chasman, L. Y. Chen, Y. D. I. Chen, E. K. Choi, S. H. Choi, I. E. Christophersen, M. K. Chung, J. W. Cole, D. Conen, J. Cook, H. J. Crijns, M. J. Cutler, S. M. Damrauer, B. R. Daniels, D. Darbar, G. Delgado, J. C. Denny, M. Dichgans, M. Dörr, E. A. Dudink, S. C. Dudley, N. Esa, T. Esko, M. Eskola, D. Fatkin, S. B. Felix, I. Ford, O. H. Franco, B. Geelhoed, R. P. Grewal, V. Gudnason, X. Guo, N. Gupta, S. Gustafsson, R. Gutmann, A. Hamsten, T. B. Harris, C. Hayward, S. R. Heckbert, J. Hernesniemi, L. J. Hocking, A. Hofman, A. R. Horimoto, J. Huang, P. L. Huang, J. Huffman, E. Ingelsson, E. G. Ipek, K. Ito, J. Jimenez-Conde, R. Johnson, J. W. Jukema, S. Kääb, M. Kähönen, Y. Kamatani, J. P. Kane, A. Kastrati, S. Kathiresan, P. Katschnig-Winter, M. Kavousi, T. Kessler, B. L. Kietselaer, P. Kirchhof, M. E. Kleber, S. Knight, J. E. Krieger, M. Kubo, L. J. Launer, J. Laurikka, T. Lehtimäki, K. Leineweber, R. N. Lemaitre, M. Li, H. E. Lim, H. J. Lin, H. Lin, L. Lind, C. M. Lindgren, M. L. Lokki, B. London, R. J. Loos, S. K. Low, Y. Lu, L. P. Lyytikäinen, P. W. Macfarlane, P. K. Magnusson, A. Mahajan, R. Malik, A. J. Mansur, G. M. Marcus, L. Margolin, K. B. Margulies, W. März, D. D. McManus, O. Melander, S. Mohanty, J. A. Montgomery, M. P. Morley, A. P. Morris, M. Müller-Nurasyid, A. Natale, S. Nazarian, B. Neumann, C. Newton-Cheh, M. N. Niemeijer, K. Nikus, P. Nilsson, R. Noordam, H. Oellers, M. S. Olesen, M. Orho-Melander, S. Padmanabhan, H. N. Pak, G. Paré, N. L. Pedersen, J. Pera, A. Pereira, D. Porteous, B. M. Psaty, S. L. Pulit, C. R. Pullinger, D. J. Rader, L. Refsgaard, M. Ribasés, P. M. Ridker, M. Rienstra, L. Risch, D. M. Roden, J. Rosand, M. A. Rosenberg, N. Rost, J. I. Rotter, S. Saba, R. K. Sandhu, R. B. Schnabel, K. Schramm, H. Schunkert, C. Schurman, S. A. Scott, I. Seppälä, C. Shaffer, S. Shah, A. A. Shalaby, J. Shim, M. B. Shoemaker, J. E. Siland, J. Sinisalo, M. F. Sinner, A. Slowik, A. V. Smith, B. H. Smith, J. G. Smith, J. D. Smith, N. L. Smith, E. Z. Soliman, N. Sotoodehnia, B. H. Stricker, A. Sun, H. Sun, J. H. Svendsen, T. Tanaka, K. Tanriverdi, K. D. Taylor, M. Teder-Laving, A. Teumer, S. Thériault, S. Trompet, N. R. Tucker, A. Tveit, A. G. Uitterlinden, P. Van Der Harst, I. C. Van Gelder, D. R. Van Wagoner, N. Verweij, E. Vlachopoulou, U. Völker, B. Wang, P. E. Weeke, B. Weijs, R. Weiss, S. Weiss, Q. S. Wells, K. L. Wiggins, J. A. Wong, D. Woo, B. B. Worrall, P. S. Yang, J. Yao, Z. T. Yoneda, T. Zeller, L. Zeng, S. A. Lubitz, K. L. Lunetta, and P. T. Ellinor. Multi-ethnic genome-wide association study for atrial fibrillation. *Nature Genetics*, 50(9):1225–1233, 2018. ISSN 15461718. doi: 10.1038/s41588-018-0133-9.

S. Sass, F. Buettner, N. S. Mueller, and F. J. Theis. A modular framework for gene set analysis integrating multilevel omics data. *Nucleic Acids Research*, 41(21):9622–9633, 2013. ISSN 03051048. doi: 10.1093/nar/gkt752.

B. Schubert, L. de la Garza, C. Mohr, M. Walzer, and O. Kohlbacher. ImmunoNodes - graphical development of complex immunoinformatics workflows. *BMC Bioinformatics*, 18:242, 2017. ISSN 14712105. doi: 10.1186/s12859-017-1667-z.

B. Schwanhäusser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, and M. Selbach. Global quantification of mammalian gene expression control. *Nature*, 473(7347):337–342, 2011. ISSN 14764687. doi: 10.1038/nature10098.

H. L. Seal. Studies in the history of probability and statistics. XV. The historical velopment of the Gauss linear model. *Biometrika*, 54(1-2):1–24, jun 1967. ISSN 00063444. doi: 10.1093/biomet/54.1-2.1. URL `https://academic.oup.com/biomet/article/54/1-2/1/331494`.

S. Seifert, S. Gundlach, O. Junge, and S. Szymczak. Integrating biological knowledge and gene expression data using pathway-guided random forests: A benchmarking study. *Bioinformatics*, 36(15):4301–4308, 2020. ISSN 14602059. doi: 10.1093/bioinformatics/btaa483.

K. Seongho. ppcor: An R Package for a Fast Calculation to Semi-partial Correlation Coefficients. *Communications for Statistical Applications and Methods*, 22(6):665–674, 2015. doi: 10.5351/CSAM.2015.22.6.665.ppcor. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4681537/pdf/nihms740182.pdf`.

A. A. Shabalin. Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, 28(10):1353–1358, 2012. ISSN 13674803. doi: 10.1093/bioinformatics/bts163.

M. I. Sigurdsson, L. Saddic, M. Heydarpour, T. W. Chang, P. Shekar, S. Aranki, G. S. Couper, S. K. Shernan, J. D. Muehlschlegel, and S. C. Body. Post-operative atrial fibrillation examined using whole-genome RNA sequencing in human left atrial tissue. *BMC Medical Genomics*, 10(1):1–11, 2017. ISSN 17558794. doi: 10.1186/s12920-017-0270-5.

G. D. Smith and G. Hemani. Mendelian randomization: Genetic anchors for causal inference in epidemiological studies. *Human Molecular Genetics*, 23(R1):89–98, 2014. ISSN 14602083. doi: 10.1093/hmg/ddu328.

L. Staerk, J. A. Sherer, D. Ko, E. J. Benjamin, and R. H. Helm. Atrial Fibrillation: Epidemiology, Pathophysiology, Clinical Outcomes. *Circulation Research*, 120(9):1501–1517, 2017. ISSN 15244571. doi: 10.1161/CIRCRESAHA.117.309732.

O. Stegle, L. Parts, R. Durbin, and J. Winn. A bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Computational Biology*, 6(5):e1000770, 2010. ISSN 1553734X. doi:

10.1371/journal.pcbi.1000770.

O. Stegle, L. Parts, M. Piipari, J. Winn, and R. Durbin. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature Protocols*, 7(3):500–507, 2012. ISSN 17542189. doi: 10.1038/nprot.2011.457.

J. D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, 100(16):9440–9445, 2003. ISSN 00278424. doi: 10.1073/pnas.1530509100.

A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, 2005. ISSN 00278424. doi: 10.1073/pnas.0506580102.

P. H. Sudmant, T. Rausch, E. J. Gardner, R. E. Handsaker, A. Abyzov, J. Huddleston, Y. Zhang, K. Ye, G. Jun, M. H. Y. Fritz, M. K. Konkel, A. Malhotra, A. M. Stütz, X. Shi, F. P. Casale, J. Chen, F. Hormozdiari, G. Dayama, K. Chen, M. Malig, M. J. Chaisson, K. Walter, S. Meiers, S. Kashin, E. Garrison, A. Auton, H. Y. Lam, X. J. Mu, C. Alkan, D. Antaki, T. Bae, E. Cerveira, P. Chines, Z. Chong, L. Clarke, E. Dal, L. Ding, S. Emery, X. Fan, M. Gujral, F. Kahveci, J. M. Kidd, Y. Kong, E. W. Lameijer, S. McCarthy, P. Flicek, R. A. Gibbs, G. Marth, C. E. Mason, A. Menelaou, D. M. Muzny, B. J. Nelson, A. Noor, N. F. Parrish, M. Pendleton, A. Quitadamo, B. Raeder, E. E. Schadt, M. Romanovitch, A. Schlattl, R. Sebra, A. A. Shabalin, A. Untergasser, J. A. Walker, M. Wang, F. Yu, C. Zhang, J. Zhang, X. Zheng-Bradley, W. Zhou, T. Zichner, J. Sebat, M. A. Batzer, S. A. McCarroll, R. E. Mills, M. B. Gerstein, A. Bashir, O. Stegle, S. E. Devine, C. Lee, E. E. Eichler, and J. O. Korbel. An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571):75–81, 2015. ISSN 14764687. doi: 10.1038/nature15394. URL `https://doi.org/10.1038/nature15394`.

K. Suhre, M. Arnold, A. M. Bhagwat, R. J. Cotton, R. Engelke, J. Raffler, H. Sarwath, G. Thareja, A. Wahl, R. K. Delisle, L. Gold, M. Pezer, G. Lauc, M. A. D. Selim, D. O. Mook-Kanamori, E. K. Al-Dous, Y. A. Mohamoud, J. Malek, K. Strauch, H. Grallert, A. Peters, G. Kastenmüller, C. Gieger, and J. Graumann. Connecting genetic risk to disease end points through the human blood plasma proteome. *Nature Communications*, 8, 2017. ISSN 20411723. doi: 10.1038/ncomms14357.

B. B. Sun, J. C. Maranville, J. E. Peters, D. Stacey, J. R. Staley, J. Blackshaw, S. Burgess, T. Jiang, E. Paige, P. Surendran, C. Oliver-Williams, M. A. Kamat, B. P. Prins, S. K. Wilcox, E. S. Zimmerman, A. Chi, N. Bansal, S. L. Spain, A. M. Wood, N. W. Morrell, J. R. Bradley, N. Janjic, D. J. Roberts, W. H. Ouwehand, J. A. Todd, N. Soranzo, K. Suhre, D. S. Paul, C. S. Fox, R. M. Plenge, J. Danesh, H. Runz, and A. S. Butterworth. Genomic atlas of the human plasma proteome. *Nature*, 558(7708):73–79, 2018. ISSN 14764687. doi: 10.1038/s41586-018-0175-2. URL `https://doi.org/10.1038/s41586-018-0175-2`.

D. Szklarczyk, A. L. Gable, K. C. Nastou, D. Lyon, R. Kirsch, S. Pyysalo, N. T. Doncheva, M. Legeay, T. Fang, P. Bork, L. J. Jensen, and C. von Mering. The STRING database in 2021: Customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research*, 49(D1):D605–D612, 2021. ISSN 13624962. doi: 10.1093/nar/gkaa1074.

A. M. Thomas, C. P. Cabrera, M. Finlay, K. Lall, M. Nobles, R. J. Schilling, K. Wood, C. A. Mein, M. R. Barnes, P. B. Munroe, and A. Tinker. Differentially expressed genes for atrial fibrillation identified by rna sequencing from paired human left and right atrial appendages. *Physiological Genomics*, 51(8):323–332, 2019. ISSN 15312267. doi: 10.1152/physiolgenomics.00012.2019.

R. Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, feb 1996. ISSN 00359246. URL `http://www.jstor.org/stable/2346178`.

O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, 2001. ISSN 13674803. doi: 10.1093/bioinformatics/17.6.520. URL `http://smi-web.`

T. Tu, S. Zhou, Z. Liu, X. Li, and Q. Liu. Quantitative proteomics of changes in energy metabolism-related proteins in atrial tissue from valvular disease patients with permanent atrial fibrillation. *Circulation Journal*, 78(4):993–1001, 2014. ISSN 13474820. doi: 10.1253/circj.CJ-13-1365.

M. van der Wijst, D. H. de Vries, H. E. Groot, G. Trynka, C. C. Hon, M. J. Bonder, O. Stegle, M. C. Nawijn, Y. Idaghdour, P. van der Harst, C. J. Ye, J. Powell, F. J. Theis, A. Mahfouz, M. Heinig, and L. Franke. The single-cell eQTLGen consortium. *eLife*, 9: 1–21, 2020. ISSN 2050084X. doi: 10.7554/eLife.52155.

A. F. van Ouwerkerk, F. M. Bosada, K. van Duijvenboden, M. C. Hill, L. E. Montefiori, K. T. Scholman, J. Liu, A. A. de Vries, B. J. Boukens, P. T. Ellinor, M. J. T. Goumans, I. R. Efimov, M. A. Nobrega, P. Barnett, J. F. Martin, and V. M. Christoffels. Identification of atrial fibrillation associated genes and functional non-coding variants. *Nature Communications*, 10:4755, 2019. ISSN 20411723. doi: 10.1038/s41467-019-12721-5. URL `https://doi.org/10.1038/s41467-019-12721-5`.

A. F. van Ouwerkerk, F. Bosada, J. Liu, J. Zhang, K. van Duijvenboden, M. Chaffin, N. Tucker, D. A. Pijnappels, P. T. Ellinor, P. Barnett, A. A. de Vries, and V. M. Christoffels. Identification of Functional Variant Enhancers Associated with Atrial Fibrillation. *Circulation Research*, 2020. ISSN 0009-7330. doi: 10.1161/circresaha.119.316006.

A. F. Van Ouwerkerk, A. W. Hall, Z. A. Kadow, S. Lazarevic, J. S. Reyat, N. R. Tucker, R. D. Nadadur, F. M. Bosada, V. Bianchi, P. T. Ellinor, L. Fabritz, J. F. Martin, W. De Laat, P. Kirchhof, I. P. Moskowitz, and V. M. Christoffels. Epigenetic and Transcrip-

tional Networks Underlying Atrial Fibrillation. *Circulation Research*, 127:34–50, 2020. ISSN 15244571. doi: 10.1161/CIRCRESAHA.120.316574.

B. J. Vilhjálmsson, J. Yang, H. K. Finucane, A. Gusev, S. Lindström, S. Ripke, G. Genovese, P. R. Loh, G. Bhatia, R. Do, T. Hayeck, H. H. Won, B. M. Neale, A. Corvin, J. T. Walters, K. H. Farh, P. A. Holmans, P. Lee, B. Bulik-Sullivan, D. A. Collier, H. Huang, T. H. Pers, I. Agartz, E. Agerbo, M. Albus, M. Alexander, F. Amin, S. A. Bacanu, M. Begemann, R. A. Belliveau, J. Bene, S. E. Bergen, E. Bevilacqua, T. B. Bigdeli, D. W. Black, R. Bruggeman, N. G. Buccola, R. L. Buckner, W. Byerley, W. Cahn, G. Cai, D. Campion, R. M. Cantor, V. J. Carr, N. Carrera, S. V. Catts, K. D. Chambert, R. C. Chan, R. Y. Chen, E. Y. Chen, W. Cheng, E. F. Cheung, S. A. Chong, C. R. Cloninger, D. Cohen, N. Cohen, P. Cormican, N. Craddock, J. J. Crowley, D. Curtis, M. Davidson, K. L. Davis, F. Degenhardt, J. Del Favero, L. E. Delisi, D. Demontis, D. Dikeos, T. Dinan, S. Djurovic, G. Donohoe, E. Drapeau, J. Duan, F. Dudbridge, N. Durmishi, P. Eichhammer, J. Eriksson, V. Escott-Price, L. Essioux, A. H. Fanous, M. S. Farrell, J. Frank, L. Franke, R. Freedman, N. B. Freimer, M. Friedl, J. I. Friedman, M. Fromer, L. Georgieva, E. S. Gershon, I. Giegling, P. Giusti-Rodrguez, S. Godard, J. I. Goldstein, V. Golimbet, S. Gopal, J. Gratten, J. Grove, L. De Haan, C. Hammer, M. L. Hamshere, M. Hansen, T. Hansen, V. Haroutunian, A. M. Hartmann, F. A. Henskens, S. Herms, J. N. Hirschhorn, P. Hoffmann, A. Hofman, M. V. Hollegaard, D. M. Hougaard, M. Ikeda, I. Joa, A. Julia, R. S. Kahn, L. Kalaydjieva, S. Karachanak-Yankova, J. Karjalainen, D. Kavanagh, M. C. Keller, B. J. Kelly, J. L. Kennedy, A. Khrunin, Y. Kim, J. Klovins, J. A. Knowles, B. Konte, V. Kucinskas, Z. A. Kucinskiene, H. Kuzelova-Ptackova, A. K. Kahler, C. Laurent, J. L. C. Keong, S. H. Lee, S. E. Legge, B. Lerer, M. Li, T. Li, K. Y. Liang, J. Lieberman, S. Limborska, C. M. Loughland, J. Lubinski, J. Lnnqvist, M. Macek, P. K. Magnusson, B. S. Maher, W. Maier, J. Mallet, S. Marsal, M. Mattheisen, M. Mattingsdal, R. W. McCarley, C. McDonald, A. M. McIntosh, S. Meier, C. J. Meijer, B. Melegh, I. Melle, R. I. Mesholam-Gately, A. Metspalu, P. T. Michie, L. Milani, V. Milanova, Y. Mokrab, D. W. Morris, O. Mors, P. B. Mortensen, K. C. Murphy, R. M. Murray, I. Myin-Germeys, B. Mller-Myhsok, M. Nelis, I. Nenadic, D. A. Nertney, G. Nestadt, K. K. Nicodemus, L. Nikitina-Zake, L. Nisenbaum, A. Nordin, E. O'Callaghan, C. O'Dushlaine, F. A. O'Neill, S. Y. Oh, A. Olincy, L. Olsen, J. Van Os, C. Pantelis, G. N. Papadimitriou, S. Papiol, E. Parkhomenko, M. T. Pato, T. Paunio, M. Pejovic-Milovancevic, D. O. Perkins, O. Pietilinen, J. Pimm, A. J. Pocklington, J. Powell, A. Price, A. E. Pulver, S. M. Purcell, D. Quested, H. B. Rasmussen, A. Reichenberg, M. A. Reimers, A. L. Richards, J. L. Roffman, P. Roussos, D. M. Ruderfer, V. Salomaa, A. R. Sanders, U. Schall, C. R. Schubert, T. G. Schulze, S. G. Schwab, E. M. Scolnick, R. J. Scott, L. J. Seidman, J. Shi, E. Sigurdsson, T. Silagadze, J. M. Silverman, K. Sim, P. Slominsky, J. W. Smoller, H. C. So, C. C. Spencer, E. A. Stahl, H. Stefansson, S. Steinberg, E. Stogmann, R. E. Straub, E. Strengman, J. Strohmaier, T. S. Stroup, M. Subramaniam, J. Suvisaari, D. M. Svrakic, J. P. Szatkiewicz, E. Sderman, S. Thirumalai, D. Toncheva, P. A. Tooney, S. Tosato, J. Veijola, J. Waddington, D. Walsh, D. Wang, Q. Wang, B. T. Webb, M. Weiser, D. B. Wildenauer, N. M. Williams, S. Williams, S. H. Witt, A. R. Wolen, E. H. Wong, B. K. Wormley, J. Q. Wu, H. S. Xi, C. C. Zai, X. Zheng, F. Zimprich, N. R. Wray, K. Stefans-

son, R. Adolfsson, O. A. Andreassen, P. M. Visscher, D. H. Blackwood, E. Bramon, J. D. Buxbaum, A. D. Børglum, S. Cichon, A. Darvasi, E. Domenici, H. Ehrenreich, T. Esko, P. V. Gejman, M. Gill, H. Gurling, C. M. Hultman, N. Iwata, A. V. Jablensky, E. G. Jonsson, K. S. Kendler, G. Kirov, J. Knight, T. Lencz, D. F. Levinson, Q. S. Li, J. Liu, A. K. Malhotra, S. A. McCarroll, A. McQuillin, J. L. Moran, B. J. Mowry, M. M. Nthen, R. A. Ophoff, M. J. Owen, A. Palotie, C. N. Pato, T. L. Petryshen, D. Posthuma, M. Rietschel, B. P. Riley, D. Rujescu, P. C. Sham, P. Sklar, D. St. Clair, D. R. Weinberger, J. R. Wendland, T. Werge, M. J. Daly, P. F. Sullivan, M. C. O'Donovan, P. Kraft, D. J. Hunter, M. Adank, H. Ahsan, K. Aittomäki, L. Baglietto, S. Berndt, C. Blomquist, F. Canzian, J. Chang-Claude, S. J. Chanock, L. Crisponi, K. Czene, N. Dahmen, I. Dos Santos Silva, D. Easton, A. H. Eliassen, J. Figueroa, O. Fletcher, M. Garcia-Closas, M. M. Gaudet, L. Gibson, C. A. Haiman, P. Hall, A. Hazra, R. Hein, B. E. Henderson, J. L. Hopper, A. Irwanto, M. Johansson, R. Kaaks, M. G. Kibriya, P. Lichtner, E. Lund, E. Makalic, A. Meindl, H. Meijers-Heijboer, B. Müller-Myhsok, T. A. Muranen, H. Nevanlinna, P. H. Peeters, J. Peto, R. L. Prentice, N. Rahman, M. J. Sánchez, D. F. Schmidt, R. K. Schmutzler, M. C. Southey, R. Tamimi, R. Travis, C. Turnbull, A. G. Uitterlinden, R. B. Van Der Luijt, Q. Waisfisz, Z. Wang, A. S. Whittemore, R. Yang, W. Zheng, S. Kathiresan, M. Pato, C. Pato, E. Stahl, N. Zaitlen, B. Pasaniuc, E. E. Kenny, M. H. Schierup, P. De Jager, N. A. Patsopoulos, S. McCarroll, M. Daly, S. Purcell, D. Chasman, B. Neale, M. Goddard, N. Patterson, and A. L. Price. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *The American Journal of Human Genetics*, 97(4):576–592, oct 2015. ISSN 0002-9297. doi: 10.1016/J.AJHG.2015.09.001.

U. Võsa, A. Claringbould, H. J. Westra, M. J. Bonder, P. Deelen, B. Zeng, H. Kirsten, A. Saha, R. Kreuzhuber, S. Yazar, H. Brugge, R. Oelen, D. H. de Vries, M. G. van der Wijst, S. Kasela, N. Pervjakova, I. Alves, M. J. Favé, M. Agbessi, M. W. Christiansen, R. Jansen, I. Seppälä, L. Tong, A. Teumer, K. Schramm, G. Hemani, J. Verlouw, H. Yaghootkar, R. Sönmez Flitman, A. Brown, V. Kukushkina, A. Kalnapenkis, S. Rüeger, E. Porcu, J. Kronberg, J. Kettunen, B. Lee, F. Zhang, T. Qi, J. A. Hernandez, W. Arindrarto, F. Beutner, J. Dmitrieva, M. Elansary, B. P. Fairfax, M. Georges, B. T. Heijmans, A. W. Hewitt, M. Kähönen, Y. Kim, J. C. Knight, P. Kovacs, K. Krohn, S. Li, M. Loeffler, U. M. Marigorta, H. Mei, Y. Momozawa, M. Müller-Nurasyid, M. Nauck, M. G. Nivard, B. W. Penninx, J. K. Pritchard, O. T. Raitakari, O. Rotzschke, E. P. Slagboom, C. D. Stehouwer, M. Stumvoll, P. Sullivan, P. A. 't Hoen, J. Thiery, A. Tönjes, J. van Dongen, M. van Iterson, J. H. Veldink, U. Völker, R. Warmerdam, C. Wijmenga, M. Swertz, A. Andiappan, G. W. Montgomery, S. Ripatti, M. Perola, Z. Kutalik, E. Dermitzakis, S. Bergmann, T. Frayling, J. van Meurs, H. Prokisch, H. Ahsan, B. L. Pierce, T. Lehtimäki, D. I. Boomsma, B. M. Psaty, S. A. Gharib, P. Awadalla, L. Milani, W. H. Ouwehand, K. Downes, O. Stegle, A. Battle, P. M. Visscher, J. Yang, M. Scholz, J. Powell, G. Gibson, T. Esko, and L. Franke. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nature Genetics*, 53(9):1300–1310, sep 2021. ISSN 15461718. doi: 10.1038/s41588-021-00913-z. URL `https://www.nature.com/articles/s41588-021-00913-z`.

B. Wang, K. L. Lunetta, J. Dupuis, S. A. Lubitz, L. Trinquart, L. Yao, P. T. Ellinor, E. J. Benjamin, and H. Lin. Integrative Omics Approach to Identifying Genes Associated With Atrial Fibrillation. *Circulation research*, 126(3):350–360, 2020. ISSN 15244571. doi: 10.1161/CIRCRESAHA.119.315179.

D. Wang, B. Eraslan, T. Wieland, B. Hallström, T. Hopf, D. P. Zolg, J. Zecha, A. Asplund, L. Li, C. Meng, M. Frejno, T. Schmidt, K. Schnatbaum, M. Wilhelm, F. Ponten, M. Uhlen, J. Gagneur, H. Hahne, and B. Kuster. A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Molecular Systems Biology*, 15(2):1–16, 2019a. ISSN 1744-4292. doi: 10.15252/msb.20188503.

Y. Wang and J.-G. Wang. Genome-Wide Association Studies of Hypertension and Several Other Cardiovascular Diseases. *Pulse*, 6:169–186, 2018. ISSN 2235-8676. doi: 10.1159/000496150.

Y. Wang, B. He, Y. Zhao, J. L. Reiter, S. X. Chen, E. Simpson, W. Feng, and Y. Liu. Comprehensive Cis-Regulation Analysis of Genetic Variants in Human Lymphoblastoid Cell Lines. *Frontiers in Genetics*, 10(September):1–12, 2019b. ISSN 16648021. doi: 10.3389/fgene.2019.00806.

Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10:57–63, 2009.

T. M. Watt, K. C. Kleeman, A. A. Brescia, E. M. Seymour, A. Kirakosyan, S. P. Khan, L. M. Rosenbloom, S. L. Murray, M. A. Romano, and S. F. Bolling. Inflammatory and Antioxidant Gene Transcripts: A Novel Profile in Postoperative Atrial Fibrillation. *Seminars in Thoracic and Cardiovascular Surgery*, 33(4):948–955, dec 2021. ISSN 1043-0679. doi: 10.1053/J.SEMTCVS.2020.11.026. URL `http://www.semthorcardiovascsurg.com/article/S1043067920304214/fulltexthttp://www.semthorcardiovascsurg.com/article/S1043067920304214/abstracthttps://www.semthorcardiovascsurg.com/article/S1043-0679(20)30421-4/abstract`.

L. C. Weng, S. H. Choi, D. Klarin, J. G. Smith, P. R. Loh, M. Chaffin, C. Roselli, O. L. Hulme, K. L. Lunetta, J. Dupuis, E. J. Benjamin, C. Newton-Cheh, S. Kathiresan, P. T. Ellinor, and S. A. Lubitz. Heritability of Atrial Fibrillation. *Circulation: Cardiovascular Genetics*, 10(6):1–7, 2017. ISSN 19423268. doi: 10.1161/CIRCGENETICS.117.001838.

A. D. Westont and L. Hood. Systems Biology, Proteomics, and the Future of Health Care: Toward Predictive, Preventative, and Personalized Medicine. *Journal of Proteome Research*, 3(2):179–196, 2004. ISSN 15353893. doi: 10.1021/pr0499693.

H. J. Westra, M. J. Peters, T. Esko, H. Yaghootkar, C. Schurmann, J. Kettunen, M. W. Christiansen, B. P. Fairfax, K. Schramm, J. E. Powell, A. Zhernakova, D. V. Zhernakova, J. H. Veldink, L. H. Van Den Berg, J. Karjalainen, S. Withoff, A. G. Uitterlinden, A. Hofman, F. Rivadeneira, P. A. Hoen, E. Reinmaa, K. Fischer, M. Nelis, L. Milani, D. Melzer, L. Ferrucci, A. B. Singleton, D. G. Hernandez, M. A. Nalls, G. Homuth, M. Nauck, D. Radke, U. Völker, M. Perola, V. Salomaa, J. Brody, A. Suchy-Dicey, S. A.

Gharib, D. A. Enquobahrie, T. Lumley, G. W. Montgomery, S. Makino, H. Prokisch, C. Herder, M. Roden, H. Grallert, T. Meitinger, K. Strauch, Y. Li, R. C. Jansen, P. M. Visscher, J. C. Knight, B. M. Psaty, S. Ripatti, A. Teumer, T. M. Frayling, A. Metspalu, J. B. Van Meurs, and L. Franke. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nature Genetics*, 45(10):1238–1243, 2013. ISSN 10614036. doi: 10.1038/ng.2756.

C. Yao, G. Chen, C. Song, J. Keefe, M. Mendelson, T. Huan, B. B. Sun, A. Laser, J. C. Maranville, H. Wu, J. E. Ho, P. Courchesne, A. Lyass, M. G. Larson, C. Gieger, J. Graumann, A. D. Johnson, J. Danesh, H. Runz, S. J. Hwang, C. Liu, A. S. Butterworth, K. Suhre, and D. Levy. Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nature Communications*, 9(1), 2018. ISSN 20411723. doi: 10.1038/s41467-018-05512-x.