Large-scale data-driven analysis to understand the genetics of Congenital Heart Disease

Dissertation

in fulfilment of the requirements for the degree of Doctor *rerum naturalium*

of the Faculty of Mathematics and Natural Sciences

at Kiel University

submitted by

Enrique Audain Martinez

Kiel, 2021

First examiner: Prof. Dr. Hinrich Schulenburg, PhD

Second examiner: Prof. Dr. Marc-Phillip Hitz, PhD

Date of the oral examination: 16.12.2021

Declaration

I, Enrique Audain Martinez, declare that:

a) Apart from my supervisors' guidance, the content and design of this thesis is all my work.

b) This thesis has not already been submitted neither partially nor wholly as part of a doctoral degree to another examining body.

c) The co-authored peer-reviewed manuscripts included in this thesis (chapters II, III and IV) are publicly available under the Creative Commons CC0 license. I have contributed substantially to the conceptualization, data analysis, data curation and writing process of the publications in chapters II, III and V; and contributed significantly to the publication presented in chapter IV on a collaborative basis.

d) This thesis has been prepared subject to the Rules of Good Scientific Practice of the German Research Foundation (DFG).

e) Prior to this thesis, I have not attempted and failed to obtain a doctoral degree.

Signature: _____

Acknowledgements

I would like to express my gratitude to everyone who made it possible to reach this achievement.

I thank my supervisors, Prof. Dr. Marc-Phillip Hitz and Prof. Dr. Hinrich Schulenburg, for their support and guidance. I thank Dr. Yasset Perez-Riverol for his former and actual guidance and friendship. I thank my colleagues Anne, Asal, Gregor, Kerstin, Kirstin and Philipp from the Zebrafish Lab at QFZ, Kiel. I thank the director of the *Klinik für angeborene Herzfehler und Kinderkardiologie,* Prof. Dr. Hans-Heiner Kramer, for having me as a PhD student. Furthermore, I thank all the collaborators who contributed to the publications supporting this work, especially Prof. Dr. Lars Allan Larsen.

I thank my former colleagues Kathya, Yamilet, Elinent and Yaretnis in Cuba for their support and friendship.

For their love and support, I would like to thank the wonderful people I have met in Germany, both friends and family, especially Alexander, Charlotte and my fiancé Wiebke.

I thank my lovely and supportive Cuban family, especially my parents Adalis and Enrique, siblings Ariannis and Allan, and Grandmother Cristina.

Thanks to all of you for accompanying me through this road.

Table of contents

Summary

Congenital Heart Disease (CHD) delineates a large group of structural defects, which can occur due to perturbations at some stage in the cardiac embryogenesis process. With a global incidence ranging from 7 to 9 cases per 1000 live births, CHD accounts for a significant fraction of new-borns deaths worldwide. Different studies have identified genetics as an essential factor underlying CHD, along with environmental factors.

The technological advances within the last years have helped improve CHD diagnosis and understand its genetic causes. Specifically, next-generation sequencing technologies (e.g., exome/genome and single-cell sequencing) have been increasingly used to discover and associate new genetic variants and genes to CHD. Nevertheless, despite the advances in our understanding of the disease, many molecular mechanisms underlying CHD remain uncertain.

Herein I present my efforts focused on discovering new genes and biological pathways altered in patients with CHD. The work is based on large CHD patient cohorts, collected and analysed as part of an international collaboration.

The adopted integrative data-driven approach in this work can roughly be grouped into two principal aims: i) the development of statistical frameworks and bioinformatics tools to analyse high-dimensional data and ii) the meta-analysis of large-scale exome sequencing data to elucidate variants and genes conferring risk of CHD.

We first developed a series of feature selection workflows and tools for analysing omics data. We later used this approach in the quality control process and filtered our large case-control exome dataset at the sample and variant level. This step is crucial to decrease the likelihood of finding false genetic associations in case-control studies.

By meta-analysing copy number variations and *de novo* variants in CHD probands, we implicated novel genes reaching genome-wide significant association with CHD and strengthened previously described associations. In addition, our analysis highlighted the central role of Notch- and calcium signalling-related pathways in the pathophysiology of CHD.

We also explored the differences between non-syndromic and syndromic CHD by analysing a large-scale exome cohort of patients. Moreover, by integrating exome and single-cell transcriptomics data, we showed relevant gene expression patterns in cardiac-specific cells and possible mechanisms underlying syndromic and non-syndromic forms of CHD.

In summary, our integrative approach, supported by the data analysis of ~15,000 CHD patients, allowed us to gain new insights into the genetic origin of CHD. Consequently, we present here a valuable resource to continue investigating the causes of CHD and pave the way to promote new studies in this area.

Zusammenfassung

Angeborene Herzfehler (AHF) sind die häufigste humane Fehlbildung mit einer weltweiten Inzidenz von 7 bis 9 Fällen pro 1000 Lebendgeburten und ursächlich für einen hohen Anteil an Kindersterblichkeit. AHF bezeichnen eine heterogene Gruppe von Fehlbindungen, die auf Grund von Störungen der Herzentwicklung auftreten. Verschiedene Studien zeigen, dass neben Umweltfaktoren auch genetische Faktoren eine wesentliche Rolle bei der Entstehung von AHF spielen.

Die technologischen Fortschritte der letzten Jahre haben dazu beigetragen, dass wesentlich häufiger eine molekulargenetische Diagnose gestellt werden kann und die genetischen Ursachen zunehmend besser verstanden werden. Insbesondere die neuen Sequenzierungstechnologien (z.B. Exom- /Genom- und Einzelzellsequenzierung) werden zunehmend eingesetzt, um neue genetische Varianten und Gene zu finden und dann mit einem AHF in einen kausalen Zusammenhang zu stellen. Trotz der Fortschritte in unserem Verständnis der molekularen Mechanismen, die den AHF zugrunde liegen, sind diese nach wie vor größtenteils ungeklärt.

Hier stelle ich meine Arbeit vor, die sich mit der Entdeckung neuer Gene und Signalwege befasst, welche in einem kausalen Zusammenhang mit der Entstehung der AHF stehen. Die Arbeit basiert auf großen AHF-Patientenkohorten, die im Rahmen internationaler Kollaborationen gesammelt und analysiert wurden.

Wir haben in dieser Arbeit einen integrativen datengestützten Ansatz verfolgt, der sich in zwei Hauptziele gliedert: i) die Entwicklung von statistischen und bioinformatischen Methoden für die Analyse hochdimensionaler Datensätze, und ii) die Meta-Analyse großer Exom-Sequenzierungsdatensätze zur Erkennung von Varianten und Genen, die in einem kausalen Zusammenhang mit der Entstehung von AHF stehen.

Wir haben zunächst eine Reihe von Methoden für die Analyse von Omics-Daten, d.h. Daten aus verschiedenen molekularbiologischen Analysen, entwickelt und diesen Ansazt bei der Qualitätskontrolle angewendet. Damit wurden die existierenden Datensätze auf Proben- und Variantenebene gefiltert. Dieser Schritt ist wichtig, um die Wahrscheinlichkeit zu verringern, dass in Fall-Kontroll-Studien falsche positive Assoziationen identifiziert werden.

Durch eine Meta-Analyse von Kopienzahlvariationen und De-Novo-Varianten bei AHF-Probanden konnten wir neue Gene identifizieren, die genomweit-signifikant mit AHF assoziiert sind. Zudem war es möglich bereits beschriebene Assoziationen zu bestätigen. Unsere bioinformatische Analyse konnte eine zentrale Rolle der Notch- und Kalzium-Signalwege in der AHF darlegen.

Weiterhin erforschten wir die Unterschiede zwischen nicht-syndromalen und syndromalen AHF durch die Analyse einer großen exomsequenzierten Fall-Kontroll-Kohorte. Durch die Integration von Exomdaten und Einzelzell-Transkriptom-Daten war es möglich relevante Genexpressionsmuster in herzspezifischen Zellen aufzuzeigen, sowie die möglichen Mechanismen, die den syndromalen und nicht-syndromalen Formen der AHF zugrunde liegen, zu identifizieren.

Insgesamt konnten wir mit unserem integrativen Ansatz, der durch die Analyse von ca. 15.000 AHF-Patientendaten gestützt wurde, neue Erkenntnisse über die Genetik der AHF und die damit verbundenen biologischen Prozesse gewinnen. Folglich haben wir hier eine wertvolle Quelle an Patientendaten für zukünftige Ursachenforschung der AHF geschaffen und den Weg für weitere Studien auf diesem Gebiet geebnet.

Chapter I. Introduction

Introduction to CHD

Congenital heart disease (CHD), the most common cause of congenital anomalies, remains one of the most prevalent global health problems. Different epidemiological studies report an estimate of 7 to 9 affected new-borns per 1,000 live births worldwide[1]. Medical improvements within the last three decades, such as surgical, interventional, and clinical intensive care, have increased survival rates dramatically. These advances have resulted in a rapidly growing number of CHD survivors reaching adulthood[2]. Also, this group of patients present an increased risk of arrhythmias, endocarditis, and heart failure, among other co-morbidities[2].

The heart is a complex organ with contributions from at least four distinct progenitor cell types: the first heart field (FHF), the second heart field (SHF), cardiac neural crest, and the proepicardial organ[3,4]. Interestingly, the exact number of distinct cardiac cell types is still not fully known. Recent estimates using novel single-cell technology point to 11 major cardiac cell types identified in adult human heart[5]. It is known that the endoderm plays an essential role in the differentiation and specification of cardiac-specific cells during embryonic development in mammals[6]. Therefore, the perturbation of these developmental processes can lead to various effects, depending on the stage of development and the nature of the perturbing factor.

In the early 1970s, Mitchell *et al*. defined CHD as “a gross structural abnormality of the heart or intrathoracic great vessels that is actually or possibly of functional significance”[7]. Thus far, more than 20 specific types of CHD have been described[8], including many complex subtypes, which range from mild to severe CHD forms. The incidence in the population varies across distinct subtypes of CHD. Mild defects such as ventricular septal defect (VSD), atrial septal defect (ASD) and patent ductus arteriosus (PDA), three of the most common lesions among affected individuals,

represent ~58% of the total burden of CHD[9]. More severe forms, such as Tetralogy of Fallot (TOF), and the hypoplastic left heart syndrome (HLHS), as well as defects resulting from abnormal left-right relationships (e.g., transposition of the great arteries (TGA)), are observed in approximately 3 to 5% of the affected individuals[9,10].

CHD can occur as part of a syndrome (syndromic CHD) along with various extra-cardiac malformations or occur isolated (non-syndromic CHD). Different studies have described the distinct genetic architecture between syndromic and non-syndromic forms of CHD[11,12]. It has also been shown that there is a larger effect of the mutations and genes associated with syndromic forms of CHD compared to non-syndromic CHD[11–13].

Despite the technological advances in the last decade and diagnostic improvements in CHD, the underlying causes of CHD often remain unclear. Numerous studies have established a contribution of both genetic and environmental factors, and epidemiological data point to genetics as a significant disease cause[10,14].

Aetiology of CHD

Environmental risk factors for CHD

Embryogenesis is a complex process that can be perturbated by external factors. For example, lack of essential nutrients or excess of toxic substances can disrupt the placental development and alter nutrient supply to the embryo. Hence leading, directly or indirectly, to abnormal development of the embryo[8].

Early studies have established that alcohol consumption during pregnancy can lead to foetal alcohol spectrum disorder[15] (FASD), a constituted risk factor for CHD. It has

been observed that a high proportion of individuals with FASD (up to ~67%) have CHD, mostly VSD, ASD and conotruncal defects[16].

Another important risk factor for CHD is diabetes mellitus[17]. Liu *et al*. observed a 4.6 and 4.2-fold increased risk of CHD in offspring of mothers with diabetes types 1 and 2, respectively[18]. Other studies have reported a higher risk of CHD (~2-fold) among mothers with pre-existing diabetes type 2 compared to mothers with gestational diabetes[19].

Like diabetes mellitus type 2, obesity is an important CHD risk factor[17]. Different population studies have reported that obese women (body mass index (BMI) > 30) were significantly more likely to have children with CHD than women with average weight (BMI: 19-24.9), with an odds ratio ranging from 1.15 to 3.4[20,21]. Both diabetes mellitus (with a higher prevalence of type 2 than type 1) and obesity are complex metabolic diseases observed frequently in the same individual. It has been hypothesised that obesity and diabetes-induced congenital malformations, such as CHD, may share a common aetiology[8,17].

The underlying mechanics by which diabetes and obesity can affect critical stages of cardiac development are not fully known. Numerous studies point to metabolic disorders associated with obesity and diabetes mellitus type 2, such as abnormalities in glucose metabolism[22], oxidative or other cellular stress[23] and deficiencies in nitric oxide signaling[24]; as potential mechanisms conferring risk for CHD.

Genetics risk factors for CHD

Several studies of families with members affected with CHD have demonstrated an important genetic component and a complex model of inheritance[25–27]. A higher recurrence of risk of CHD has also been observed in first-degree relatives compared

to second- and third-degree relatives[25,27]. These reports suggest a polygenic model of inheritance rather than monogenic. A recent large-scale study of families with recurrent CHD found different co-occurrence patterns among sub-types of cardiac malformations[28].

An increased risk of CHD has been observed among twins with a greater concordance of CHD in monozygotic compared to dizygotic twins[29]. Moreover, an increased recurrence of related forms of CHD among siblings has been reported. For example, two of the most common forms of CHD, ASD and heterotaxia, were observed with a recurrence of 3.4 and 79.1 in the Danish national cohort study 1, respectively[30]. These observations are consistent with the prevalence of different CHD subtypes (e.g., ASD and VSD) reported in a recent meta-analysis summarising 260 studies[9].

Other rare Mendelian forms of CHD, such as severe mitral valve prolapse[31] (sMVP) and bicuspid aortic valve[32] (BAV), have been previously described. Studies also indicate an increased incidence of CHD in populations with high levels of consanguinity, which suggests a role for recessive genetic contributions[33].

Despite these earlier studies implicating genetic causes as an important risk factor for CHD, a large proportion of CHD occurs in families with no apparent history of CHD. These observations have paved the way for more recent studies, which suggest that a significant proportion of these cases can be attributable to *de novo* genetic events, including copy number variants (CNVs), single nucleotide variants (SNVs) and more complex variations[11,12,34,35]. These studies point to a significant genetic contribution to CHD and have expanded our understanding of its genetic origin. Nevertheless, the contribution of specific genetic mechanisms to distinct CHD subtypes as well as the contribution of several genetic loci (e.g., single loci with significant effect, combined effect of a few loci, digenic or polygenic effects) is not well described thus far.

Variant spectrum associated with CHD phenotypes

Chromosomal abnormalities

Aneuploidies, defined as “a chromosome number that deviates from a multiple of the haploid set”[36], were the earliest genetic causes associated with CHD. Hartman *et al.* estimate the proportion of CHD associated with chromosomal abnormalities ranging from 9 to 18%[37].

Up to 98% of foetuses with CHD and a cytogenetic abnormality show an extra-cardiac defect. Unfortunately, numerous genes are altered in individuals with aneuploidy, and a single gene cause is often not identifiable[38].

CHD has been linked to diverse forms of aneuploidy with a significant incidence. For example, earlier studies have observed that 35 to 50% of live-born infants with trisomy 21[39], 60 to 80% with trisomy 13 and trisomy 18[40], and 33% with monosomy of chromosome X[41], present CHD.

Recently, the introduction of non-invasive prenatal testing (NIPT) routine in a large number of individuals has improved the detection rate of foetal aneuploidies, such as trisomy 13, 18, 21 and sex chromosome aneuploidy, and its subsequence association with CHD[42–44].

Copy Number Variations

CNVs are a type of structural variation that alter the number of copies of specific regions of DNA, which can be deleted or duplicated. Such chromosomal deletions and duplications can involve thousands of nucleotides, ranging in size from ~1 Kilo-base pairs (Kbps) up to multiple Mega-base pairs (Mbps), and can be inherited or spontaneously arise *de novo*[45].

GnomAD-SV[46] is a reference atlas for genomic structural variations (SVs) from deep whole-genome sequencing in ~15,000 human samples aggregated as part of the gnomAD database, and has helped to extend our understanding of the diversity of mutational patterns among structural variations. These structural variations are thought to contribute approximately to 25% of all rare loss-of-function variants in each genome. Among the SVs described in this gnomAD-SV, CNVs (comprising both duplications and deletions) contributed to ~50% of variability[46].

Over the past 15 years, different studies have allowed us to enhance our understanding of the role of CNVs in cardiac morphogenesis and related disorders, as well as our knowledge about its relevance for clinical practice[45,47]. These findings have also strengthened the evidence that rare CNVs represent a considerable source of the genetic variation contributing to CHD[48–50].

Several recognisable clinical syndromes have been associated with CHD. Del22q11.2 (OMIM 611867), a deletion of ~3 Mb, represents the most common human microdeletion and has been associated with a broad phenotypic spectrum, including CHD, palate abnormalities, hypocalcemia, immunodeficiency, characteristic facial features, neurodevelopmental abnormalities, learning disabilities and psychiatric disorders. It is also known as DiGeorge Syndrome (OMIM 188400) and Velocardio-facial syndrome (OMIM 192430).

Other well-characterised CNVs associated with CHD include del8p23, which involves the cardiac transcription factor GATA4 (OMIM 600576). Individuals with this syndrome manifest not only CHD but also developmental delay[51]. Williams-Beuren syndrome (OMIM 194050), caused by a deletion of 1.5 to 1.8 Mb on chromosome 7q11.23, has been associated with cardiac diseases consisting of supravalvar aortic and pulmonary stenosis and results from haploinsufficiency for *ELN*[52,53]. Another well-studied

syndrome is Jacobsen Syndrome (OMIM 147791), resulting from the deletion of 11q24-25. Individuals affected with this syndrome manifest various extra-cardiac and cardiac malformations, including VSD and ASD[54,55]. The analysis of larger cohorts of patients with CHD has recently found several recurrent CNVs associated with CHD. These CNVs have been found affecting loci such as 1q21.1 (OMIM 274000), 16p13.3 (OMIM 613458), 15q11.2 (OMIM 615656), and 2p13.3 (OMIM 608748) among others[56,57].

Single Nucleotide Variants

Different reports suggest that most CHD is sporadic, with only 2.2% of patients with CHD having affected first-degree relatives[27,28]. The stable incidence of CHD, despite its low reproductive potential, suggests a vital role of *de novo* variants (DNVs) causing CHD[35]. The advance of next-generation sequencing technologies and their applications have allowed gaining deeper insights into the role of DNVs in CHD[11,12]. Unlike inherited variants, DNVs have been exposed to less evolutionary selection and are, on average, more deleterious. It has been reported that *de novo* single nucleotide variants occur with a rate of ~1.6 per exome and ~65 per genome[58].

DNVs account for at least 20% of CHD. Although these variants have been notably more associated with syndromic forms of CHD (those with extra-cardiac and, for example, neurodevelopmental abnormalities)[11,35], there is a small but considerable contribution to non-syndromic forms of CHD (CHD not associated with a known syndrome)[11,13]. In particular, the contribution of DNVs to isolated forms of CHD have become more clearly described by analysing larger cohorts of patients[11,12,59].

Further, the introduction of genome-wide association analysis (GWAS) has allowed us to assess the role of common variants in CHD. Unlike rare pathogenic variants,

which are supposed to have a larger effects size, the association of common loci with CHD has been underreported thus far. Nevertheless, an increasing number of GWAS studies of large cohorts have enhanced our understanding of the contribution of common genomic variations to CHD[60–63].

Next-Generation Sequencing diagnostic improvements and capabilities

Widespread utilisation of next-generation sequencing technologies within the last decade has substantially improved our understanding of the genetic mechanisms underlying complex diseases such as CHD. The advances in high-throughput technologies, such as exome/genome sequencing (ES/GS), have allowed identifying genomic variants (e.g., DNVs, variants with marked reduced penetrance and somatic alterations) and disease associations with higher sensitivity compared to traditional genomic methods (e.g., array-based comparative genomic hybridisation)[64,65]. Also, the advances in these technologies, the development of sophisticated statistical methods, biological databases and bioinformatic tools have played an essential role in studying heterogeneous diseases such as CHD[66].

Exome and genome sequencing (ES/GS)

Traditional Sanger sequencing of individual genes is a relatively time-consuming approach that demands considerable effort. Gene panel-based approaches have overcome some of the limitations of Sanger sequencing, but they require prior knowledge of potential disease-causing genes and are limited to a relatively small number of candidate genes[65].

In contrast, exome and genome sequencing technologies have experienced a notable improvement over the past years. In particular, the efficiency, depth of coverage, high-throughput nature and exponentially decreasing costs have allowed for rapid sequencing and analysis with high accuracy. ES has been extensively used to identify disease-causing mutations in protein-coding regions of the genome. Recently, an increasing number of GS studies has been focused on understanding the contribution of the non-coding regions of the genome[67,68], as well as the impact of common variants[61,62].

Thus, the analysis of genomic data resulting from exome or genome sequencing and GWAS has substantially contributed to progress in our understanding of the biology underlying numerous diseases. In particular, the study of rare genetic diseases, of which molecular diagnosis is challenging due to its low incidence, has markedly been benefited from ES/GS[12,68].

Adaption and progress of statistical methods and tools to analyse large-scale genomics datasets

The progress of sequencing technologies has greatly impacted the amount and complexity of the data produced in biomedical research. Consequently, algorithms and bioinformatic tools traditionally used for the analysis of genomic data are becoming obsolete. At the same time, advances and improvements in high-throughput technologies have led to the development of novel computational approaches for analysing complex and large biological datasets[66,69].

As described by Poplin et al.[70], a typical bioinformatics pipeline for analysing high-throughput exome and genome sequencing data can be summarised as follows:

First, raw read data stored in FASTQ files are converted to recalibrated, analysis-ready reads (shorts fragment of DNA sequence). This step usually involves computational tools such as SAMtools[71] and the Genome Analysis Toolkit[72] (GATK). Second, a mapping tool (e.g., Burrows-Wheeler Aligner, BWA) is used to align the reads to a reference sequence (e.g., human reference genome GRCh38). The aligned reads are mainly stored in Binary Sequence Alignment/Map format (BAM/CRAM files). Next, a "calling variant" step is performed on "active genome regions" (regions that differ considerably from the genome reference), where reads from these "active regions" are split into k-mers (overlapping subsequence) and reassembled into candidate haplotypes using de-Bruijn-like graphs. Then follows a step that computes the read's likelihoods of being derived from each haplotype. Lastly, raw genotype likelihoods are calculated across all samples and used to call raw variants in the cohort being analysed[70]. A more detailed description of the overall process can be found on the GATK website (https://gatk.broadinstitute.org/hc/en-us).

Recently, an increased number of studies have been focusing on analysing large cohorts of samples (usually thousands of samples). These studies demonstrated that the 'joint calling' of the samples offer a powerful tool to solve several shortcomings inherent to processing sequencing data[46,73]. These shortcomings are mainly associated with technical artefacts (e.g., different exome/genome capture products). In particular, the process of joint-calling has facilitated the development of machine learning-based approaches such as the Variant Quality Score Recalibration (VQSR). This tool fits a Gaussian mixture model on the set of known 'true' variants (training set) to estimate the probability that a putative variant is a true genetic variant versus a sequencing or data processing artefact[70].

Variant annotation: tools and resources

Due to the massive amount of data produced in high-throughput sequencing experiments, there is a need for robust computational tools to aid the prioritisation of variants across transcripts and manage the complexities of variant analysis. This is a critical step in analysing and interpreting genomic data, and numerous tools have been developed to overcome these issues. Bioinformatic tools such as the Variant Effect Predictor[74] (VEP), ANNOVAR[75] and SnpEff[76] aim to perform annotations and analysis of most types of genomic variations in both coding and non-coding regions of the genome. The variants annotation and its posterior prioritisation is a crucial step in disease investigation and population studies.

An important component of these tools is the source of transcript annotation and variants in protein-coding or non-coding regions. Thus, when analysing human samples, for example, GENCODE[77] and Reference Sequence[78] (RefSeq) at the National Centre for Biotechnology Information (NCBI), are the two major sources of Homo Sapiens genome annotation information.

Variant and gene prioritisation tools

Most of the variations within a human genome are benign. They are supposed to have no damaging impact leading to a specific genetic disorder. Therefore, identifying pathogenic variants given the vast candidate pool of variants in a human exome or genome is a challenging problem that has led to the development of diverse variant prioritisation tools[79].

Even though the output from annotation tools (e.g., VEP) offer the first level of information, assigning pathogenicity of a genetic variant and its clinical significance

falls along a spectrum, ranging from variants being almost certainly pathogenic to variants being almost certainly benign[80].

For this proposal, traditional approaches have been based on conservation and protein structure information to predict the consequence of a missense variant change on the function of the protein[81,82]. More powerful techniques have recently been developed that extend the scope of variant prioritisation and improve accuracy[79,83–85]. These tools have evolved to predict the probability of loss-of-function of on single genetic variant (e.g., LOFTEE[73]) and prioritise coding regions (e.g., CCRs[86]).

Computational frameworks for processing large-scale genomic datasets

Like other research fields producing a substantial amount of data, genomics has been experiencing (still ongoing) a notable year-over-year data growth. The improvements of sequencers regarding throughput and an increasing number of dedicated sequencing facilities have accelerated the acquisition of large volumes of genomic data. As a result, public repositories to store raw sequencing data, such as Sequence Read Archive (SRA) and European Nucleotide Archive (ENA), have been doubling in size approximately every two years. Cloud computing has recently emerged as a solution to manage large computational resources for processing genomic datasets efficiently. Nevertheless, the cloud computing model raises potential issues regarding patients' data privacy and security, currently under discussion by the scientific community and lawmakers[87].

The analysis of larger and complex datasets with existing computational tools has become infeasible. This fact has accelerated the need to adapt existing computational frameworks or develop new ones to overcome the issues associated with processing genomic data at scale[88].

In this context, distributed computing systems have played a key role in building efficient computational frameworks for analysing big data, including genomic data. Notability Spark, a unified analytics engine for big data processing, has been adopted as a key technology in developing computational tools for analysing big genomic data at scale. Spark consists of a programming model based on dependency graphs and operates on Resilient Distributed Datasets (RDDs), a fault-tolerant collection of objects that can be processed in parallel across a computer cluster[69,89].

ADAM (https://github.com/bigdatagenomics/adam) and Hail (https://hail.is), both computational frameworks built on top of Spark, have emerged as an efficient solution for processing large-scale genomic datasets. Hail is a python library specialised in the processing of large-scale genomics datasets developed by the Broad Institute (USA) Genetics group and has been recently adopted to develop bioinformatics pipelines in large genetic projects such as gnomAD[73]. It has also facilitated the discovery of new variants and genes associated with rare genetic disorders, including cases with CHD[90–93].

In this thesis, we have therefore adopted an integrative data-driven approach to extend our understanding of the genetic CHD causes. To overcome such challenges, we followed two principal aims: i) the development of statistical frameworks and bioinformatics tools to analyse high-dimensional data and (ii) the meta-analysis of large-scale exome sequencing data to elucidate variants and genes conferring risk of CHD.

Overview of the thesis

The present thesis is distributed in six chapters, including the introduction chapter (Chapter I). We highlight in Chapter II the application of machine learning-based methods for analysing high-dimensional data. Specifically, we emphasise the development and application of feature selection techniques for analysing gene and protein expression datasets. Also, we have applied later the proposed techniques for inferring population ancestry in large-scale ES cohorts. Ultimately, the proper matching-ancestry in case-control association studies is a condition required for avoiding spurious gene-diseases associations.

In Chapter III, we present one of the largest integrative meta-analyses of genomics variants in the field of CHD thus far. By aggregating data from ~200 different studies, we analysed the joint contribution of CNV and DNV to CHD by studying ~10,000 CHD probands. Moreover, the conducted data-driven approach considers gene expression at different human developmental stages, and system genetics methods show the potentialities of integrative approaches to study candidate genes associated with complex diseases such as CHD.

In Chapter IV, we describe a study of families with the concurrence of CHD. A system genetics integrative approach allowed us to identify deficiencies in calcium signalling as a cause of CHD. The results could be replicated in an independently sequenced exome CHD cohort.

Chapter V introduces a large-scale case-control association study involving ~57,000 exomes. The analysis focuses on discovering novel CHD causing genes in both syndromic and non-syndromic forms of CHD.

Finally, Chapter VI summarises the overall discussion of relevant topics in the field of congenital heart disease, the main results of this thesis and future directions in the development of diagnostics tools, data analysis, and integrative multi-omics approaches to enhance our knowledge of CHD.

References

1. van der Linde, D. *et al.* Birth Prevalence of Congenital Heart Disease Worldwide. *J. Am. Coll. Cardiol.* **58**, 2241–2247 (2011).
2. van der Bom, T. *et al.* The changing epidemiology of congenital heart disease. *Nat. Rev. Cardiol.* **8**, 50–60 (2011).
3. Kirby, M. L. & Waldo, K. L. Neural crest and cardiovascular patterning. *Circ. Res.* **77**, 211–5 (1995).
4. Rickert-Sperling, S., Kelly, R. G. & Driscoll, D. J. *Congenital Heart Diseases: The Broken Heart*. *Congenital Heart Diseases: The Broken Heart: Clinical Features, Human Genetics and Molecular Pathways* (Springer Vienna, 2016). doi:10.1007/978-3-7091-1883-2
5. Litviňuková, M. *et al.* Cells of the adult human heart. *Nature* **588**, 466–472 (2020).
6. Waldo, K. *et al.* A novel role for cardiac neural crest in heart development. *J. Clin. Invest.* **103**, 1499–1507 (1999).
7. Mitchell, S. C., Korones, S. B. & Berendes, H. W. Congenital heart disease in 56,109 births. Incidence and natural history. *Circulation* **43**, 323–32 (1971).
8. Kalisch-Smith, J. I., Ved, N. & Sparrow, D. B. Environmental risk factors for congenital heart disease. *Cold Spring Harb. Perspect. Biol.* **12**, a037234 (2020).
9. Liu, Y. *et al.* Global birth prevalence of congenital heart defects 1970-2017: Updated systematic review and meta-analysis of 260 studies. *Int. J. Epidemiol.* **48**, 455–463 (2019).
10. Zaidi, S. & Brueckner, M. Genetics and Genomics of Congenital Heart Disease. *Circ. Res.* **120**, 923–940 (2017).
11. Sifrim, A. *et al.* Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing. *Nat. Genet.* **48**, 1060–5 (2016).
12. Jin, S. C. *et al.* Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. *Nat. Genet.* (2017). doi:10.1038/ng.3970
13. Wessels, M. W. & Willems, P. J. Genetic factors in non-syndromic congenital heart malformations. *Clin. Genet.* **78**, 103–23 (2010).
14. Triedman, J. K. & Newburger, J. W. Trends in Congenital Heart Disease: The Next Decade. *Circulation* **133**, 2716–33 (2016).
15. Jones, K. L. & Smith, D. W. Recognition of the fetal alcohol syndrome in early infancy. *Lancet (London, England)* **302**, 999–1001 (1973).
16. Burd, L. *et al.* Congenital heart defects and fetal alcohol spectrum disorders. *Congenit. Heart Dis.* **2**, 250–5 (2007).
17. Helle, E. & Priest, J. R. Maternal Obesity and Diabetes Mellitus as Risk Factors for Congenital Heart Disease in the Offspring. *J. Am. Heart Assoc.* **9**, e011541 (2020).
18. Liu, S. *et al.* Association between maternal chronic conditions and congenital heart defects: A population-based cohort study. *Circulation* **128**, 583–589 (2013).
19. Hoang, T. T., Marengo, L. K., Mitchell, L. E., Canfield, M. A. & Agopian, A. J. Original Findings and Updated Meta-Analysis for the Association between Maternal Diabetes and Risk for Congenital Heart Disease Phenotypes. *American Journal of Epidemiology* **186**, 118–128 (2017).

20. Mills, J. L., Troendle, J., Conley, M. R., Carter, T. & Druschel, C. M. Maternal obesity and congenital heart defects: A population-based study. *Am. J. Clin. Nutr.* **91**, 1543–1549 (2010).
21. Moore, L. L., Singer, M. R., Bradlee, M. L., Rothman, K. J. & Milunsky, A. A prospective study of the risk of congenital defects associated with maternal obesity and diabetes mellitus. *Epidemiology* **11**, 689–694 (2000).
22. Ohuchi, H. *et al.* High prevalence of abnormal glucose metabolism in young adult patients with complex congenital heart disease. *Am. Heart J.* **158**, 30–39 (2009).
23. Hobbs, C. A., Cleves, M. A., Zhao, W., Melnyk, S. & James, S. J. Congenital heart defects and maternal biomarkers of oxidative stress. *Am. J. Clin. Nutr.* **82**, 598–604 (2005).
24. Hrstka, S. C. L., Li, X., Nelson, T. J., Jeanne, L. & Olson, T. M. NOTCH1-Dependent Nitric Oxide Signaling Deficiency in Hypoplastic Left Heart Syndrome Revealed Through Patient-Specific Phenotypes Detected in Bioengineered Cardiogenesis. *Stem Cells* **35**, 1106–1119 (2017).
25. Peyvandi, S. *et al.* Risk of congenital heart disease in relatives of probands with conotruncal cardiac defects: an evaluation of 1,620 families. *Am. J. Med. Genet. A* **164A**, 1490–5 (2014).
26. Corone, P., Bonaiti, C., Feingold, J., Fromont, S. & Berthet-Bondet, D. Familial congenital heart disease: How are the various types related? *Am. J. Cardiol.* **51**, 942–945 (1983).
27. Øyen, N. *et al.* Recurrence of congenital heart defects in families. *Circulation* **120**, 295–301 (2009).
28. Ellesøe, S. G. *et al.* Familial co-occurrence of congenital heart defects follows distinct patterns. *Eur. Heart J.* **39**, 1015–1022 (2018).
29. Wang, X. *et al.* Influence of genes and the environment in familial congenital heart defects. *Mol. Med. Rep.* **9**, 695–700 (2014).
30. Øyen, N. *et al.* Recurrence of discordant congenital heart defects in families. *Circ. Cardiovasc. Genet.* **3**, 122–128 (2010).
31. Dina, C. *et al.* Genetic association analyses highlight biological pathways underlying mitral valve prolapsed. *Nat. Genet.* **47**, 1206–1211 (2015).
32. Garg, V. *et al.* Mutations in NOTCH1 cause aortic valve disease. *Nature* **437**, 270–4 (2005).
33. Shieh, J. T. C., Bittles, A. H. & Hudgins, L. Consanguinity and the risk of congenital heart disease. *American Journal of Medical Genetics, Part A* **158 A**, 1236–1241 (2012).
34. Soemedi, R. *et al.* Contribution of global rare copy-number variants to the risk of sporadic congenital heart disease. *Am. J. Hum. Genet.* **91**, 489–501 (2012).
35. Homsy, J. *et al.* De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. *Science* **350**, 1262–6 (2015).
36. Orr, B., Godek, K. M. & Compton, D. Aneuploidy. *Curr. Biol.* **25**, R538–R542 (2015).
37. Hartman, R. J. *et al.* The contribution of chromosomal abnormalities to congenital heart defects: A population-based study. *Pediatr. Cardiol.* **32**, 1147–1157 (2011).
38. Wimalasundera, R. C. & Gardiner, H. M. Congenital heart disease and aneuploidy. *Prenat. Diagn.* **24**, 1116–22 (2004).
39. Benhaourech, S., Drighil, A. & Hammiri, A. El. Congenital heart disease and

Down syndrome: various aspects of a confirmed association. *Cardiovasc. J. Afr.* **27**, 287–290 (2016).
40. Peterson, J. K., Kochilas, L. K., Catton, K. G., Moller, J. H. & Setty, S. P. Long-Term Outcomes of Children With Trisomy 13 and 18 After Congenital Heart Disease Interventions. *Ann. Thorac. Surg.* **103**, 1941–1949 (2017).
41. Silberbach, M. *et al.* Cardiovascular Health in Turner Syndrome: A Scientific Statement From the American Heart Association. *Circulation. Genomic and precision medicine* **11**, e000048 (2018).
42. Biró, O., Rigó, J. & Nagy, B. Noninvasive prenatal testing for congenital heart disease–cell-free nucleic acid and protein biomarkers in maternal blood. *Journal of Maternal-Fetal and Neonatal Medicine* **33**, 1044–1050 (2020).
43. Petersen, A. K. *et al.* Positive predictive value estimates for cell-free noninvasive prenatal screening from data of a large referral genetic diagnostic laboratory. *Am. J. Obstet. Gynecol.* **217**, 691.e1-691.e6 (2017).
44. Hu, H. *et al.* Noninvasive prenatal testing for chromosome aneuploidies and subchromosomal microdeletions/microduplications in a cohort of 8141 single pregnancies. *Hum. Genomics* **13**, 14 (2019).
45. Costain, G., Silversides, C. K. & Bassett, A. S. The importance of copy number variation in congenital heart disease. *NPJ genomic Med.* **1**, 16031 (2016).
46. Collins, R. L. *et al.* A structural variation reference for medical and population genetics. *Nature* **581**, 444–451 (2020).
47. Edwards, J. J. & Gelb, B. D. Genetics of congenital heart disease. *Current Opinion in Cardiology* **31**, 235–241 (2016).
48. Thienpont, B. *et al.* Submicroscopic chromosomal imbalances detected by array-CGH are a frequent cause of congenital heart defects in selected patients. *Eur. Heart J.* **28**, 2778–2784 (2007).
49. Silversides, C. K. *et al.* Rare Copy Number Variations in Adults with Tetralogy of Fallot Implicate Novel Risk Gene Pathways. *PLoS Genet.* **8**, (2012).
50. Hitz, M. P. *et al.* Rare Copy Number Variants Contribute to Congenital Left-Sided Heart Disease. *PLoS Genet.* **8**, e1002903 (2012).
51. Tomita-Mitchell, A., Maslen, C. L., Morris, C. D., Garg, V. & Goldmuntz, E. GATA4 sequence variants in patients with congenital heart disease. *J. Med. Genet.* **44**, 779–783 (2007).
52. Hinton, R. B. *et al.* Elastin haploinsufficiency results in progressive aortic valve malformation and latent valve disease in a mouse model. *Circ. Res.* **107**, 549–557 (2010).
53. Thomas Collins, R. Cardiovascular disease in Williams syndrome. *Current Opinion in Pediatrics* **30**, 609–615 (2018).
54. Ye, M. *et al.* Deletion of ETS-1, a gene in the Jacobsen syndrome critical region, causes ventricular septal defects and abnormal ventricular morphology in mice. *Hum. Mol. Genet.* **19**, 648–656 (2009).
55. Grossfeld, P. D. *et al.* The 11q terminal deletion disorder: A prospective study of 110 cases. *Am. J. Med. Genet.* **129A**, 51–61 (2004).
56. Hanchard, N. A. *et al.* Assessment of large copy number variants in patients with apparently isolated congenital left-sided cardiac lesions reveals clinically relevant genomic events. *Am. J. Med. Genet. A* **173**, 2176–2188 (2017).
57. Mlynarski, E. E. *et al.* Rare copy number variants and congenital heart defects in the 22q11.2 deletion syndrome. *Hum. Genet.* **135**, 273–285 (2016).
58. Jónsson, H. *et al.* Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* **549**, 519–522 (2017).

59. Al Turki, S. *et al.* Rare variants in NR2F2 cause congenital heart defects in humans. *Am. J. Hum. Genet.* **94**, 574–585 (2014).
60. Agopian, A. J. *et al.* Genome-Wide Association Studies and Meta-Analyses for Congenital Heart Defects. *Circ. Cardiovasc. Genet.* **10**, e001449 (2017).
61. Hanchard, N. A. *et al.* A genome-wide association study of congenital cardiovascular left-sided lesions shows association with a locus on chromosome 20. *Hum. Mol. Genet.* **25**, 2331–2341 (2016).
62. Lin, Y. *et al.* Association analysis identifies new risk loci for congenital heart disease in Chinese populations. *Nat. Commun.* **6**, 8082 (2015).
63. Lahm, H. *et al.* Congenital heart disease risk loci identified by genome-wide association study in European patients. *J. Clin. Invest.* **131**, (2021).
64. Blue, G. M. *et al.* Targeted next-generation sequencing identifies pathogenic variants in familial congenital heart disease. *J. Am. Coll. Cardiol.* **64**, 2498–2506 (2014).
65. Lahaye, S. *et al.* Utilization of Whole Exome Sequencing to Identify Causative Mutations in Familial Congenital Heart Disease. *Circ. Cardiovasc. Genet.* **9**, 320–329 (2016).
66. Povysil, G. *et al.* Rare-variant collapsing analyses for complex traits: guidelines and applications. *Nat. Rev. Genet.* **20**, 747–759 (2019).
67. Turro, E. *et al.* Whole-genome sequencing of patients with rare diseases in a national health system. *Nature* **583**, 96–102 (2020).
68. Short, P. J. *et al.* De novo mutations in regulatory elements in neurodevelopmental disorders. *Nature* **555**, 611–616 (2018).
69. Ferraro Petrillo, U., Sorella, M., Cattaneo, G., Giancarlo, R. & Rombo, S. E. Analyzing big datasets of genomic sequences: Fast and scalable collection of k-mer statistics. *BMC Bioinformatics* **20**, 138 (2019).
70. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* 201178 (2017). doi:10.1101/201178
71. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–9 (2009).
72. McKenna, A. *et al.* The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
73. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
74. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
75. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, (2010).
76. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin).* **6**, 80–92 (2012).
77. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).
78. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).
79. Eilbeck, K., Quinlan, A. & Yandell, M. Settling the score: variant prioritization and Mendelian disease. *Nat. Rev. Genet.* **18**, 599–612 (2017).

80. Li, M. M. *et al.* Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer: A Joint Consensus Recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. *J. Mol. Diagn.* **19**, 4–23 (2017).
81. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature Methods* **7**, 248–249 (2010).
82. Sim, N. L. *et al.* SIFT web server: Predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* **40**, (2012).
83. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).
84. Samocha, K. E. *et al.* Regional missense constraint improves variant deleteriousness prediction. *bioRxiv* 148353 (2017). doi:10.1101/148353
85. Qi, H. *et al.* MVP predicts the pathogenicity of missense variants by deep learning. *Nat. Commun.* **12**, 510 (2021).
86. Havrilla, J. M., Pedersen, B. S., Layer, R. M. & Quinlan, A. R. A map of constrained coding regions in the human genome. *Nat. Genet.* **51**, 88–95 (2019).
87. Langmead, B. & Nellore, A. Cloud computing for genomic data analysis and collaboration. *Nat. Rev. Genet.* **19**, 208–219 (2018).
88. Stephens, Z. D. *et al.* Big data: Astronomical or genomical? *PLoS Biol.* **13**, (2015).
89. Zaharia, M. *et al.* Apache spark: A unified engine for big data processing. *Commun. ACM* **59**, 56–65 (2016).
90. Farhan, S. M. K. *et al.* Exome sequencing in amyotrophic lateral sclerosis implicates a novel gene, DNAJC7, encoding a heat-shock protein. *Nat. Neurosci.* **22**, 1966–1974 (2019).
91. Feng, Y. C. A. *et al.* Ultra-Rare Genetic Variation in the Epilepsies: A Whole-Exome Sequencing Study of 17,606 Individuals. *Am. J. Hum. Genet.* **105**, 267–282 (2019).
92. De Lillo, A. *et al.* Phenome-wide association study of TTR and RBP4 genes in 361,194 individuals reveals novel insights in the genetics of hereditary and wildtype transthyretin amyloidoses. *Hum. Genet.* **138**, 1331–1340 (2019).
93. van Walree, E. S. *et al.* Germline variants in HEY2 functional domains lead to congenital heart defects and thoracic aortic aneurysms. *Genet. Med.* **23**, 103–110 (2021).

Chapter II. Accurate and fast feature selection workflow for high-dimensional omics data

The original peer-reviewed publication presented in this chapter (pages 33-52) is publicly available at *https://doi.org/10.1371/journal.pone.0189875*.

PLOS ONE

RESEARCH ARTICLE

Accurate and fast feature selection workflow for high-dimensional omics data

Yasset Perez-Riverol[1]*, Max Kuhn[2], Juan Antonio Vizcaíno[1], Marc-Phillip Hitz[3,4,5,6], Enrique Audain[3,4,5]*

1 European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom, **2** RStudio Inc., Boston, MA, United States of America, **3** Department of Congenital Heart Disease and Pediatric Cardiology, Universitätsklinikum Schleswig–Holstein Kiel, Kiel, Germany, **4** German Center for Cardiovascular Research (DZHK), Berlin, Germany, **5** Department of Human Genetics, University Medical Center Schleswig-Holstein (UKSH), Kiel, Germany, **6** Wellcome Trust Sanger Institute, Cambridge, United Kingdom

* yperez@ebi.ac.uk (YPR); enrique.audain@uksh.de (EA)

Citation: Perez-Riverol Y, Kuhn M, Vizcaíno JA, Hitz M-P, Audain E (2017) Accurate and fast feature selection workflow for high-dimensional omics data. PLoS ONE 12(12): e0189875. https://doi.org/10.1371/journal.pone.0189875

Editor: Quan Zou, Tianjin University, CHINA

Received: June 27, 2017

Accepted: December 4, 2017

Published: December 20, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: YP-R is supported by BBSRC 'PROCESS' grant (BB/K01997X/1). JAV acknowledges the Wellcome Trust (grant number WT101477MA) and EMBL core funding. EA and MPH are supported by DZHK (German Center for Cardiovascular Research), partner sites: Kiel, Germany. The funder provided support in the form of salaries for authors [MK], but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The

Abstract

We are moving into the age of 'Big Data' in biomedical research and bioinformatics. This trend could be encapsulated in this simple formula: $D = S * F$, where the volume of data generated (D) increases in both dimensions: the number of samples (S) and the number of sample features (F). Frequently, a typical omics classification includes redundant and irrelevant features (e.g. genes or proteins) that can result in long computation times; decrease of the model performance and the selection of suboptimal features (genes and proteins) after the classification/regression step. Multiple algorithms and reviews has been published to describe all the existing methods for feature selection, their strengths and weakness. However, the selection of the correct FS algorithm and strategy constitutes an enormous challenge. Despite the number and diversity of algorithms available, the proper choice of an approach for facing a specific problem often falls in a 'grey zone'. In this study, we select a subset of FS methods to develop an efficient workflow and an R package for bioinformatics machine learning problems. We cover relevant issues concerning FS, ranging from domain's problems to algorithm solutions and computational tools. Finally, we use seven different proteomics and gene expression datasets to evaluate the workflow and guide the FS process.

Introduction

The term 'Big Data' is often used to describe the huge volumes of information produced by modern systems such as mobile devices, tracking tools and sensors [1, 2]. In biomedical research, the growth of high-throughput (omics) technologies has resulted in an exponential growth in the dimensionality and sample size. This increase has two major directions: i) the number of samples processed, powered by novels machines (i.e. sequencers and mass spectrometers); and ii) the features, attributes and variables collected alongside each sample [3]. This high-dimensional environment becomes a challenge to many modelling tasks used in bioinformatics, ranging from sequence analysis to spectral analyses as well as literature mining.

specific roles of these authors are articulated in the 'author contributions' section.

Competing interests: The authors have declared that no competing interests exist. Even though one of the authors (MK) is affiliated to a commercial company (RStudio Inc.), this does not alter our adherence to PLOS ONE policies on sharing data and materials.

Abbreviations: CM, Correlation Matrix; FS, Feature Selection; ML, Machine Learning; PCA, Principal Component Analysis; RFE, Recursive Feature Elimination; RMSE, Root Mean Square Error; RF, Random Forest; SVM, Support Vector Machine; TNBC, Triple-Negative Breast Cancer; X2, Univariate Correlation.

Reducing data complexity is therefore crucial for data analysis tasks, knowledge inference using machine learning (ML) algorithms, and data visualization [4–6].

The 'curse of dimensionality' (term first introduced by Bellman in 1957) [7] described the problem caused by the exponential increase in volume associated with adding extra dimensions to an Euclidean space. In this context, the typical bioinformatics problem involves both: relevant and redundant features. Therefore, a Feature Selection (FS) approach becomes a crucial and non-trivial task because: i) it provides a deeper insight into the underlying processes that are the foundation of the data; ii) it improves the performance (CPU-time and memory) of the ML step, by reducing the number of variables; and iii) it produces better model results avoiding overfitting. However, a FS algorithm brings an important decision in any ML workflow (e.g. classification of protein/gene expression profiles): are there redundant features (e.g. proteins or genes) in the dataset that are irrelevant and/or redundant for the biological study?

The most-common attempt to address the FS problem (the so-called univariate filtering approach) is to use a variable ranking method to filter out the least promising variables before using a multivariate method [8]. These methods have been used extensively in computational biology for cancer classification using microarray data [9, 10]. However, correlation filters could prompt some loss of relevant features that are meaningless by themselves but that can be useful in combination. To overcome this effect, a set of algorithms has been proposed to combine the original variables into a new and smaller subset of features, such as Principal Component Analysis (PCA) and Linear Discriminant Analysis. In PCA [11], new orthogonal features (latent variables or principal components) are obtained by maximizing the variation of the original features. The number of the latent features (factors) can be much lower than the number of original features, so that the data can be visualized in a much lower-dimensional space. As correlation filters, PCA methods can reduce the number of variables by looking into the feature dependencies without taking into account the final learning model. In 1997, a powerful strategy emerged that combines a FS algorithm with a learning/classification step: the so-called *wrapper* methods [12]. These wrapper approaches (e.g. *forward selection* and *backward elimination*) can use the prediction performance of a given ML approach to assess the relative usefulness of different subsets of variables. An exhaustive search can be performed if the number of variables is not too large.

Due to the diversity of FS methods available, it is hard to choose the correct approach needed to accomplish a specific task beforehand (e.g. regression or classification). In 2007, Saeys and co-workers published an introduction to FS in bioinformatics [3]. Also, several reviews have focused on the application in computational biology of particular methods such as PCA [13, 14] or Support Vector Machines (SVM) [15]. However, most of this work has been done to describe current methods in isolation and not to evaluate how they could be combined. In this manuscript, we developed a FS workflow and an R package for high-dimensional omics data analysis. The workflow combined univariate/multivariate correlation filters with wrapper feature backward elimination and it was applied to regression and classification problems. We benchmarked the individual steps of the described workflow, highlighting the optimal steps in different scenarios, using seven different omics datasets. Finally, we discuss major challenges when applying the described workflow to classification problems of high-dimensional omics data.

Materials and methods

Transcriptomic dataset of breast tumor samples (Dataset 1)

We first used a gene expression dataset (GEO (Gene Expression Omnibus) accession number: GSE5325) from *Saal et al.* [16], which has already been extensively studied before [13]. The

authors performed a study using microarrays to measure the expression of 27,648 genes in 105 breast tumor samples. The dataset includes the estrogen receptor alpha status (0 = negative, 1 = positive), a transcription factor recognized as being important for stimulating the growth of a large proportion of breast cancers and used to explore co-expression [17].

High-resolution isoelectric focusing proteomics dataset (Dataset 2)

The second dataset is the result of an electrophoresis experiment on peptide samples [18]. A total of 7,391 peptides were identified in 12 fractions, where each fraction corresponded to an experimental isoelectric point. This dataset has been used before to develop a ML model that can accurately predict the theoretical isoelectric points for peptides and proteins based on the amino acid sequence properties [5, 19].

Triple-Negative Breast Cancer (TNBC) dataset (Dataset 3)

A third dataset containing protein quantification data using a label free technique was included [20]. The dataset assembles a panel of 44 (including samples and technical replicates) human breast cell lines and clinical tumors for analyzing the proteomics landscape of TNBC. The studied cell lines cover mesenchymal-, luminal-, and basal-like subtypes, as well as three receptor-positive and one non-tumorigenic cell lines. Thus, the idea behind including this dataset was to evaluate the ability of the proposed FS workflow to classify subtypes of cellular lines.

Transcriptomics analysis of left ventricles of mouse hearts (Dataset 4)

A fourth dataset included the results of a transcriptomics analysis of left ventricles of mouse hearts subjected to an isoproterenol challenge [21]. In the study, the authors utilized expression arrays from left ventricular (LV) tissues, with and without an isoproterenol treatment, to understand the genetic control of gene expression and its relationship with heart failure. Then, the issue arising here suggests a binary classification problem where the researcher could be interested in, in order to know the optimal feature subset which could best discriminate between both classes (treated and non-treated samples).

Expression data from normal and prostate tumor tissues (Datasets 5, 6, and 7)

Recently, Li *et al*. have used several gene expression datasets to benchmark different FS algorithms [22]. From the original microarray datasets, we have selected three of those datasets (GEO accession number: GSE6919), to compare the FS workflow with the results obtained by *Li et al*. **Note 1** (**S1 File**) summarizes the main characteristics of the datasets described previously.

Workflow R-package

An R-package has been developed to reproduce the proposed workflow (https://github.com/enriquea/feseR). For its development five main R packages were used: i) **Caret** [23] (**C**lassification **A**nd **RE**gression **T**raining) (http://topepo.github.io/caret), containing a set of functions that attempt to streamline the process for creating predictive models; ii) **randomForest** [24], a package enabling Random Forest analysis (https://cran.r-project.org/web/packages/randomForest/); iii) **prcomp**, a native function included in the R package *stats*; iv) ***Kernlab*** [25] (https://cran.r-project.org/web/packages/kernlab/), which provides the user with basic kernel functionality (e.g., computing a kernel matrix), along with some utility functions,

commonly used in kernel-based methods; and v) the **FSelector** package [26] (https://cran.r-project.org/web/packages/FSelector/), which offers algorithms for filtering attributes (e.g. chi-squared, information gain, and linear correlation).

We have used the current FS workflow and R-package in combination for two different ML (regression/classification) problems. Six of the datasets represent classification of (protein/gene) expression profiles and the last one a regression problem for the accurate estimation of the isoelectric point of peptides and proteins. In the following sections, we discuss the results of combining the different steps of the FS workflow depending of the ML problem.

Results and discussion

A good feature subset can be defined as one that contains features highly correlated with (predictive of) outcome, yet uncorrelated (independent) with (not predictive of) each other. Nevertheless, the existing diversity of FS methods makes it challenging to choose the correct one for the task at hand (**S1 File, Note 2**). Fig 1 represents the proposed overall workflow to perform FS in high-dimensional omics big data. First, a univariate correlation filter can be used before applying any wrapper approach, to determine the relation between each feature and the class or predicted variable. Then, a second filtering step (Correlation Matrix (CM) or PCA), can follow, in order to determine the dependencies between the different dataset features. Finally, backward elimination is achieved by wrapping a ML method, such as Random Forest and SVM around each example.

Removing irrelevant features: Univariate correlation filtering

The univariate correlation filtering step removes all features that are not directly related to their class variables. When we applied this approach to **Dataset 1** it removed those genes with a non-correlated expression to the presence or absence of estrogen receptor alpha, reducing the number of genes from 8,534 (only those genes showing expression in all samples were considered) to 1,697. In **Dataset 2**, we used the univariate filter to remove features (amino acid properties) unrelated with the isoelectric point. Fig 2A shows the high-correlation found among the original 545 physicochemical peptides properties considered for the 7,391 peptides. We implemented a univariate correlation filter to remove all features that were not correlated with the isoelectric point (correlation coefficient $< = 0.30$), reducing the number of variables to 89 features. When we extended the analysis to the remaining benchmarking datasets, we observed that, in general, univariate correlation filtering removed more than 80% of the original features that were not related to the predicted variable. As previously discussed by other

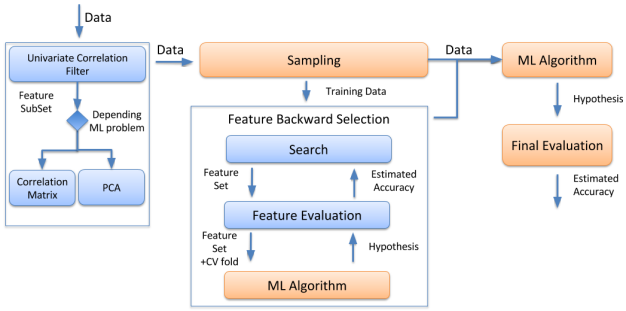


Fig 1. Proposed workflow for FS including a filtering step with univariate and/or multivariate approaches, followed by a wrapper approach (recursive feature elimination).

https://doi.org/10.1371/journal.pone.0189875.g001

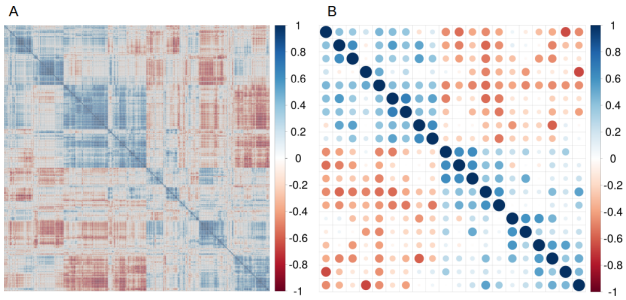


Fig 2. (A) Correlation matrix for the 544 physicochemical (features) of the 7,391 peptides (samples) included in Dataset 2; (B) the final 20 variables after the correlation-matrix filtering steps.

https://doi.org/10.1371/journal.pone.0189875.g002

authors [8], univariate correlation filtering should be always applied at early stages of any classification and/or regression process. However, univariate correlation filtering can only be used to study the relationship of each feature with a class variable, but cannot be applied to find the relationships among them. For this reason, a multivariate step (e.g. correlation matrix) was used (Fig 1) to remove the redundancy among highly-correlated features (correlation coefficient $> = 0.75$).

Reducing feature complexity: CM or PCA

We implemented two different strategies (depending on the classification or regression problem) to reduce the number of variables, while keeping most of the original and relevant information: CM and PCA. **Dataset 2** is a good example of a dataset containing regression related problems. In this particular case, the aim was to predict more accurately the isoelectric point of peptides and proteins, using other physicochemical features of the peptides. Therefore, the final model should be based on, or be correlated to, the original features (because they would be used in the future to make a predictor that could be applied for other datasets). One of the simplest and most powerful filtering approaches to remove feature redundancy, while keeping original features, is the use of a CM filter. For example, peptides properties such as aromatic rings, bond and carbon atom counts are strongly correlated [5, 27]. Therefore, any of these variables could be used as a proxy for all the others. It should be noted that several features clustered together, suggesting a high-redundancy in the feature set. By applying the CM filter, it is possible to remove those that are redundant (or irrelevant) and to keep only a reduced feature set for subsequent analysis steps. The present workflow keeps only 20 variables (out of the original 545 features, see Fig 2B) for the final ML step (Fig 1). The current approach also reuses the final model in new datasets because the filtering steps preserve the original variables by only removing the redundant ones.

Opposite to **Dataset 2**, the other datasets constitute good examples of classification related problems. In addition to the CM filter approach, we implemented and studied the use of Principal Component Analysis (PCA) as a multivariate filter to reduce the number of features. PCA reduces the dimensionality of the data while retaining most of the variation in the predictor variables [13]. Thus, by using a few components, PCA can represent each sample by using relatively few (new) variables instead of (potentially) thousands of them. Fig 3 shows the PCA performed in **Dataset 1**. The proportion of the variation present in all genes is encompassed within each of the principal components, with the first few components representing most of it (Fig 3A). The cumulative variance analysis shows that most of the variance is contained in the first 30 principal components (75%), where only 76 components reach a 95% of variance

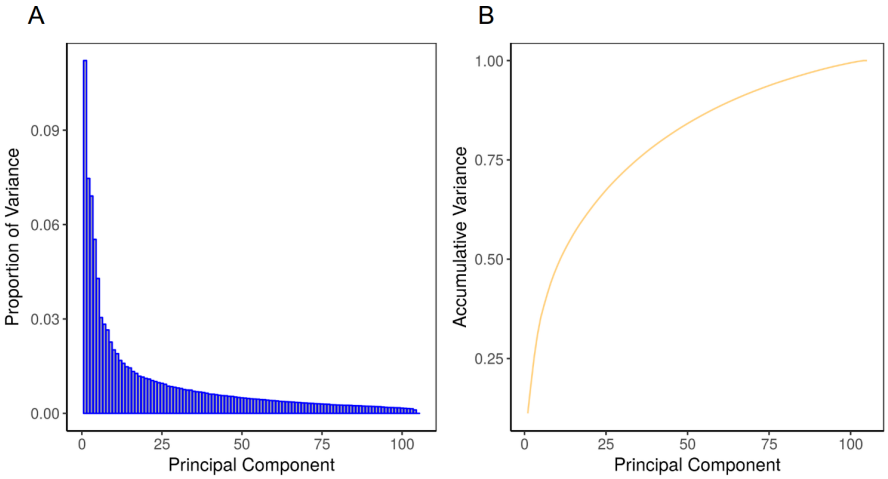


Fig 3. (A) Proportion of variance and (B) cumulative variance of principal components for the analysis of Dataset 1.

https://doi.org/10.1371/journal.pone.0189875.g003

(Fig 3B) and 104 components are enough to retain all the original variance. This number of variables is 10-fold smaller than the original 1,697 features obtained after applying the univariate correlation filter.

When the number of variables is larger than the number of samples, PCA can reduce the dimensionality of the samples to, at most, the number of samples, without losing information [13, 28]. We obtained the same results when PCA was applied to the other relevant datasets (**Dataset 3** to **Dataset 7**, those with a classification problem, S2 File). However, since the principal components are linear combinations of the original data, it is not obvious how model parameter estimates can relate back to the original variables. Thus, this method is not suitable for problems where it is required to keep the primary information (e.g. in the case of regression problems, **Dataset 2**).

Optimizing the feature selection: Wrapper recursive feature elimination

All filtering FS approaches previously shown (e.g. correlation-based or PCA) are relatively easy to implement and computationally fast. Therefore, these algorithms represent a suitable choice in the first stage of any given FS pipeline. However, wrapper methods should be used in the last steps to find the "optimal" feature subsets, by iteratively selecting features based on classifier performance (Fig 1). The wrapper methods should be combined with cross-validation steps to improve the final results [12, 29]. These cross-validation steps can be used to assess the results of the learning analysis (e.g. regression or classification) and help to generalize these steps to an independent dataset. The goal of cross-validation is to define a dataset to "test" the model in the training phase (i.e., the validation dataset), in order to limit problems like overfitting [29]. In the proposed workflow, we used a recursive feature elimination (backward elimination) approach in combination with two ML models (Random Forest and SVM) to systematically increase each ML step. The number of cross-validation iterations should be evaluated in detail because it could significantly increase the running time without improving the performance of the model prediction.

We implemented the wrapper backward elimination step in combination with the SVM radial kernel, in order to predict the isoelectric point using **Dataset 2**. Table 1 shows the performance (regarding running time and model prediction accuracy) of the feature workflow for

Table 1. Benchmark of the SVM regression model for Dataset 2 applying different FS methods (SVM), no feature selection, (X2) univariate correlation alone, (CM) correlation matrix filtering, (RFE) and wrapper feature elimination. The figures indicated using the prefixes CV3, CV7 and CV10 correspond to the number of interactions in the cross-validation steps during the RFE feature selection.

	R^2	RMSE	Time (min)	# Features
SVM	0.97	0.88	6.8	545
X2-CM-SVM	0.98	0.57	0.5	28
RFE-SVM-CV3	0.98	0.32	35	4
RFE-SVM-CV7	0.98	0.32	115	4
RFE-SVM-CV10	0.98	0.32	168	4
X2-CM-RFE-SVM-CV3	0.98	0.33	11	2
X2-CM-RFE-SVM-CV7	0.98	0.34	36	2
X2-CM-RFE-SVM-CV10	0.98	0.34	48.1	2

https://doi.org/10.1371/journal.pone.0189875.t001

Dataset 2. We benchmarked all the FS combinations with the SVM model by removing each of them. Applying the SVM model alone (**SVM**) without FS or cross-validation helps to predict the isoelectric point with a high root-mean-square error (RMSE) of 0.88. In contrast, when both correlation filters (**X2-CM-SVM**) were applied, RMSE and running time decreased to 0.57 and 0.50 min, respectively. When the complete workflow (**X2-CM-RFE-SVM-CV3**) was used RMSE decreased to 0.33 (Table 1). It should be noted that when pre-filtering was applied (**RFE-SVM-CV3**), RMSE decreased to 0.32 and two new variables were added to the SVM model. However, this improvement in performance (e.g. low RMSE) decreased the overall efficiency of the workflow by increasing the execution time three-fold. Also, we observed no changes where the number of cross-validation steps was increased (Table 1).

Wrapper backward elimination step provided a powerful method to optimize the final subset of variables in response to the regression SVM model. Fig 4 shows the final results of the isoelectric point prediction (**Dataset 2**) for all FS combinations. Backward selection in combination with the cross-validation step enables a better estimation of the variable prediction (isoelectric point) in the regions where less experimental evidences exist (basic pH range). This workflow has been used in a recent approach to predict the isoelectric point and it has proven to predict the isoelectric point more accurately than any other algorithm so far. A similar implementation was applied to the remaining datasets (1, 3–7) where a Random Forest model was wrapped around, using a recursive approach to evaluate the performance and the variable weight following different FS workflows. We first evaluated the Random Forest approach for FS without any filtering and parameter tuning as discussed before by *Díaz-Uriarte et al.* [30]. In addition, four recursive feature elimination methods, wrapped with Random Forest, were combined as follows: **RFE-RF** without any pre-filtering step (i.e. other FS methods), PCA combined with **RFE-RF**, univariate correlation filtering (**X2**) combined with **RFE-RF,** and finally, all methods were used sequentially: **X2-PCA-RFE-RF** or **X2-CM-RFE-RF**.

Fig 5 shows the performance evaluation (for the expression datasets 1, 3–7) of each complete FS combination (**X2-PCA-RFE-RF** and **X2-CM-RFE-RF**) and the random forest classification without FS step. We use the approach previously reported by *Pochet et al.* [31], where 20-fold randomized test data were used to summarize the accuracy in the prediction (see detailed description in S2 File). Also, we kept a 10-fold internal cross-validation step in all implementations of recursive feature selection trials. The results shown that when any of the full FS approaches are applied the average accuracy is higher compare with the results when not FS is used (red box plots). Only, in **Dataset 3** the workflow using PCA is less efficient than the random forest without FS step which can be related with the low number of samples analyzed (44). Importantly, even when RF perform very well it retains all the original features on each making difficult to decided which features are more relevant for the classification (S1

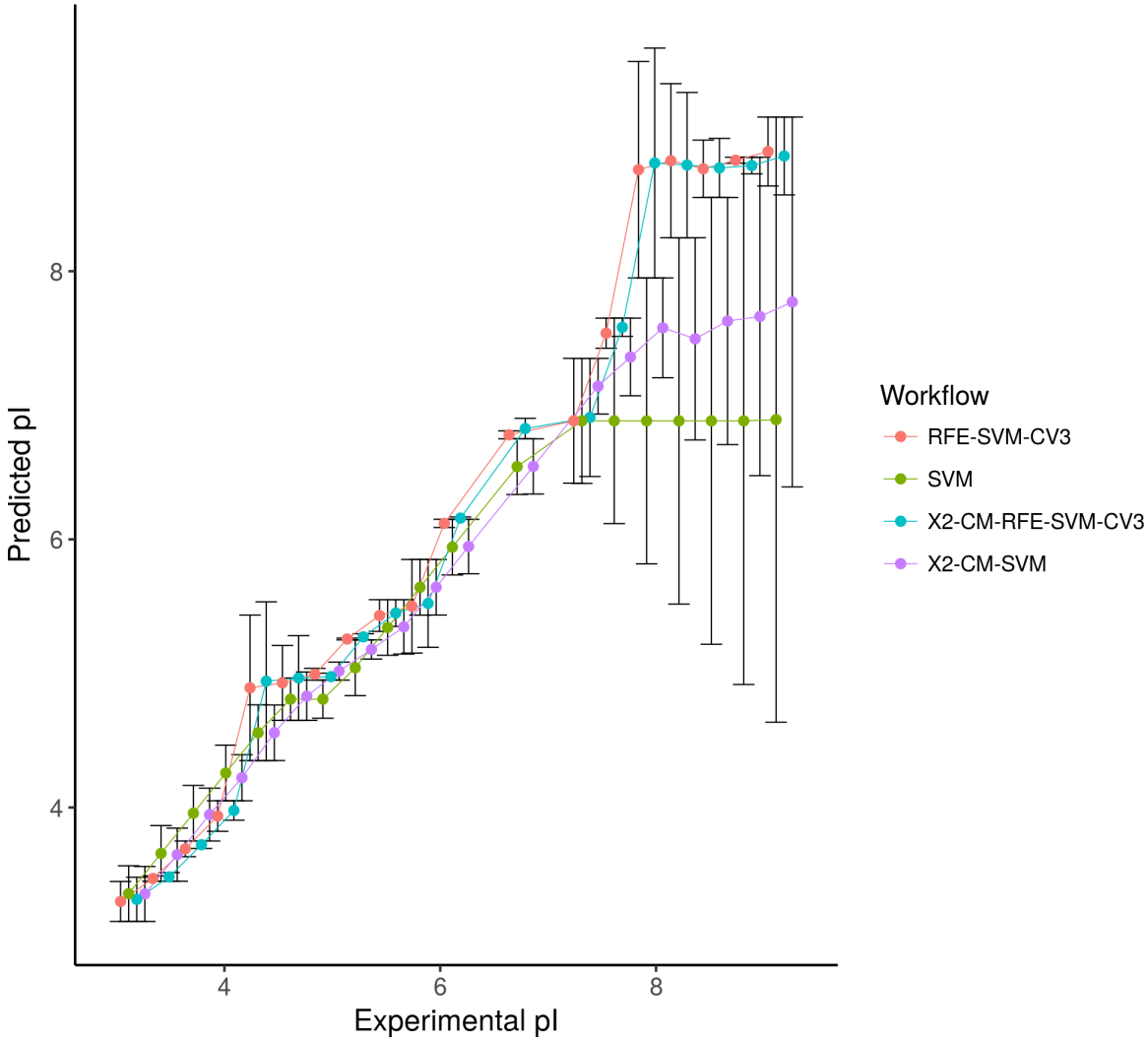


Fig 4. Error plot of predicted isoelectric point vs the experimental isoelectric point (Dataset 2): (SVM) applying FS or cross-correlation step; (X2-CM-SVM) adding correlation filters as the only steps for feature selection; (RFE-SVM-CV3) recursive feature elimination, three interactions of cross-validation combined with SVM; (X2-CM-RFE-SVM-CV3) considering the full FS workflow.

https://doi.org/10.1371/journal.pone.0189875.g004

File, Table 2). Both FS workflows reduce the number of variables in all cases in more than 90% (S1 File, Table 2), with average accuracy always above 70% (Fig 5). Because both workflow shows similar performance and some users may want to select PCA (less variables) or CM (original features), the R-package allows to define which multivariate option use during the FS.

Table 2 summarizes the benchmark metrics (accuracy, standard deviation, number of final features and time) for each evaluated FS workflow (in **Dataset 1**). While all methods kept the accuracy in the range 83–88%, when all methods were combined (proposed workflow) a lower standard deviation was obtained. Using a Random Forest model without FS, the classification process was faster than in the case of any other combination, keeping all the relevant features (1,969 of them). Including PCA and Recursive Feature Elimination (**PCA-RFE-RF**), we observed a strong feature reduction (7–10 components) and a better standard deviation (5.4).

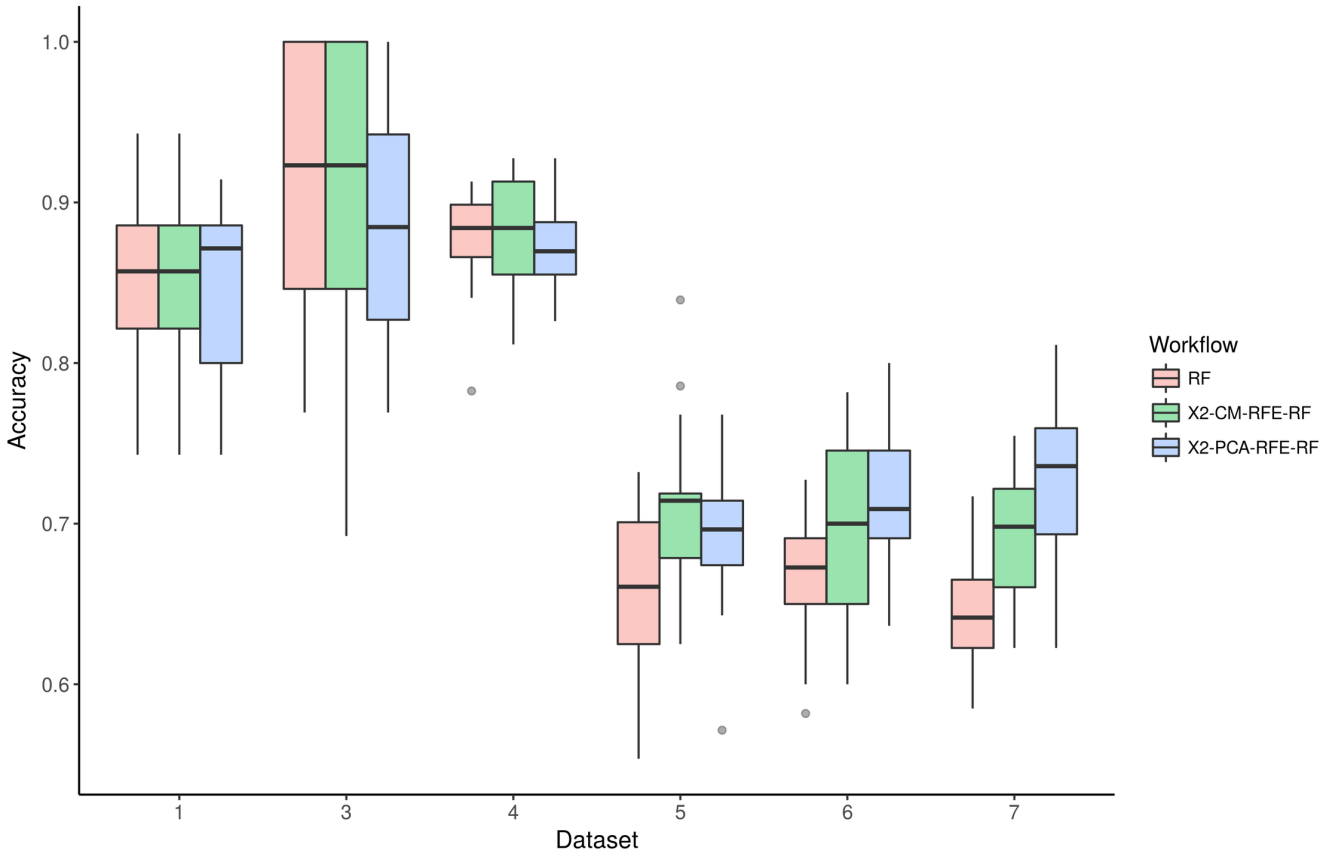


Fig 5. Accuracy vs. feature selection combination for expression datasets (1, 3, 4, 5, 6 and 7). (**RF**) Random Forest without previous feature selection step; (**X2-CM-RFE-RF**), random forest classification after the feature selection step using univariate correlation filter with matrix correlation and recursive feature elimination; (**X2-PCA-RFE-RF**), random forest classification after the feature selection step using univariate correlation filter with principal component analysis and recursive feature elimination. All methods include an internal cross-validation 10-fold step. All accuracy metrics were estimated following the approach previously reported by *Pochet et al.* [31], where 20-fold randomized test data were used to summarize the accuracy of the FS combination.

https://doi.org/10.1371/journal.pone.0189875.g005

Selecting a univariate correlation filter (**X2-RFE-RF**), a lowest standard deviation was obtained (3.6).

Fig 6 visualizes the results of the Random Forest classification algorithm without (Fig 6A, Fig 6C and Fig 6E) and with (Fig 6B, Fig 6D and Fig 6F) a FS step; for Datasets 1, 3, and 4,

Table 2. Benchmarking of the random forest model (classification) for Dataset 1, when different FS methods are applied: (RF) random forest only, (RFE) wrapper recursive feature elimination with 10-times internal cross-validation, (PCA) principal component analysis, (X2) univariate correlation filtering or (CM) correlation matrix filter. Each method is applied 20 times with randomized and class-balanced training datasets. The accuracy values provided correspond to the average value.

	Accuracy (%)	SD	Time (min)	# features
RF	83.46	8.1	1.46	1969
RFE-RF	84.61	6.3	15.83	30
PCA-RFE-RF	83.43	5.4	3.12	10
X2-RFE-RF	87.04	3.6	4.92	25
X2-PCA-RFE-RF	88.21	4.5	3.51	8
X2-CM-RFE-RF	85.01	5.7	6.35	8

https://doi.org/10.1371/journal.pone.0189875.t002

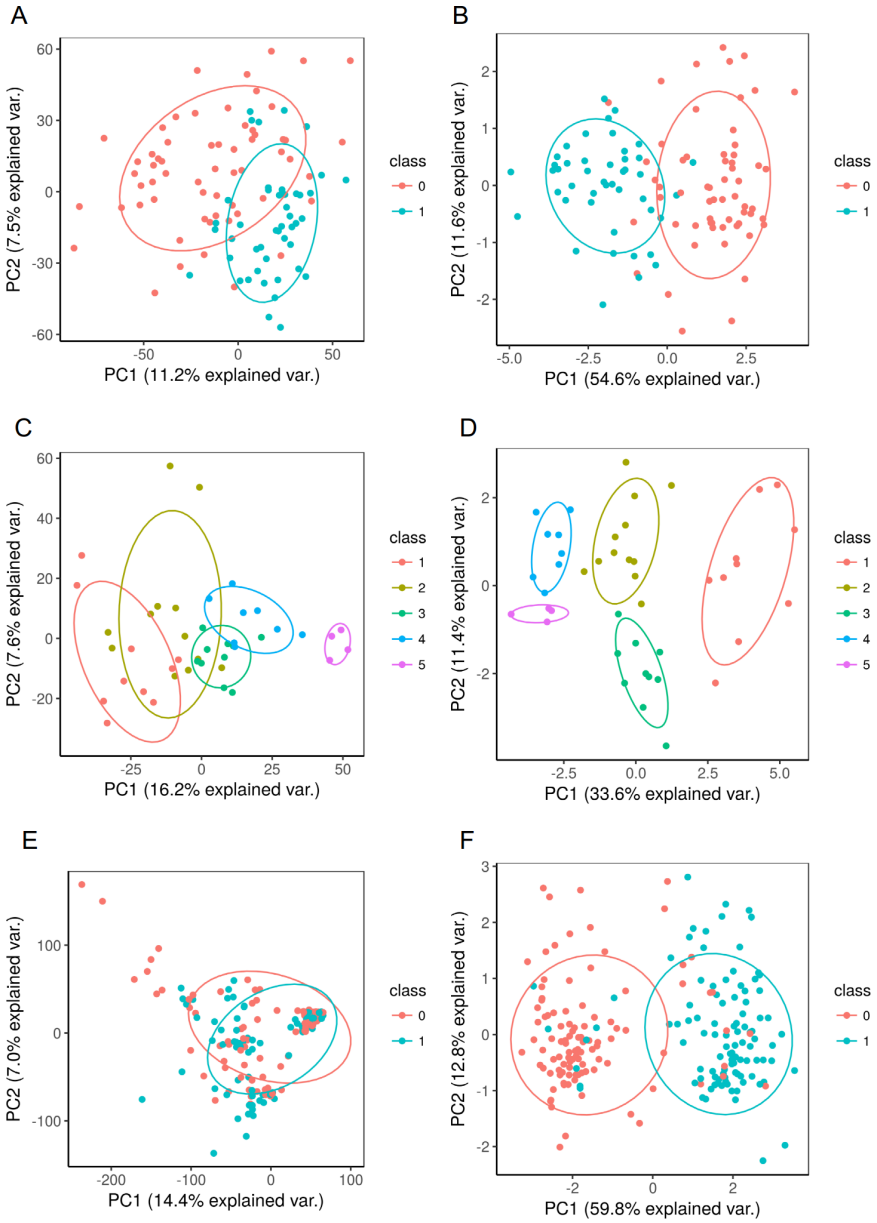


Fig 6. Visualization of the classification process using the first two principal components (PC1 and PC2) from the original data before (A, C, E) and after (B, D, F), to apply the following FS workflow: Univariate correlation (X2) with correlation matrix filter (CM) follow by Recursive Feature Elimination (RFE) wrapped with random forest (RF). The figure shows the classes distribution for **Dataset 1** (A, B), **Dataset 3** (C, D) and **Dataset 4** (E, F).

https://doi.org/10.1371/journal.pone.0189875.g006

respectively. The results show that the remaining features obtained allow to 'discriminate' between the different samples classes or groups (see detailed description in S2 File). It can be concluded that for those classification problems where the original features are needed, the PCA step could be removed without sacrificing general performance (accuracy, standard deviation, or CPU time). In contrast, univariate correlation filtering FS steps had a key impact on

Table 3. Performance comparison between the proposed approach (X2-PCA-RFE-RF) and the method reported by Li *et al.* [22]. The computer used in the original manuscript was an Intel(R) Core(TM) i5-4690 @ 3.5 GHz CPU, with 16 GB of RAM. In this study, we used an Intel(R) Core(TM) i5-4200 @ 2.5 GHz CPU, with 16 GB of RAM.

Dataset	Method	Accuracy	Variables	Runtime (min)
GSE6919/GPL8300	Current Workflow	0.77	35	8.50
	Li *et al.*	0.72	92	74.30
GSE6919/GPL92	Current Workflow	0.80	5	9.11
	Li *et al.*	0.73	174	71.50
GSE6919/GPL93	Current Workflow	0.81	6	12.00
	Li *et al.*	0.71	121	68.60

https://doi.org/10.1371/journal.pone.0189875.t003

the final results of the Random Forest model by increasing the performance in all the studied combinations. As we pointed out earlier, PCA 'obfuscates' the primary information, and thus, can potentially result in problems. When it is desirable to keep the "initial nature" of the variables, filtering methods (e.g. univariate correlation filter) exhibit a good performance (**Tables 1 and 2**) with a considerable lower number of features.

Summary of the benchmarking process

We have demonstrated the impact of the FS workflow in the classification and/or regression results as well as in the performance of the ML algorithm (CPU time and memory). Finally, we applied the same FS workflow to gene expression data from normal and prostate tumor tissues (**Datasets 5, 6 and 7**), and compared them with the results obtained by *Li et al.* [22], who used a similar approach on the same datasets (see Table 9 in [22]). Even though we observe a slight improvement in the classification accuracy in these three datasets (**Table 3**), the most notable differences were found in the number of features obtained by the final models and in the total runtime, using a similar computational platform. Thus, the results from the comparison reinforce our previous observations and validate the effectiveness of the FS workflow proposed in this manuscript. Another comparison was performed using the recently published tool based on maximum relevance–maximum distance (MRMD, http://lab.malab.cn/soft/MRMD/index_en.html) by *Zou et. al.* [32] (**Table 3**, **S1 File**). In general, we observed that both methods were comparable regarding the accuracy of the classification. However, some notable differences arose considering the number of the optimal (final) variables and the runtime. The proposed FS workflow performed better than MRMD for the analyzed datasets, by selecting in all cases less than 10% of the variables, at more than 80% reduction of the compute time.

Conclusions

FS selection algorithms are playing a major role to select correct variables for different classification and regression problems. Nevertheless, choosing the appropriate algorithm (or combination of algorithms) is not a trivial task. Different studies have highlighted methods to perform FS, but unfortunately, a thorough comparison including proper benchmarking is still lacking. Another major challenge remains: how to efficiently combine different FS methods to improve the final results. The developed FS workflow shown in this manuscript combines major strengths of univariate filtering methods, with CM and PCA strategies, as well as recursive feature elimination in two well-known learning problems: classification and regression. When univariate filtering was used in both types of problems the number of features was reduced by 80% without compromising the accuracy of the final model, and decreasing the CPU time of the learning model steps. The introduction of a wrapper method (recursive feature elimination) in combination with the learning model improved the accuracy in both

cases. If the wrapper method is applied without a previous filtering step, the CPU-time becomes too high. Finally, we demonstrated that the use of an intermediate FS step to remove redundancy between variables and features can significantly increase the accuracy of the learning model. This can be achieved by transforming the original variables into new components (retaining most of the variability in the original values) using PCA or by removing redundant highly correlated variables.

Large efforts have taken place in recent years to adopt individual FS methods. However, in our opinion, a multiple FS step workflow offers more promising results. Future developments should focus on other fields where the number of samples is growing considerably (e.g. clinical genomics, text and literature mining), and on the combination of heterogeneous datasets from different sources.

Supporting information

S1 File. Supplementary Information 1.
(DOCX)

S2 File. Supplementary Information 2.
(PDF)

Author Contributions

Conceptualization: Yasset Perez-Riverol.

Formal analysis: Yasset Perez-Riverol, Enrique Audain.

Software: Enrique Audain.

Supervision: Yasset Perez-Riverol.

Visualization: Enrique Audain.

Writing – original draft: Yasset Perez-Riverol, Enrique Audain.

Writing – review & editing: Max Kuhn, Juan Antonio Vizcaíno, Marc-Phillip Hitz.

References

1. Lynch C. Big data: How do your data grow? Nature. 2008; 455(7209):28–9. https://doi.org/10.1038/455028a PMID: 18769419.
2. Perez-Riverol Y, Bai M, da Veiga Leprevost F, Squizzato S, Park YM, Haug K, et al. Discovering and linking public omics data sets using the Omics Discovery Index. Nature biotechnology. 2017; 35 (5):406–9. https://doi.org/10.1038/nbt.3790 PMID: 28486464.
3. Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007; 23(19):2507–17. https://doi.org/10.1093/bioinformatics/btm344 PMID: 17720704.
4. Barbu A, She Y, Ding L, Gramajo G. Feature Selection with Annealing for Computer Vision and Big Data Learning. IEEE Trans Pattern Anal Mach Intell. 2017; 39(2):272–86. https://doi.org/10.1109/TPAMI.2016.2544315 PMID: 27019473.
5. Perez-Riverol Y, Audain E, Millan A, Ramos Y, Sanchez A, Vizcaino JA, et al. Isoelectric point optimization using peptide descriptors and support vector machines. Journal of proteomics. 2012; 75(7):2269–74. https://doi.org/10.1016/j.jprot.2012.01.029 PMID: 22326964.
6. Wang R, Perez-Riverol Y, Hermjakob H, Vizcaino JA. Open source libraries and frameworks for biological data visualisation: A guide for developers. Proteomics. 2014. https://doi.org/10.1002/pmic.201400377 PMID: 25475079.
7. Bellman R. Dynamic programming and Lagrange multipliers. Proceedings of the National Academy of Sciences. 1956; 42(10):767–9.

8. Michalak K, Kwaśnicka H. Correlation-based feature selection strategy in classification problems. International Journal of Applied Mathematics and Computer Science. 2006; 16:503–11.
9. Wang Y, Tetko IV, Hall MA, Frank E, Facius A, Mayer KF, et al. Gene selection from microarray data for cancer classification—a machine learning approach. Computational biology and chemistry. 2005; 29 (1):37–46. https://doi.org/10.1016/j.compbiolchem.2004.11.001 PMID: 15680584.
10. Wang Y, Makedon F, Pearlman J. Tumor classification based on DNA copy number aberrations determined using SNP arrays. Oncol Rep. 2006; 15 Spec no.:1057–9. PMID: 16525700.
11. Jolliffe I. Principal component analysis: Wiley Online Library; 2002.
12. Kohavi R, John GH. Wrappers for feature subset selection. Artificial intelligence. 1997; 97(1–2):273–324.
13. Ringner M. What is principal component analysis? Nature biotechnology. 2008; 26(3):303–4. https://doi.org/10.1038/nbt0308-303 PMID: 18327243.
14. Yeung KY, Ruzzo WL. Principal component analysis for clustering gene expression data. Bioinformatics. 2001; 17(9):763–74. PMID: 11590094.
15. Yang ZR. Biological applications of support vector machines. Brief Bioinform. 2004; 5(4):328–38. PMID: 15606969.
16. Saal LH, Johansson P, Holm K, Gruvberger-Saal SK, She QB, Maurer M, et al. Poor prognosis in carcinoma is associated with a gene expression signature of aberrant PTEN tumor suppressor pathway activity. Proceedings of the National Academy of Sciences of the United States of America. 2007; 104 (18):7564–9. https://doi.org/10.1073/pnas.0702507104 PMID: 17452630; PubMed Central PMCID: PMCPMC1855070.
17. Duffy MJ. Estrogen receptors: role in breast cancer. Crit Rev Clin Lab Sci. 2006; 43(4):325–47. https://doi.org/10.1080/10408360600739218 PMID: 16769596.
18. Perez-Riverol Y, Sanchez A, Ramos Y, Schmidt A, Muller M, Betancourt L, et al. In silico analysis of accurate proteomics, complemented by selective isolation of peptides. Journal of proteomics. 2011; 74 (10):2071–82. https://doi.org/10.1016/j.jprot.2011.05.034 PMID: 21658481.
19. Audain E, Ramos Y, Hermjakob H, Flower DR, Perez-Riverol Y. Accurate estimation of isoelectric point of protein and peptide based on amino acid sequences. Bioinformatics. 2016; 32(6):821–7. https://doi.org/10.1093/bioinformatics/btv674 PMID: 26568629.
20. Lawrence RT, Perez EM, Hernandez D, Miller CP, Haas KM, Irie HY, et al. The proteomic landscape of triple-negative breast cancer. Cell Rep. 2015; 11(4):630–44. https://doi.org/10.1016/j.celrep.2015.03.050 PMID: 25892236; PubMed Central PMCID: PMCPMC4425736.
21. Wang JJ, Rau C, Avetisyan R, Ren S, Romay MC, Stolin G, et al. Genetic Dissection of Cardiac Remodeling in an Isoproterenol-Induced Heart Failure Mouse Model. PLoS genetics. 2016; 12(7):e1006038. https://doi.org/10.1371/journal.pgen.1006038 PMID: 27385019; PubMed Central PMCID: PMCPMC4934852.
22. Li S, Oh S. Improving feature selection performance using pairwise pre-evaluation. BMC bioinformatics. 2016; 17:312. https://doi.org/10.1186/s12859-016-1178-3 PMID: 27544506; PubMed Central PMCID: PMCPMC4992252.
23. Kuhn M. Caret package. Journal of Statistical Software. 2008; 28(5):1–26.
24. Liaw A, Wiener M. Classification and regression by randomForest. R news. 2002; 2(3):18–22.
25. Zeileis A, Hornik K, Smola A, Karatzoglou A. kernlab-an S4 package for kernel methods in R. Journal of statistical software. 2004; 11(9):1–20.
26. Romanski P, Kotthoff L, Kotthoff ML. Package 'FSelector'. URL http://cran/r-project.org/web/packages/FSelector/index.html; 2013.
27. Audain E, Sanchez A, Vizcaino JA, Perez-Riverol Y. A survey of molecular descriptors used in mass spectrometry based proteomics. Current topics in medicinal chemistry. 2014; 14(3):388–97. PMID: 24304317.
28. Chambers SE, Hoskins PR, Haddad NG, Johnstone FD, McDicken WN, Muir BB. A comparison of fetal abdominal circumference measurements and Doppler ultrasound in the prediction of small-for-dates babies and fetal compromise. Br J Obstet Gynaecol. 1989; 96(7):803–8. PMID: 2669932.
29. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC bioinformatics. 2006; 7:91. https://doi.org/10.1186/1471-2105-7-91 PMID: 16504092; PubMed Central PMCID: PMCPMC1397873.
30. Diaz-Uriarte R, Alvarez de Andres S. Gene selection and classification of microarray data using random forest. BMC bioinformatics. 2006; 7:3. https://doi.org/10.1186/1471-2105-7-3 PMID: 16398926; PubMed Central PMCID: PMCPMC1363357.

31. Pochet N, De Smet F, Suykens JA, De Moor BL. Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction. Bioinformatics. 2004; 20 (17):3185–95. https://doi.org/10.1093/bioinformatics/bth383 PMID: 15231531.

32. Zou Q, Zeng J, Cao L, Ji R. A novel features ranking metric with application to scalable visual and bioinformatics data classification. Neurocomputing. 2016; 173:346–54.

Supplementary Information

Table of Contents

Note 1: Datasets

Table 1: Description of the genomics and proteomics datasets used for benchmarking the feature selection (FS) methods developed in this manuscript.

# dataset	GEO accession numbers	Technology	Samples	Features	# Classes	Description	ML problem	References
1	GSE5325	Transcriptomics	106	8534	2	Analysis of breast cancer tumor samples using 2-color cDNA microarrays.	Classification	[1]
2	-	Proteomics	7000	545	-	Peptides fractionated by IPG-HPLC and analyzed by Mass spectrometry.	Regression	[2, 3]
3	-	Proteomics	44	15525	5	Triple-Negative Breast Cancer (TNBC) proteome. Label-free deep proteome analysis of 44 (samples and technical replicates) human breast specimens.	Classification	[4]
4	GSE48760	Transcriptomics	208	25697	2	Transcriptomics analysis of left ventricles of mouse subjected to an isoproterenol challenge.	Classification	[5]
5	GSE6919 /GPL8300	Transcriptomics	171	12558	4	Expression data from normal and prostate tumor tissues.	Classification	[6, 7]
6	GSE6919 /GPL92	Transcriptomics	168	12553	4	Expression data from normal and prostate tumor tissues.	Classification	[6, 7]
7	GSE6919 /GPL93	Transcriptomics	165	12626	4	Expression data from normal and prostate tumor tissues.	Classification	[6, 7]

Note 2: Feature Selection Approaches

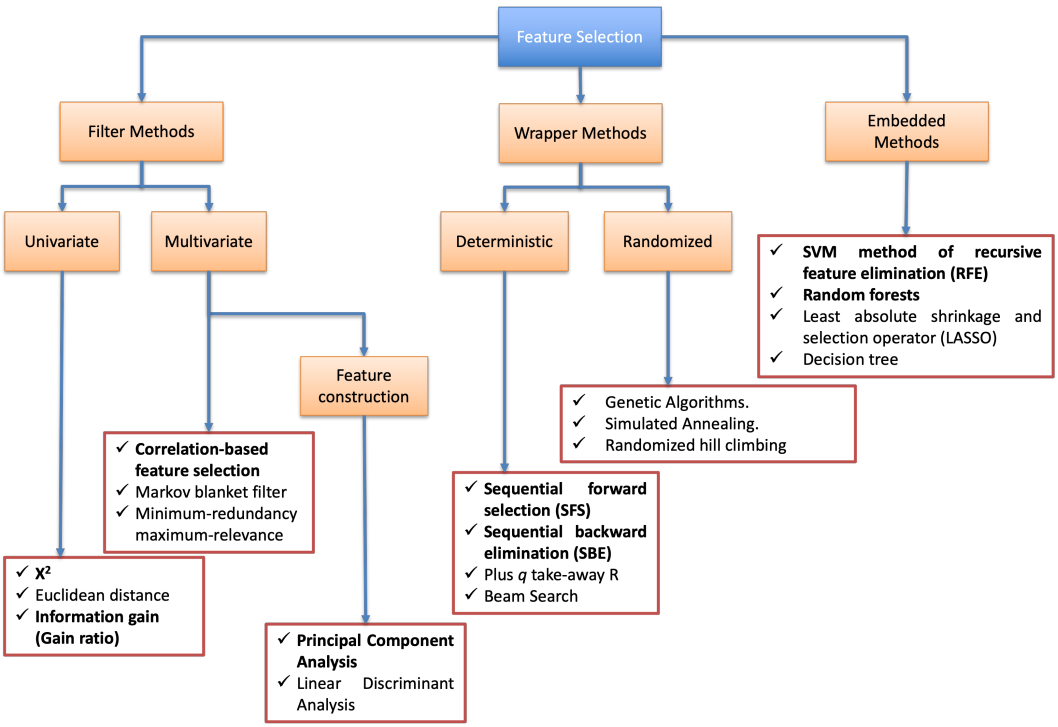


Figure 1: Schema of the existing Feature Selection Approaches. We highlighted (in **bold** letters) all the algorithms that can be combined using the proposed workflow in this study.

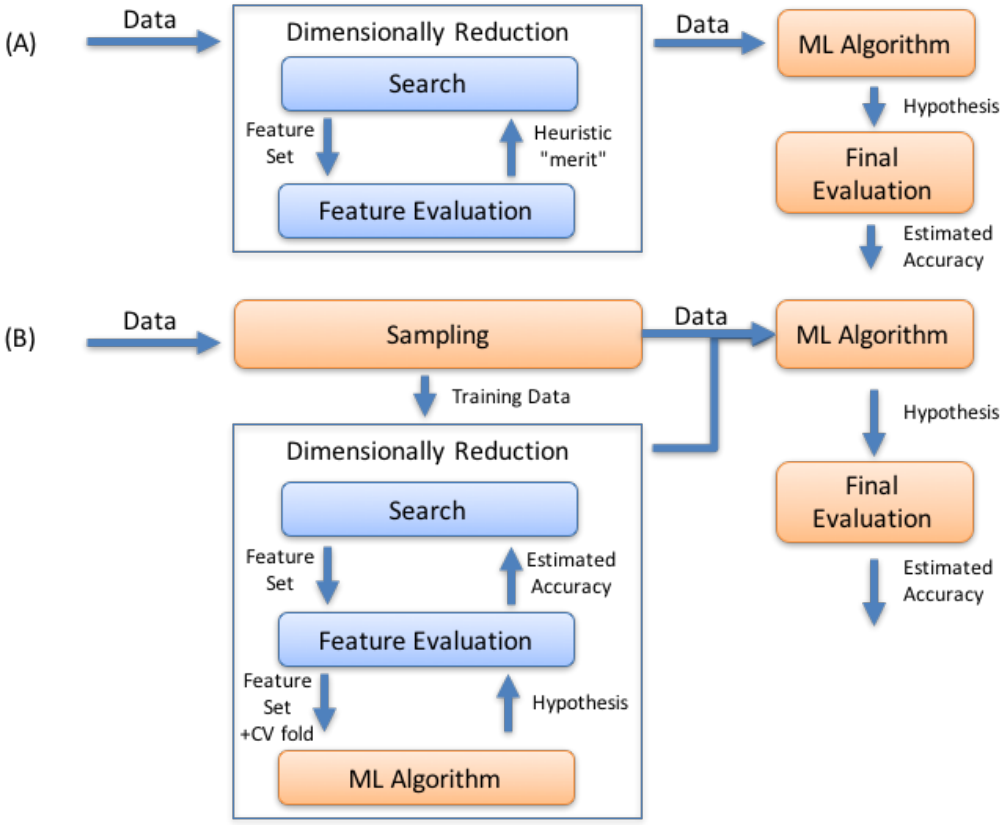


Figure 2: Diagram of Feature Selection (FS) Filtering (A) and (B) Wrapper approaches. In the filtering approaches the dimension reduction occurs before the machine learning step. In contrast, in the Wrapper approaches the machine learning model (ML Algorithm) is used in combination with the FS step to improve the selection process. The proposed workflow combines the strengths of both techniques.

Note 3: Number of Final Features after the FS method step

Table 2: Number of final features after the FS step for the proposed workflow. Three different combinations are shown: The Random Forest classification algorithm without a feature selection step (**RF);** the combination of the univariate correlation filter with the matrix correlation and recursive feature elimination before the Random Forest classification step (**X2-CM-RFE-RF**); and finally, the combination of univariate correlation filter with principal component analysis, before the Random Forest classification step (**X2-PCA-RFE-RF**).

Workflow	**GSE5325** (Dataset 1)	**TNBC** (Dataset 3)	**GSE48760** (Dataset 4)	**GSE6919 /GPL8300** (Dataset 5)	**GSE6919 /GPL92** (Dataset 6)	**GSE6919 /GPL93** (Dataset 7)
RF	1,874	1,419	4,117	4,803	4,761	4,941
X2-CM-RFE-RF	8	20	7	50	45	50
X2-PCA-RFE-RF	8	6	5	35	5	6

Note 4: Benchmark of the R workflow proposed with the MRMD tool

Table 3. Performance evaluation on TNBC and GSE5325 datasets for MRMD tool and the proposed FS workflow(s). MRMD: Maximum-Relevance-Maximum-Distance, X2: Univariate Correlation filter, CM: Multivariate Correlation Matrix, RFE: Recursive Feature Elimination, RF: Random Forest, PCA: Principal Component Analysis, GI: Gain Information filter.

	TNBC			**GSE5325**		
	Acc. Max.	**# Features**	**Runtime (min)**	**Acc. Max.**	**# Features**	**Runtime**
MRMD	0.98	2689	18.46	0.95	409	57.53
X2-CM-RFE-RF	1.00	20	3.03	0.94	8	6.39
X2-PCA-RFE-RF	0.96	6	1.99	0.91	7	5.26
GI-CM-RFE-RF	0.98	30	1.67	0.94	20	2.62
GI-PCA-RFE-RF	0.96	8	1.01	0.91	4	2.65

References

1. Saal LH, Johansson P, Holm K, Gruvberger-Saal SK, She QB, Maurer M, et al. Poor prognosis in carcinoma is associated with a gene expression signature of aberrant PTEN tumor suppressor pathway activity. Proceedings of the National Academy of Sciences of the United States of America. 2007;104(18):7564-9. doi: 10.1073/pnas.0702507104. PubMed PMID: 17452630; PubMed Central PMCID: PMCPMC1855070.
2. Perez-Riverol Y, Audain E, Millan A, Ramos Y, Sanchez A, Vizcaino JA, et al. Isoelectric point optimization using peptide descriptors and support vector machines. Journal of proteomics. 2012;75(7):2269-74. doi: 10.1016/j.jprot.2012.01.029. PubMed PMID: 22326964.
3. Perez-Riverol Y, Sanchez A, Ramos Y, Schmidt A, Muller M, Betancourt L, et al. In silico analysis of accurate proteomics, complemented by selective isolation of peptides. Journal of proteomics. 2011;74(10):2071-82. doi: 10.1016/j.jprot.2011.05.034. PubMed PMID: 21658481.
4. Lawrence RT, Perez EM, Hernandez D, Miller CP, Haas KM, Irie HY, et al. The proteomic landscape of triple-negative breast cancer. Cell Rep. 2015;11(4):630-44. doi: 10.1016/j.celrep.2015.03.050. PubMed PMID: 25892236; PubMed Central PMCID: PMCPMC4425736.
5. Wang JJ, Rau C, Avetisyan R, Ren S, Romay MC, Stolin G, et al. Genetic Dissection of Cardiac Remodeling in an Isoproterenol-Induced Heart Failure Mouse Model. PLoS genetics. 2016;12(7):e1006038. doi: 10.1371/journal.pgen.1006038. PubMed PMID: 27385019; PubMed Central PMCID: PMCPMC4934852.
6. Chandran UR, Ma C, Dhir R, Bisceglia M, Lyons-Weiler M, Liang W, et al. Gene expression profiles of prostate cancer reveal involvement of multiple molecular pathways in the metastatic process. BMC Cancer. 2007;7:64. doi: 10.1186/1471-2407-7-64. PubMed PMID: 17430594; PubMed Central PMCID: PMCPMC1865555.
7. Li S, Oh S. Improving feature selection performance using pairwise pre-evaluation. BMC bioinformatics. 2016;17:312. doi: 10.1186/s12859-016-1178-3. PubMed PMID: 27544506; PubMed Central PMCID: PMCPMC4992252.

Chapter III. Integrative analysis of genomic variants reveals new associations of candidate haploinsufficient genes with congenital heart disease

The original peer-reviewed publication presented in this chapter (pages 54-77) is publicly available at *https://doi.org/10.1371/journal.pgen.1009679.*

PLOS GENETICS

RESEARCH ARTICLE

Integrative analysis of genomic variants reveals new associations of candidate haploinsufficient genes with congenital heart disease

Enrique Audain[1,2☯], Anna Wilsdon[3☯], Jeroen Breckpot[4], Jose M. G. Izarzugaza[5], Tomas W. Fitzgerald[6], Anne-Karin Kahlert[1,2,7], Alejandro Sifrim[8,9], Florian Wünnemann[10], Yasset Perez-Riverol[11], Hashim Abdul-Khaliq[12], Mads Bak[13,14], Anne S. Bassett[15,16], Woodrow D. Benson[17], Felix Berger[18], Ingo Daehnert[19], Koenraad Devriendt[4], Sven Dittrich[20], Piers EF Daubeney[21], Vidu Garg[22,23,24,25], Karl Hackmann[7], Kirstin Hoff[1,2], Philipp Hofmann[1,2], Gregor Dombrowsky[1,2], Thomas Pickardt[26], Ulrike Bauer[26], Bernard D. Keavney[27,28], Sabine Klaassen[29,30,31], Hans-Heiner Kramer[1,2], Christian R. Marshall[32,33], Dianna M. Milewicz[34], Scott Lemaire[35], Joseph S. Coselli[36], Michael E. Mitchell[36], Aoy Tomita-Mitchell[36], Siddharth K. Prakash[34], Karl Stamm[36], Alexandre F. R. Stewart[37], Candice K. Silversides[15], Reiner Siebert[38,39], Brigitte Stiller[40], Jill A. Rosenfeld[17], Inga Vater[39], Alex V. Postma[41,42], Almuth Caliebe[39], J. David Brook[3], Gregor Andelfinger[43], Matthew E. Hurles[44], Bernard Thienpont[4,45], Lars Allan Larsen[13]*, Marc-Phillip Hitz[1,2,39,44]*

1 Department of Congenital Heart Disease and Pediatric Cardiology, University Hospital of Schleswig-Holstein, Kiel, Germany, **2** German Center for Cardiovascular Research (DZHK), Kiel, Germany, **3** School of Life Sciences, University of Nottingham, University Park, Nottingham, United Kingdom, **4** Centre for Human Genetics, Katholieke Universiteit Leuven, Leuven, Belgium, **5** Department of Health Technology, Technical University of Denmark, Lyngby, Denmark, **6** European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Cambridge, United Kingdom, **7** Institute for Clinical Genetics, Faculty of Medicine Carl Gustav Carus, TU Dresden, Dresden, Germany, **8** Department of Human Genetics, University of Leuven, KU Leuven, Leuven, Belgium, **9** Sanger Institute-EBI Single-Cell Genomics Centre, Wellcome Trust Sanger Institute, Hinxton, United Kingdom, **10** Montreal Heart Institute, Université de Montréal, Québec, Canada, **11** European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom, **12** Clinic for Pediatric Cardiology—University Hospital of Saarland, Homburg (Saar), Germany, **13** Department of Cellular and Molecular Medicine, University of Copenhagen, Copenhagen, Denmark, **14** Department of Clinical Genetics, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark, **15** Toronto Congenital Cardiac Centre for Adults, and Division of Cardiology, Department of Medicine, University Health Network, Toronto, Canada, **16** Department of Psychiatry, University of Toronto, Toronto, Canada, **17** Department of Pediatrics, Medical College of Wisconsin, Milwaukee, Wisconsin, United States of America, **18** Department of Congenital Heart Disease—Pediatric Cardiology, German Heart Center Berlin, Berlin, Germany, **19** Department of Pediatric Cardiology and Congenital Heart Disease, Heart Center, University of Leipzig, Leipzig, Germany, **20** Department of Pediatric Cardiology, University Hospital Erlangen, Friedrich-Alexander-University Erlangen-Nürnberg (FAU), Erlangen, Germany, **21** Division of Paediatric Cardiology, Royal Brompton Hospital, London, United Kingdom, **22** The Heart Center, Nationwide Children's Hospital, Columbus, Ohio, United States of America, **23** Department of Molecular Genetics, The Ohio State University, Columbus, Ohio, United States of America, **24** Center for Cardiovascular Research, Nationwide Children's Hospital, Columbus, Ohio, United States of America, **25** Department of Pediatrics, The Ohio State University, Columbus, Ohio, United States of America, **26** Competence Network for Congenital Heart Defects, Berlin, Germany, **27** Division of Cardiovascular Sciences, School of Medical Sciences, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester, United Kingdom, **28** Division of Evolution & Genomic Sciences, School of Biological Sciences, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester, United Kingdom, **29** Experimental and Clinical Research Center (ECRC), a joint cooperation between the Charité Medical Faculty and the Max-Delbrück-Center for Molecular Medicine (MDC), Berlin, Germany, **30** Charité—Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Department of Pediatric Cardiology, Berlin, Germany, **31** DZHK (German Centre for Cardiovascular Research), partner site Berlin, Berlin, Germany, **32** The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, Canada,

OPEN ACCESS

Citation: Audain E, Wilsdon A, Breckpot J, Izarzugaza JMG, Fitzgerald TW, Kahlert A-K, et al. (2021) Integrative analysis of genomic variants reveals new associations of candidate haploinsufficient genes with congenital heart disease. PLoS Genet 17(7): e1009679. https://doi.org/10.1371/journal.pgen.1009679

Editor: Anthony B. Firulli, Indiana University Purdue University at Indianapolis, UNITED STATES

Received: December 3, 2020

Accepted: June 23, 2021

Published: July 29, 2021

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pgen.1009679

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files. The data used in this study have

been already published. The reference for each individual study is shown in S2 and S3 Tables. The assembled DNV dataset used in this study is provided in the S14 Table. In addition, we have provided a BED file with the CNV dataset (S15 Table).

Funding: This study was supported by the German Center for Cardiovascular Research (DZHK) partner sites Berlin, Kiel; the Competence Network for Congenital Heart Defects, National Register for Congenital Heart Defects and KinderHerz (E.A. and M.P.H). BK is supported by a British Heart Foundation Personal Chair (CH/13/2/30154). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: The Department of Molecular and Human Genetics at Baylor College of Medicine receives revenue from clinical genetic testing conducted at Baylor Genetics Laboratories. M.E.H. is a co-founder of, consultant to and holds shares in Congenica, a genetics diagnostic company.

33 Genome Diagnostics, Department of Paediatric Laboratory Medicine, The Hospital for Sick Children, Toronto, Canada, **34** Department of Internal Medicine, McGovern Medical School, University of Texas Health Science Center at Houston, Houston, Texas, United States of America, **35** Michael E. DeBakey Department of Surgery, Baylor College of Medicine, Houston, Texas, United States of America, **36** Department of Surgery, Division of Cardiothoracic Surgery, Medical College of Wisconsin, Milwaukee, Wisconsin, United States of America, **37** Ruddy Canadian Cardiovascular Genetics Centre, University of Ottawa Heart Institute, Ottawa, Canada, **38** Institute of Human Genetics, University Hospital Ulm, Ulm, Germany, **39** Department of Human Genetics, University Medical Center Schleswig-Holstein (UKSH), Kiel, Germany, **40** Department of Congenital Heart Disease and Pediatric Cardiology, University Heart Center Freiburg—Bad Krozingen, Freiburg, Germany, **41** Department of Medical Biology, Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands, **42** Department of Clinical Genetics, Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands, **43** Cardiovascular Genetics, Department of Pediatrics, Centre Hospitalier Universitaire Saint-Justine Research Centre, Université de Montréal, Montreal, Canada, **44** Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, United Kingdom, **45** Laboratory of Translational Genetics, Department of Human Genetics, KU Leuven, Leuven, Belgium

☯ These authors contributed equally to this work.
* larsal@sund.ku.dk (LAL); Marc-Phillip.Hitz@uksh.de (MPH)

Abstract

Numerous genetic studies have established a role for rare genomic variants in Congenital Heart Disease (CHD) at the copy number variation (CNV) and *de novo* variant (DNV) level. To identify novel haploinsufficient CHD disease genes, we performed an integrative analysis of CNVs and DNVs identified in probands with CHD including cases with sporadic thoracic aortic aneurysm. We assembled CNV data from 7,958 cases and 14,082 controls and performed a gene-wise analysis of the burden of rare genomic deletions in cases versus controls. In addition, we performed variation rate testing for DNVs identified in 2,489 parent-offspring trios. Our analysis revealed 21 genes which were significantly affected by rare CNVs and/or DNVs in probands. Fourteen of these genes have previously been associated with CHD while the remaining genes (*FEZ1*, *MYO16*, *ARID1B*, *NALCN*, *WAC*, *KDM5B* and *WHSC1*) have only been associated in small cases series or show new associations with CHD. In addition, a systems level analysis revealed affected protein-protein interaction networks involved in Notch signaling pathway, heart morphogenesis, DNA repair and cilia/centrosome function. Taken together, this approach highlights the importance of re-analyzing existing datasets to strengthen disease association and identify novel disease genes and pathways.

Author summary

Congenital heart disease (CHD) is the most common congenital anomaly and represents a major global health burden. Multiple studies have identified a key genetic component contributing to the aetiology of CHD. However, despite the advances in the field of CHD within the last three decades, the genetic causes underlying CHD are still not fully understood. Herein we have assembled a large patient CHD cohort and performed a data-driven meta-analysis of genomic variants in CHD. This analysis has allowed us to strengthen the disease association of known CHD genes, as well as identifying novel haploinsufficient CHD candidate genes.

Introduction

Congenital Heart Disease (CHD) accounts for a large fraction of foetal and infant deaths, with incidence rates ranging from 7–9 per 1000 live births [1]. Within the last 30 years, survival rates have substantially increased due to improvements in surgical, interventional and clinical intensive care resulting in a rapidly growing number of CHD survivors reaching adulthood [2]. Nevertheless, there is still increased morbidity and mortality in individuals with CHD, resource utilization is high especially among severely affected patients, and importantly, the underlying etiology remains unclear for the majority of cases.

CHD is multifactorial, with both environmental and genetic risk factors [3,4]. Familial aggregation of CHD including Thoracic aortic aneurysm (TAA), as well as a large proportion of genomic copy number variants (CNVs) and *de novo* intragenic variations (DNVs) in probands with CHD suggest a strong genetic component. An estimated 4–20% of CHD cases are due to rare CNVs, suggesting that a significant part of CHD is caused by gene-dosage defects [5]. Recently, exome sequencing in large cohorts has been used to identify novel disease genes and strengthen known disease associations through the demonstration of an excess of *de novo* protein truncating variants (PTV) and rare inherited loss-of-function (LOF) variants in probands with CHD [6,7].

Overlaying both CNVs and PTVs has been used to define novel CHD relevant disease genes in contiguous gene disorders [8,9]. Following this principle, we have performed a genome-wide integrative meta-analysis of published and publicly available datasets of CNVs and DNVs identified in probands with CHD. This analysis, which is one of the larger meta-analyses of genomic variants in CHD so far, strengthens the disease association of known CHD genes and identifies novel haploinsufficient CHD candidate genes.

Results

Cohort description and workflow

We assembled a cohort with 7,958 cases (comprising both non-syndromic CHD, syndromic CHD and TAA cases) and 14,082 controls (**S1 Table**). Of the total of cases, 777 (~10%) were diagnosed with Thoracic Aortic Aneurysm (TAA). An overview of the sources used to assemble the present cohort is listed in **S2 Table** (for CHD cases) and **S3 Table** (for controls). We applied a set of quality control filters to our assembled CNV data before performing case-control association tests (Materials and Methods). In addition, common CNVs (minor allele frequency (MAF) in controls > 0.01) were excluded from the analysis. After filtering, 6,746 cases and 14,024 controls remained for further downstream analysis. Furthermore, we built a dataset of *de novo* variations (DNVs) identified in 2,489 probands with CHD from parent-offspring trios [6,7].

CNV burden test of known CHD genes

Haploinsufficiency has been shown to cause a reasonable proportion of CHD [5]. Thus, genes known to be associated with CHD and genes which are intolerant for LOF variations should be deleted more often in probands with CHD than in controls. To test this hypothesis, we performed a CNV burden test using sets of genes known to be involved in CHD. In addition, we included genes known to be associated with developmental disorders, a curated list of known haploinsufficient disease genes, autosomal recessive disease genes and genes predicted to be intolerant to LOF variations (based on the observed/expected LOF ratio from gnomAD [10]). The burden test was performed using a logistic regression framework [11] (implemented in

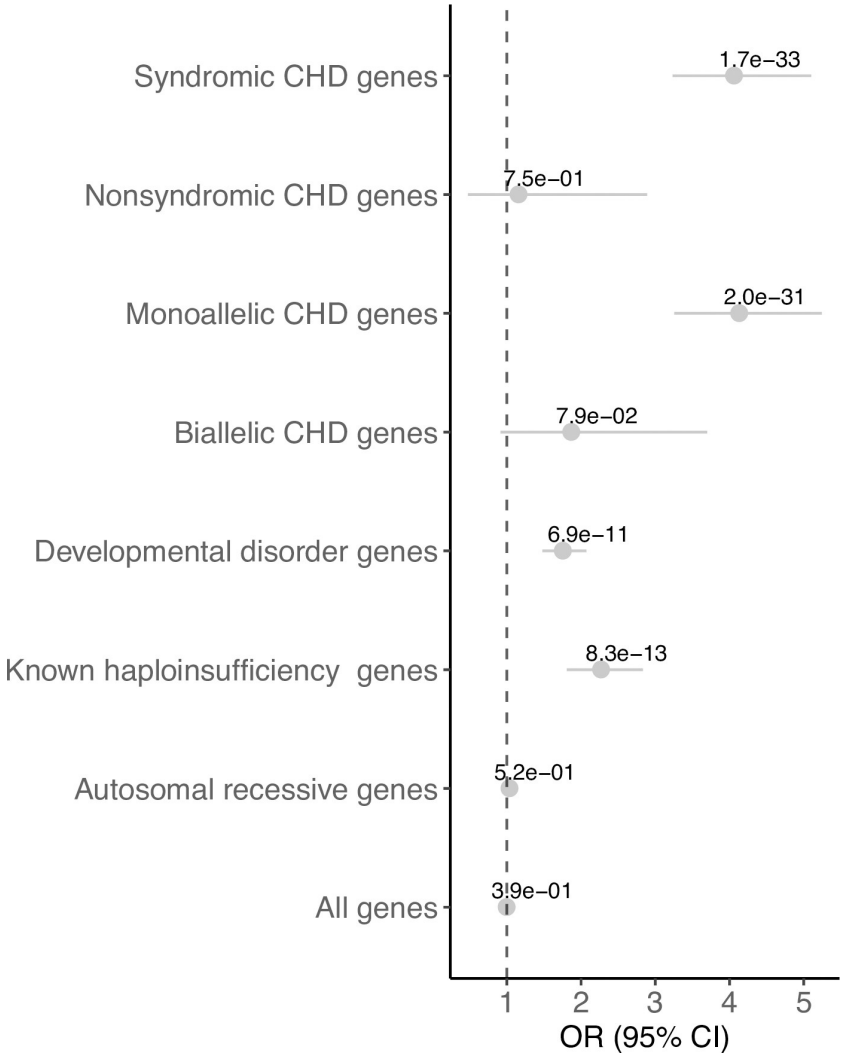


Fig 1. CNV burden test on known gene sets. The forest plot shows the odds ratio (dots), the 95% confidence intervals indicating the certainty about the OR (interrupted line) and the *P-value* in the indicated gene sets.

https://doi.org/10.1371/journal.pgen.1009679.g001

PLINK v1.7). Fig 1 and S4 Table summarize the results from the burden test on the different gene sets: known CHD genes (grouped in syndromic, non-syndromic, monoallelic and biallelic), developmental disorder genes, haploinsufficiency disease genes, autosomal recessive genes and all protein-coding genes. We tested all protein-coding genes to address the possibility that the analyses could be biased by differences in the CNV rate within the case and control groups, since we have assembled our cohort from different datasets. We did not observe genome-wide (all tested protein-coding genes) enrichment ($P = 0.39$, $OR = 0.99$) nor enrichment in the autosomal recessive gene set ($P = 0.52$, $OR = 1.03$) when comparing rare CNV deletions in cases vs controls. In contrast, the analysis revealed significant differences in the burden of CNV deletions between cases and controls for the set of haploinsufficiency genes ($P = 8.29 \times 10^{-13}$, $OR = 2.27$). As expected, our analysis revealed significant enrichment for the set of known

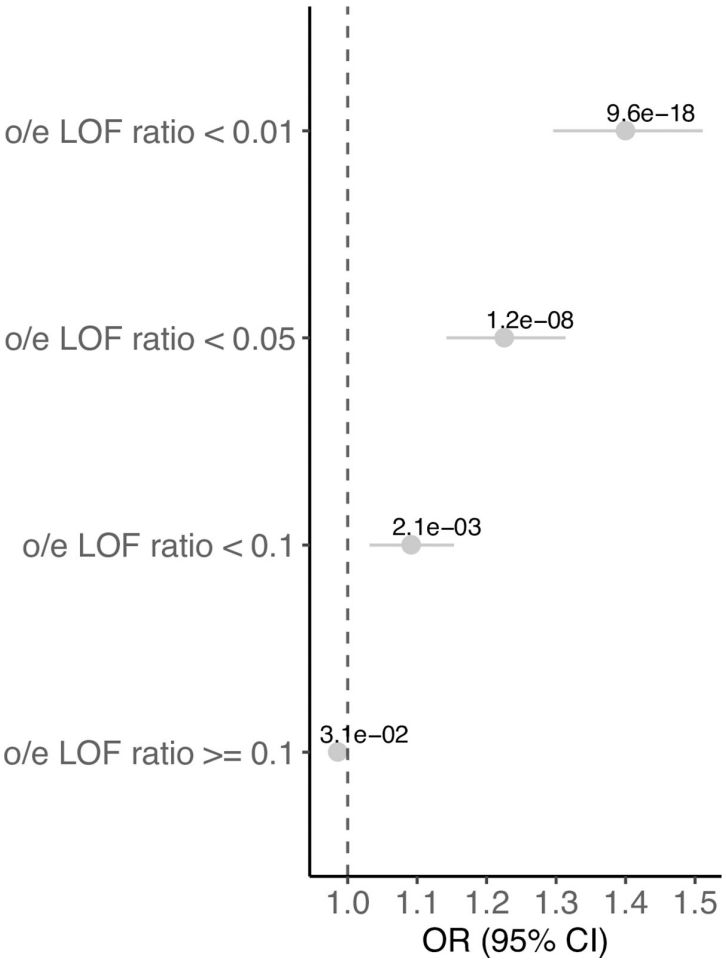


Fig 2. CNV burden test on constraint LOF genes at different observed/expected LOF ratio thresholds. The forest plot shows the odds ratio (dots), the 95% confidence intervals indicating the certainty about the OR (interrupted line) and the *P-value* in the indicated gene sets.

https://doi.org/10.1371/journal.pgen.1009679.g002

CHD genes, which is mainly explained by the contribution of monoallelic CHD genes ($P = 2.04 \times 10^{-31}$, $OR = 4.13$) and syndromic CHD gene set ($P = 1.66 \times 10^{-33}$, $OR = 4.06$). Unlike the monoallelic and syndromic CHD gene sets, no significant enrichment was found for the nonsyndromic ($P = 0.75$, $OR = 1.16$) and biallelic ($P = 0.08$, $OR = 1.87$) CHD gene sets. Our analysis revealed a moderate enrichment of rare CNVs in the developmental disorder gene set ($P = 6.90 \times 10^{-11}$, $OR = 1.75$).

When the regression-based analysis was performed at different levels of the observed/expected LOF ratio (*oeLOF*) constraint metric (Fig 2 and S5 Table), we observed the higher enrichment toward the most LOF constrained genes ($oeLOF < 0.01$, $P = 9.55 \times 10^{-18}$, $OR = 1.40$) and still a moderate enrichment for genes with $oeLOF < 0.1$ ($P = 0.002$, $OR = 1.09$). No enrichment was observed in the set with *oeLOF* ratio $> = 0.1$ ($P = 0.03$, $OR = 0.99$). Based on these results we conclude that haploinsufficiency causes a significant component of CHD.

Genome-wide identification of haploinsufficiency candidate disease genes for CHD

To perform a systematic, genome-wide identification of potential haploinsufficient CHD disease genes and loci, we analysed the CNV burden of 19,969 protein-coding genes (GENCODE v19). To this end, we compared the number of rare CNV deletions (MAF < 0.01) among cases and controls for each gene, and identified genes with significant CNV burden using a permutation test (significance level of adjusted $P < 0.05$, see Materials and Methods). If a CNV spanned two or more genes, all affected protein-coding genes were considered in the analysis. The distributions of rare CNV deletions in CHD cases across all 22 human autosomes is shown in Fig 3. Significant candidate genes had a median number of 12 overlapping CNVs in cases, compared to a median of 0 overlapping CNVs in controls (S1A Fig). Because CNVs can be large chromosomal aberrations, multiple genes were affected by some of the CNVs. In total, 528 genes (Sheet A in S6 Table) reached significance (Permutation test, *P* adjusted < 0.05). These 528 genes encompass a total of 63 loci (Sheet B in S6 Table, highlighted in magenta in

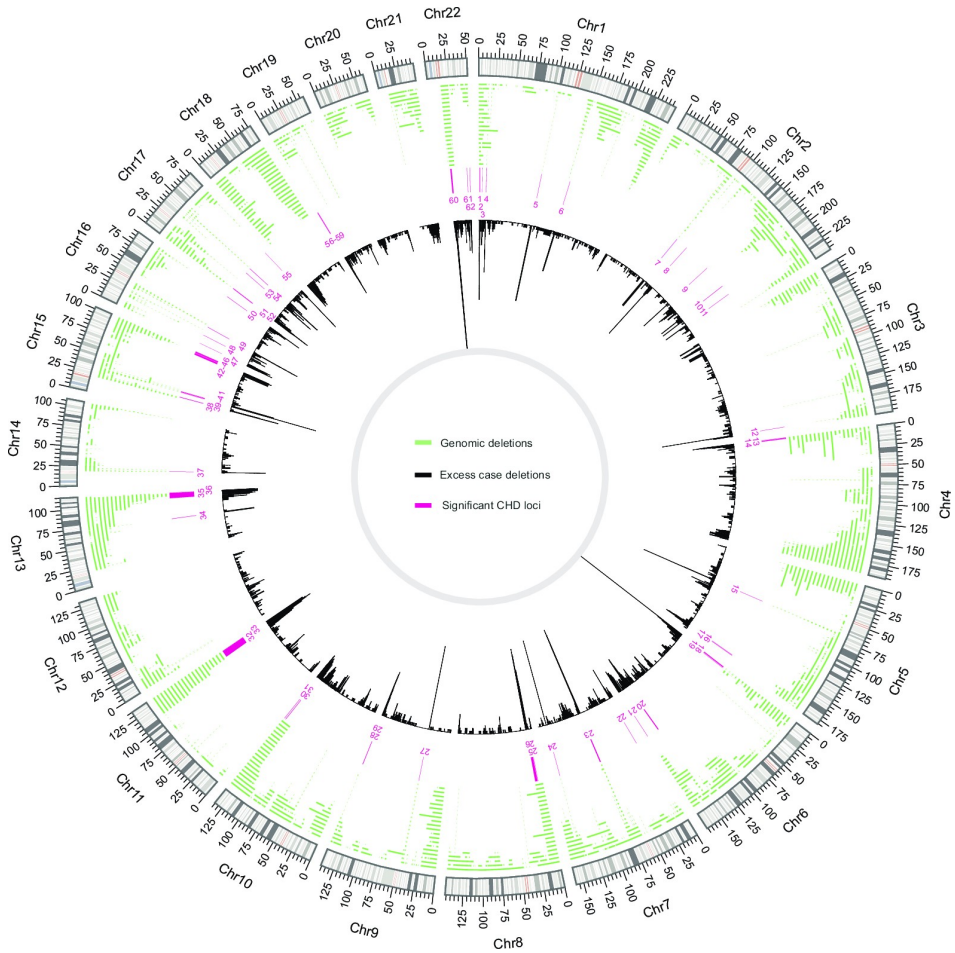


Fig 3. CNV deletion distribution across the 22 autosomes. The plot shows the distribution of rare CNV deletions (green track) in CHD cases, the differences between the overlapping CNV deletions in cases and controls (black track) and highlight the location of the 63 significant loci discovered (in magenta).

https://doi.org/10.1371/journal.pgen.1009679.g003

Fig 3). The sizes of these loci range from 558 bp to 10.5 Mbp, with a median value of 243 Kbp (S1B Fig). The number of genes per locus ranged from 1 to 48, with a median value of 3 (S1C Fig). Only 16 loci contained a single gene (Sheet B in S6 Table).

In addition, we tested previously described CNV deletion syndrome regions (https://decipher.sanger.ac.uk/disorders#syndromes/overview) associated with developmental disorders and/or CHD for enrichment in our analysis (Materials and Methods). We found eight of these regions enriched in the dataset (S7 Table), with the 16p11.2-p12.2 locus being the region with the largest number of deletions in cases (n = 230).

Shared genetic architecture of CHD and TAA

We independently performed a genome-wide test without the TAA cases to evaluate its impact on CHD. As expected, most of the genes (447 out of 528) remained significant after removing the contribution of TAA cases, since ~90% of the cases in the analyzed CNV cohort were CHD. Ten genes were significantly enriched independently when analyzing CHD and TAA cases, while 61 were significantly enriched only in TAA cases (S8 Table).

De novo variation analysis

To identify an independent set of haploinsufficient CHD candidate genes, we combined *de novo* variations identified in two large-scale CHD case-control studies [6,7] and performed a gene-based *de novo* variation (DNV) burden test [12]. We analysed a total of 4,195 rare DNVs within 2,534 genes in the patient cohort. After classifying every variant into functional groups (Materials and Methods) 526 of these variants were predicted to be protein-truncating and 2,647 were missense. We evaluated for potential differences of the DNV rates between cohorts (see Materials and Methods). Comparison of the rate of each variant type across the groups was non-significant ($P > 0.05$, Poisson test, S9 Table).

We used two available statistical methods, Mupit [12] and DeNovoWEST [13], which test the significance of observed DNV at gene level, by comparing the number of observed variations with the number of expected variations (based on a sequence-dependent variation recurrence rate, see Materials and Methods). While Mupit focuses on enrichment of protein-truncating DNVs specifically, the DeNovoWEST test incorporates missense constraint information at variant level and applies a unified severity scale at variant level based on the empirically-estimated positive predictive value of being pathogenic. Based on the complementary results of both tests [13], we reported the minimal observed DNV *p-value* (P_{dnv}) per gene.

We identified 14 genes significantly enriched in the DNV analysis ($P < 0.05$ after Bonferroni correction for multiple testing, S10 Table). All of these genes were affected by at least two constrained non-synonymous DNV (nsDNV) and show significant overlap with 11/14 (78.6%) of the genes being known CHD disease genes. *CHD7* (OMIM 214800) was the most significant haploinsufficient gene ($P = 2.84 \times 10^{-26}$) with 18 nsDNVs identified in the patient cohort. Other highly enriched genes for nsDNV—*KMT2D* (OMIM 147920), *KMT2A* (OMIM 605130), *NSD1* (OMIM 117550), *TAB2* (OMIM 614980), and *ADNP* (OMIM 615873)—have been previously associated with different types of neurodevelopmental disorders with co-occurrence of CHD. In the case of *KDM5B* (OMIM 618109), it has only been described in the context of a recessive neurodevelopmental phenotype with cases presenting ASD (Atrial septal defects) [14,15].

We next evaluated the distribution of o/e LOF ratio at different levels of DNV enrichment (genes were split based on P_{dnv}). Since the o/e ratio of LOF variation in each gene is strongly affected by its length, we instead used the 90% upper bound of its confidence interval (termed LOEUF), which keeps the direct estimate of the o/e ratio and allows to distinguish small genes

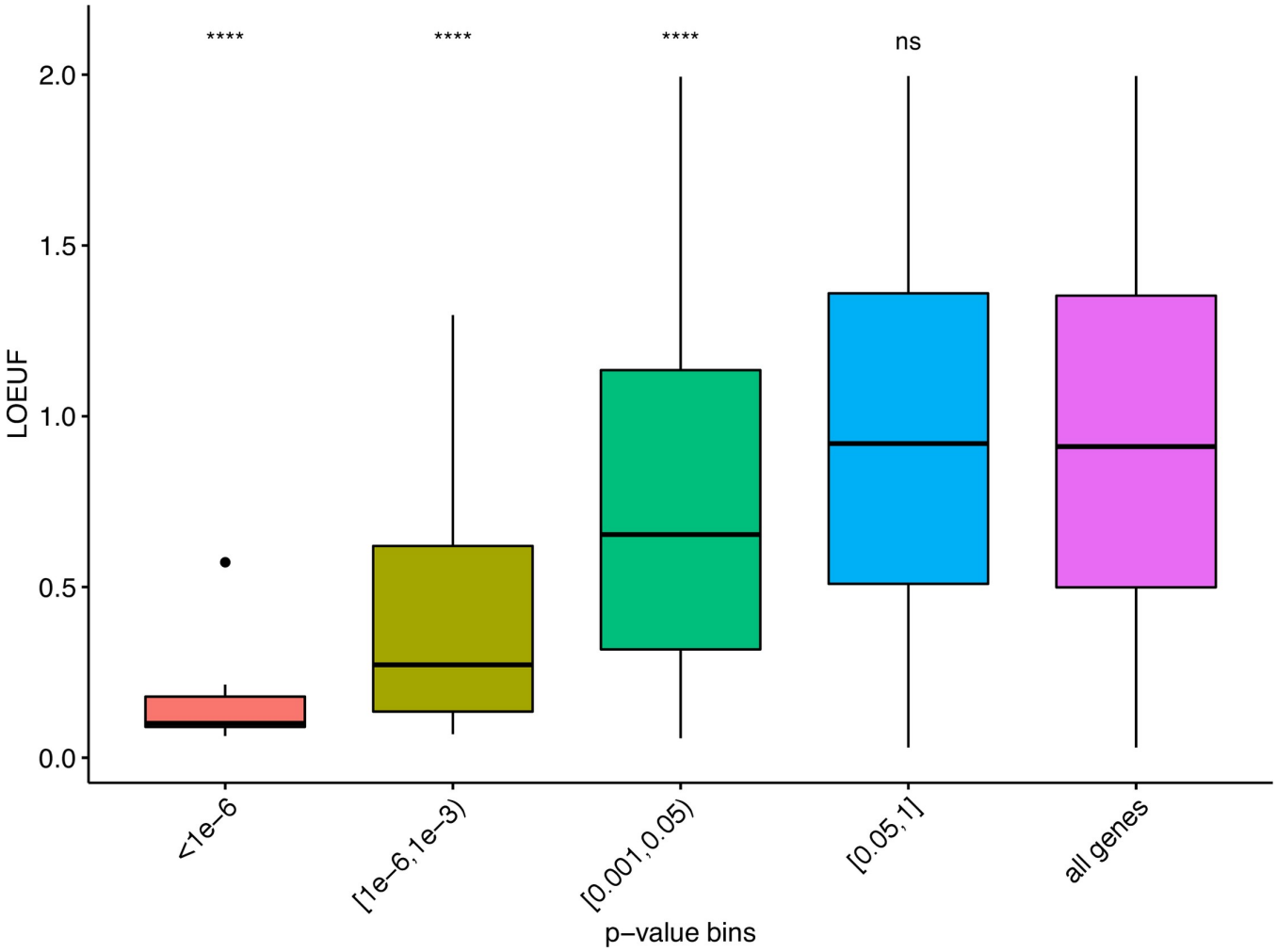


Fig 4. Comparison of the distribution of LOEUF metric at different level of significance of nsDNV-enriched genes. X-axis denotes the P-values from the DNV analysis (binned). Y-axis denotes the o/e LOF ratio upper bound fraction (LOEUF). All groups were compared against the LOEUF distribution of all protein-coding genes (purple). Differences between the distributions were tested using a two-sided Wilcoxon rank sum test. ****: P<0.0001, ns: non-significant.

https://doi.org/10.1371/journal.pgen.1009679.g004

from large genes, as suggested by Karczewski *et al* [10]. We observed that the genes with higher enrichment for nsDNV (lower P_{dnv}) show a significant decreased LOEUF compared to the mean of all protein-coding genes (Fig 4).

Integration of DNV and CNV results

To identify high confidence haploinsufficient CHD disease genes, we performed a joint analysis integrating the results from the CNV and the DNV analysis. We combined the results from both analyses (P_{dnv} and P_{cnv}) using the Fisher combine method. We demonstrated that both enriched genes for DNV and CNV deletions are significantly represented among LOF constraint genes (measured by the o/e LOF ratio). Therefore, we applied a Bonferroni multiple testing correction using independent hypothesis weighting [16] (IHW) by incorporating the gene o/e LOF ratio, as a measure of haploinsufficiency (S2 Fig). Our analysis revealed 21 genes that were significantly enriched for CNV deletions and/or non-synonymous DNV (Table 1).

Table 1. Top 21 significant genes arising from both the permutation-based test and the DNV rate-based test. Cases/Controls: Number of cases and controls carrying CNV deletions overlapping the gene in the CNV analysis. P_{cnv}: p-value from the CNV permutation test. nsDNV: Number of constrained non-synonymous variations in the *de novo* analysis. P_{dnv}: p-value from the DNV analysis. Significant: The analysis where the gene was significant (dnv: DNV analysis, cnv: CNV analysis, both: Both analysis, none: Non-significant neither DNV nor CNV analysis). *metaP*: combined p-value (P_{dnv} and P_{cnv}) using the Fisher method. P_{ihw}: Bonferroni corrected p-value using independent hypothesis weighting (IHW) and LOEUF metric as covariate. LOEUF: o/e LOF ratio upper bound fraction from gnomAD. *All the 21 genes were significant after combining their p-values and applying Bonferroni correction. [1]Evidence is from mouse models [24,62].

	CNV			DNV			Combined			
Gene	case	control	P_{cnv}	nsDNV	P_{dnv}	*Significant	*metaP*	P_{ihw}	LOEUF	Known CHD
CHD7	6	1	6.80E-03	18	2.84E-26	dnv	1.25E-26	8.05E-23	0.076	Yes
KMT2D	0	0	1.00E+00	18	1.32E-25	dnv	7.67E-24	4.93E-20	0.103	Yes
NSD1	1	1	5.63E-01	12	1.00E-14	dnv	1.90E-13	2.14E-09	0.095	Yes
KMT2A	0	0	1.00E+00	7	1.00E-14	dnv	3.32E-13	1.86E-09	0.065	Yes
NOTCH1	10	24	1.00E+00	7	1.00E-14	dnv	3.32E-13	2.14E-09	0.097	Yes
TAB2	12	0	1.00E-04	5	3.46E-09	both	1.03E-11	5.75E-08	0.098	Yes
ANKRD11	13	0	1.00E-04	3	2.32E-05	cnv	4.85E-08	2.72E-04	0.107	Yes
WHSC1	11	0	1.00E-04	3	8.73E-05	cnv	1.71E-07	9.96E-04	0.119	No[1]
ADNP	0	0	1.00E+00	4	9.94E-09	dnv	1.93E-07	1.13E-03	0.123	Yes
DYRK1A	4	0	1.43E-02	4	9.46E-07	dnv	2.59E-07	1.64E-03	0.214	Yes
NALCN	10	1	1.00E-04	3	1.76E-04	cnv	3.32E-07	6.83E-03	0.522	No
ELN	30	0	1.00E-04	2	1.77E-04	cnv	3.34E-07	7.50E-03	0.871	Yes
WAC	7	0	4.00E-04	3	1.31E-04	none	9.33E-07	5.44E-03	0.084	No
RBFOX2	1	0	3.45E-01	4	1.59E-07	dnv	9.72E-07	6.25E-03	0.194	Yes
KANSL1	94	110	2.00E-04	2	3.38E-04	none	1.19E-06	6.92E-03	0.238	Yes
MYO16	13	2	1.00E-04	2	8.38E-04	cnv	1.45E-06	1.74E-02	0.272	No
MED13L	2	1	2.70E-01	4	4.58E-07	dnv	2.09E-06	1.22E-02	0.064	Yes
KDM5B	0	0	1.00E+00	4	1.45E-07	dnv	2.43E-06	4.97E-02	0.572	No
GATA6	0	0	1.00E+00	5	1.80E-07	dnv	2.98E-06	1.92E-02	0.174	Yes
ARID1B	4	0	1.31E-02	4	1.48E-05	none	3.20E-06	1.87E-02	0.102	No
FEZ1	10	0	1.00E-04	2	2.22E-03	cnv	3.62E-06	4.30E-02	0.414	No

https://doi.org/10.1371/journal.pgen.1009679.t001

A gene was included in the final set of haploinsufficient CHD disease genes if it reached a significant corrected $metaP < 0.05$ (after Bonferroni adjustment with IHW).

Subclassification of CHD phenotypes

We performed a further analysis based on specific CHD subtypes in addition to the collective analysis of all CHD phenotypes. The analysis focused on simplex CHD cases, within two main categories: Conotruncal ($N_{CNV} = 873$, $N_{DNV} = 234$) and LVOTO ($N_{CNV} = 594$, $N_{DNV} = 351$). We only included cases with a clear phenotypic description and without any overlapping phenotypic features between the two categories (LVOTO/Conotruncal). The conotruncal group consisted mostly of Tetralogy of Fallot (TOF), Truncus arteriosus and Transposition of great arteries (TGA), whereas LVOTO mainly constituted Aortic stenosis (AS) including Bicuspid aortic valve disease (BAV), Coarctation and Hypoplastic left heart syndrome (HLHS).

As described above, we performed the same integrative approach for the LVOTO and conotruncal groups to identify CHD subtype-specific genes. Our analysis showed four significant genes (**S11 Table**). Three are observed in LVOTO (*KMT2D*, *KMT2A* and *TAB2*) and a single gene showed significant enrichment in the conotruncal subtype (*NSD1*).

Significant CHD genes are highly and/or differentially expressed in the heart

We next evaluated the expression pattern of the 21 significant genes (Bonferroni corrected $metaP < 0.05$) in the heart using RNA-Seq data from human tissues at different developmental

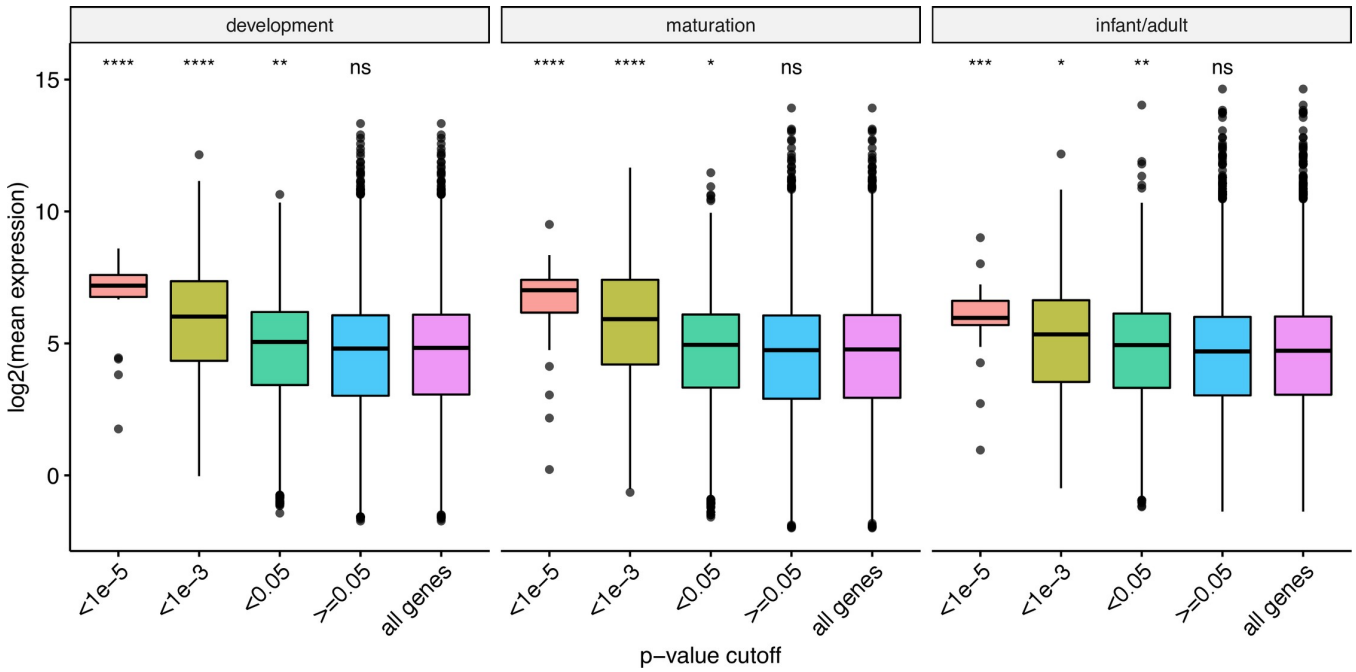


Fig 5. Comparison of the mean expression (heart) distribution at different *metaP* cut-offs. Panels show three different heart development stages: early development, maturation and infant/adult. X-axis denotes the combined p-value from DNV and CNV analysis (*metaP*, at different cut-offs). Y-axis denotes the genes' mean expression in the heart (log scale). The 21 significant candidate CHD genes (Table 1) are contained in the fraction with the higher expression (red box). Differences between the distributions were tested using a two-sided Wilcoxon rank sum test (reference group: all genes). ****: P<0.0001, ***: P<0.001, **: P<0.01, *: P<0.05, ns: non-significant.

https://doi.org/10.1371/journal.pgen.1009679.g005

time points [17]. We stratified the analysis based on stages of heart development (see Materials and Methods). Our analysis revealed that the most significant genes (*metaP* $< 1 \times 10e^{-5}$) show significantly increased mean expression in the heart ($P < 0.0001$, Wilcox test) at different developmental stages (development, maturation and infant/adult), compared to all protein-coding genes (Fig 5). Moreover, 18 out of 21 genes fall in the in the top quartile of heart expression in both developmental and maturation stages (S3 Fig). To complement our expression analysis, we compared gene expression during human heart development with expression in two other mesodermal organs: kidney and liver. This allowed identification of genes with significant changes in its expression levels during crucial heart developmental stages, which would have not been possible when focusing on expression levels alone (Materials and Methods). We found that 17 out of 21 CHD candidate genes are differentially expressed in the heart ($R^2 > 0.50$, Bonferroni corrected $P < 0.01$) when compared to its expression levels in kidney and/or liver. Interestingly, the three genes (*FEZ1*, *NALCN* and *MYO16*) which are not among the highly expressed genes, were found to be significantly differentially expressed during heart development compared to kidney and/or liver (S12 Table).

CNV/DNVs burden of specific protein complexes

CNVs and DNVs can affect heart development either through haploinsufficiency of a single gene, or through its combined impact on the function of several genes. Indeed, oligogenic models have been implicated in CHD, and proteins acting in the same complex or pathway are known to be encoded in genomic clusters [18,19]. We therefore conducted a systems-level analysis to identify global mechanisms by which haploinsufficiency might promote CHD. In

particular, we assessed the combined effect of CNVs and DNVs with respect to human protein-protein interactions (PPIs). The InWeb and ConsesusPathDB databases provides ranked information about experimentally determined physical interactions and, therefore, serves as a proxy to understand the functional effects of CNV/DNVs on human protein complexes (Materials and Methods). The genes with Benjamini–Hochberg adjusted *metaP* < 0.05 (n = 492 genes) were used as seeds to build a PPI network from the data available in InWEb and ConsessusPathDB. No additional interections were considered. The final network consisted of 164 proteins and 290 interactions (**S4 Fig**). A total of 10 overlapping sub-clusters within this network were identified using the in-built clustering algorithm implemented in GeNets [20] (Materials and Methods). Gene-ontology (GO) enrichment analysis suggested that four out of these ten sub-clusters are enriched for genes involved in Notch signaling pathway, cardiocyte differentiation, DNA repair and centrosome function (**Fig 6**). All the four clusters accommodate more CNV deletions in CHD cases compared to controls. Six out of the ten sub-clusters did not show significant enrichment for any particular biological process.

Discussion

We performed a meta-analysis of rare genomic variants in a cohort of 10,447 CHD probands, which provides a useful resource for interpreting CNVs and DNVs identified in patients with CHD. We implemented a statistical approach which allows the integration of different types of genomic variants to discover novel genes associated with CHD. Our data-driven integrative analysis took into account three major criteria at the genomic level: a) gene enrichment for DNVs, b) gene enrichment for CNV deletions and c) gene intolerance for LOF variations. Our analysis identified 21 significant haploinsufficient CHD genes. Fourteen of these are known CHD genes, and the remaining seven genes have not previously been associated with CHD (**Table 1**).

To further strengthen associations, we made use of a newly published human transcriptome atlas covering different developmental, maturation and adult stages in numerous organs [17]. Similar to previous results [7], our analysis highlights that the majority of the 21 significant genes are highly expressed during critical stages of heart development. Unlike earlier studies [7,21] which did not address the importance of expression changes over time, we evaluated the differential expression patterns of genes by comparing levels of expression in the heart, kidney and liver at different time points in development. This analysis allowed us to strengthen disease association for genes not falling under the high expression group and highlight the critical importance of all 21 genes independently of the genomic approach. This aspect is complemented by the fact that the majority of genes (14/21) were already known to cause CHD. To further strengthen disease associations, spatiotemporal expression at single-cell resolution during critical cardiac developmental timepoints and analyses of animal models with targeted mutation in the candidate disease genes is warranted. This could strengthen disease association further and provide pathophysiological information.

Among the 21 likely haploinsufficient disease genes for which the combined analyses showed enrichment (Bonferroni corrected *metaP* < 0.05), 14 genes (*CHD7*, *KMT2D*, *KMT2A*, *NOTCH1*, *NSD1*, *TAB2*, *ANKRD11*, *ADNP*, *DYRK1A*, *RBFOX2*, *KANSL1*, *ELN*, *MED13L* and *GATA6*) are well-established CHD genes, and our data confirms this association. To the best of our knowledge, association between CHD and seven genes (*KDM5B*, *WHSC1*, *WAC*, *NALCN*, *ARID1B*, *FEZ1* and *MYO16)* had either not been established, or had been reported in small cases studies or a single individual only.

KDM5B is not an established CHD gene thus far, although one patient with compound heterozygous frameshift variants had an ASD [14]. While some have argued against the

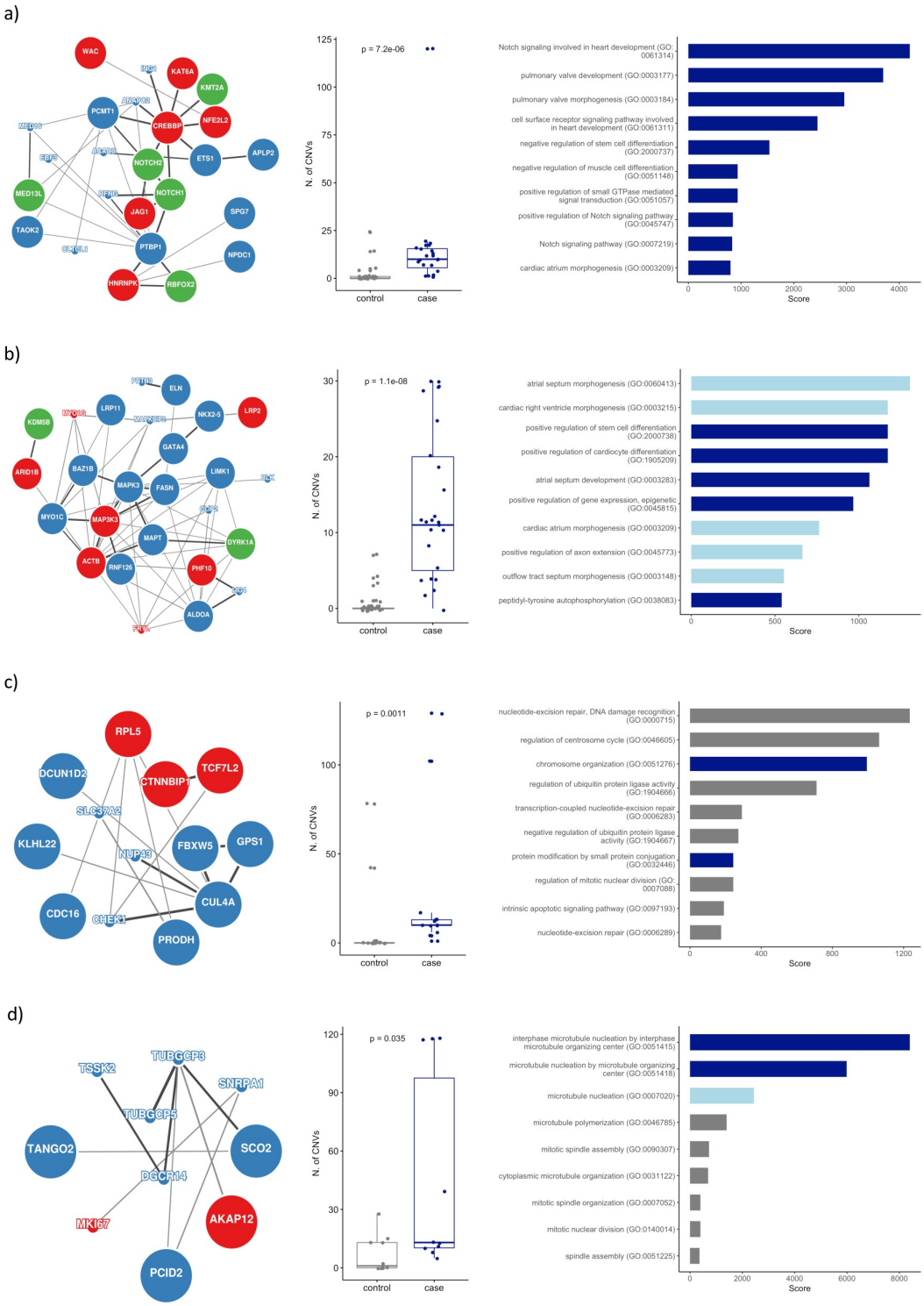


Fig 6. Identification of functional networks enriched for proteins encoded by genes affected by CNVs and/or DNVs associated with CHD. The protein-protein interaction networks (a-d, for clusters 1, 3, 8 and 9 respectively) were identified using GeNets (S4 Fig). Proteins are shown as nodes, interactions as edges. Enrichment for CNVs (blue) and DNVs (green) are highlighted. Proteins

with no specific enrichment for CNV and/or DNVs but with B-H adjusted *metaP* < 0.05 are highlighted in red. The size of the circles denotes if the genes was found significantly highly and/or differentially expressed in the heart (large circles: significant expression; small circles: non-significant). The distribution of CHD case-CNVs and control-CNVs are shown for each cluster. Significant difference in the CNV distribution was calculated using a Wilcox rank sum test. The horizontal bar plots show the top ten GO enriched terms for each cluster (output from Enrichr tool). X-axis in the horizontal bar plot denotes the combined score from Enrichr, which is computed by multiplying the log-transformed p-value and the z-score. Bar color encoded the GO biological process significant level (dark blue: FDR < 5%, light blue: FDR 5–10%, grey: FDR > 10%).

https://doi.org/10.1371/journal.pgen.1009679.g006

haploinsufficiency of the gene [22], our analysis suggests *KDM5B* as a plausible haploinsufficient CHD gene. Additional functional studies are warranted to confirm its role in CHD.

A recent CNV meta-analysis [23] based on non-syndromic CHD patients found that duplication of *WHSC1* (also known as *NSD2*) is a possible cause of CHD. However, haploinsufficiency of *WHSC1* has not previously been associated with CHD. In support of its role in CHD, *Whsc1* has been reported to cause heart malformations in mouse models [24]. In addition, *WHSC1* is known to interact with *NKX2*.5 [24]. In spite of this, the low incidence of CHD in individuals with Wolf-Hirschhorn syndrome suggests that haploinsufficiency of *WHSC1* alone does not cause CHD.

Heterozygous truncating variations in *WAC*, as well as CNV deletions involving this gene, have been recently associated with the DeSanto-Shinawi syndrome, a rare neurodevelopmental disorder characterized by global developmental delay [25,26]. Furthermore, in two non-consanguineous unrelated individuals with heart malformations, among other disorders [27], microdeletions at 10p11.23-p12.1 (overlapping *ARMC4*, *MPP7*, *BAMBI* and *WAC*) were identified. Despite these isolated reports, no definite association between *WAC* and CHD has been established.

DNVs in *NALCN* have been reported to cause a dominant condition characterized by multiple features including developmental delay, congenital contractures of the limbs and face and hypotonia [28,29]. However, among the phenotypes observed, CHD have been not described thus far.

Heterozygous variation of *ARID1B* is a frequent cause of intellectual disability [30,31]. A recent analysis of 143 patients with *ARID1B* variations showed that individuals display a spectrum of clinical characteristics. Congenital heart defects were observed in 19.5% of the patients [32].

FEZ1 is a neurodevelopmental gene, which has been associated with schizophrenia [33]. *Fez1* has been reported to be regulated by *Nkx2-5* in heart progenitors in mice, suggesting a possible role in heart development [34].

MYO16 (*NYAP3*) encodes an unconventional myosin protein, involved in regulation of neuronal morphogenesis [35]. We have not found an association between *MYO16* and heart development in the literature.

Although, several genes have been shown to be altered in syndromic and non-syndromic cases with CHD and TAA (e.g. *HEY2* [36], *MYH11* and *NOTCH1* [37]), among the 10 genes significant in our analysis for TAA, CHD and the combined scenario, none has been reported previously to be associated with either CHD or TAA. Given the limited data size and only accessing CNV calls from TAA cases, future studies looking at CNV and DNV in both phenotypes are required to establish stronger genotype-phenotype correlation to better understand a possibly shared genetic architecture for the two disease entities.

Also, our study did not identify strong signals associated with specific CHD subtypes. We identified four genes with significant enrichment (adjusted *metaP* < 0.05) within the two evaluated CHD subtypes (LVOTO and conotruncal defects). Three were associated with LVOTO, and the contribution was mainly from DNV. *KMT2D* was enriched in LVOTO, which is

consistent with the reported spectrum of CHD in patients with Kabuki syndrome, where a large proportion of individuals have LVOTO type CHD [38,39]. PDA and septal defects predominate in Wiedemann-Steiner Syndrome (*KMT2A*). However, aortic insufficiency and BAV have all been reported [40] suggesting that LVOTO might form part of the phenotypic spectrum. *TAB2* was also enriched in LVOTO. Mutations in *TAB2* are associated with a wide range of cardiac phenotypes [41–43]. Deletions at 6q24 causing haploinsufficiency of *TAB2* have been associated with outflow tract abnormalities, including LVOTO [44,45], which might point to a role of *TAB2* LVOTO pathogenesis. *NSD1* was the only gene identified as significant among conotruncal defects. The significance of this is unclear as septal defects predominate in Sotos Syndrome. However, patients in previous studies were ascertained focusing on Sotos syndrome (OMIM 117550) rather than CHD. Given the current size of each CHD subgroup and the low number of CNV and DNV events in these genes, these results cannot be considered conclusive. More precise phenotypic descriptions for each individual are necessary to increase the genotype-phenotype correlation for individual genes and improve the identification of genetic subnetworks, along with larger sample sizes.

In addition to the gene-centered analysis, we also applied a systems-level analysis in order to identify potential novel pathophysiological mechanisms affected by haploinsufficiency. In this approach, we took advantage of GeNets [20], a computational framework for the analysis of protein-protein interactions, developed for the interpretation of genomic data. Our analysis allowed us to identify PPI clusters enriched for genes affected by CNVs and/or DNV in patients. Furthermore, GO enrichment analysis suggested distinct biological functions for four of these clusters.

Cluster 1 (**Fig 6A**) contains proteins involved in the Notch signaling pathway. Our data corroborate previous studies that confirm the central role of Notch pathway in the pathophysiology of CHD [46] and highlights the shared contribution of CNVs and DNVs within the cluster. Cluster 3 (**Fig 6B**) contains proteins driving essentials processes in the development of the heart such as atrial septum and cardiac right ventricle morphogenesis as well as proteins playing significant role in the positive regulation of gene expression. These mechanisms has been well studied elsewhere [47]. Interestingly, three out of the seven candidate novel CHD genes (*WAC*, *ARID1B* and *KDMB5*) were found to be contributing to these two clusters. Cluster 8 (**Fig 6C**) showed enrichment for processes related with chromosome organization and DNA repair. Association between DNA repair and CHD is not well established thus far. Cluster 9 (**Fig 6D**) was found to be associated with microtubule organizing function. This biological process has been not described in the context of CHD, although an earlier report [48] describes complex CHD among the phenotypes in individuals with 15q11.2 deletion syndrome, which involves the tubulin gamma complex protein 5.

Given the heterogenous data sources and the complex inheritance patterns often observed in patients with CHD, our study has limitations. Firstly, the patient data was collected from almost 200 different published sources, and in many cases it was only possible to obtain data from CNV calls which had already been suggested to be pathogenic. Thus, we are aware that our patient data are incomplete because genome-wide CNV data are missing from a large part of the patient cohort and re-emphasizes already established associations. This is not the case for controls, for which genome-wide data was used. As a direct consequence, even though the difference between the rates of CNV deletions in controls and cases decreased dramatically after applying a quality control filtering step, a slight difference remained between both cohorts. The lack of collected CNV data spanning sex chromosomes limited the analysis only to autosomal chromosomes. In addition, the distribution of CNVs that overlap known microdeletion syndromes such as DiGeorge syndrome and Williams syndrome is overrepresented in the dataset. Similarly, the degree of phenotyping varied across the different studies, and

often only basic phenotypic terms relating to CHD were available. This made it impossible to refine the diagnosis to a precise phenotypic class of CHD in many individuals.

Previous research has shown that the chances of finding a genetic cause of CHD is higher in syndromic, rather than non-syndromic CHD [6,7]. Therefore, it is not surprising that known CHD genes were enriched in this cohort. To identify non-syndromic causes of CHD, it is important to take into account previous findings, which have shown a significant excess of apparently deleterious inherited PTVs in unaffected parents [6]. To help address the challenges of identifying non-syndromic CHD genes, we have used an integrative approach, which has allowed us to look for novel CHD associations in a larger sample size, in a binary fashion. This will help to facilitate future studies.

Variable expression and reduced penetrance are common features in CHD, including in even well-established conditions such as Noonan Syndrome [49]. Most cases of CHD occur as a one off in the family and when recurrences do happen, CHD subtypes are more likely to be discordant than concordant [50]. This suggests that other modifying factors may be at play, including genetic, or environmental factors, or both. This presents a further challenge in identifying genetic causes of CHD. Moving forward, detailed phenotypic descriptions using a standardized system, and better data sharing strategies (including primary data), will facilitate further gene discovery and improved genotype-phenotype correlation in CHD subgroups.

In summary, we have performed an integrative analysis of CNVs, DNVs, o/e LOF ratio and expression during heart development amongst more than 10,000 CHD patients. Our analyses identify seven potential disease genes and mechanisms with novel association with CHD and strengthen previously reported associations.

Materials and methods

Ethics statement

This project was based on unidentifiable data and did not require approval by science ethics committees in Denmark or Christian-Albrechts-Universität in Kiel.

Cohort description

Our cohort contains 7,958 CHD cases and 14,082 controls (see summary at **S1 Table**). Data from both affected and unaffected individuals were collected from 190 different CNV studies (**S2 Table**). Most of the CNV data included in the present study were assembled from public repositories, data available from literature as well as clinical data (see **S2 and S3 Tables** for a more detailed description). We sampled all available and accessible studies as of February 2018 cited in PubMed. We have focused on studies incorporating Caucasian samples (the larger sampled population) to decrease population effects. Given no primary data were available for most of the studies, we cannot exclude the possibility that non-Caucasian samples were included. Both non-syndromic and syndromic CHD cases were included and CNVs were mapped to the human genome build NCBI37/hg19. Phenotype information was reviewed and if possible standardised across studies, to ensure consistency and accuracy. We used as controls re-analysed samples from the Wellcome Trust Case Control Consortium 2, the Genetic Association Network (GAIN) and the Ottawa Heart Institute (**S3 Table**). In addition, we built a dataset from the two largest DNV studies in CHD published thus far, which include a total of 2,489 parent-offspring trios [6,7].

Determining the gene sets

The compiled highly confident CHD gene list consisted of genes with plausible disease-causing mutations in an interpretable functional region of a gene reported in association with CHD in

three or more unrelated individuals. Genes, where fewer than three reports of CHD exist, were also considered if there was available solid functional evidence, such as a mouse model that displays CHD or *in silico* evidence. CHD genes were then coded as either non-syndromic (isolated CHD), or syndromic based on published phenotypes. The list of genes associated with developmental disorders was derived from the Developmental Disorder Genotype to Phenotype (DDG2P) list maintained by Decipher and the European bioinformatics Institute [51]. All genes were annotated with either monoallelic or biallelic as appropriate, based on the published literature (**S13 Table**).

CNV analysis

Only autosomal CNVs were included in the analysis. All the CNV boundaries were determined using genome build NCBI37/hg19. For the CNVs provided in hg18, we used the Assembly Converter (https://www.ensembl.org/Homo_sapiens/Tools/AssemblyConverter) build on CrossMap (http://crossmap.sourceforge.net/) to convert samples to NCBI37/hg19. Also, an extra validation step of all CNV boundaries was performed using the R-package *BSgenome.Hsapiens.UCSC.hg19*. Smaller and longer CNVs were filtered out by applying a size cut-off of 5 Kb and 20 Mb as lower and upper limit, respectively. It has been demonstrated before that smaller and larger CNV calls tend to have a high rate of false positives [52,53]. We removed CNVs overlapping more than 50% of telomeres, centromeres and segmental duplication regions. In addition, we computed the internal CNV frequencies by counting the number of relative overlaps (>50% reciprocal overlap) on the CNVs control subset divided by the total number of controls. The internal MAF was computed for deletions and duplications subsets separately. Only CNVs with a minor allele frequency (MAF) < 0.01 in controls and overlapping ten or more CNV platform calling probes (Affymetrix Array 6.0 and Illumina Human660W-Quad) were considered for downstream analysis. Our analysis was focused only on CNV deletions. The distributions of the number of CNV deletions per individual within the case and control groups were compared (two-sided Wilcoxon rank sum test) to evaluate the impact of the quality control filtering step (**S5 Fig**). After filtering, 6,746 cases (3, 929 harbouring CNV deletions) and 14,024 controls (12,585 harbouring CNV deletions) remained for further analysis. A region-based permutation test (using PLINK version 1.07, test '*—cnv-test-region—mperm 10000*') was used on the filtered set to perform a case-control association analysis. For the gene-based permutation analysis, we reported both the 'point-wise' empirical p-value (EMP1) and the empirical adjusted p-value (EMP2), which controls the family-wise error rate (FWER) (http://zzz.bwh.harvard.edu/plink/perm.shtml). In addition to the gene-centered permutation testing, a similar region-based permutation analysis was performed to access enrichment in known CNV deletion syndromes. All CNVs deletions passing QC filtering overlapping these regions were considered in the analysis. The region genomic coordinates and syndrome descriptions were downloaded from the *Database of genomic variation and phenotype in humans using Ensembl resources* (Decipher, https://decipher.sanger.ac.uk/disorders#syndromes/overview).

CNV burden test on gene sets

A logistic regression-based burden test ('cnv-enrichment-test' in PLINK v1.7) [11] was performed on different gene sets (known CHD genes (non-syndromic/syndromic/biallelic/monoallelic), developmental disorder genes, known haploinsufficient disease genes [54], autosomal recessive disease genes [55,56] and low observed/expected LOF ratio genes, **S13 Table**) using the rare CNV deletions passing the quality control and filtering stage. For every gene set examined, the binary phenotype (CHD case or control) was regressed on the number of genes

disrupted by one or more CNVs. The averaged CNV size and the number of segments per individual were used as covariates into the model to control for potential differences between cases and controls as suggested by Raychaudhuri *et al* [11]. In addition, the PLINK implementation of this test was slightly modified by including a third (categorical) covariate, the sample study ID, since we have assembled the CNV data from different sources.

DNV analysis

The assembled DNV dataset (Sheet A in **S14 Table**) was re-annotated using the Variant Effect Predictor (VEP version 90) tool. All the DNVs included in this study were validated with the VariantValidator tool [57] (Sheet B in **S14 Table**). Based on the VEP annotation, we classified every variation into three major functional groups as follows: a) Protein truncation variant (stop_gained, splice_acceptor, splice_donor, frameshift, initiator_codon, start_lost, conserved_exon_terminus), b) missense variant (stop_lost, missense, inframe_deletion, inframe_insertion, coding_sequence, protein_altering) and c) silent variant (synonymous). Variants with minor allelic frequency (MAF) > 0.01 in gnomAD database were excluded from the analysis. The rates of rare DNVs (MAF < 0.01) in both DNV studies [6,7] were compared (Poisson test) for different variant consequence groups (PTV, missense and synonymous). No significant differences were found between the DNV rates for any of the evaluated groups (**S9 Table**). *De novo* variation recurrence significance testing was performed to evaluate the impact of DNVs at gene level using the Mupit tool [12]. By default, Mupit uses the sequence-specific variation rate published by Samocha *et al* [58]. A second test, DeNovoWEST [13], was used to assess gene-wise *de novo* variation enrichment. DeNovoWEST assigns a variation severity score (based on the variant consequences and the CADD score) to all classes of variants as a proxy of its deleteriousness. For each tested gene, the minimal *p-value* obtained from Mupit and DeNovoWEST was reported (P_{dnv}). The corrected *P* value was computed using the Bonferroni method with n = 18,272.

Inferring differentially and highly expressed genes

Differentially expressed genes (DEGs) were identified by comparing the gene expression profile in heart to kidney and liver at matched time points. We used maSigPro R-package [59] for inferring genes with dynamic temporal profiles from time-course transcriptomic data as previously described by Cardoso-Moreira *et al* [17]. As the input for maSigPro, we used the count per million matrix (CPM, output from EdgeR package) hosted in ArrayExpress (E-MTAB-6814). Genes which did not reach a CPM > 0.5 in at least five samples were excluded from the analysis. We ran maSigPro on the time scale measured in days post-conception using defaults parameters and only included time points with at least two biological replicates. A gene was selected as DEG if the R^2 (goodness-of-fit) parameter was higher than 0.50 and Bonferroni corrected $P < 0.01$. The R^2 parameter distinguish genes with clear expression trends from genes with 'flat' expression profile. **S12 Table** lists the final DEGs identified in the heart ($R^2 > 0.50$). To assess the gene expression levels in the heart, the RPKM matrix was used. Gene expression levels were averaged among samples in the different development stages of the heart as follow: early development (4wpc-8wpc), maturation (9wpc-20wpc), infant/adult (newborn-adulthood). Genes were ranked based on the computed mean expression.

Identification of CNV/DNV enriched PPI sub-clusters

A protein-protein interaction network was constructed using the GeNets framework [20] and the information from InWeb [60] and ConsensusPathDB [61] protein-protein interaction databases. Nodes in the network correspond to proteins whereas edges represent their physical

interactions. The network was strictly seeded with 492 candidate genes, those with a significant adjusted *metaP* < 0.05 (Benjamini-Hochberg's false discovery rate, FDR). The PPI network was partitioned into overlapping sub-clusters using the in-built clustering method described in GeNets [20]. Only statistically significant sub-clusters (p-value < 0.05, permutation test) with at least 5 proteins were considered for further analysis. Finally, Gene Ontology enrichment analysis (Biological Process database 2018) of each identified sub-cluster was performed using the enrichr tool (https://maayanlab.cloud/Enrichr/).

Supporting information

S1 Fig. Distribution of overlapping CNVs, size and genes in 63 CHD loci. A) CNVs per gene. Overlapping CNVs for each of the 528 significant candidate genes are shown as box-and-whiskers plots. Statistically significant difference was observed between the two distributions (Mann-Whitney test, ***: P<0.001). B) Size of loci in kilobase-pairs (kbp). C) Number of genes per locus. Median values are shown above each box.
(TIF)

S2 Fig. Statistical framework to discover novel candidate CHD genes by integrating DNV and CNV deletions. The workflow follows four major steps: 1) Data aggregation and quality control of both DNV data and CNV data, 2) DNV rate-based enrichment testing and CNV deletions case/control association analysis at gene level are performed independently, 3) the results are combined using the Fisher method and 4) P-values are Bonferroni corrected using the Independent Hypothesis Weighting method (IHW). As independent covariate for the IHW method, the o/e LOF ratio upper bound fraction (LOEUF) was used.
(TIF)

S3 Fig. Heart expression pattern of the 21 significant genes at different heart development stages. Panels show three different heart development stages: early development (red), maturation (green) and infant/adult (blue). The x-axis denotes the percentile rank of heart expression in the heart. The y-axis denotes the o/e LOF ratio upper bound fraction (LOEUF) from gnomAD. Dashed lines denote the threshold for highly expressed genes (expression rank > = 0.75) and highly LOF constrained genes (LOEUF < = 0.30).
(TIF)

S4 Fig. The functional network enriched for proteins encoded by genes affected by CNVs and/or DNVs associated with CHD. Ten sub-clusters were identified using GeNets. Proteins are shown as nodes, interactions as edges. Enrichment for CNVs (blue), DNVs (green) or both independently (purple) are highlighted. Proteins with no specific enrichment for CNV and/or DNVs but with B-H adjusted *metaP* < 0.05 are highlighted in red. The size of the circles denotes if the gene was found significantly highly and/or differentially expressed in the heart (large circles: significant expression; small circles: non-significant).
(TIF)

S5 Fig. Distribution of the number of CNV deletions per individual in both control and CHD case cohorts before (A) and after (B) applying the quality control filtering approach. Differences between the distributions were tested using a two-sided Wilcoxon rank sum test. ****: P<0.0001.
(TIF)

S1 Table. Number of probands in the CNV cohort. Stratified by CHD cases/controls and CNV type (deletion/duplication).
(XLSX)

S2 Table. Sources of the CNV-case cohorts used in this study.
(XLSX)

S3 Table. Sources of the CNV-control cohorts used in this study.
(XLSX)

S4 Table. Gene set-based logistical regression enrichment CNV analysis.
(XLSX)

S5 Table. Gene set-based logistical regression enrichment CNV analysis stratified by observed/expected LOF ratio (from gnomAD).
(XLSX)

S6 Table. Sheet A) Gene-based case/control CNV-deletions permutation testing (PLINK results). Sheet B) Significant locus (Locus ID and contributing genes).
(XLSX)

S7 Table. Case/control CNV permutation testing on known deletion syndrome (PLINK results).
(XLSX)

S8 Table. CNV case/control permuatation testing output from PLINK. The table shows the 528 significant genes combining both CHD and TAA cases (CHD+TAA), the contribution of only CHD cases (only CHD) and the contribution of only TAA cases (only TAA).
(XLSX)

S9 Table. Comparision of the DNV rates, stratified by variant consequences, between two independent cohorts.
(XLSX)

S10 Table. Gene-based DNV analysis.
(XLSX)

S11 Table. Meta-analysis of CNV/DNV stratified by CHD sub-types (Conotruncal and LVOTO). Table shows the four genes with Bonferroni corrected metaP < 0.05. Cases/Controls: Number of cases and controls carrying CNV deletions overlapping the gene in the CNV analysis. *p_cnv*: p-value from the CNV permutation test. nsDNV: Number of constrained non-synonymous variations in the *de novo* analysis. *p_dnv*: p-value from the DNV analysis. *metaP*: combined p-value (P_{dnv} and P_{cnv}) using the Fisher method. adj meta*P*: Bonferroni corrected p-value using independent hypothesis weighting (IHW) and LOEUF metric from gnomad as covariate.
(XLSX)

S12 Table. Differentially expressed genes in the heart compared to kidney and liver.
(XLSX)

S13 Table. List of gene sets used in the CNV enrichment analysis.
(XLSX)

S14 Table. Sheet A) List of the *de novo* variants analized in this study. Sheet B) Validation results of the DNVs from the VarinatValidator tool.
(XLSX)

S15 Table. Rare CNV deletions (MAF 0.01) used in this study. CNVs bounderies are determinated in genome build hg19.
(XLSX)

Acknowledgments

We express our gratitude to the patients and their families for their participation in the analysed studies. We would like to thank the Genetic Association Information Network (GAIN) and dbGAP for making the data available. We would like to thank the Wellcome Trust Case Control Consortium (WTCCC) for making the data accessible https://www.wtccc.org.uk/info/access_to_data_samples.html. We used data from the the Deciphering Developmental Disorders (DDD) study. The DDD study presents independent research commissioned by the Health Innovation Challenge Fund, a parallel funding partnership between the Wellcome Trust and the UK Department of Health, and the Wellcome Trust Sanger Institute. The views expressed in this publication are those of the author(s) and not necessarily those of the Wellcome Trust or the UK Department of Health. We would like to thank the Pediatric Cardiac Genomics Consortium (PCGC) and dbGAP for making the data publicly available. This study makes use of data generated by the DECIPHER community. A full list of centres who contributed to the generation of the data is available from http://decipher.sanger.ac.uk and via email from decipher@sanger.ac.uk. We thanks to Margarida C. Moreira and the Kaessmann Lab by making accessible the human RNA-Seq data and the support for the data analysis. We thanks Rasmus Wernersson and Federico de Masi for facilitating the use of the protein-protein interaction database, InWeb, in this work. We thanks the Lage Lab and Taibo Li for their support with GeNets. And finally, we would like to thank all data submitters and collaborators for their contributions.

Author Contributions

Conceptualization: Enrique Audain, Matthew E. Hurles, Bernard Thienpont, Lars Allan Larsen, Marc-Phillip Hitz.

Data curation: Enrique Audain, Anna Wilsdon, Jeroen Breckpot, Jose M. G. Izarzugaza, Anne-Karin Kahlert, Hashim Abdul-Khaliq, Mads Bak, Anne S. Bassett, Woodrow D. Benson, Felix Berger, Ingo Daehnert, Koenraad Devriendt, Sven Dittrich, Piers EF Daubeney, Vidu Garg, Karl Hackmann, Kirstin Hoff, Philipp Hofmann, Gregor Dombrowsky, Thomas Pickardt, Bernard D. Keavney, Sabine Klaassen, Christian R. Marshall, Dianna M. Milewicz, Scott Lemaire, Joseph S. Coselli, Michael E. Mitchell, Aoy Tomita-Mitchell, Siddharth K. Prakash, Karl Stamm, Alexandre F. R. Stewart, Candice K. Silversides, Reiner Siebert, Brigitte Stiller, Jill A. Rosenfeld, Inga Vater, Alex V. Postma, Almuth Caliebe, Lars Allan Larsen, Marc-Phillip Hitz.

Formal analysis: Enrique Audain, Jose M. G. Izarzugaza, Alejandro Sifrim, Lars Allan Larsen, Marc-Phillip Hitz.

Funding acquisition: Hans-Heiner Kramer, Matthew E. Hurles.

Investigation: Anna Wilsdon, Jeroen Breckpot.

Methodology: Tomas W. Fitzgerald, Alejandro Sifrim, Florian Wünnemann, Yasset Perez-Riverol, Gregor Andelfinger.

Resources: Thomas Pickardt, Ulrike Bauer, Hans-Heiner Kramer.

Writing – original draft: Enrique Audain, Anna Wilsdon, Jill A. Rosenfeld.

Writing – review & editing: Enrique Audain, Jeroen Breckpot, Tomas W. Fitzgerald, Anne-Karin Kahlert, Florian Wünnemann, Yasset Perez-Riverol, Hashim Abdul-Khaliq, Mads Bak, Anne S. Bassett, Woodrow D. Benson, Felix Berger, Ingo Daehnert, Koenraad Devriendt, Vidu Garg, Karl Hackmann, Kirstin Hoff, Philipp Hofmann, Gregor Dombrowsky, Bernard D. Keavney, Sabine Klaassen, Hans-Heiner Kramer, Christian R. Marshall, Dianna M. Milewicz, Scott Lemaire, Joseph S. Coselli, Michael E. Mitchell, Aoy Tomita-Mitchell, Siddharth K. Prakash, Karl Stamm, Alexandre F. R. Stewart, Candice K. Silversides, Reiner Siebert, Brigitte Stiller, Inga Vater, Alex V. Postma, Almuth Caliebe, J. David Brook, Gregor Andelfinger, Matthew E. Hurles, Bernard Thienpont, Lars Allan Larsen, Marc-Phillip Hitz.

References

1. van der Linde D, Konings EEM, Slager MA, Witsenburg M, Helbing WA, Takkenberg JJM, et al. Birth Prevalence of Congenital Heart Disease Worldwide. J Am Coll Cardiol. 2011; 58: 2241–2247. https://doi.org/10.1016/j.jacc.2011.08.025 PMID: 22078432
2. van der Bom T, Zomer AC, Zwinderman AH, Meijboom FJ, Bouma BJ, Mulder BJM. The changing epidemiology of congenital heart disease. Nat Rev Cardiol. 2011; 8: 50–60. https://doi.org/10.1038/nrcardio.2010.166 PMID: 21045784
3. Warnes CA, Botto L, Correa A, Britt AE, Elixson M, Jenkins KJ, et al. Noninherited Risk Factors and Congenital Cardiovascular Defects: Current Knowledge. Circulation. 2007; 115: 2995–3014. https://doi.org/10.1161/CIRCULATIONAHA.106.183216 PMID: 17519397
4. Zaidi S, Brueckner M. Genetics and Genomics of Congenital Heart Disease. Circ Res. 2017; 120: 923–940. https://doi.org/10.1161/CIRCRESAHA.116.309140 PMID: 28302740
5. Andersen TA, Troelsen KDLL, Larsen LA. Of mice and men: Molecular genetics of congenital heart disease. Cell Mol Life Sci. 2014; 71: 1327–1352. https://doi.org/10.1007/s00018-013-1430-1 PMID: 23934094
6. Sifrim A, Hitz M-P, Wilsdon A, Breckpot J, Turki SH Al, Thienpont B, et al. Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing. Nat Genet. 2016; 48: 1060–5. https://doi.org/10.1038/ng.3627 PMID: 27479907
7. Jin SC, Homsy J, Zaidi S, Lu Q, Morton S, DePalma SR, et al. Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. Nat Genet. 2017. https://doi.org/10.1038/ng.3970 PMID: 28991257
8. Tan TY, Gonzaga-Jauregui C, Bhoj EJ, Strauss KA, Brigatti K, Puffenberger E, et al. Monoallelic BMP2 Variants Predicted to Result in Haploinsufficiency Cause Craniofacial, Skeletal, and Cardiac Features Overlapping Those of 20p12 Deletions. Am J Hum Genet. 2017; 101: 985–994. https://doi.org/10.1016/j.ajhg.2017.10.006 PMID: 29198724
9. Glessner JT, Bick AG, Ito K, Homsy J, Rodriguez-Murillo L, Fromer M, et al. Increased frequency of de novo copy number variants in congenital heart disease by integrative analysis of single nucleotide polymorphism array and exome sequence data. Circ Res. 2014; 115: 884–896. https://doi.org/10.1161/CIRCRESAHA.115.304458 PMID: 25205790
10. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. bioRxiv. 2019; 531210. https://doi.org/10.1101/531210
11. Raychaudhuri S, Korn JM, McCarroll SA, International Schizophrenia Consortium, Altshuler D, Sklar P, et al. Accurately assessing the risk of schizophrenia conferred by rare copy-number variation affecting genes with brain function. Allison DB, editor. PLoS Genet. 2010; 6: e1001097. https://doi.org/10.1371/journal.pgen.1001097 PMID: 20838587
12. McRae JF, Clayton S, Fitzgerald TW, Kaplanis J, Prigmore E, Rajan D, et al. Prevalence and architecture of de novo mutations in developmental disorders. Nature. 2017; 542: 433–438. https://doi.org/10.1038/nature21062 PMID: 28135719
13. Kaplanis J, Samocha KE, Wiel L, Zhang Z, Kevin J, Eberhardt RY, et al. Integrating healthcare and research genetic data empowers the discovery of 49 novel developmental disorders. bioRxiv. 2019; 797787. https://doi.org/10.1101/797787
14. Faundes V, Newman WG, Bernardini L, Canham N, Clayton-Smith J, Dallapiccola B, et al. Histone Lysine Methylases and Demethylases in the Landscape of Human Developmental Disorders. Am J Hum Genet. 2018; 102: 175–187. https://doi.org/10.1016/j.ajhg.2017.11.013 PMID: 29276005

15. Martin HC, Jones WD, McIntyre R, Sanchez-Andrade G, Sanderson M, Stephenson JD, et al. Quantifying the contribution of recessive coding variation to developmental disorders. Science. 2018; 362: 1161–1164. https://doi.org/10.1126/science.aar6731 PMID: 30409806

16. Ignatiadis N, Klaus B, Zaugg JB, Huber W. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. Nat Methods. 2016; 13: 577–80. https://doi.org/10.1038/nmeth.3885 PMID: 27240256

17. Cardoso-Moreira M, Halbert J, Valloton D, Velten B, Chen C, Shao Y, et al. Gene expression across mammalian organ development. Nature. 2019; 571: 505–509. https://doi.org/10.1038/s41586-019-1338-5 PMID: 31243369

18. Priest JR, Osoegawa K, Mohammed N, Nanda V, Kundu R, Schultz K, et al. De Novo and Rare Variants at Multiple Loci Support the Oligogenic Origins of Atrioventricular Septal Heart Defects. Jiao K, editor. PLoS Genet. 2016; 12: e1005963. https://doi.org/10.1371/journal.pgen.1005963 PMID: 27058611

19. Reuter MS, Jobling R, Chaturvedi RR, Manshaei R, Costain G, Heung T, et al. Haploinsufficiency of vascular endothelial growth factor related signaling genes is associated with tetralogy of Fallot. Genet Med. 2019; 21: 1001–1007. https://doi.org/10.1038/s41436-018-0260-9 PMID: 30232381

20. Li T, Kim A, Rosenbluh J, Horn H, Greenfeld L, An D, et al. GeNets: a unified web platform for network-based genomic analyses. Nat Methods. 2018; 15: 543–546. https://doi.org/10.1038/s41592-018-0039-6 PMID: 29915188

21. Sevim Bayrak C, Zhang P, Tristani-Firouzi M, Gelb BD, Itan Y. De novo variants in exomes of congenital heart disease patients identify risk genes and pathways. Genome Med. 2020; 12: 9. https://doi.org/10.1186/s13073-019-0709-8 PMID: 31941532

22. Lebrun N, Mehler-Jacob C, Poirier K, Zordan C, Lacombe D, Carion N, et al. Novel KDM5B splice variants identified in patients with developmental disorders: Functional consequences. Gene. 2018; 679: 305–313. https://doi.org/10.1016/j.gene.2018.09.016 PMID: 30217758

23. Fotiou E, Williams S, Martin-Geary A, Robertson DL, Tenin G, Hentges KE, et al. Integration of Large-Scale Genomic Data Sources With Evolutionary History Reveals Novel Genetic Loci for Congenital Heart Disease. Circ Genomic Precis Med. 2019; 12: 442–451. https://doi.org/10.1161/CIRCGEN.119.002694 PMID: 31613678

24. Nimura K, Ura K, Shiratori H, Ikawa M, Okabe M, Schwartz RJ, et al. A histone H3 lysine 36 trimethyltransferase links Nkx2-5 to Wolf-Hirschhorn syndrome. Nature. 2009; 460: 287–291. https://doi.org/10.1038/nature08086 PMID: 19483677

25. DeSanto C, K D'Aco, Araujo GC, Shannon N, DDD Study, Vernon H, et al. WAC loss-of-function mutations cause a recognisable syndrome characterised by dysmorphic features, developmental delay and hypotonia and recapitulate 10p11.23 microdeletion syndrome. J Med Genet. 2015; 52: 754–61. https://doi.org/10.1136/jmedgenet-2015-103069 PMID: 26264232

26. Wentzel C, Rajcan-Separovic E, Ruivenkamp CAL, Chantot-Bastaraud S, Metay C, Andrieux J, et al. Genomic and clinical characteristics of six patients with partially overlapping interstitial deletions at 10p12p11. Eur J Hum Genet. 2011; 19: 959–964. https://doi.org/10.1038/ejhg.2011.71 PMID: 21522184

27. Okamoto N, Hayashi S, Masui A, Kosaki R, Oguri I, Hasegawa T, et al. Deletion at chromosome 10p11.23-p12.1 defines characteristic phenotypes with marked midface retrusion. J Hum Genet. 2012; 57: 191–6. https://doi.org/10.1038/jhg.2011.154 PMID: 22258158

28. Chong JX, McMillin MJ, Shively KM, Beck AE, Marvin CT, Armenteros JR, et al. De novo mutations in NALCN cause a syndrome characterized by congenital contractures of the limbs and face, hypotonia, and developmental delay. Am J Hum Genet. 2015; 96: 462–73. https://doi.org/10.1016/j.ajhg.2015.01.003 PMID: 25683120

29. Fukai R, Saitsu H, Okamoto N, Sakai Y, Fattal-Valevski A, Masaaki S, et al. De novo missense mutations in NALCN cause developmental and intellectual impairment with hypotonia. J Hum Genet. 2016; 61: 451–5. https://doi.org/10.1038/jhg.2015.163 PMID: 26763878

30. Hoyer J, Ekici AB, Endele S, Popp B, Zweier C, Wiesener A, et al. Haploinsufficiency of ARID1B, a member of the SWI/SNF-A chromatin-remodeling complex, is a frequent cause of intellectual disability. Am J Hum Genet. 2012; 90: 565–572. https://doi.org/10.1016/j.ajhg.2012.02.007 PMID: 22405089

31. Fitzgerald TW, Gerety SS, Jones WD, Van Kogelenberg M, King DA, McRae J, et al. Large-scale discovery of novel genetic causes of developmental disorders. Nature. Nature Publishing Group; 2015. pp. 223–228. https://doi.org/10.1038/nature14135 PMID: 25533962

32. van der Sluijs PJ, Jansen S, Vergano SA, Adachi-Fukuda M, Alanay Y, AlKindy A, et al. The ARID1B spectrum in 143 patients: from nonsyndromic intellectual disability to Coffin-Siris syndrome. Genet Med. 2019; 21: 1295–1307. https://doi.org/10.1038/s41436-018-0330-z PMID: 30349098

33. Kang E, Burdick KE, Kim JY, Duan X, Guo JU, Sailor KA, et al. Interaction between FEZ1 and DISC1 in regulation of neuronal development and risk for schizophrenia. Neuron. 2011; 72: 559–71. https://doi.org/10.1016/j.neuron.2011.09.032 PMID: 22099459

34. Barth JL, Clark CD, Fresco VM, Knoll EP, Lee B, Argraves WS, et al. Jarid2 is among a set of genes differentially regulated by Nkx2.5 during outflow tract morphogenesis. Dev Dyn. 2010; 239: 2024–33. https://doi.org/10.1002/dvdy.22341 PMID: 20549724

35. Yokoyama K, Tezuka T, Kotani M, Nakazawa T, Hoshina N, Shimoda Y, et al. NYAP: a phosphoprotein family that links PI3K to WAVE1 signalling in neurons. EMBO J. 2011; 30: 4739–54. https://doi.org/10.1038/emboj.2011.348 PMID: 21946561

36. van Walree ES, Dombrowsky G, Jansen IE, Mirkov MU, Zwart R, Ilgun A, et al. Germline variants in HEY2 functional domains lead to congenital heart defects and thoracic aortic aneurysms. Genet Med. 2021; 23: 103–110. https://doi.org/10.1038/s41436-020-00939-4 PMID: 32820247

37. Lindsay ME, Dietz HC. The genetic basis of aortic aneurysm. Cold Spring Harb Perspect Med. 2014; 4. https://doi.org/10.1101/cshperspect.a015909 PMID: 25183854

38. Digilio MC, Gnazzo M, Lepri F, Dentici ML, Pisaneschi E, Baban A, et al. Congenital heart defects in molecularly proven Kabuki syndrome patients. Am J Med Genet Part A. 2017; 173: 2912–2922. https://doi.org/10.1002/ajmg.a.38417 PMID: 28884922

39. Yuan SM. Congenital heart defects in Kabuki syndrome. Cardiology Journal. Cardiol J; 2013. pp. 121–124. https://doi.org/10.5603/CJ.2013.0023 PMID: 23558868

40. Baer S, Afenjar A, Smol T, Piton A, Gérard B, Alembik Y, et al. Wiedemann-Steiner syndrome as a major cause of syndromic intellectual disability: A study of 33 French cases. Clin Genet. 2018; 94: 141–152. https://doi.org/10.1111/cge.13254 PMID: 29574747

41. Permanyer E, Laurie S, Blasco-Lucas A, Maldonado G, Amador-Catalan A, Ferrer-Curriu G, et al. A single nucleotide deletion resulting in a frameshift in exon 4 of TAB2 is associated with a polyvalular syndrome. Eur J Med Genet. 2020;63. https://doi.org/10.1016/j.ejmg.2020.103854 PMID: 31981616

42. Ackerman JP, Smestad JA, Tester DJ, Qureshi MY, Crabb BA, Mendelsohn NJ, et al. Whole Exome Sequencing, Familial Genomic Triangulation, and Systems Biology Converge to Identify a Novel Nonsense Mutation in TAB2-encoded TGF-beta Activated Kinase 1 in a Child with Polyvalvular Syndrome. Congenit Heart Dis. 2016; 11: 452–461. https://doi.org/10.1111/chd.12400 PMID: 27452334

43. Ritelli M, Morlino S, Giacopuzzi E, Bernardini L, Torres B, Santoro G, et al. A recognizable systemic connective tissue disorder with polyvalvular heart dystrophy and dysmorphism associated with TAB2 mutations. Clin Genet. 2018; 93: 126–133. https://doi.org/10.1111/cge.13032 PMID: 28386937

44. Thienpont B, Zhang L, Postma A V, Breckpot J, Tranchevent LC, Van Loo P, et al. Haploinsufficiency of TAB2 Causes Congenital Heart Defects in Humans. Am J Hum Genet. 2010; 86: 839–849. https://doi.org/10.1016/j.ajhg.2010.04.011 PMID: 20493459

45. Cheng A, Neufeld-Kaiser W, Byers PH, Liu YJ. 6q25.1 (TAB2) microdeletion is a risk factor for hypoplastic left heart: A case report that expands the phenotype. BMC Cardiovasc Disord. 2020;20. https://doi.org/10.1186/s12872-020-01328-0 PMID: 31952508

46. Chapman G, Moreau JLM, I P E, Szot JO, Iyer KR, Shi H, et al. Functional genomics and gene-environment interaction highlight the complexity of congenital heart disease caused by Notch pathway variants. Hum Mol Genet. 2020; 29: 566–579. https://doi.org/10.1093/hmg/ddz270 PMID: 31813956

47. Tomita-Mitchell A, Maslen CL, Morris CD, Garg V, Goldmuntz E. GATA4 sequence variants in patients with congenital heart disease. J Med Genet. 2007; 44: 779–83. https://doi.org/10.1136/jmg.2007.052183 PMID: 18055909

48. Doornbos M, Sikkema-Raddatz B, Ruijvenkamp CAL, Dijkhuizen T, Bijlsma EK, Gijsbers ACJ, et al. Nine patients with a microdeletion 15q11.2 between breakpoints 1 and 2 of the Prader-Willi critical region, possibly associated with behavioural disturbances. Eur J Med Genet. 2009; 52: 108–115. https://doi.org/10.1016/j.ejmg.2009.03.010 PMID: 19328872

49. Roberts AE, Allanson JE, Tartaglia M, Gelb BD. Noonan syndrome. The Lancet. Elsevier B.V.; 2013. pp. 333–342. https://doi.org/10.1016/S2213-8587(13)70153-0 PMID: 24703051

50. Øyen N, Poulsen G, Boyd HA, Wohlfahrt J, Jensen PKA, Melbye M. Recurrence of congenital heart defects in families. Circulation. 2009; 120: 295–301. https://doi.org/10.1161/CIRCULATIONAHA.109.857987 PMID: 19597048

51. Wright CF, Fitzgerald TW, Jones WD, Clayton S, McRae JF, Van Kogelenberg M, et al. Genetic diagnosis of developmental disorders in the DDD study: A scalable analysis of genome-wide research data. Lancet. 2015; 385: 1305–1314. https://doi.org/10.1016/S0140-6736(14)61705-0 PMID: 25529582

52. Zhang D, Qian Y, Akula N, Alliey-Rodriguez N, Tang J, Bipolar Genome Study, et al. Accuracy of CNV Detection from GWAS Data. PLoS One. 2011; 6: e14511. https://doi.org/10.1371/journal.pone.0014511 PMID: 21249187

53. Singh T, Walters JTR, Johnstone M, Curtis D, Suvisaari J, Torniainen M, et al. The contribution of rare variants to risk of schizophrenia in individuals with and without intellectual disability. Nat Genet. 2017; 49: 1167–1173. https://doi.org/10.1038/ng.3903 PMID: 28650482

54. Rehm HL, Berg JS, Brooks LD, Bustamante CD, Evans JP, Landrum MJ, et al. ClinGen—The clinical genome resource. N Engl J Med. 2015; 372: 2235–2242. https://doi.org/10.1056/NEJMsr1406261 PMID: 26014595

55. Blekhman R, Man O, Herrmann L, Boyko AR, Indap A, Kosiol C, et al. Natural selection on genes that underlie human disease susceptibility. Curr Biol. 2008; 18: 883–9. https://doi.org/10.1016/j.cub.2008.04.074 PMID: 18571414

56. Berg JS, Adams M, Nassar N, Bizon C, Lee K, Schmitt CP, et al. An informatics approach to analyzing the incidentalome. Genet Med. 2013; 15: 36–44. https://doi.org/10.1038/gim.2012.112 PMID: 22995991

57. Freeman PJ, Hart RK, Gretton LJ, Brookes AJ, Dalgleish R. VariantValidator: Accurate validation, mapping, and formatting of sequence variation descriptions. Hum Mutat. 2018; 39: 61–68. https://doi.org/10.1002/humu.23348 PMID: 28967166

58. Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM, et al. A framework for the interpretation of de novo mutation in human disease. Nat Genet. 2014; 46: 944–950. https://doi.org/10.1038/ng.3050 PMID: 25086666

59. Nueda MJ, Tarazona S, Conesa A. Next maSigPro: updating maSigPro bioconductor package for RNA-seq time series. Bioinformatics. 2014; 30: 2598–602. https://doi.org/10.1093/bioinformatics/btu333 PMID: 24894503

60. Li T, Wernersson R, Hansen RB, Horn H, Mercer J, Slodkowicz G, et al. A scored human protein-protein interaction network to catalyze genomic interpretation. Nat Methods. 2017; 14: 61–64. https://doi.org/10.1038/nmeth.4083 PMID: 27892958

61. Herwig R, Hardt C, Lienhard M, Kamburov A. Analyzing and interpreting genome data at the network level with ConsensusPathDB. Nat Protoc. 2016; 11: 1889–907. https://doi.org/10.1038/nprot.2016.117 PMID: 27606777

62. Dilg D, Saleh RNM, Phelps SEL, Rose Y, Dupays L, Murphy C, et al. HIRA is required for heart development and directly regulates Tnni2 and Tnnt3. Fraidenraich D, editor. PLoS One. 2016; 11: e0161096. https://doi.org/10.1371/journal.pone.0161096 PMID: 27518902

Chapter IV. Systems genetics analysis identifies calcium-signalling defects as novel cause of congenital heart disease

The original peer-reviewed publication presented in this chapter (pages 79-91) is publicly available at *https://doi.org/10.1186/s13073-020-00772-z*.

Izarzugaza et al. Genome Medicine (2020) 12:76
https://doi.org/10.1186/s13073-020-00772-z

Genome Medicine

RESEARCH Open Access

Systems genetics analysis identifies calcium-signaling defects as novel cause of congenital heart disease

Jose M. G. Izarzugaza[1†], Sabrina G. Ellesøe[2†], Canan Doganli[3†], Natasja Spring Ehlers[1], Marlene D. Dalgaard[1,4], Enrique Audain[5], Gregor Dombrowsky[5], Karina Banasik[2], Alejandro Sifrim[6,7], Anna Wilsdon[8], Bernard Thienpont[7], Jeroen Breckpot[7,9], Marc Gewillig[10], Competence Network for Congenital Heart Defects, Germany, J. David Brook[8], Marc-Phillip Hitz[5,6,11], Lars A. Larsen[3*] and Søren Brunak[2*]

Abstract

Background: Congenital heart disease (CHD) occurs in almost 1% of newborn children and is considered a multifactorial disorder. CHD may segregate in families due to significant contribution of genetic factors in the disease etiology. The aim of the study was to identify pathophysiological mechanisms in families segregating CHD.

Methods: We used whole exome sequencing to identify rare genetic variants in ninety consenting participants from 32 Danish families with recurrent CHD. We applied a systems biology approach to identify developmental mechanisms influenced by accumulation of rare variants. We used an independent cohort of 714 CHD cases and 4922 controls for replication and performed functional investigations using zebrafish as in vivo model.

Results: We identified 1785 genes, in which rare alleles were shared between affected individuals within a family. These genes were enriched for known cardiac developmental genes, and 218 of these genes were mutated in more than one family. Our analysis revealed a functional cluster, enriched for proteins with a known participation in calcium signaling. Replication in an independent cohort confirmed increased mutation burden of calcium-signaling genes in CHD patients. Functional investigation of zebrafish orthologues of *ITPR1*, *PLCB2*, and *ADCY2* verified a role in cardiac development and suggests a combinatorial effect of inactivation of these genes.

Conclusions: The study identifies abnormal calcium signaling as a novel pathophysiological mechanism in human CHD and confirms the complex genetic architecture underlying CHD.

Keywords: Congenital heart disease, Genetics, Whole exome sequencing, Developmental biology, Systems biology, Calcium signaling

* Correspondence: larsal@sund.ku.dk; soren.brunak@cpr.ku.dk
†Jose M. G. Izarzugaza, Sabrina G. Ellesøe and Canan Doganli contributed equally to this work.
[3]Department of Cellular and Molecular Medicine, University of Copenhagen, Blegdamsvej 3A, DK-2200 Copenhagen, Denmark
[2]Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Blegdamsvej 3A, DK-2200 Copenhagen, Denmark
Full list of author information is available at the end of the article

Background

Congenital heart disease (CHD) represents malformations of the heart or intra-thoracic vessels, which affect cardiac function and occur in almost 1% of live births [1]. Most patients survive until adulthood, and the number of adults with CHD has increased to above 3 million in Europe and the USA alone [2]. These patients are challenged by serious cardiovascular complications, which require specialized care, and their children have significantly increased CHD risk [3].

Although the specific cause of CHD is unknown for most patients, genetic factors contribute significantly to the etiology (reviewed in [3, 4]).

Families presenting with recurrent cases suggest that mutations with large effects segregate with CHD in such families. Recent targeted next generation sequencing (NGS) efforts have shown that causative mutations may be identified in one third of CHD families [5]. In the majority of familial cases, the unidentified pathogenic variants may be explained by a "burden of genetic variation" model, which hypothesizes that CHD may occur if the developing embryo is exposed to a certain burden of rare and common genetic variants, possibly in combination with epigenetic, environmental, or stochastic effects [4].

Genomic information obtained by NGS analyses may further clarify genetic and molecular mechanisms causing CHD and thus contribute to improved health care and counseling of the patients and families. However, because of the complex genetic architecture of CHD, more sophisticated methods for interpretation of genetic variation need to be developed before such information may be translated into clinical use.

Here, we performed whole exome sequencing (WES) in a cohort of 32 Danish families in which several family members presented with CHD. Utilizing a systems biology approach, we discovered recurrent mutation of genes involved in calcium signaling. Loss-of-function analyses of three of the genes confirmed their crucial role in embryonic heart development.

Methods

Patient material

CHD families were identified through the Danish National Patient Registry (DNPR) and contacted by letter. Clinical diagnoses were validated by manual review of the patient files, and detailed pedigrees were constructed based on interview with patients (Additional file 1: Fig. S1). There was no known consanguinity in the families. All family members presenting with CHD were non-syndromic, except III.1 in family 545, who had a diagnosis of Turner syndrome. Cardiac malformations are listed below affected family members in Additional file 1: Fig. S1. The age at diagnosis of affected family members ranged from 1 day to 65 years (median age, 1 year; Q1, 29 days; Q3, 8 years). Clinical evaluation for CHD was not performed on asymptomatic family members. Termination of pregnancy was performed in fewer than 5 cases (data not shown).

DNA samples were obtained from a cohort consisting of 90 individuals in 32 families. Affected family members were not screened for 22q11 deletions, but index patients were screened for mutations in the *NKX2-5* gene prior to WES. A frameshift mutation (c.112delG) was found to segregate with ASD in a single family [6]. This family was excluded from the cohort prior to exome sequencing.

Relatedness of family members was determined using the algorithm implemented in VCFtools. Pairwise relatedness of all 90 individuals is shown in Additional file 1: Fig. S2A. Relatedness between individuals not belonging to the same family peaks around zero, confirming that the families are unrelated. Individuals within families have positive relatedness score, confirming that they are related (Additional file 1: Fig. S2B).

The self-reported ethnicity of the individuals in our cohort was Danish. To verify the ethnic homogeneity of the 90 samples, we used the algorithm for detection of outliers implemented in PLINK. For each individual in our cohort, we calculated the population distance to the 20 nearest neighbors using an identity-by-state (IBS) matrix based on 208,163 variants. These distances were compared to the mean of the population in terms of standard deviations (*Z*-scores). This test showed that no samples had *Z*-scores below − 2.5 confirming the ethnic homogeneity of the samples (Additional file 1: Fig. S3). Principal component analysis (PCA) was performed using FlashPCA (https://github.com/gabraham/flashpca). PCA was based on a selected set of common (MAF > 5%) high-quality genetic variants that overlap with the 1000 Genomes data [7]. The Danish samples which were analyzed together with reference samples of different Super Populations from 1000 Genomes appear similar to European populations (Additional file 1: Fig. S4). We identified one outlier sample, but this particular individual had two siblings with similar type of CHD and was not excluded from the study.

Whole exome sequencing

Ninety exomes corresponding to 79 patients and 11 obligate carriers were sequenced. Capture followed the protocol corresponding to an Agilent Sure Select exome v4 kit, and 100-bp paired-end sequencing reads were generated on an Illumina Hiseq 2000 machine. Both exome capture and sequencing were performed at BGI Europe's facility in Copenhagen. The average sequencing coverage was 91.8×. Eighty-six percent of the samples had a sequence coverage of 30× or higher in 80% of the exome. Cumulative sequencing coverage is shown in

Additional file 1: Fig. S5A. Mean number of sequencing reads per sample was 64,396,134 (range 47,821,039–85, 327,021). The distribution of sequencing reads is shown in Additional file 1: Fig. S5B.

Variants were stored in a mySQL database to facilitate filtering and comparison between both individuals and families. Population-wide allele frequencies were used to remove variants with a minor allele frequency higher than 1%. Finally, variants with unclear impact on gene function were removed (see definitions below).

Bioinformatics pipeline for variant calling

Bioinformatics analysis of the sequencing data followed standard practices in the field and included assessment of data quality with FastQC; mapping to the human reference genome (hg19/Grhc37) with BWA mem [8]; removal of PCR duplicates, local indel realignment, base quality score recalibration, and variant calling with HaplotypeCaller were performed with GATK [9]. Only variants with Phred scores ≥ 30 were considered. The functional consequences of variants were assessed with Ensembl's Variant Effect Predictor (VEP) tool [10]. This step included the prediction of the pathogenicity of identified variants with both SIFT [11] and Polyphen-2 [12]. Variants were stored in a mySQL database to facilitate filtering and comparison between both individuals and families.

Filtering of variants

Following genotype calling, variants were filtered to meet a number of sequencing quality requirements prior to consideration. The variants should correspond to single nucleotide events both in the reference and alternative alleles, be supported by read depth of at least 30 at the genomic position, and have a Phred call quality greater than 30. Similarly, variants were required to be present in all affected members of a family, either in heterozygosity or in homozygosity.

Population-wide allele frequencies were used to further filter the candidate variants under the prerogative that the observed incidence of CHD is not coherent with highly frequent variants being causative factors. Variants were filtered from our analysis if they were observed with a minor allele frequency (MAF) higher than 1% in either any of the main populations (AFR, AMR, ASN, CEU) comprised by the 1000 Genomes (1000G) project [7], the catalog of Exome Annotation Consortium [13] (ExAC), or its European (non-Finnish) subpopulation. Similarly, we exploited the Danish ancestry of the patients for further reduction of the number of candidate variants. First, we discarded from further analyzing those variants with MAFs higher than 1% in 2000 Danish exomes [14]. Second, variants present in 5% of the alleles of the 100 parents included in the Genome Denmark cohort [15, 16] were excluded in further analyses. This step implied a lift-over [17] of variants between the coordinates of reference genomes hg38 and hg19; in cases where multiple variants remapped to the same genomic position, the highest allele frequency for each possible allele was considered.

A final filtering exploited the functional consequences of variants according to affected genomic elements. Variants with unclear impact on gene function were disregarded. These included all noncoding variants (except intronic variants in splice sites), stop codon variants where the stop codon is retained, and nonsense-mediated decay (NMD) transcript variants.

We did not filter variants according to in silico prediction of functional consequence (e.g., Polyphen or SIFT scores).

Enrichment of known CHD genes

We curated two lists of known CHD genes. A list of 829 genes which cause CHD when mutated in mice was derived from data in the Mouse Genome Informatics database (Additional file 1: Table S1). A list of 144 human CHD disease genes was curated from the literature (Additional file 1: Table S2). Statistical significance of overlaps was calculated using one-tailed Fisher's exact test, with a significance level of $p < 0.05$.

Definition of enriched protein-protein interaction modules

Protein-protein interactions (PPIs) were obtained from InWeb (InWeb5.5rc3), a scored high-confidence PPI database which contains 87% of human proteins and more than 500,000 interactions [18]. PPIs were represented as a network; nodes represent proteins and edges represent interactions between these proteins. InWeb contains protein interactions for 1310 of the 1785 candidate genes identified in our families. We used data from InWeb to generate a PPI network of these 1310 genes and their first-degree interactors. The PPI network was pruned to include only high-confidence relationships; we disregarded high-throughput yeast-2-hybrid experiments ("Matrix" interactions according to InWeb's terminology) and interactions with a confidence below 0.1. After confidence filtering, the PPI network included a total of 8186 genes and 29,463 high-quality interactions.

The PPI network was partitioned into overlapping modules (clusters) with ClusterOne [19] using the confidence score derived from InWeb to weight the clustering. Clusters were considered significant when below a corrected (Benjamini-Hochberg FDR method) one-sided p value of 0.05 and a minimum of five proteins. Highly connected clusters of proteins identified in this fashion constitute a proxy for functionally related protein complexes.

Enrichment of candidate genes was determined for each cluster by permutation analysis ($k = 10e4$) using random clusters of comparable size. Correction for multiple testing was performed using the procedure of Bonferroni-Holm. A one-sided p value of 0.05 was used as significance level to identify clusters significantly enriched for candidate genes.

Probability of being loss-of-function intolerant (pLI) values of genes encoding the 27 proteins in the calcium-signaling module, 829 and 144 known CHD genes from mouse models and patients, respectively, was obtained from the ExAC database. The distributions were compared to all 18,225 genes listed in the ExAC database using a Kruskal-Wallis one-way analysis of variance on ranks. Significance level was 0.05.

Replication

WES data from a previously published, independent cohort was used for replication, please see Sifrim et al. for details [20]; WES data from a total of 714 non-syndromic CHD cases and 4922 controls of European ancestry was analyzed.

Control samples were participants in the INTERVAL randomized controlled trial which were recruited with the active collaboration of NHS Blood and Transplant England, which has supported fieldwork and other elements of the trial. DNA extraction and genotyping were funded by the National Institute of Health Research (NIHR), the NIHR BioResource, and the NIHR Cambridge Biomedical Research Centre. The academic coordinating center for INTERVAL was supported by core funding from the NIHR Blood and Transplant Research Unit in Donor Health and Genomics, UK Medical Research Council (G0800270), British Heart Foundation (SP/09/002), and NIHR Research Cambridge Biomedical Research Centre.

Replication was performed by comparing the number of cases and controls harboring very rare (MAF < 0.001) variants in any of the genes *ADCY2*, *ADCY5*, *CACNA1D*, *CACNA1H*, *CACNA1I*, *CACNA1S*, *GRIA4*, *ITPR1*, *NFAT5*, and *PLCB2*. Data from *CACNA1F* was missing; thus, this gene was excluded from the analysis. The MAF cutoff (0.001) was chosen by performing a burden test for synonymous variants in 10,000 random sets of genes of the same size as the set above (two-sided Fisher's exact test). For two different scenarios (MAF < 0.01 and MAF < 0.001), we observed a better match between expected vs. observed p values using MAF < 0.001 (Additional file 1: Fig. S12).

Synonymous variants, protein altering variants (PAV) (inframe_insertion, start_lost, stop_retained, stop_lost, missense, inframe_deletion, protein_altering, start_retained), and protein truncating variants (PTV) (stop_gained, splice_donor, splice_acceptor, frameshift) were identified using the Variant Effect Predictor tool (VEP API v90). Quality control and filtering were performed using Hail 0.2. In brief, both samples and variants were included for further analysis if they met the following filtering criteria: call rate ≥ 0.95, genotype quality average ≥ 25, and depth average ≥ 15. In addition, only genotypes (heterozygous) with allelic balance (AB) within the range 0.25–0.75 were retained in the analysis. Deleterious variants were identified by assigning a Combined Annotation Dependent Depletion (CADD) score [21] and a score based on regional missense constraint (MPC score) [22]. MPC score > 2 was used to identify pathogenic variants. Previous analysis of de novo variants identified in 5620 cases with neurodevelopmental disorders shoved that variants with MPC > 2 have a rate 5.79 times higher in cases than controls. Statistical significance was calculated using two-tailed Fisher's exact test, with a significance level of $p < 0.05$.

To test the specificity of the observed burden of rare variants, we created 10,000 random sets of 10 genes with the same size distribution as the gene set of 10 calcium-signaling genes, which were analyzed in the replication study. For each of the 10,000 random gene sets, we counted the number of cases with pathogenic variants (MPC score > 2).

Zebrafish maintenance and microinjection

AB/TL zebrafish strain (obtained from the Zebrafish International Resource Center) was used in the experiments. Embryos were maintained and staged as previously described [23, 24]. 1.5, 3, and 6 ng of adcy2a-SP-MO (5′-GGATGAGGGTAACTCACCTGACATT-3′), itpr1b-SP-MO (5′-GTGCATAAACGCGGCCTTACCTCGA-3′), plcb2-SP-MO (5′-CTGTAGTTTCTGTTCACCTCATCAG-3′), and standard-MO (5′-CCTCTTACCTCAGTTACAATTTATA-3′) and 0.5, 1, and 2 ng of adcy2a-SP-MO2 (5′-CCCCCAGTCTCCAAACACTCACCAG-3′), itpr1b-SP-MO2 (5′-CCAGACTGTAGACAAGAGAGACATG-3′), and plcb2-SP-MO2 (5′-TGTGGTAAAGGATACTCCACCCAGT-3′) (Gene Tools, LLC) were injected into one-cell stage embryos and harvested at 48 hpf. Embryos were imaged under Zeiss AxioZoom V16 (Carl Zeiss, Brock Michelsen A/S, Denmark).

In order to verify SP-MO efficiencies, embryos injected with SP-MOs were collected at 48 hpf and RNA was isolated by QIAzol reagent (Qiagen). cDNA synthesis was performed using iScript Reverse Transcription Supermix (Bio-Rad). SP-MO knockdown was assessed by PCR using gene-specific primers (Additional file 1: Table S3) spanning the SP-MO-targeted exon.

Whole-mount in situ hybridization

Digoxigenin (DIG)-labeled anti-sense *myl7* and *mef2cb* [25] riboprobes were synthesized from linearized pGEM-

T easy and pCMV-SPORT6.1 vectors respectively using the DIG RNA labeling mix (Roche) and the T7 RNA polymerase (Roche). Embryos collected at 48 hpf were raised in the presence of 0.2 mM 1-phenyl-2-thiourea (PTU) upon gastrulation for optical clarity [24]. For analysis of *myl7* and *mef2cb* expression, embryos at 48 hpf and 10 somite stages, respectively, were dechorionated and fixed in freshly prepared 4% paraformaldehyde in phosphate-buffered saline (PBS, pH 7.4) overnight. Whole-mount in situ hybridization was performed as previously described with minor modifications [26]. Embryos were imaged under Zeiss AxioZoom V16 (Carl Zeiss, Brock Michelsen A/S, Denmark). The staining intensity of *mef2cb* riboprobe was measured as integrated density (IntDen, ImageJ software-NIH, USA), and mean values from three embryos of each group were plotted relative to wild-type value. Data were shown as mean ± std dev.

Real-time quantitative RT-PCR

Total RNA from pools of 50 zebrafish embryos was extracted using TRIzol (Ambion) and a RNeasy mini kit (Qiagen) and used to synthesize random-primed cDNA (SuperScript II Reverse Transcriptase, Invitrogen). A Brilliant III Ultra-Fast SYBR® Green QPCR Master Mix (Agilent Technologies) was used for cDNA amplification. Samples were analyzed using a 7500 fast real-time PCR system (Applied Biosystems). Data were normalized to the average expression of housekeeping genes *actb1* and *rpl13a*. RT-PCR primers are listed in Additional file 1: Table S3.

Results

Exome sequencing data analysis

We performed whole exome sequencing of 90 individuals belonging to 32 multiplex CHD families (Additional file 1: Figs. S1-S4). To identify potentially disease-causing genes, we analyzed rare variants shared by affected family members. The analysis was performed after removing the variants that are likely sequencing artifacts, that occupy genomic elements with a mild functional consequence, or that can be found widespread in the general population (see the "Methods" section, workflow shown in Additional file 1: Fig. S6). We identified 3698 rare inherited variants in 1785 genes, denoted candidate disease genes (CDGs) hereafter. The number of CDGs per family ranged from 56 to 507 (Additional file 1: Fig. S7).

Recurrent candidate disease genes

To test if particular candidate genes were overrepresented, we calculated the fraction of CDGs shared by all possible pairs of families in the cohort (Fig. 1a, Additional file 1: Fig. S8). Pairs of families only share a small fraction of their CDGs (median = 0.05, 72.3% of pairwise values < 0.2).

To analyze this in more detail, we calculated the number of families with rare inherited variants in each of the 1785 CDGs. None of the 1785 CDGs was mutated in more than seven families (Fig. 1b, Additional file 1: Fig. S9) after disregarding genes that rely on the accumulation of variants to exert their biological function, like hypermutated genes in the MHC machinery or genes encoding olfactory receptors.

We identified 218 CDGs shared between two or more families. For variants shared between two and three families, less than 20% were identical across families (Fig. 1c, d). We propose that these variants might partially explain the origin of the disease. For example, three different rare alleles of *DNAH5* were identified in affected members of families 489, 732, and 1121, where patients presented with ASD, VSD, and outflow tract defects. Mutation of *DNAH5* is associated with primary ciliary dyskinesia (PCD, OMIM #608644) [27]. A subset of PCD patients present with CHD [28]. Another example of a CDG where rare variants were found in more than one family is *KMT2D*, which encodes a histone methyl transferase involved in heart development [29]. Mutation of *KMT2D* is associated with Kabuki syndrome (OMIM #147920), a rare developmental disorder which includes CHD as part of a wide phenotypical spectrum [30].

CDGs are enriched for known disease genes

We expect that a subgroup of variants in our 1785 CDGs could be causative mutations leading to CHD. In such a scenario, a diverse but limited number of genomic variants, acting by the accumulation of additive effects, could explain the occurrence of CHD in individual families. This etiological diversity at the genomic level would hinder detection with traditional association analysis, but we hypothesized that the CDGs would be enriched for known CHD disease genes. To test this, we calculated the overlap between the 1785 CDGs in our families and curated lists of genes known to cause CHD in mouse models and patients, when mutated (Additional file 1: Table S1 and S2). We observed significant enrichment of known CHD genes among CDGs affecting five or fewer families (Fig. 2a, b). When only variants scored pathogenic by SIFT/Polyphen-2 are considered, this enrichment increases Additional file 1: Fig. S10). In order to validate the CDG list and explore the possibility of a selection bias in our sets of known CHD genes, we produced 10,000 random gene sets (one for each of the two sets) and ascertained the overlap with the CDGs. This analysis corroborated the statistical significance of the observations (p value< 0.0001 for both sets).

Distribution of known CHD genes across families is shown in Additional file 1: Fig. S11 and Additional file 1: Table S4. Individual families often present with rare mutations in more than one known CHD gene, suggesting a

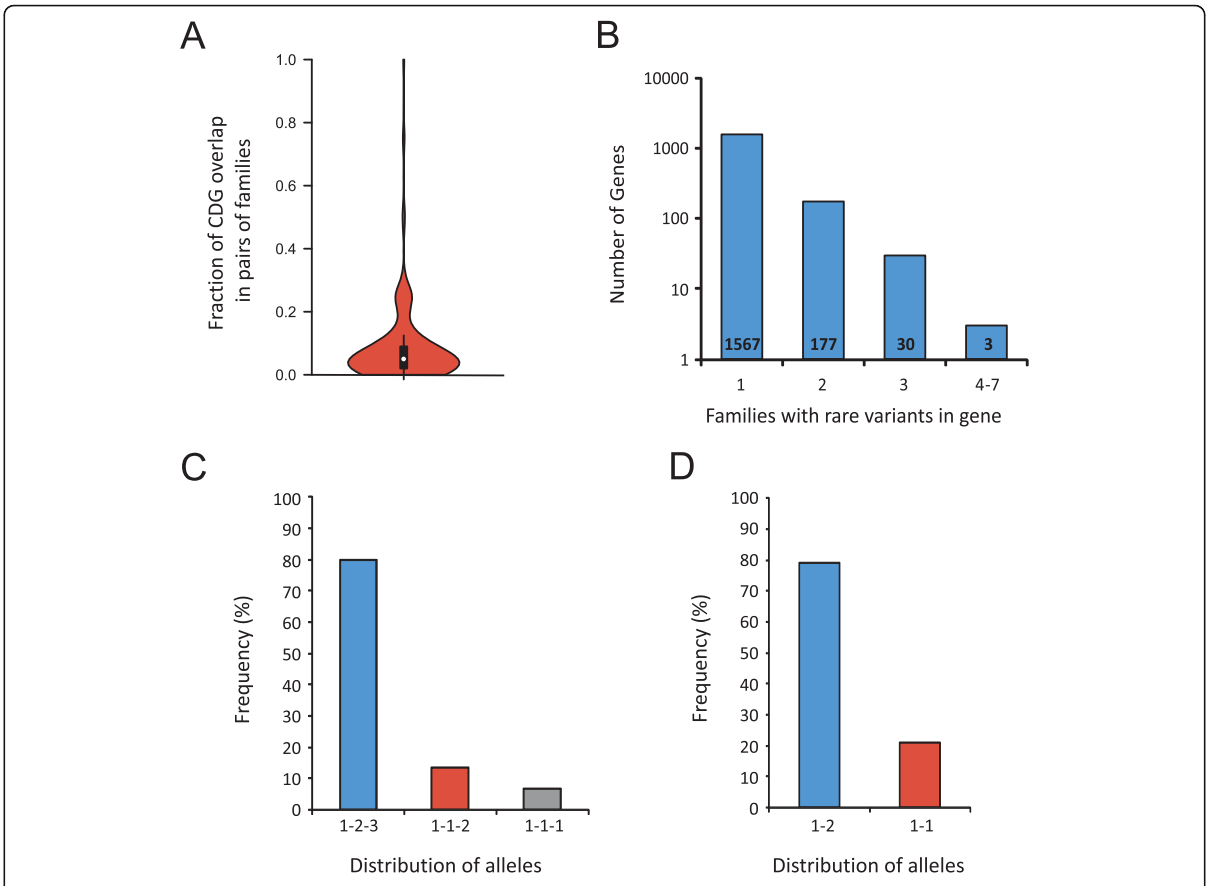


Fig. 1 Distribution of candidate disease genes and variants across families. **a** Overlap between CDGs in pairs of families. **b** Distribution of CDGs across families. The number of CDGs found in one, two, three, and 4–7 families is shown. **c** Distribution of alleles in CDGs found in three families (1-2-3, three different alleles; 1-1-2, two different alleles; 1-1-1, same allele found in all three families). **d** Distribution of alleles in CDGs shared in two families (1-2, different alleles; 1-1, same allele).

substantial fraction of the observed heart defects might be explained by a combination of rare mutations inherited within the families. Under the hypothesis that CHD runs within individual families as the result of such a combined effect from mutations in several developmental genes, we would expect that families present significant enrichment of mutations harbored by known CHD genes. Enrichment of CHD genes from mouse models per family was determined using a permutation test (n = 10,000). Affected individuals from 23 families share rare mutations in known CHD genes. In 19 of these 23 families (78.3%), enrichment of known CHD genes among the CDGs is statistically significant at $p < 0.05$ (Fig. 2c).

For comparison, we filtered our variants following the strategy, recently applied by Bayrak et al. [31]. In this approach, we only included variants, shared between at least two affected individuals per family and with population MAF < 0.001 (across all GenomAD populations and in the GenomAD European, not Finnish population). In addition, genes with a Gene Damage Index (http://lab.rockefeller.edu/casanova/GDI) prediction of "high" were removed from the list of CDGs. With this approach, we identified a total of 719 variants in 577 genes. When we compared with our curated lists of CHD genes in mouse models and patients (Additional file 1: Table S1 and S2), we identified 36 CDGs overlapping with mouse CHD genes and five CDGs overlapping with human CHD genes. However, the overlaps were not statistically significant ($p = 0.1406$ and 0.8178, respectively, Fisher's exact test), suggesting that very stringent filtering removes a significant number of causative variants.

Rare variants affect functional modules in a protein-protein interaction network

We further investigated whether CHD was caused by the disruption of cooperative protein functionality at the systems level rather than at the individual gene level. A

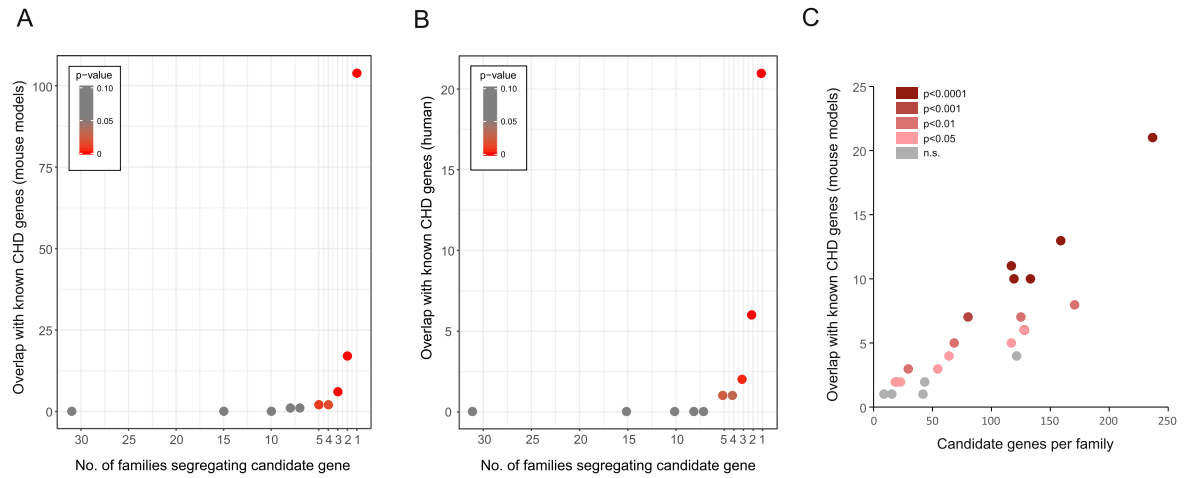


Fig. 2 Enrichment of CHD genes in candidate disease genes. Overlap between CDGs and known CHD genes from mouse models (**a**) and patients (**b**). The number of overlapping genes is plotted against the number of families the CDGs were found in. **c** The number of overlapping genes (mouse models) per family. Statistical significance of the overlap is indicated by color code (red colors, significant; gray color, not significant (n.s.))

protein-protein interaction network was generated using 1310 seed genes (the subset of the 1785 CDGs for which InWeb has recorded protein interactions). Their first-degree neighbors and their interactions from the current version of InWeb (InWeb5.5rc3) [18] were added as well, summing up to a total of 8186 proteins and 29,463 interactions. Using the graph clustering algorithm ClusterOne [19], we identified 230 significant PPI clusters considering a threshold p value of 0.05 and a minimum number of five genes. Of these, 25 clusters presented two or more genes mutated in at least one of the families (Additional file 1: Table S5). By permutation analysis ($k = 10{,}000$), we identified two clusters accommodating more CDGs than expected by chance (p values of 0.039 and 0.0033, respectively).

One of these corresponds to a small module, encoded by of five genes (*INSL1*, *RXFP2*, *RLN1*, *RLN2*, *RXFP1*) where the three first are CDGs. Due to the small size and low connectivity of this cluster, we decided to focus on the second cluster.

The second significant cluster corresponds to a highly interconnected group of 27 proteins, encoded by eleven CDGs and their first-degree interaction partners (Fig. 3a). These CDGs present mutations in ten different families. Affected individuals from three families shared rare mutations in more than one of the eleven genes, and two genes (*ITPR1* and *CACNA1S*) were each mutated in two different families (Additional file 1: Table S6).

The 27 proteins in the cluster constitute calcium channels or signal transduction enzymes such as adenylate cyclase that interacts with calcium-dependent protein kinases. A Gene Ontology term enrichment analysis using AmiGO [32] for the 27 proteins in the cluster showed enrichment in biological processes involved in calcium signaling (Additional file 1: Table S7).

To investigate the functional importance of the genes encoding proteins in the cluster, we compared the probability of being loss-of-function intolerant (pLI) of these 27 genes with known CHD genes and all 18,225 genes listed in the Exome Aggregation Consortium database. The distributions of pLI scores of known CHD genes have median pLI values of 0.63 and 0.86, respectively, and the distributions suggest that more than half of known CHD genes are intolerant against loss-of-function mutations (Fig. 3b). The distribution of pLI scores of the 27 genes encoding proteins in the significant cluster is skewed towards the higher values of the distribution (median = 0.98), suggesting intolerance to loss-of-function mutations and supporting the hypothesis that these genes might potentially play an active role in the etiology of CHD.

Replication

Using WES data from an independent cohort of 714 CHD cases and 4922 controls [20], we tested the mutation burden of the same calcium-signaling gene set, as we found mutated in our families (genes listed in Additional file 1: Table S6). In this gene set, we analyzed the distribution of CADD and MPC variant scores between CHD cases and controls, and observed significant different score distributions of rare (MAF < 0.001) variants, predicted to alter the gene products (i.e., PAV and PTV) (Fig. 4). Variants with MPC score above 2 (MPC2 variants) have been shown to be

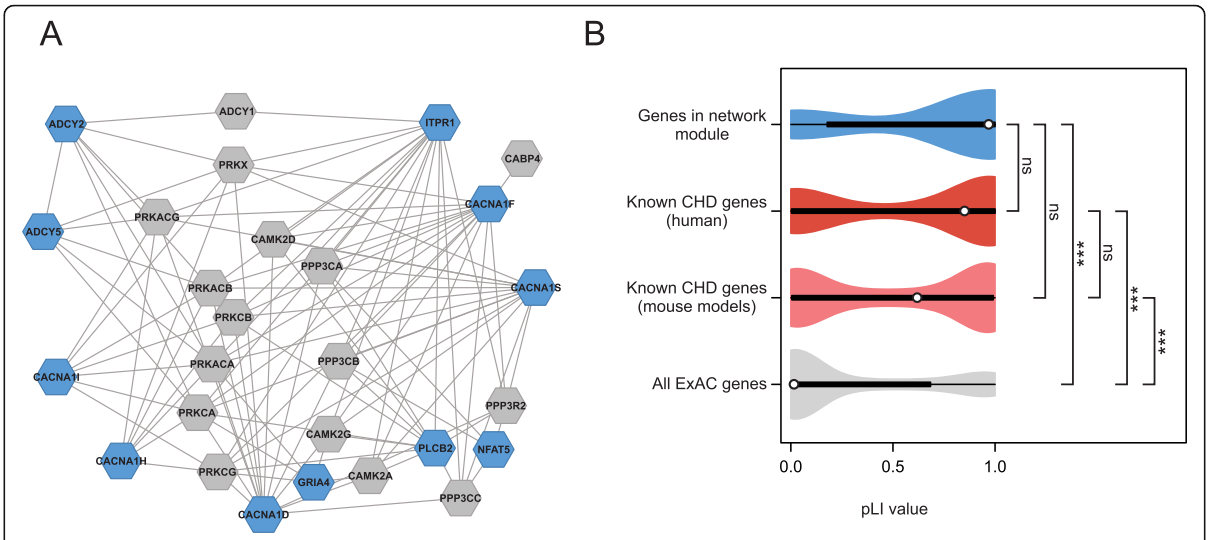


Fig. 3 Identification of a calcium-signaling network affected by rare mutations identified in CHD families. **a** Network module of CDGs (blue) and their interaction partners (gray). The module accommodates more CDGs than expected by chance (adjusted *p* value 0.0033). Proteins are shown as hexagons; protein interactions are shown with lines. **b** Violin plots of distributions of pLI scores in genes encoding the 27 proteins in the network module (upper, blue), known CHD genes from patients and mouse models (middle, red and pink, respectively), and all 18,225 genes listed in ExAC with a calculated pLI score (lower, gray).***$p < 0.001$. ns, not significant (Kruskal-Wallis one-way analysis of variance on ranks)

significantly associated with pathogenesis [22]. Thus, to test for burden of pathogenic mutations, we calculated the number of cases and controls harboring MPC2 variants in the calcium-signaling gene set. We observed more than 2-fold enrichment (OR 2.68, *p* value 3.7e−04) of such variants in CHD cases (Additional file 1: Table S8). To test if this enrichment was specific for the calcium-signaling gene set, we created 10,000 random gene set with the same size distribution as the calcium genes. For each random set, we counted the number of CHD cases harboring MPC2 variants. Only four of the random gene sets were found with more MPC2 variants in cases than the calcium gene set (Additional file 1: Fig. S13).

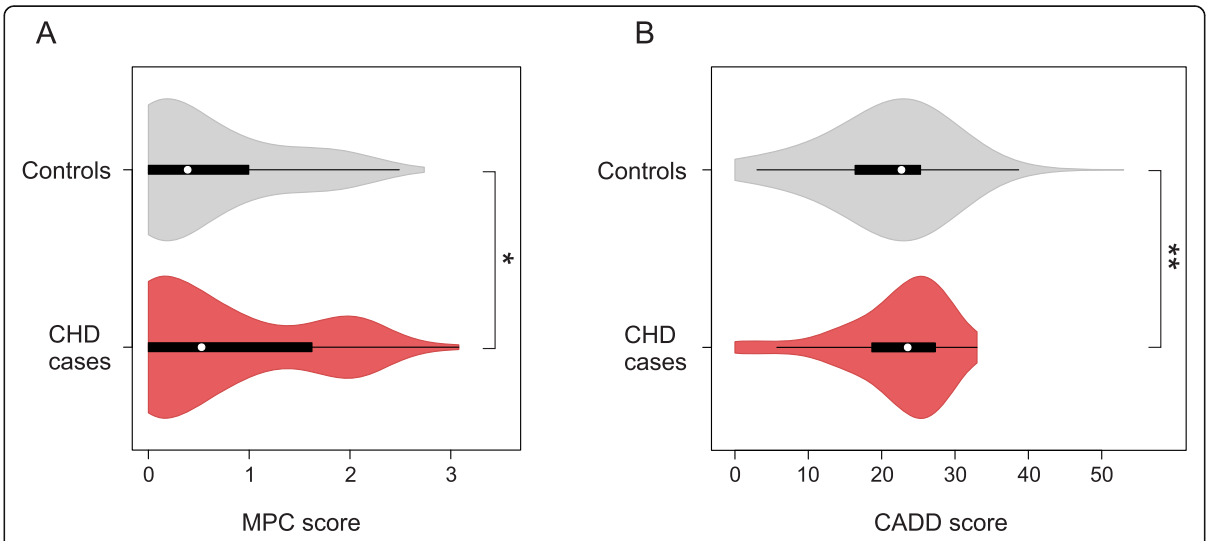


Fig. 4 Distribution of MPC and CADD scores of rare variants in 714 CHD cases and 4922 controls. Protein altering and truncating variants (PAV and PTV) with MAF < 0.001 identified in the genes *ADCY2*, *ADCY5*, *CACNA1D*, *CACNA1H*, *CACNA1I*, *CACNA1S*, *GRIA4*, *ITPR1*, *NFAT5*, and *PLCB2* were scored using MPC score [22] (**a**) or CADD score [21] (**b**). $N_{CHD} = 136$ variants. $N_{Controls} = 982$ variants. Difference between median values of controls and cases was determined using a Mann-Whitney rank-sum test. **$p < 0.01$, *$p < 0.05$

These results confirm an increased burden of pathogenic mutations in genes involved in calcium signaling among CHD patients.

Knockdown of *adcy2a*, *itpr1b*, and *plcb2* causes cardiac malformations in zebrafish

We used zebrafish as a model to address the functional relevance in heart development, of genes within the identified cluster. In the cluster, 11 out of 27 genes were mutated in the families we assessed. From these 11 genes, we assessed the consequence of loss-of-function for three zebrafish genes, *adcy2a*, *itpr1b*, and *plcb2*. These genes are orthologues to the human genes *ADCY2*, *ITPR1*, and *PLCB2*, encoding calcium-signaling proteins, of which roles in cardiac development have been elusive.

We found that knockdown of either of the three genes by morpholino oligonucleotides gave rise to abnormal morphology or laterality of the heart (Fig. 5a). The abnormal morphology included aberrant atrioventricular canal (AVC) formation where the ring structure of AVC was hindered, mis-looping of the heart, and narrowing in the atrium or ventricle. Laterality defects reflected straight or reversed hearts. Injection of *adcy2a*-MO, *itpr1b*-MO, and *plcb2*-MO resulted in cardiac defects in 53%, 73%, and 66% of embryos, respectively (Fig. 5b).

In addition to the cardiac defects, knockdown of *itpr1b* caused edema in the brain, whereas curved tail was seen by knockdown of *plcb2* (Fig. 5a).

Knockdown of more than one gene by injections of both efficient and sub-efficient doses of MO resulted in an increase of the number of embryos with heart defects, suggesting a combinatorial effect (Fig. 5b, Additional file 1: Fig. S14).

We showed that the morpholinos work efficiently causing splicing defects (Additional file 1: Fig. S15), and we validated the specificity of the morpholinos, by using a second set of morpholinos and co-injection of in vitro expressed mRNAs encoding wild-type proteins (Fig. 5b, Additional file 1: Fig. S15). Unfortunately, we were not able to clone the cDNA of *itpr1b*, presumably due to the large size of the transcript (8.4 kbp coding region), but co-injection of wild-type *adcy2a* and *plcb2* mRNA resulted in partial rescue of the cardiac phenotype, confirming the specificity of the morpholinos.

To investigate effects of gene knockdown at the molecular level, we analyzed the expression of *mef2c*, a transcription factor which is involved in cardiac morphogenesis and myogenesis and known to be a specific

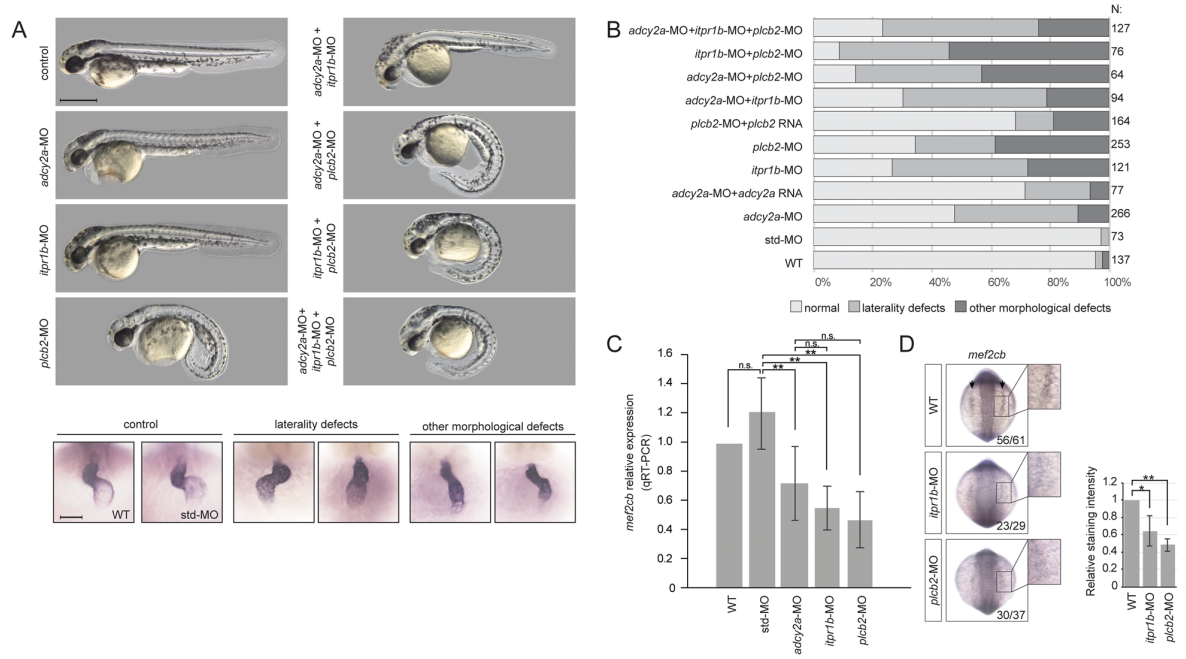


Fig. 5 Functional validation of candidate genes in zebrafish. **a** Phenotype of controls and morphants. Single genes and combinations of genes targeted by splicing morpholinos are indicated on the left. Upper panels: gross appearance of 48 hpf zebrafish embryos. Lower panel: examples of cardiac phenotypes of 48 hpf wt, control, and morphant embryos. Hearts were visualized by ISH with a probe against the cardiac marker *myl7*. **b** Quantification of phenotypes in wt, controls, morphants, and mRNA rescued morphants. Note the combinatorial effects on cardiac phenotypes when more than one gene is affected. **c**, **d** Expression of *mef2cb* in 10 somite stage zebrafish embryos, analyzed by qRT-PCR (**c**) and ISH (**d**). ISH staining intensity was quantified and analyzed using Student's *T* test. $*p < 0.05$, $**p < 0.01$

target for Ca^{2+}-dependent control of gene expression in cardiomyocytes [33, 34]. In zebrafish, two copies of *mef2c* exist. We analyzed the expression of *mef2ca* and *mef2cb* in wild-type, control, and morphant embryos at 10 somite stages, where both genes are expressed in the bilateral heart fields at the anterior lateral plate mesoderm. We observed significantly reduced expression of *mef2cb* in *adcy2a*, *itpr1b*, and *plcb2* morphants (Fig. 5c, d).

Discussion

Interpretation of variants detected in exomes from CHD patients is challenging due to the extreme heterogeneity which characterize CHD [4]. Analysis of complex networks has previously been applied successfully to interpret genetic variants [31, 35–38]. Such analyses often use stringent variant filtering criteria to identify candidate genes, followed by network analysis to interpret biological function. Here, we present an alternative approach, based on lenient filtering of variants, followed by stringent network analysis using a high-confidence human PPI network [18].

We analyzed the exomes of 90 individuals in 32 multiplex CHD families for rare variants, which were shared among affected individuals within a family. Considering that CHD rarely has a monogenic cause, but may be caused by a burden of rare and common genetic variants, we selected an arbitrary filtering threshold of MAF above 0.01 and we did not filter the variants for in silico predictions of consequence. We expect that our filtering has removed a significant amount of neutral variants, while keeping possible variants with medium to strong effects. However, it is likely that we also have removed a number of causative variants by applying this criterion, as well as our focus on the exome limits the number of possible causative variants identified in our study.

We identified 1785 candidate disease genes. The large number of candidate genes is partly a consequence of our non-stringent filtering strategy and underlines the challenge in interpreting variants in CHD patients, even in multiplex families. However, our analysis showed that the 1785 CDGs were enriched for genes known to cause heart defects and that application of very stringent filtering conditions removes a significant part of causative variants.

Approximately 12% of the CDGs were found in more than one family, but none of them was represented in more than seven families, and on average, less than 10% of the CDGs found in a family were shared with another family. Thus, our data support that CHD is an extremely heterogeneous disorder and consistent with a model, where several rare genetic variants contribute to the pathogenesis in the individual patient. However, we do not expect that all of our CDGs are true CHD disease genes; thus, a significant proportion of the rare variants, which are shared by affected individuals, may be of limited functional consequence.

To identify pathophysiological mechanisms associated with CHD, we integrated PPI data in our analysis of CDGs and discovered that eleven of the CDGs converge in a functional cluster of genes, which encode proteins involved in calcium signaling. Analysis of an independent case-control cohort confirmed increased mutation burden of this calcium-signaling gene set in CHD patients, thus supporting that defects in calcium signaling are associated with CHD.

Intracellular calcium plays essential roles in physiology and pathophysiology of the adult heart. In the healthy heart, intracellular calcium fluxes control cardiomyocyte contraction and mutations in calcium-handling genes may cause arrhythmia [39]. In addition, calcium also modulates transcription in cardiomyocytes through a complex signaling network, and abnormal calcium handling and signaling are part of the pathophysiology in congestive heart failure [33, 40].

Defects in calcium signaling have not previously been associated with CHD in humans, but animal studies implicate an important role of calcium signaling in heart development. Targeted deletion of *Nfatc1*is embryonic lethal in mice and causes malformation of the cardiac valves, outflow tract, and ventricular septum [41, 42]. Overexpression of activated calcineurin rescues cardiac developmental defects in calreticulin-deficient mice [43]. Pharmacological blockade of L-type calcium channels during embryonic development causes heart defects in mice [44]. However, epidemiological studies show that therapeutic doses of calcium channel blockers have very low teratogenicity [45, 46].

Two of the 27 proteins within the cluster we identified were previously implicated in cardiac development and CHD. Knockout of *Nfat5* is embryonic lethal in mice and results in reduced compaction of cardiomyocytes in the ventricular wall and trabeculae [47]. Double knockout of *Itpr1* and *Itpr3* results in aberrant development of the outflow tract and right ventricle, while mice defective of only one of the two genes develop normally [48]. Likewise, double knockout of *Itpr1* and *Itpr2* is embryonic lethal and results in cardiac malformation [49], together suggesting a redundant role of IP3 receptors in mammalian heart development.

Analysis of embryonic mice has shown that NFAT5 and IP3 receptors are expressed in the heart during embryonic development, but also in several other tissues [47–49]. Similarly, many well-characterized CHD genes are also expressed in extra-cardiac embryonic tissues, and some of these genes have been associated with both isolated and syndromic forms of CHD [3, 50]. Thus, future detailed genotype-phenotype analyses of CHD

disease genes may be useful for further dissection of embryonic developmental mechanisms.

Functional validation of *ADCY2*, *ITPR1*, and *PLCB2* in a zebrafish model confirmed a critical role during embryonic heart development. We observed defects in cardiac morphology and laterality in a significant proportion of embryos injected with morpholinos targeting *adcy2a*, *itpr1b*, and *plcb2*. Importantly, we observed more embryos with cardiac defects when combinations of two and all three genes were knocked down, supporting that the three genes interact functionally in heart development. Specificity of the morpholinos used in the experiments was confirmed by rescue experiments and further corroborated by significant reduction in the expression of *mef2cb*, a zebrafish orthologue of *Mef2c*, which is a well-established target of calcium signaling in cardiomyocytes [48, 51]. Thus, the results of the zebrafish experiments support the hypothesis that expression and interaction of the three genes play important roles in heart development. However, it is important to note that depletion of the gene products precipitated by morpholino injections in zebrafish does not represent a true model of the molecular pathology of CHD in affected family members. Firstly, the majority of mutations, which we identified in calcium-signaling genes, were missense mutations and carriers were all heterozygous. Thus, we do not know if the effects of the mutations were haploinsufficiency or gain-of-function. And second, it is possible that factors other than calcium-signaling genes may play a role in the molecular pathology of individual patients.

The cardiac malformations, present in our family cohort and replication cohort, were unselected with respect to severity or anatomical similarity. Thus, our data indicate that calcium-signaling defects are associated with both familial and sporadic CHD and unrelated to specific groups of malformations.

We have recently shown that specific malformations co-occur in families, suggesting that specific developmental programs may be responsible for certain groups of heart malformations [52]. We suggest that WES or whole genome sequencing of cohorts of families selected for specific malformations, combined with systems-based analysis, as presented here, may be a useful strategy for identification of pathophysiological mechanisms in CHD.

Conclusions

Our data support a model where CHD is caused by a combinatorial effect of rare and common genetic variants. Our systems-level analysis of rare genetic variants, shared by affected individuals in families, and functional analysis in zebrafish identified defects in calcium signaling as a novel pathophysiological mechanism in CHD.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13073-020-00772-z.

Additional file 1: Supplementary Figures and Tables. Pedigrees of 32 Danish multiplex CHD families (**Fig. S1**). Relatedness of the 90 individuals included in the study (**Fig. S2**). Homogeneity of the cohort (**Fig. S3**). Principle component analysis (**Fig. S4**). Sequencing coverage and number of reads (**Fig. S5**). Overview of the sequencing and data analysis processes (**Fig. S6**). Number of CDGs per family when all rare variants (left) or only high severity variants (right) were considered (**Fig. S7**). Overlap between CDGs in pairs of families (**Fig. S8**). Families with rare inherited variants in CDGs (**Fig. S9**). Overlap between the 1,785 CDGs in our families and a curated list of 829 genes known to cause CHD in mice (**Fig. S10**). Distribution of CHD genes across families (**Fig. S11**). Quantile-quantile plots (**Fig. S12**). Distribution of pathogenic mutations in random gene-sets (**Fig. S13**). Injection of sub-efficient doses of MOs and quantification of heart phenotypes in WT, controls and morphants injected with second set of MOs (**Fig. S14**). Efficiency of the splice blocking morpholinos (MOs) used against *adcy2a*, *itpr1b* and *plcb2* (**Fig. S15**). Human orthologues to 829 genes known to be associated with CHD in mouse models (**Table S1**). A list of 144 Human CHD disease genes (**Table S2**). Primers used in the study (**Table S3**). Variants identified in known human CHD genes (**Table S4**). Significant protein-protein interaction clusters with at least two CDGs (**Table S5**). Rare calcium signaling gene variants shared among affected individuals in multiplex CHD families (**Table S6**). Gene ontology term enrichment of 27 genes within the cluster shown in Fig. 3a (**Table S7**). Replication using WES data from 714 CHD cases and 4922 controls (**Table S8**).

Acknowledgements

The mef2cb probe plasmid was a kind gift of Dr. Yaniv Hinits, King's College London. We thank Matthew E. Hurles, Wellcome Trust Sanger Institute, for providing access to published WES control data. Control samples were participants in the INTERVAL randomized controlled trial.

Authors' contributions

SB and LAL designed and supervised the project. SGE established the family study cohort. MPH, JB, MG, DB, and Competence Network for Congenital Heart Defects, Germany, established the secondary patient cohort. JMGI, SGE, LAL, EA, MPH, AS, AW, KB, BT, NSE, MD, and GD generated and/or analyzed the data. CD performed the zebrafish experiments. All authors contributed to writing the manuscript and approved the final manuscript.

Funding

This work is supported by The Danish National Advanced Technology Foundation (The Genome Denmark platform, grant 019-2011-2), The Novo Nordisk Foundation (grants NNF14CC0001 and NNF12OC0001790), Aase og Ejnar Danielsens Fond, Børnehjertefonden, The Danish Heart Association, Dagmar Marshalls fond, Arvid Nilssons Fond, Oda og Hans Svenningsens Fond, Eva & Henry Frænkels Mindefond, Kong Christian Den Tiendes Fond, The A.P. Møller Foundation for the Advancement of Medical Sciences, The Lundbeck Foundation (R209-2015-2604), and Villum Fonden. JB is supported by the Van de Werf fund for cardiovascular research and a clinical research fund of UZ Leuven. AS is supported by the FWO (Postdoctoral Fellow number 12W7318N). The Clinical Academic Group in Precision Diagnostics in Cardiology under Greater Copenhagen Health Science Partners is also acknowledged.

Availability of data and materials

A list of rare variants, shared between affected family members, which our results and conclusions are based on, is available upon request. Individual exome sequencing data cannot be shared due to concerns over patient privacy. Other data generated or analyzed during this study are included in the main paper or its additional files.

Ethics approval and consent to participate

This project was approved by the Danish Data Protection Agency (2009-41-3570) and the Danish National Board of Health (H-D-2009-070). The study conformed to the principles of the Helsinki Declaration. Health records were

accessed only after explicit consent by the patients. All zebrafish research was approved by and conducted under license from the Danish Animal Experiments Inspectorate.

Consent for publication
Consent for publication was obtained from the participants.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Health Technology, Technical University of Denmark, Kemitorvet, DK-2800 Kgs. Lyngby, Denmark. [2]Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Blegdamsvej 3A, DK-2200 Copenhagen, Denmark. [3]Department of Cellular and Molecular Medicine, University of Copenhagen, Blegdamsvej 3A, DK-2200 Copenhagen, Denmark. [4]DTU Multi Assay Core (DMAC), Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark. [5]Department of Congenital Heart Disease and Pediatric Cardiology, Universitätsklinikum Schleswig–Holstein Kiel, Kiel, Germany. [6]Wellcome Trust Sanger Institute, Cambridge, UK. [7]Centre for Human Genetics, Katholieke Universiteit Leuven, Leuven, Belgium. [8]School of Life Sciences, University of Nottingham, Nottingham, UK. [9]Genetics and Genome Biology, Hospital for Sick Children, Toronto, ON, Canada. [10]Pediatric Cardiology Unit, University Hospitals Leuven, Leuven, Belgium. [11]Institute of Human Genetics, Christian-Albrechts-University Kiel & University Hospital Schleswig-Holstein, Campus Kiel, Kiel, Germany.

Received: 27 November 2019 Accepted: 7 August 2020
Published online: 28 August 2020

References
1. van der Linde D, Konings EE, Slager MA, Witsenburg M, Helbing WA, Takkenberg JJ, et al. Birth prevalence of congenital heart disease worldwide: a systematic review and meta-analysis. J Am Coll Cardiol. 2011;58(21):2241–7.
2. Warnes CA. Adult congenital heart disease: the challenges of a lifetime. Eur Heart J. 2017;38(26):2041–7.
3. Pierpont ME, Brueckner M, Chung WK, Garg V, Lacro RV, McGuire AL, et al. Genetic basis for congenital heart disease: revisited: a scientific statement from the American Heart Association. Circulation. 2018;138(21):e653–711.
4. Blue GM, Kirk EP, Giannoulatou E, Sholler GF, Dunwoodie SL, Harvey RP, et al. Advances in the genetics of congenital heart disease: a clinician's guide. J Am Coll Cardiol. 2017;69(7):859–70.
5. Blue GM, Kirk EP, Giannoulatou E, Dunwoodie SL, Ho JW, Hilton DC, et al. Targeted next-generation sequencing identifies pathogenic variants in familial congenital heart disease. J Am Coll Cardiol. 2014;64(23):2498–506.
6. Ellesoe SG, Johansen MM, Bjerre JV, Hjortdal VE, Brunak S, Larsen LA. Familial atrial septal defect and sudden cardiac death: identification of a novel NKX2-5 mutation and a review of the literature. Congenit Heart Dis. 2016; 11(3):283–90.
7. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68–74.
8. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
9. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(5):491–8.
10. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics. 2010;26(16):2069–70.
11. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31(13):3812–4.
12. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248–9.
13. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016; 536(7616):285–91.
14. Lohmueller KE, Sparso T, Li Q, Andersson E, Korneliussen T, Albrechtsen A, et al. Whole-exome sequencing of 2,000 Danish individuals and the role of rare coding variants in type 2 diabetes. Am J Hum Genet. 2013;93(6):1072–86.
15. Besenbacher S, Liu S, Izarzugaza JM, Grove J, Belling K, Bork-Jensen J, et al. Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. Nat Commun. 2015;19(6):5969.
16. Maretty L, Jensen JM, Petersen B, Sibbesen JA, Liu S, Villesen P, et al. Sequencing and de novo assembly of 150 genomes from Denmark as a population reference. Nature. 2017;548(7665):87–91.
17. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, et al. The UCSC Genome Browser Database: update 2006. Nucleic Acids Res. 2006; 34(Database issue):D590–8.
18. Li T, Wernersson R, Hansen RB, Horn H, Mercer J, Slodkowicz G, et al. A scored human protein-protein interaction network to catalyze genomic interpretation. Nat Methods. 2017;14(1):61–4.
19. Nepusz T, Yu H, Paccanaro A. Detecting overlapping protein complexes in protein-protein interaction networks. Nat Methods. 2012;9(5):471–2.
20. Sifrim A, Hitz MP, Wilsdon A, Breckpot J, Turki SH, Thienpont B, et al. Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing. Nat Genet. 2016;48(9):1060–5.
21. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46(3):310–5.
22. Samocha KE, Kosmicki JA, Karczewski KJ, O'Donnell-Luria AH, Pierce-Hoffman E, MacArthur DG, Neale BM, Daly MJ. Regional missense constraint improves variant deleteriousness prediction. BioRxiv. https://www.biorxiv.org/content/10.1101/148353v1.
23. Kimmel CB, Ballard WW, Kimmel SR, Ullmann B, Schilling TF. Stages of embryonic development of the zebrafish. Dev Dyn. 1995;203(3):253–310.
24. Hinits Y, Pan L, Walker C, Dowd J, Moens CB, Hughes SM. Zebrafish Mef2ca and Mef2cb are essential for both first and second heart field cardiomyocyte differentiation. Dev Biol. 2012;369(2):199–210.
25. Thisse C, Thisse B. High-resolution in situ hybridization to whole-mount zebrafish embryos. Nat Protoc. 2008;3(1):59–69.
26. Monnich M, Borgeskov L, Breslin L, Jakobsen L, Rogowski M, Doganli C, et al. CEP128 localizes to the subdistal appendages of the mother centriole and regulates TGF-beta/BMP signaling at the primary cilium. Cell Rep. 2018; 22(10):2584–92.
27. Zariwala MA, Knowles MR, Omran H. Genetic defects in ciliary structure and function. Annu Rev Physiol. 2007;69:423–50.
28. Kennedy MP, Omran H, Leigh MW, Dell S, Morgan L, Molina PL, et al. Congenital heart disease and other heterotaxic defects in a large cohort of patients with primary ciliary dyskinesia. Circulation. 2007;115(22):2814–21.
29. Ang SY, Uebersohn A, Spencer CI, Huang Y, Lee JE, Ge K, et al. KMT2D regulates specific programs in heart development via histone H3 lysine 4 di-methylation. Development. 2016;143(5):810–21.
30. Bogershausen N, Wollnik B. Unmasking Kabuki syndrome. Clin Genet. 2013; 83(3):201–11.
31. Sevim BC, Zhang P, Tristani-Firouzi M, Gelb BD, Itan Y. De novo variants in exomes of congenital heart disease patients identify risk genes and pathways. Genome Med. 2020;12(1):9–0709.
32. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S. AmiGO: online access to ontology and annotation data. Bioinformatics. 2009;25(2):288–9.
33. Dewenter M, von der Lieth A, Katus HA, Backs J. Calcium signaling and transcriptional regulation in cardiomyocytes. Circ Res. 2017;121(8):1000–20.
34. Lin Q, Schwarz J, Bucana C, Olson EN. Control of mouse cardiac morphogenesis and myogenesis by transcription factor MEF2C. Science. 1997;276(5317):1404–7.
35. Gustafsson M, Nestor CE, Zhang H, Barabási AL, Baranzini S, Brunak S, et al. Modules, networks and systems medicine for understanding disease and aiding diagnosis. Genome Med. 2014;6(10):82.
36. Glessner JT, Bick AG, Ito K, Homsy J, Rodriguez-Murillo L, Fromer M, et al. Increased frequency of de novo copy number variants in congenital heart disease by integrative analysis of single nucleotide polymorphism array and exome sequence data. Circ Res. 2014;115(10):884–96.
37. Liu X, Yagi H, Saeed S, Bais AS, Gabriel GC, Chen Z, et al. The complex genetics of hypoplastic left heart syndrome. Nat Genet. 2017;49(7):1152–9.
38. Lage K, Greenway SC, Rosenfeld JA, Wakimoto H, Gorham JM, Segre AV, et al. Genetic and environmental risk factors in congenital heart disease functionally converge in protein networks driving heart development. Proc Natl Acad Sci U S A. 2012;109(35):14035–40.

39. Landstrom AP, Dobrev D, Wehrens XHT. Calcium signaling and cardiac arrhythmias. Circ Res. 2017;120(12):1969–93.
40. Houser SR, Piacentino V III, Weisser J. Abnormalities of calcium cycling in the hypertrophied and failing heart. J Mol Cell Cardiol. 2000;32(9):1595–607.
41. Ranger AM, Grusby MJ, Hodge MR, Gravallese EM, de la Brousse FC, Hoey T, et al. The transcription factor NF-ATc is essential for cardiac valve formation. Nature. 1998;392(6672):186–90.
42. de la Pompa JL, Timmerman LA, Takimoto H, Yoshida H, Elia AJ, Samper E, et al. Role of the NF-ATc transcription factor in morphogenesis of cardiac valves and septum. Nature. 1998;392(6672):182–6.
43. Guo L, Nakamura K, Lynch J, Opas M, Olson EN, Agellon LB, et al. Cardiac-specific expression of calcineurin reverses embryonic lethality in calreticulin-deficient mouse. J Biol Chem. 2002;277(52):50776–9.
44. Porter GA Jr, Makuck RF, Rivkees SA. Intracellular calcium plays an essential role in cardiac development. Dev Dyn. 2003;227(2):280–90.
45. Davis RL, Eastman D, McPhillips H, Raebel MA, Andrade SE, Smith D, et al. Risks of congenital malformations and perinatal events among infants exposed to calcium channel and beta-blockers during pregnancy. Pharmacoepidemiol Drug Saf. 2011;20(2):138–45.
46. Weber-Schoendorfer C, Hannemann D, Meister R, Eléfant E, Cuppers-Maarschalkerweerd B, Arnon J, et al. The safety of calcium channel blockers during pregnancy: a prospective, multicenter, observational study. Reprod Toxicol. 2008;26(1):24–30.
47. Mak MC, Lam KM, Chan PK, Lau YB, Tang WH, Yeung PK, et al. Embryonic lethality in mice lacking the nuclear factor of activated T cells 5 protein due to impaired cardiac development and function. PLoS One. 2011;6(7):e19186.
48. Nakazawa M, Uchida K, Aramaki M, Kodo K, Yamagishi C, Takahashi T, et al. Inositol 1,4,5-trisphosphate receptors are essential for the development of the second heart field. J Mol Cell Cardiol. 2011;51(1):58–66.
49. Uchida K, Aramaki M, Nakazawa M, Yamagishi C, Makino S, Fukuda K, et al. Gene knock-outs of inositol 1,4,5-trisphosphate receptors types 1 and 2 result in perturbation of cardiogenesis. PLoS One. 2010;5(9):10.
50. Andersen TA, Troelsen KL, Larsen LA. Of mice and men: molecular genetics of congenital heart disease. Cell Mol Life Sci. 2014;71(8):1327–52.
51. Faustino RS, Behfar A, Groenendyk J, Wyles SP, Niederlander N, Reyes S, et al. Calreticulin secures calcium-dependent nuclear pore competency required for cardiogenesis. J Mol Cell Cardiol. 2016;92:63–74.
52. Ellesoe SG, Workman CT, Bouvagnet P, Loffredo CA, McBride KL, Hinton RB, et al. Familial co-occurrence of congenital heart defects follows distinct patterns. Eur Heart J. 2018;39(12):1015–22.

Publisher's Note

Chapter V. Exome analysis of 4,747 congenital heart disease cases and 52,881 controls

Exome analysis of 4,747 congenital heart disease cases and 52,881 controls

Enrique Audain[1,2 #], Anna Wilsdon[3 #], Gregor Dombrowsky[1,2], Alejandro Sifrim[4], Yasset Perez-Riverol[5], Allan Daily[6], Pavlos Antoniou[6], Philipp Hofmann[1,2], Anne-Karin Kahlert[1,2], Ulrike Bauer[7], Thomas Pickardt[7], Anselm Uebing[1,2], Hans-Heiner Kramer[1,2], Vivek Iyer[6], Lars Allan Larsen[8], J. David Brook[3*], Matthew E. Hurles[6 *], Marc-Phillip Hitz[1,2,6 *]

# These authors contributed equally to this work.

* Joint corresponding authors.

1 Department of Congenital Heart Disease and Pediatric Cardiology, University Hospital of Schleswig- Holstein, Kiel, Germany

2 German Center for Cardiovascular Research (DZHK), Kiel, Germany

3 School of Life Sciences, University of Nottingham, University Park, Nottingham, United Kingdom

4 Department of Human Genetics, University of Leuven, KU Leuven, Leuven, Belgium

5 European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom

6 Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, United Kingdom

7 Competence Network for Congenital Heart Defects, Berlin, Germany

8 Department of Cellular and Molecular Medicine, University of Copenhagen, Copenhagen, Denmark

ABSTRACT

Several studies have demonstrated the value of large-scale human exome and genome data to maximise gene discovery in rare diseases. Following this approach, we have aggregated and analysed the exomes of 4,747 cases and 52,881 controls to identify genes conferring substantial risk for congenital heart disease (CHD). We identified both rare loss-of-function and missense coding variants in ten genes with a significant genome-wide association as a likely cause of CHD (Bonferroni adjusted $P < 0.05$) and additionally four genes with a significant association at a false discovery rate (*FDR)* threshold of 5%. Furthermore, by independently analysing syndromic and non-syndromic CHD probands, we highlighted distinct genetic contributions to these different forms of CHD. Moreover, by meta-analysing the exome data with single-cell transcriptomics human heart data, we identified cardiac-specific cells as well as putative biological processes underlying the pathogenesis of CHD. In summary, our findings strengthen previous associations and identified novel genes contributing to the aetiology of CHD.

INTRODUCTION

Congenital Heart Disease (CHD) is a global health problem, affecting ~1-2% of live births worldwide[1]. However, despite advances in our understanding of its disease aetiology in recent years, a significant proportion of CHD cases remain unexplained, suggesting that its genetic causes, and other risk factors, remain poorly understood[2,3]. Recent advances in exome and genome sequencing technologies have opened up new avenues to study and gain new insights into the genetic and epigenetic mechanisms underlying rare diseases such as CHD[4,5]. Specifically, accessibility to larger sequenced cohorts of patients has increased the possibility of discovering novel genetic variants and genes associated with CHD[6,7].

Previous studies have defined the association of inherited and *de novo* variations as a cause of CHD[5–7]. Furthermore, these earlier studies have highlighted the differences between the genetic architecture of syndromic (with extracardiac malformations) and non-syndromic (isolated) CHD[6,7]. Continuing collaboration between the scientific community and healthcare teams has driven efforts to integrate and analyse larger cohorts of patients and has demonstrated the potential of this approach to uncover novel variants and genes associated with CHD[8].

Following on from these earlier efforts, we present an exome analysis of 4,747 CHD cases and 52,881 controls. Thus, we assembled a CHD case-control cohort which allowed us to expand our understanding of the disease and unravel new genetic associations. Moreover, since this cohort on hand contains one of the largest groups of non-syndromic CHD cases (n=2,929) studied so far, it has helped refine our understanding of the genetics underlying non-syndromic CHD.

We supplemented our analysis by considering expression patterns of significant and suggestive disease genes for syndromic and non-syndromic CHD at single-cell resolution. We meta-analysed data obtained in the case-control study in the context of a unique publicly available human heart single-cell dataset[9]. This complementary analysis identified putative biological processes enriched for genes differentially expressed in cardiac-specific cells, which also showed an increased burden of constrained non-synonymous variants in CHD cases compared to controls. Besides, we showed the differences of these processes between the distinct CHD categories.

Taken together, our findings allowed us to identify associations of ten genome-wide significant (*Bonferroni adjusted P < 0.05*) genes and an additional four genes at *FDR* 5% with CHD.

RESULTS

Cohort description and analysis workflow.

We combined and analysed the exomes of 4,747 CHD cases (aCHD, refers to all CHD cases) and 52,881 controls. CHD cases were further classified in syndromic CHD (sCHD, individuals with extracardiac malformations or neurodevelopmental disability, n=1,818) and non-syndromic CHD (nsCHD, individuals with isolated CHD, n=2,929). All samples and genetics variants were subjected to a sequence of quality control steps to obtain a final cohort of unrelated and matched-ancestry individuals as well as a set of high confidence variants for downstream analysis (see **Methods**).

We evaluated the distribution pattern of high-confidence loss-of-function (hcLOF) and missense constrained variants (missC) across a spectrum of LOF and missense constrained genes (**Methods**). In addition, we performed gene-based burden testing

to identify genes with a high risk of conferring CHD as well as the expression pattern at single-cell resolution and putative biological processes associated with the disease. **Figure 1** summarises the workflow followed in this study to discover novel associations with CHD.

Distinct pattern of loss-of-function constrained genes identified between sCHD and nsCHD.

Previous studies have suggested a larger effect of loss-of-function (LOF) variations leading to syndromic forms of CHD compared to non-syndromic forms[6,7]. To determinate if this holds true in this cohort on hand, we evaluated the mutational burden in sCHD and nsCHD compared to controls across the spectrum of loss-of-function constrained genes. Following the approach proposed by the gnomAD consortium[10], we binned 19,923 protein-coding genes (~1,900 genes per bin) based on the genes observed/expected LOF ratio upper fraction (termed LOEUF) and applied a logistic regression model (see **Methods**) per bin (i.e., gene-set). This allowed us to assess enrichment of three different functional categories of variant (hcLOF, missC and synonymous), stratified by CHD probands (aCHD, sCHD and nsCHD).

The highest enrichment was observed towards the most LOF constrained genes (bin 1) for hcLOF variants (**Figure 2**), with major contribution of sCHD cases (OR = 2.27, $P < 2 \times 10^{-16}$) compared to nsCHD (OR = 1.52, $P = 1.2 \times 10^{-13}$). A moderate enrichment was observed for missC variations, suggesting that this class of variants could have a similar (although smaller) functional impact compared to hcLOF variants. Although reduced in magnitude, this same pattern was also observed in the set of genes in the second LOEUF constraint bin, whereas no enrichment was observed towards less

LOEUF constraint bins (**Figure 2**). No enrichment was observed across the evaluated bins when the set of synonymous variants was used as a negative control set (**Figure 2)**.

When the same analysis was performed across the gene missense constraint spectrum, assessed by the observed/expected missense ratio upper fraction metric (termed MOEUF), a similar pattern as described above was observed (**Supplemental Figure 1**).

These results demonstrate a larger effect of hcLOF compared to missC variants across the LOEUF and MOEUF spectrum, with the major contribution observed in sCHD compared with nsCHD. Nevertheless, the results suggest that both hcLOF and missC variants are important genetic components contributing to CHD development.

Gene-based enrichment analysis.

To identify genes that confer a significant risk of CHD, we performed a case-control burden analysis by aggregating rare variants (MAF < 0.1%) at the gene level. It has been demonstrated that collapsing variants within genomic regions (e.g., genes) increases the power to discover new associations at low allele frequencies[11]. Following this principle, we conducted a Fisher exact test to identify genes with a significant burden of mutations in CHD cases compared to controls, evaluated independently for sCHD and nsCHD.

Like earlier comparable case-control exome studies[12–14], the burden test was performed separately for hcLOF (P_{lof}) and missC (P_{miss}), and the minimal *p-value* observed per gene between these two variant categories was selected as the study-wide *p-value* (*P*). hcLOF variants were defined using the LOFTEE tool, whereas missense constraint variants were defined based on different missense

deleteriousness prediction scores (see **Methods, Supplemental Figure 2**). Ten genes were identified with significant *P* after correcting for multiple testing using the Bonferroni method (**Table 1**). In addition, four genes showed significant associations with CHD at *FDR* 5%. Moreover, the evaluation of the set of synonymous variants showed a similar distribution of expected vs observed p-values, suggesting no genomic inflation of the test statistic (**Supplemental Figure 3**).

Notably, *KMT2A* (OMIM 159555) showed the highest enrichment among syndromic forms of CHD (**Figure 3a**). On the other hand, *NOTCH1* (OMIM 190198) was the most commonly mutated gene when evaluating the non-syndromic forms of CHD (**Figure 3b**) and warranted further investigation (companion manuscript).

Other genes reaching a significant level of association include *NSD1* (OMIM 606681), *TAB2* (OMIM 605101), *KAT6A* (OMIM 601408), *PTPN11* (OMIM 176876), *SMAD4* (OMIM 600993), *FLT4* (OMIM 136352), and the X-linked gene *BCOR* (OMIM 300485). They have all been previously described in the context of CHD, and our results corroborate these early findings.

The association of *PBX1* (OMIM 176310), *CTCF* (OMIM 604167) and *KAT6B* (OMIM 605880) with CHD (**Table 1**) has been previously reported in small cases series, and our results reinforce these earlier associations.

HCAR1 (OMIM 606923) and *SHOX2* (OMIM 602504) have not previously been associated with CHD at a genome-wide level. However, both genes were significantly associated at *FDR* 5% with nsCHD (**Figure 3b**).

Differentially expressed genes in cardiac-specific cells show a distinct enrichment pattern in syndromic and non-syndromic CHD.

Next, we investigated the mutational burden of differentially expressed genes (DEGs) in cardiac-specific cells by comparing aCHD cases vs controls, as well as considering sCHD and nsCHD independently. To this end, we meta-analysed the exome data with a publicly available human heart transcriptomic dataset generated from early developmental stages of the human heart (6.5 and 7 weeks-post-conception). Using the logistic regression framework mentioned above, we performed gene-set enrichment analysis on 15 distinct cardiac cell clusters reported by Asp *et al.*[9] Both hcLOF and missC mutations were evaluated independently and stratified further by proband CHD status (aCHD, sCHD and nsCHD) (**Figure 4**).

Five cardiac-specific cell clusters were found significantly enriched (Bonferroni adjusted $P < 0.05$) for hcLOF variations when analysing aCHD probands vs controls (**Figure 4**): Smooth muscle cells (C5), Cardiac neural crest cells (C14), Epicardium-derived cells (C3), Capillary endothelium (C0) and Atrial cardiomyocytes (C7). Enrichment of hcLOF variants for DEGs in Smooth muscle cells (cluster 5) showed significant contribution for both sCHD and nsCHD. Cardiac neural crest cells (C14) and Atrial cardiomyocytes (C7) mainly showed significant contributions for sCHD, whereas the cluster of Capillary endothelium cells was observed significantly enriched when analysing nsCHD vs controls (**Figure 4**).

A similar enrichment pattern was observed when analysing the set of missC variants (**Supplemental Figure 4**). In addition to the clusters C0, C5, C7 and C14; which were also found significantly enriched for hcLOF variants; two other cardiac-specific cell clusters showed significant burden of missC variants in CHD cases (aCHD) compared to controls: Endothelium/pericytes cells (C10) and Fibroblast cells (C2).

The synonymous variants set was used as a negative control to determine whether this method was appropriate in this context. No enrichment was found for any clusters evaluated in the distinct scenarios (Bonferroni adjusted $P > 0.05$, **Supplemental Figure 5**).

Despite the limited number of time points in embryogenesis analysed, these results provide valuable evidence regarding the possible mechanisms involved in the pathogenesis of CHD.

Gene Ontology (GO) enrichment analysis.

To provide additional supporting evidence for our previous findings, we performed Gene Ontology (GO) enrichment analysis to link the enriched DEGs in cardiac-specific cell clusters to biological processes. Specifically, we analysed the set of DEGs with an unadjusted $P < 0.01$ (Fisher Exact test) in the case-control burden analysis within the cell clusters showing enrichment in either aCHD, sCHD or nsCHD analysis (**Figure 5**).

Among the DEGs in cardiac-specific cells evaluated with the Enrichr tool[15] (see **Methods**), four clusters showed at least one GO term with *FDR* < 1%. Cluster 7 (atrial cardiomyocytes, **Figure 5a**) was mainly associated with biological processes involved in developing cardiac muscle tissue, and the observed signal was driven chiefly by *NKX2-5, MYH6, MYOCD, PKP2, BMP7, ANKRD1* and *ACTC1.* DE genes in cluster 0 (capillary endothelium, **Figure 5b**) showed enrichment for vasculogenesis, with contribution from *KDR*, *NOTCH1* and *RASIP1*. Cluster 5 (smooth muscle cells, **Figure 5c**) was associated with extracellular matrix organisation processes, with a noteworthy contribution of genes that contain a collagen-like domain (e.g., *COL14A1* and *COL1A2*), as well as *ELN* and *FBN2.* DEGs in cluster 10 (endothelium and pericyte

cells, **Figure 5d**), which demonstrated the higher enrichment of missC variants (**Supplemental Figure 4**), also showed enrichment of biological process involved in the cellular response to vascular endothelial growth factor stimulus and the regulation of cell migration as part of sprouting angiogenesis. *DLL4, FLT4, KDR, MEOX2* and *NOTCH1* all contributed to this cluster.

DISCUSSION

In this study, we aggregated 57,628 human exomes and conducted both a gene- and gene-set centred case-control burden analysis to improve our understanding of the genetic causes of CHD. After quality control at the sample and variant level, we leveraged a comprehensive CHD case-control cohort with unrelated and ancestry-matched individuals. Specifically, the availability of sub-phenotype descriptions allowed us to explore the differences between syndromic and non-syndromic forms of CHD.

By exploiting gene-level constraint information from an external resource[10], we investigated the contribution and the properties of loss-of-function and missense constraint variants independently for all CHD cases, as well as syndromic and non-syndromic CHD. Like earlier comparable studies[6,7], our results revealed a higher contribution of LOF variants to CHD compared to missense variants, confirming that this type of variation captures the largest size effect. Subsequently, the analysis of syndromic cases revealed a higher burden of LOF mutations when compared with the non-syndromic cohort. This effect was mainly a result of the contribution of genes with

a higher intolerance to loss-of-function variations. This same pattern was also observed when analysing the genes by missense constraint.

By performing a gene-based case-control burden analysis, we assessed the contribution to CHD at the gene level. Our analysis revealed ten genes that reached genome-wide significant levels of association with CHD (*NSD1*[16], *TAB2*[17], *KAT6A*[18], *PTPN11*[19], *CTCF*[20], *SMAD4*[21], *FLT4*[22], *NOTCH1*[23,24], *BCOR*[25] and *KMT2A*[26]). Previous studies have established a possible role of these genes as a cause of CHD, and our results confirm this association (**Table 1**). Furthermore, four candidate genes (*PBX1, SHOX2, KAT6B, HCAR1*) were found contributing to both syndromic and non-syndromic CHD at *FDR* 5%.

PBX1 has been primarily associated with congenital abnormalities of the kidney and urinary tract (CAKUT)[27]; however, previous studies have reported isolated cases carrying *de novo* missense variations leading to syndromic CHD[27,28]. In line with these early reports, our analysis revealed a significant burden of missense constrained variants in *PBX1* in syndromic CHD patients (**Table 1**). It has also been demonstrated that deficiency of *Pbx1* impacts branchial arch artery patterning and results in the failure of cardiac outflow tract septation[29]. Interestingly, this gene was also found differentially expressed in Epicardium and Smooth muscle cells (**Supplemental Figure 6**). Thus, our findings suggest that *PBX1* contributes significantly to syndromic forms of CHD.

Our analysis found that in the nsCHD cohort, *SHOX2* was significantly enriched (at *FDR* 5%) for missC variants (**Table 1**). Recent studies in animal models have

demonstrated that the *Shox2* null mice are embryonic-lethal[30]. Cardiovascular defects identified in these mice included an abnormally low heartbeat rate, a severely hypoplastic Sinoatrial Node (SAN), hypoplastic or absent sinus valves[30], and other atrial abnormalities (e.g., enlarged atrial chamber and thinner atrial wall). Subsequently, *SHOX2* has been described as playing a key role in developing the Sinoatrial Node[30,31]. In addition, *SHOX2* was identified as a significant DEG in atrial cardiomyocytes (**Figure 5a, Supplemental Figure 6**), providing further supporting evidence of its role in heart development, most likely by regulating the activity of *NXK2-5*[30,32] and *TBX5*[33]. These results imply that *SHOX2* is a plausible novel non-syndromic CHD gene.

Truncating variants in the Lysine Acetyltransferase 6B gene (*KAT6B*) have been associated with Say–Barber–Biesecker–Young–Simpson Syndrome (SBBYSS, OMIM 603736) and Genitopatellar Syndrome (GTPTS, OMIM 606170). Heart defects have been reported as part of the phenotypic spectrum of SBBYSS[34]. In a recent study of 32 individuals with *KAT6B* disorder, 47% showed cardiovascular anomalies, mainly atrial septal defects, ventricular septal defects, and patent ductus arteriosus. Our results have identified that *KAT6B* was differentially expressed in the cluster of atrial cardiomyocytes cells (**Supplemental Figure 6**), which suggest a possible role in the early cardiac development program. Our analysis extends previous findings associating loss-of-function variations in *KAT6B* to sCHD.

The Hydroxycarboxylic Acid Receptor 1 (*HCAR1*) does not appear to have been associated with CHD thus far, and our findings suggest this gene may be a novel

candidate CHD gene. It was not differentially expressed in any of the cardiac-specific cell clusters analysed.

By meta-analysing the genomic data with heart single-cell transcriptomic data, we investigated the pattern of expression of DEGs for aCHD, sCHD and nsCHD in a range of cardiac-specific cells. Using Gene Ontology enrichment as a complementary analysis, we identified key gene markers and biological processes associated with CHD. Unlike previous studies[7,35], which focused on whole heart bulk-RNA sequencing data, the use of transcriptomic data at a single-cell resolution significantly enhanced our work. It allowed the analysis of candidate gene expression patterns to specific cardiac cell clusters important for early cardiac development.

This analysis revealed an interesting expression pattern for sCHD and nsCHD in different cardiac cell clusters. Also, we demonstrated that missense constrained variants could have a similar functional impact compared to loss-of-function variants, although to a lesser degree. For instance, the significant enrichment for sCHD in cardiac neural crest cells (cNCCs) suggests a broader contribution of patients affected by syndromic occurrences, not limited to heart development only. Perturbations in the cNCCs migration process can lead to a wide spectrum of human cardio-craniofacial syndromes, including DiGeorge Syndrome (22q11.2 Deletion Syndrome, OMIM 188400) and CHARGE (OMIM 214800). The enrichment observed in capillary endothelium and pericyte cells for nsCHD, associated with the vasculogenesis process, suggests that the phenotypic occurrence in these patients is limited to the cardiovascular system rather than affecting a broader spectrum of cells.

Whilst the results are promising, they are limited because the currently available human heart single-cell map[9] is incomplete (e.g., only a few early developmental time points). Therefore, future studies integrating mouse and human single-cell heart and whole-embryo data are warranted.

In summary, we meta-analysed ~57,000 exomes and complemented this with the study of transcriptomic data at single-cell resolution. This study has enabled us to strengthen previously described associations with CHD, discover novel candidate genes, and provide a deeper understanding of the pathophysiological mechanisms underlying CHD.

METHODS

Cohort description

To create a comprehensive CHD case-control cohort, exome sequencing data from multiple individuals was combined in a unique reference dataset. CHD cases were mainly sequenced as part of an initiative from the German Competence Network for Congenital Heart Defects, the Deciphering Developmental Disorder (DDD) project and the University of Nottingham (UK); controls were sequenced as part of the UK Biobank (UKBB). Samples from the UKBB dataset with phenotype description labelled as Schizophrenia (SCZ), bipolar disorder (BP) or developmental delay (DD) were excluded from the analysis. Accordingly, a small fraction of samples in the UKBB cohort (127 samples), labelled as CHD cases, were included in the analysis. In total, we assembled an exome dataset consisting of 57,628 samples (4,747 CHD cases and

52,881 controls). The assembled cohort was processed using the same computational pipelines as described below.

Alignment and variant calling

CRAM-level data for all previously and newly sequenced samples were realigned to the human genome build GRCh38 using the BWA tool (version 0.7). Variants were jointly called using the Genome Analysis Toolkit (GATK, version 4.1), following the Broad Institute best-practice guidelines for germline single nucleotide variants (SNVs) and short insertions/deletions (indels). Briefly, HaplotypeCaller was used in GVCF mode to process samples individually, such that every position in the genome was assigned with a likelihood of being or not being a variant. The GenomicsDB (https://github.com/Intel-HLS/GenomicsDB) tool was used then to import and merge the per-sample GVCF genotype data. Samples were then jointly genotyped for high confidence alleles using the GenotypeGVCFs tool. The Variant Quality Score Recalibration (VQSR) in GATK was applied independently for SNVs and indels to assess variant call accuracy. The complete process was executed using standard pipelines from the Human Genetics Informatics (HGI) unit from the Wellcome Trust Sanger Institute (WTSI).

To perform scalable downstream analysis of the sequencing data, the multi-sample cohort-VCF generated from the previous step was imported into Hail 0.2 (https://hail.is), a python-like library for analysing genomic data at scale, using the function *hl.import_vcf.* Subsequence sample- and variant-level quality control (QC) was performed using the Hail framework (see below), following mainly the workflows proposed by the gnomAD project[10], otherwise explicitly specified. The Hail-based

pipelines (https://github.com/enriquea/wes_chd_ukbb) used in this study are publicly available on GitHub.

Sample QC

Hard filters. To compute sample QC metrics, a set of high-confidence variants was defined by applying the following criteria: (i) bi-allelic, (ii) variants with high call-rate (> 0.99) across all samples in the call set and (iii) common single nucleonic variants (allelic frequency > 0.1%). The individual's chromosomal sex was inferred by calculating the inbreeding coefficient (*F-stat*) on chromosome X over the set of variants described above. The *hl.impute_sex* Hail function was used to perform the computation. This approach adopts the same implementation as the PLINK tool (v1.7). In addition, the coverage of the chromosome Y (normalized to chromosome 20) was used with the *F* stat to define the sample sex as follow: *male*: $F > 0.6$ and normalized Y coverage > 0.1, *female*: $F < 0.4$ and normalized Y coverage < 0.1. Samples with values outside these ranges were labelled as sex unspecific (**Supplemental data, Figure S1**). Samples were marked as failing hard filters if: a) chromosomal sex was unspecific, b) exhibited sample-specific low call rates (< 0.85) and c) mean coverage on chromosome 20 was equal to zero. **Table S1** (**Supplemental data**) summarises the number of samples affected per hard filter.

Inferring population ancestry. The 1000 Genomes Phase 3 sequence data aligned to the human genome build GRCh38 (European Variation Archive (EVA) accession: PRJEB30460) was used to impute the global ancestry within the samples in the exome sequencing cohort. Both datasets were first merged based on locus and reference/alternate alleles. After merging, the Hail function *hl.hwe_normalized_pca*

was used to compute the ten first principal components on the subset of the well-behaved variants, defined as described above (see Hard filters section). A total of ~76,000 variants were included in the final set.

The set of 2,548 samples with known ancestry (from the 1000 Genomes Phase 3 dataset) was leveraged to build a random forest-based classifier using the top 15 computed principal components (PCs) as input features. Two-thirds of these samples were used as a *training dataset* and the remainder used as a *test dataset*. This step was combined with a recursive feature (a.k.a principal components) elimination procedure to define the optimal combination of PCs achieving the highest accuracy in the classification on the test data. In addition, a 10-fold cross-validation step was used for tuning the model parameters as previously described[36].

The model achieving the highest accuracy (>0.97) was then used to predict the ancestry of the remaining samples (discovery dataset with unknown ancestry). Each sample was broadly assigned to one of European (EUR), American (AMR), African (AFR), East Asian (EAS) or South Asian (SAS) population labels if random forest probability (*p*) > 0.8. Samples failing this threshold were labelled as OTHER. **Figure S2** and **Table S2** (**Supplemental data**) summarise the ancestry inference process results. The implemented approach showed high accuracy in classifying samples with reported ethnicity from the UK Biobank cohort (**Supplemental data, Table S3**).

Inferring sample relatedness. The *hl.pc_relate* function from Hail was used to compute the relatedness between samples. Relatedness was computed among samples passing the hard filters. A variant was considered for inferring relatedness if it met the following criteria: 1) protein-coding exonic variant, 2) autosomal, 3) bi-allelic single nucleotide variants (SNVs), 4) call rate across samples > 95%, 5) allele frequency

(internal) > 1% and 6) LD-pruned with a cut-off at *r2* = 0.1. After running *hl.pc_relate*, Hail's *hl.maximal_independent_set* function was used to select the largest set of samples with no pair of samples related at the second-degree relatedness or closer (kinship coefficient > 0.125), prioritising cases over controls. This process filtered out a total of 3,782 samples (either twin/duplicated or first-degree relatives).

Platform inference. Detailed capture platform meta-data information was missing for a fraction of the samples within the assembled cohort (~20%). To impute a platform for these samples, we adopted the data-driven approach proposed by gnomAD[10]. In brief, a list of the known exome capture intervals across multiple exome capture products was compiled for imputing samples platforms (including Agilent Sure Select All Exons products (version 2 to 5) and IDT xGEN). Only bi-allelic variants falling within these regions were included in the analysis. A sample per interval call-rate matrix was computed by considering the set of biallelic variants within each interval. The call-rate values were further discretised as non-called (0) and called (1) by applying a call-rate cut-off at 0.25 and principal component analysis performed on the discrete matrix. The top seven principal components (variance explained higher than 98%) were used as input for HDBSCAN (https://hdbscan.readthedocs.io), an unsupervised clustering method that allowed us to group and assign generic sample platform labels. **Figure S3** shows the samples projected onto principal components two and three. This method assigned the platform accurately for 100% of the samples in the UK Biobank (those with known platform labels), demonstrating the validity of this approach.

Platform- and population-specific outliers filtering. Sample ancestry and capture platform are two of the most frequent cofounders when analysing exome sequencing

data. Thus, we computed a set of sample quality control metrics stratified by population and platform to detect sample outliers. Specifically, we computed the number of deletions, the number of insertions, the number of SNVs, the ratio of deletions to insertions, the ratio of transitions to transversions, and the ratio of heterozygous to homozygous variants using the Hail function *hl.sample_qc*. A sample was marked as an outlier and filtered out if the value for a given QC metric was four median absolute deviations (MAD) from its median. **Table S4 (Supplemental data)** summarises the number of samples detected as outliers per evaluated QC metric.

Final sample QC and evaluation. After applying the above sample QC steps and filtering out the samples without approval for analysis, our cohort consisted of 49,308 samples (**Supplemental data, Table S5**). At this stage, multi-allelic variants were split using the Hail function *hl.split_multi_hts,* and the dataset was filtered to high-quality genotypes. Genotypes were defined as high-quality if: a) dept of coverage >= 10, b) genotype quality >= 20 and c) genotype allele balance of heterozygotes > 0.20.
In addition, we evaluated the per sample distribution of the depth of coverage (DP) and genotype quality (GQ) stratified by case/control and male/female status. Our analysis revealed a comparable distribution of these metrics between cases/controls (**Supplemental data, Figure S4**) and male/females (**Supplemental data, Figure S5**). Mean DP values ranged between 20-35X (recommended cut-off is >10X) whereas GQ values ranged between 50-80 (recommended cut-off is >20).

Variant QC

To define a set of high-quality variants for downstream analysis, we then applied several QC steps to the variants present in samples passing the sample QC process.

Hard filters. We followed the variant QC scheme proposed by Karczewski *et al.*[10], where variants were flagged as failing hard filters if they showed a) an excess of heterozygotes (inbreeding coefficient < -0.3) and b) an absence of at least one sample with a high-quality genotype (allele-count zero, as defined above).

RF model. A random forest (RF) model was trained and applied to distinguish true variations from potential false positives[10]. Positive training sets were downloaded from gnomAD repository (gs://gcp-public-data--gnomad/truth-sets/hail-0.2). Variants failing traditional GATK hard filters (QD < 2 or FS > 60 or MQ < 30) were used as a negative training set. Allele- and site-specific sequencing quality metrics were used as features for training the model (**Supplemental data, Table S6**). Features were imputed using its median where the value was missing. The chromosome 20 (test set) was left out of the training process for evaluation proposes. The final RF model achieved an accuracy >0.97 on this set of variants (test set). A variant was filtered out if the RF probability of being false positive was higher than 0.8.

VQSR filter. In addition to the proposed RF model, we applied the conventional GATK Variant Quality Score Recalibration (VQSR) as a complementary approach to filter out low-quality variants. We used the recommended annotations and training datasets as suggested by the GATK best practices (https://gatkforums.broadinstitute.org/gatk). Both SNVs and indels were excluded if they failed the VQSR filter, according to the default settings. This allowed us to identify a fraction of variants that were likely false positives that passed the RF filter (**Supplemental data, Figure S6**).

Coverage. Finally, we defined a variant as passing the QC if it a) was covered by the major capture platforms used in the assembled cohort (different versions of Agilent Sure Select All Exome and IDT xGen panel 1) and b) showed coverage of 10X or more in at least the 90% of the samples in the gnomAD genome dataset (version 3.1.0).

Table S7 (Supplemental data) summarises the number of variants affected by each applied filter and the final number of variants considered for further analysis.

Variant annotation

The cohort-VCF file was annotated using the Variant Effect Predictor tool (API version 94) with the flag *--everything*. The most severe variant consequence per protein-coding transcript was considered. The variant consequence severity was set based on the severity rank from Ensembl (https://www.ensembl.org), which prioritise variants as follows: protein-truncating > protein-altering > synonymous variants. The VEP tool functionalities were extended by using the plug-ins CADD (version 1.6) and dbNSFP[37] (version 4.1a) to annotate different missense variant pathogenicity scores (CADD[38], MPC[39], REVEL[40] and MVP[41]).

Defining a set of loss-of-function and missense constraint variants

We enriched the dataset for high confidence loss-of-function (hcLOF) variants and missense constrained (missC) variants. hcLOF variants were annotated as indicated by the LOFTEE tool (https://github.com/konradjk/loftee) with its default parameters and included stop-gained, essential splice, and frameshift variants. To define a set of missC variants, we evaluated four state-of-art pathogenicity prediction scores: a) CADD[38], MPC[39], REVEL[40] and MVP[41]. Specifically, the performance of these scores

was assessed by classifying benign and pathogenic missense variants (accessed through the ClinVar database, https://www.ncbi.nlm.nih.gov/clinvar) in the context of known CHD genes. In brief, receiver operating characteristic (ROC) analysis was conducted for benign and pathogenic variants within known CHD genes. The analysis was further stratified by splitting the gene set into LOF constraint (LOEUF < 0.35) and LOF non-constraint (LOEUF >= 0.35) genes. A score was defined as a 'good predictor' if achieved an area-under-curve (AUC) > 90% in both evaluated scenarios. Three of these scores (CADD, REVEL and MVP) met this criterion. A missense variant was defined as missC if it was predicted as likely deleterious by at least two of these scores based on the optimal threshold suggested by the ROC analysis (**Supplemental Figure 2**).

Defining rare variants

Variants were filtered based on the cohort-specific allelic frequency ('internal' AF) as well as using external datasets. A variant was defined as rare if AF was lower than 0.001 (MAF 0.1%) in the gnomAD database[10] (both exomes v2.1.1 and genomes v3.0.0), the RUMC cohort[42], as well as AFs from an *in-house* German exome sequencing cohort.

Gene-set enrichment analysis

Generation of gene sets. Gene set-level association analysis was performed to assess whether an excess of the possible pathogenic variants was enriched for a particular category of genes (as described below). This procedure was executed for the following gene sets:

a) LOEUF gene bins: Constraint loss-of-function (LOF) metrics per protein-coding genes were accessed through gnomAD resource[10]. Genes were ranked by their observed/expected LOF mutation ratio upper fraction (termed LOUEF), and ten bins with an equal number of genes (~1,900 genes per bin) were defined. Lower values of LOEUF (e.g., bins 1 and 2) denotes most LOF constrained genes.

b) MOEUF gene bins: Similar as described above for LOEUF genes, but genes were binned based on its observed/expected missense mutation ratio upper fraction (termed MOEUF).

c) Differentially expressed genes (DEGs) in cardiac-specific cells: DEGs identified in 15 distinct cardiac cell clusters reported by Asp *et al.*[9] In brief, genes were determined as significantly differentially expressed in a particular cardiac cell cluster if the averaged log-fold change (logFC) > 0 (upregulated) at FDR 1%.

Gene set-based association analysis. For each sample within the filtered dataset, we generate a Minimal Allele Count (MAC) metric by aggregating high confidence Genotypes (DP >= 10, GQ >= 20 and allelic balance heterozygous > 0.2) across the genes within the gene set. Then, a burden logistic regression test was performed using CHD case/control status as response and the five first ancestry principal component and sex as covariates using the Hail function *hl.logistic_regression_rows*. The analysis was stratified at the sample and variant level. At the sample level, the data was divided based on the syndromic status; three categories were tested: aCHD (all CHD cases vs control), nsCHD (non-syndromic CHD cases vs control) and sCHD (syndromic CHD cases vs control). At variant level, three different groups were evaluated based on the predicted severity of the variants: hcLOF (most severe), missC and synonymous. The synonymous variant set was used as a negative control set at the variant level to

evaluate for potential artefacts. The odds ratio (exp (beta coefficient)), 95% confidence interval and *p-value* metrics were used to evaluate significant enrichment.

Gene-based burden testing

We performed case-control gene-centred burden test analysis to assess genes with significant association with CHD. Fisher Exact test was performed independently for rare (MAF 0.1%) hcLOF and missC variants. To define the significant study-wide *p-value*, the minimal *p-value* (*P*) per gene between these two categories was chosen. The analysis was further stratified by syndromic status to assess the distinct contribution of these categories to CHD. A gene was defined as genome-wide significant if it reached a Bonferroni corrected $P < 0.05$ and suggested significant if *FDR* < 5%. In addition, the set of synonymous variants was used as a negative control set since no difference between cases/control is expected on this set of variations (quantile-quantile plots, **Supplemental Figure 3**).

Gene Ontology enrichment analysis

The R-package *Enrichr* (with the *Biological_Process_2018* database) was used to perform Gene Ontology (GO) enrichment analysis. The analysis was conducted on the differentially expressed genes (DEGs) in cardiac-specific cell clusters, which also showed unadjusted $P < 0.01$ (Fisher Exact test) from the case-control burden analysis. The evaluated DEGs were previously reported by Asp *et al.*[9] with no additional processing. GO terms with only one overlapping gene were filtered out. A biological process term was considered significant if *FDR* < 1% as reported by the Enrichr tool[15].

Table 1. Top 21 genes in the case-control burden analysis using the Fisher Exact test stratified by syndromic status (sCHD and nsCHD). A total of 16,351 genes were tested per variant type (hcLOF and missC). Analysis: sCHD or nsCHD vs controls. Consequence: denotes the consequence group with the minimal p-value (*P*). sCHD: number of syndromic cases (heterozygous). nsCHD: number of non-syndromic cases (heterozygous). Controls: number of controls (heterozygous). *P*: the minimal p-value per gene between P_{lof} and P_{miss}. P adj (FDR): Adjusted minimal p-value (*P*) using the B-H method with n = 2*16,351. P adj (Bonferroni): Adjusted minimal p-value (*P*) using the Bonferroni method with n = 2*16,351. In bold are highlighted the ten genes with *Bonferroni adjusted P* < 0.05.

Genes	Analysis	Consequence	sCHD	nsCHD	Controls	*P*	*P* adj (FDR)	*P* adj (Bonferroni)
KMT2A	sCHD	hcLOF	8	0	0	9.76E-13	3.19E-08	3.19E-08
SMAD4	sCHD	missC	11	3	16	2.47E-10	4.04E-06	8.09E-06
NOTCH1	nsCHD	hcLOF	2	7	0	8.48E-10	2.77E-05	2.77E-05
PTPN11	sCHD	missC	11	5	25	8.78E-09	9.57E-05	2.87E-04
TAB2	sCHD	hcLOF	5	1	0	3.13E-08	2.56E-04	1.02E-03
NSD1	sCHD	hcLOF	5	1	1	1.83E-07	1.20E-03	5.98E-03
BCOR	sCHD	hcLOF	4	0	0	9.93E-07	4.06E-03	3.25E-02
KAT6A	sCHD	hcLOF	4	1	0	9.93E-07	4.06E-03	3.25E-02
PBX1	sCHD	missC	6	3	6	7.73E-07	4.06E-03	2.53E-02
FLT4	nsCHD	hcLOF	0	5	0	3.32E-07	5.43E-03	1.09E-02
CTCF	sCHD	missC	4	1	1	4.84E-06	1.58E-02	1.58E-01
KAT6B	sCHD	hcLOF	4	1	1	4.84E-06	1.58E-02	1.58E-01
SHOX2	nsCHD	missC	1	10	21	1.81E-06	1.98E-02	5.93E-02
HCAR1	nsCHD	missC	2	9	18	4.40E-06	3.60E-02	1.44E-01
ADNP	sCHD	hcLOF	3	0	0	3.15E-05	6.44E-02	1.00E+00
CHD7	sCHD	hcLOF	3	0	0	3.15E-05	6.44E-02	1.00E+00
EP300	sCHD	hcLOF	3	1	0	3.15E-05	6.44E-02	1.00E+00
KMT2D	sCHD	hcLOF	3	0	0	3.15E-05	6.44E-02	1.00E+00
KRT25	sCHD	missC	8	3	31	2.51E-05	6.44E-02	8.19E-01
QRICH1	sCHD	hcLOF	3	0	0	3.15E-05	6.44E-02	1.00E+00
SLC38A9	nsCHD	missC	0	6	6	1.19E-05	7.78E-02	3.89E-01

References

1. van der Linde, D. *et al.* Birth Prevalence of Congenital Heart Disease Worldwide. *J. Am. Coll. Cardiol.* **58**, 2241–2247 (2011).
2. van der Bom, T. *et al.* The changing epidemiology of congenital heart disease. *Nat. Rev. Cardiol.* **8**, 50–60 (2011).
3. Zaidi, S. & Brueckner, M. Genetics and Genomics of Congenital Heart Disease. *Circ. Res.* **120**, 923–940 (2017).
4. Homsy, J. *et al.* De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. **350**, 1262–1266 (2015).
5. Izarzugaza, J. M. G. *et al.* Systems genetics analysis identifies calcium-signaling defects as novel cause of congenital heart disease. *Genome Med.* **12**, 76 (2020).
6. Sifrim, A. *et al.* Distinct genetic architectures for syndromic and non-syndromic congenital heart defects identified by exome sequencing. *Nat. Genet.* **48**, 1060–5 (2016).
7. Jin, S. C. *et al.* Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. *Nat. Genet.* (2017). doi:10.1038/ng.3970
8. Audain, E. *et al.* Integrative analysis of genomic variants reveals new associations of candidate haploinsufficient genes with congenital heart disease. *PLOS Genet.* **17**, e1009679 (2021).
9. Asp, M. *et al.* A Spatiotemporal Organ-Wide Gene Expression and Cell Atlas of the Developing Human Heart. *Cell* **179**, 1647-1660.e19 (2019).
10. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
11. Povysil, G. *et al.* Rare-variant collapsing analyses for complex traits: guidelines and applications. *Nat. Rev. Genet.* **20**, 747–759 (2019).
12. Guo, M. H., Plummer, L., Chan, Y.-M., Hirschhorn, J. N. & Lippincott, M. F. Burden Testing of Rare Variants Identified through Exome Sequencing via Publicly Available Control Data. *Am. J. Hum. Genet.* **103**, 522–534 (2018).
13. Flannick, J. *et al.* Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. *Nature* **570**, 71–76 (2019).
14. Singh, T., Neale, B. M. & Daly, M. J. Exome sequencing identifies rare coding variants in 10 genes which confer substantial risk for schizophrenia on behalf of the Schizophrenia Exome Meta-Analysis (SCHEMA) Consortium*. *medRxiv* 2020.09.18.20192815 (2020). doi:10.1101/2020.09.18.20192815
15. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90-7 (2016).
16. Tatton-Brown, K. *et al.* Genotype-Phenotype Associations in Sotos Syndrome: An Analysis of 266 Individuals with NSD1 Aberrations. *Am. J. Hum. Genet.* **77**, 193–204 (2005).
17. Thienpont, B. *et al.* Haploinsufficiency of TAB2 Causes Congenital Heart Defects in Humans. *Am. J. Hum. Genet.* **86**, 839–849 (2010).
18. Urreizti, R. *et al.* Five new cases of syndromic intellectual disability due to KAT6A mutations: widening the molecular and clinical spectrum. *Orphanet J. Rare Dis.* **15**, 44 (2020).
19. Sarkozy, A. *et al.* Correlation between PTPN11 gene mutations and congenital heart defects in Noonan and LEOPARD syndromes. *J. Med. Genet.* **40**, 704–8 (2003).

20. Gregor, A. *et al.* De novo mutations in the genome organiser CTCF cause intellectual disability. *Am. J. Hum. Genet.* **93**, 124–131 (2013).
21. Lin, A. E. *et al.* Gain-of-function mutations in SMAD4 cause a distinctive repertoire of cardiovascular phenotypes in patients with Myhre syndrome. *Am. J. Med. Genet. A* **170**, 2617–31 (2016).
22. Reuter, M. S. *et al.* Haploinsufficiency of vascular endothelial growth factor related signaling genes is associated with tetralogy of Fallot. *Genet. Med.* **21**, 1001–1007 (2019).
23. Kerstjens-Frederikse, W. S. *et al.* Cardiovascular malformations caused by NOTCH1 mutations do not keep left: Data on 428 probands with left-sided CHD and their families. *Genet. Med.* **18**, 914–923 (2016).
24. Garg, V. *et al.* Mutations in NOTCH1 cause aortic valve disease. *Nature* **437**, 270–274 (2005).
25. Fan, Z. *et al.* BCOR regulates mesenchymal stem cell function by epigenetic mechanisms. *Nat. Cell Biol.* **11**, 1002–1009 (2009).
26. Baer, S. *et al.* Wiedemann-Steiner syndrome as a major cause of syndromic intellectual disability: A study of 33 French cases. *Clin. Genet.* **94**, 141–152 (2018).
27. Arts, P. *et al.* Paternal mosaicism for a novel PBX1 mutation associated with recurrent perinatal death: Phenotypic expansion of the PBX1-related syndrome. *Am. J. Med. Genet. A* **182**, 1273–1277 (2020).
28. Alankarage, D. *et al.* Functional characterisation of a novel PBX1 de novo missense variant identified in a patient with syndromic congenital heart disease. *Hum. Mol. Genet.* **29**, 1068–1082 (2020).
29. CP, C. *et al.* Pbx1 functions in distinct regulatory networks to pattern the great arteries and cardiac outflow tract. *Development* **135**, 3577–3586 (2008).
30. Espinoza-Lewis, R. A. *et al.* Shox2 is essential for the differentiation of cardiac pacemaker cells by repressing Nkx2-5. *Dev. Biol.* **327**, 376–385 (2009).
31. Munshi, N. V. Gene regulatory networks in cardiac conduction system development. *Circulation Research* **110**, 1525–1537 (2012).
32. Yang, T., Huang, Z., Li, H., Wang, L. & Chen, Y. P. Conjugated activation of myocardial-specific transcription of Gja5 by a pair of Nkx2-5-Shox2 co-responsive elements. *Dev. Biol.* **465**, 79–87 (2020).
33. Puskaric, S. *et al.* Shox2 mediates Tbx5 activity by regulating Bmp4 in the pacemaker region of the developing heart. *Hum. Mol. Genet.* **19**, 4625–4633 (2010).
34. Gannon, T. *et al.* Further delineation of the KAT6B molecular and phenotypic spectrum. *Eur. J. Hum. Genet.* **23**, 1165–1170 (2015).
35. Sevim Bayrak, C., Zhang, P., Tristani-Firouzi, M., Gelb, B. D. & Itan, Y. De novo variants in exomes of congenital heart disease patients identify risk genes and pathways. *Genome Med.* **12**, 9 (2020).
36. Perez-Riverol, Y., Kuhn, M., Vizcaíno, J. A., Hitz, M. P. & Audain, E. Accurate and fast feature selection workflow for high-dimensional omics data. *PLoS One* **12**, e0189875 (2017).
37. Liu, X., Li, C., Mou, C., Dong, Y. & Tu, Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human non-synonymous and splice-site SNVs. *Genome Med.* **12**, 103 (2020).
38. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).

39. Samocha, K. E. *et al.* Regional missense constraint improves variant deleteriousness prediction. *bioRxiv* 148353 (2017). doi:10.1101/148353
40. Ioannidis, N. M. *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am. J. Hum. Genet.* **99**, 877–885 (2016).
41. Qi, H. *et al.* MVP predicts the pathogenicity of missense variants by deep learning. *Nat. Commun.* **12**, 510 (2021).
42. Kaplanis, J. *et al.* Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature* **586**, 757–762 (2020).

Main figures

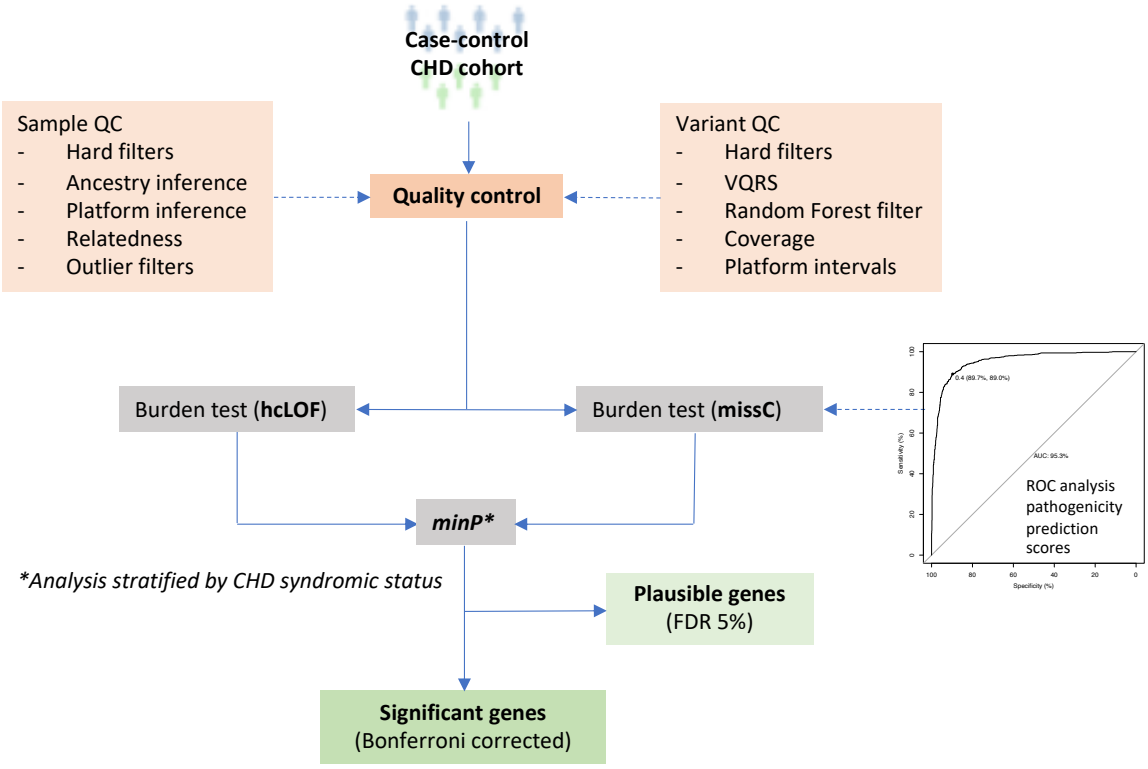


Figure 1. Analysis workflow for disease gene discovery. Quality control processes were conducted at sample and variant levels. Gene-based case-control burden testing (Fisher's Exact test) was performed for high-confidence loss-of-function (hcLOF) and missense constrained variants (missC) independently. The per gene minimal *p-value* (*P*) from both analyses was set as the study-wide *p-value*, corrected for multiple testing using the Bonferroni and B-H methods. The burden analysis was stratified by syndromic status vs control.

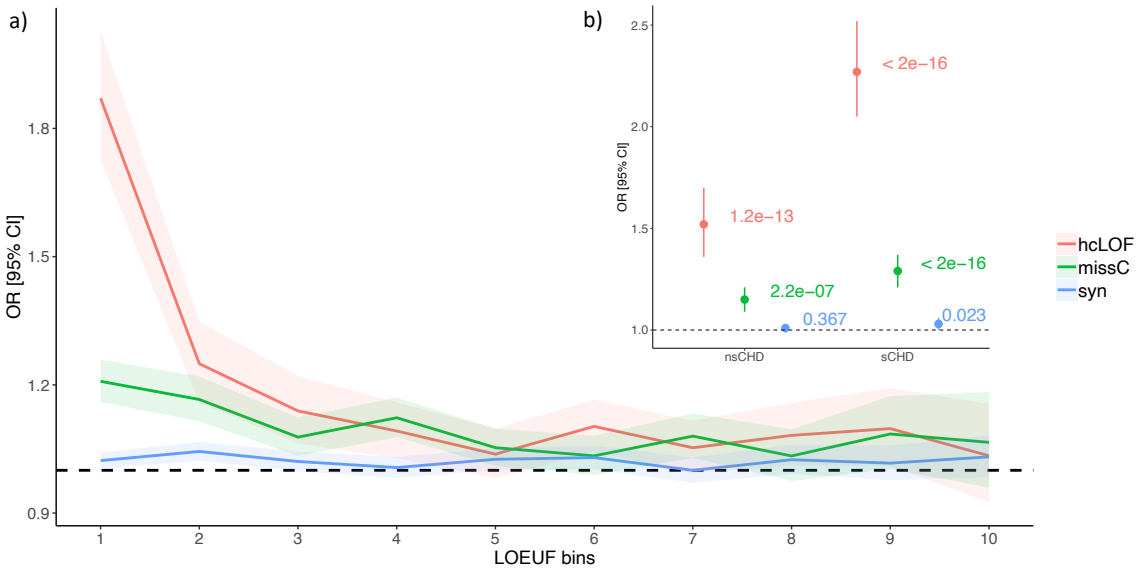


Figure 2. Enrichment analysis across the LOF constraint gene spectrum. Protein-coding genes were binned based on the LOEUF metric as proposed by gnomAD. Every bin contains ~1,900 genes. Top bins (1, 2) contain genes with the highest intolerance to loss-of-function. a) Enrichment analysis comparing aCHD vs controls. b) Enrichment analysis stratified by syndromic status (sCHD and nsCHD) vs controls in the top constraint LOF bin (1). The x-axis indicates the constraint bins; the y-axis shows the Odd Ratios (OR) and the 95% confidence interval.

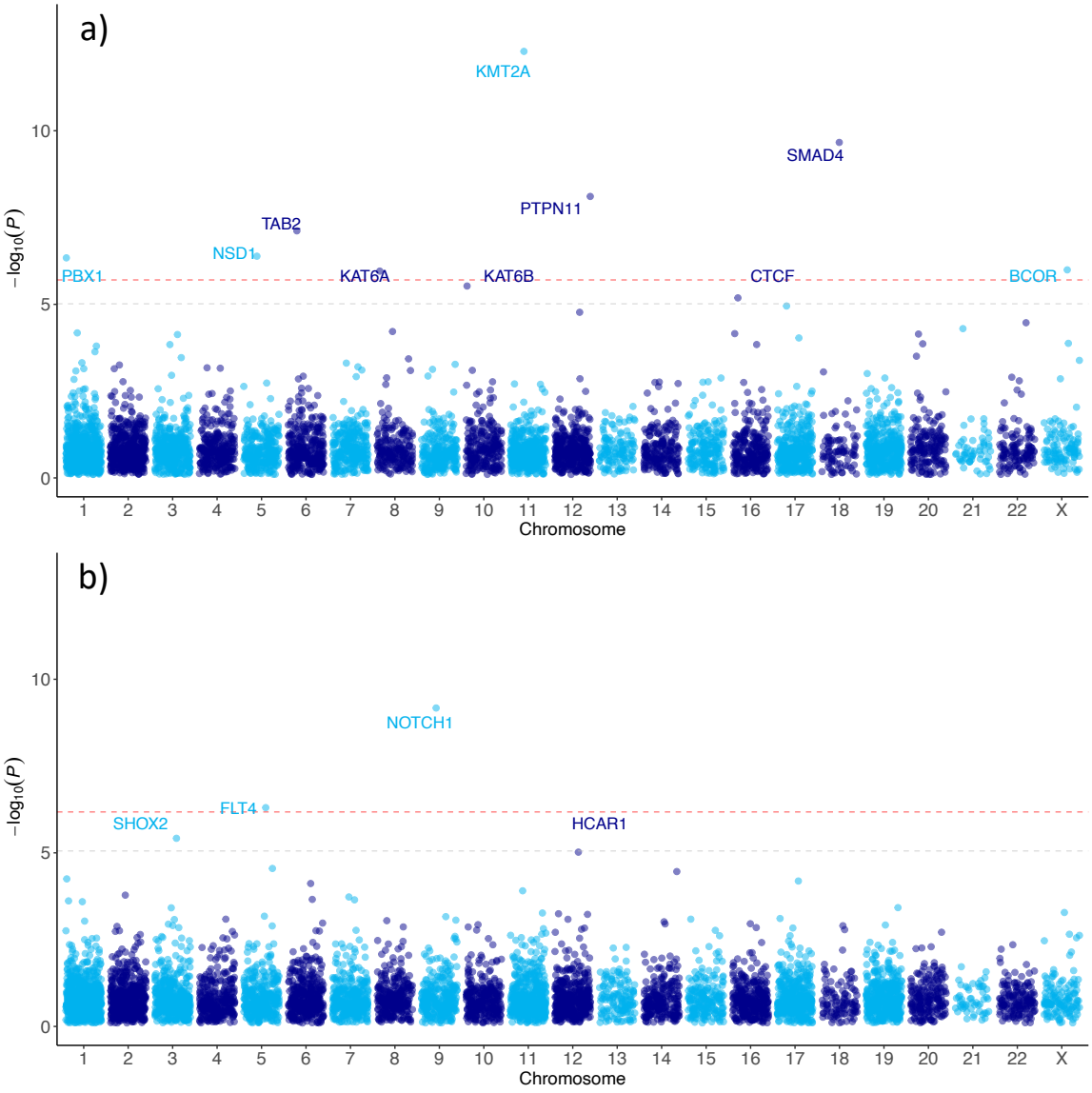


Figure 3. Log-transformed minimal p-value (*P*) per gene (y-axis) against its chromosomal location (x-axis). Red dashed line denotes the threshold for genes reaching exome-wide significance (Bonferroni adjusted $P < 0.05$); grey dashed line marks the threshold for genes reaching suggestive exome-wide significance (FDR 5%). a) Burden analysis of sCHD vs controls; b) burden analysis of nsCHD vs controls.

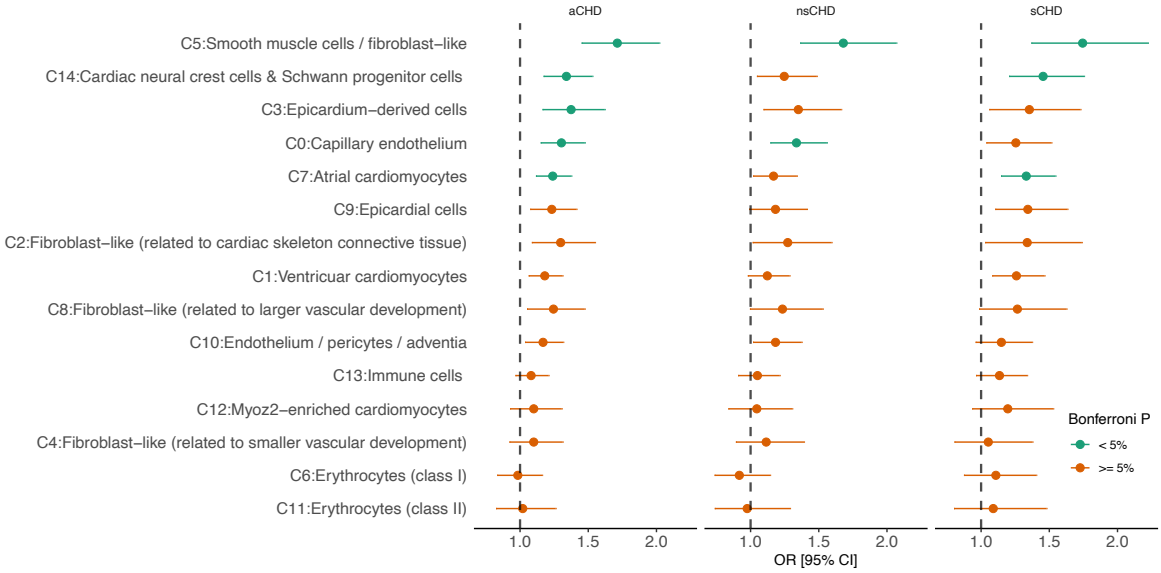
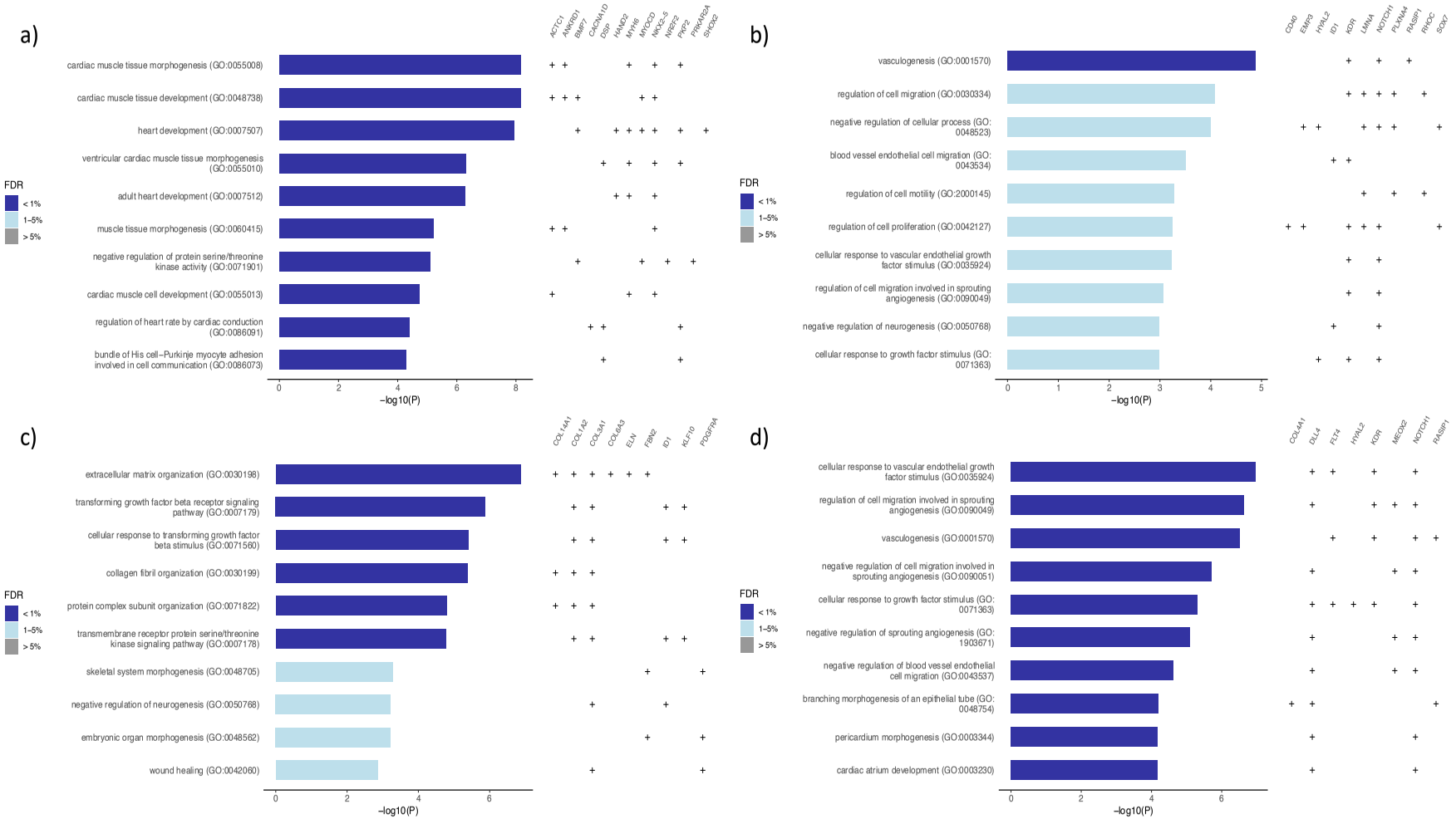


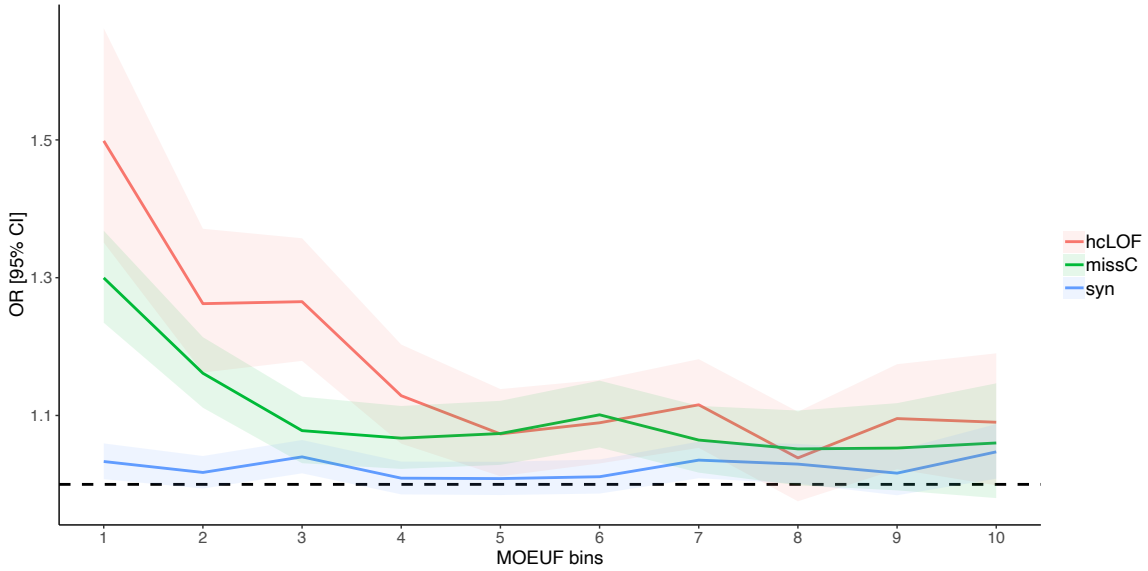
Figure 4. Logistic regression-based enrichment analysis of differentially expressed genes (DEGs) in cardiac-specific cell clusters for high-confidence loss-of-function variants (hcLOF). The analysis was stratified by syndromic status (aCHD, sCHD and nsCHD). The x-axis denotes the Odd Ratio (OR) and the 95% confidence interval. *P-values* were adjusted using the Bonferroni method (0.05 / 45 tests) to assess significant enrichment. Cardiac cell clusters C0, C3, C5, C7 and C14, show significant enrichment when analysing aCHD vs controls. The enrichment observed in clusters C7 and C14 showed a major contribution of sCHD. In comparison, cluster C0 provided the major contribution to nsCHD.



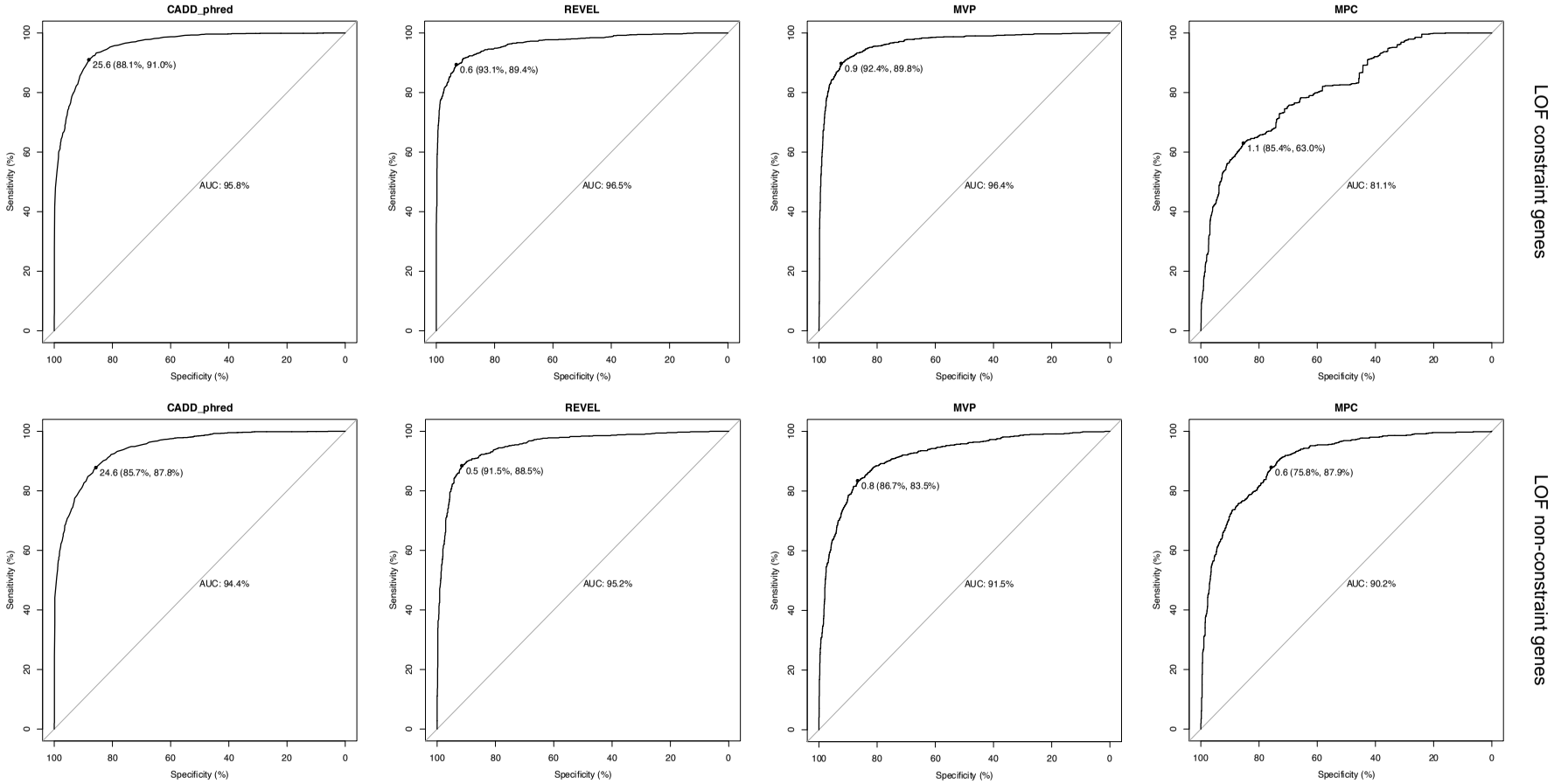
(**Figure 5**. Legend on next page)

Figure 5. Gene Ontology (GO) enrichment analysis of DEGs in cardiac-specific cells with unadjusted $P < 0.01$ from the case-control burden analysis. a) C7: atrial cardiomyocytes cells, b) C0: capillary endothelium, c) C5: smooth muscle cells and d) C10: endothelium and pericytes cells. Only clusters with at least one GO term with *FDR* < 1% are shown. For every GO term, the overlapping DE genes (+) are shown.

Supplemental figures

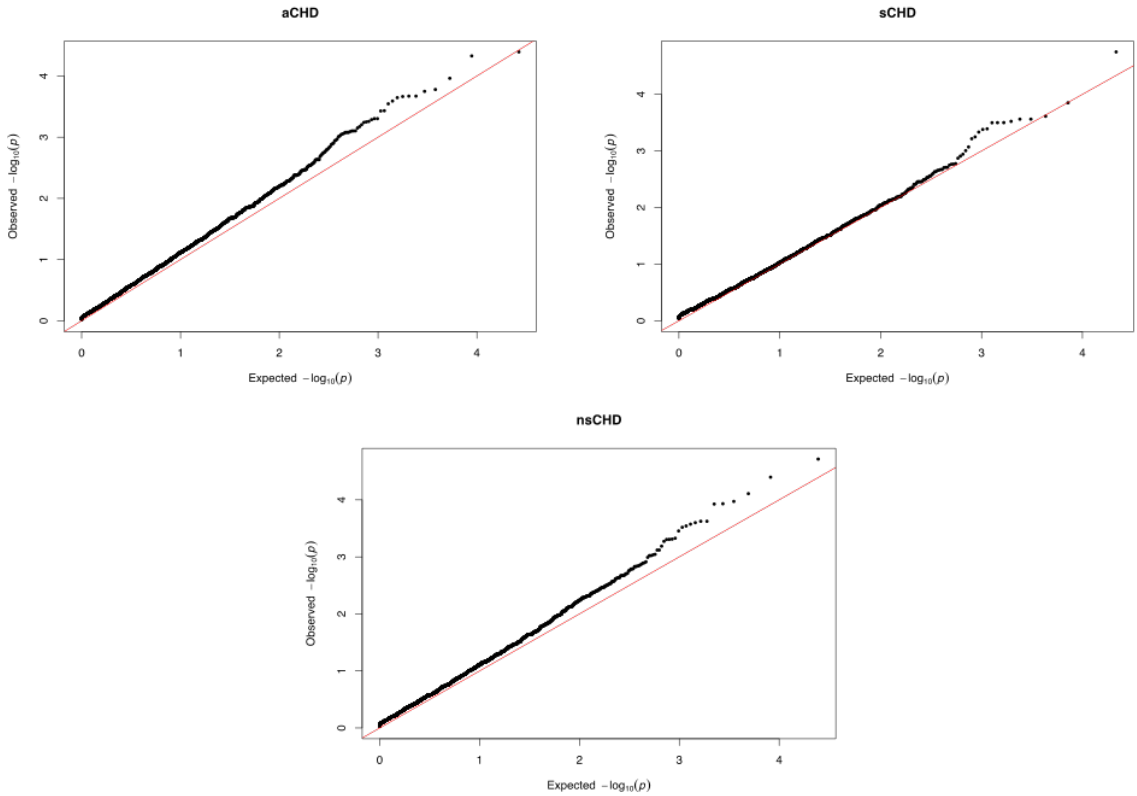


Supplemental Figure 1. Enrichment analysis across the missense constraint gene spectrum. Protein-coding genes were binned based on the MOEUF metric as proposed by gnomAD. Every bin contains ~1,900 genes. Top bins (1, 2) contain the genes with the highest intolerance to missense variation. Enrichment analysis per bin for aCHD are shown. The x-axis indicates the constraint bins; the y-axis shows the Odd Ratios (OR) and the 95% confidence interval.

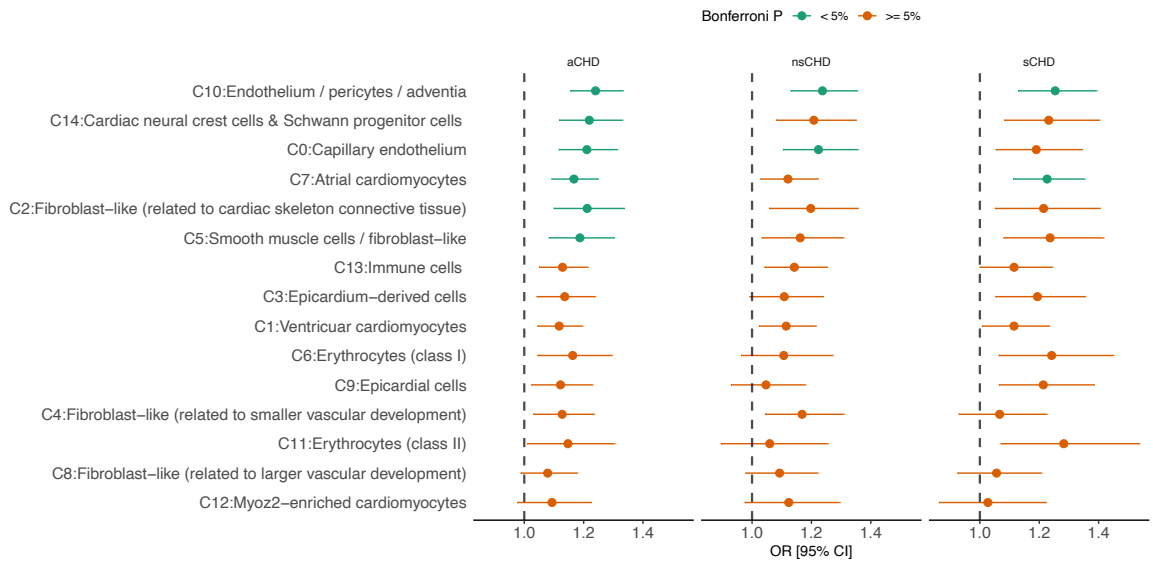


(**Supplemental Figure 2.** Legend on next page)

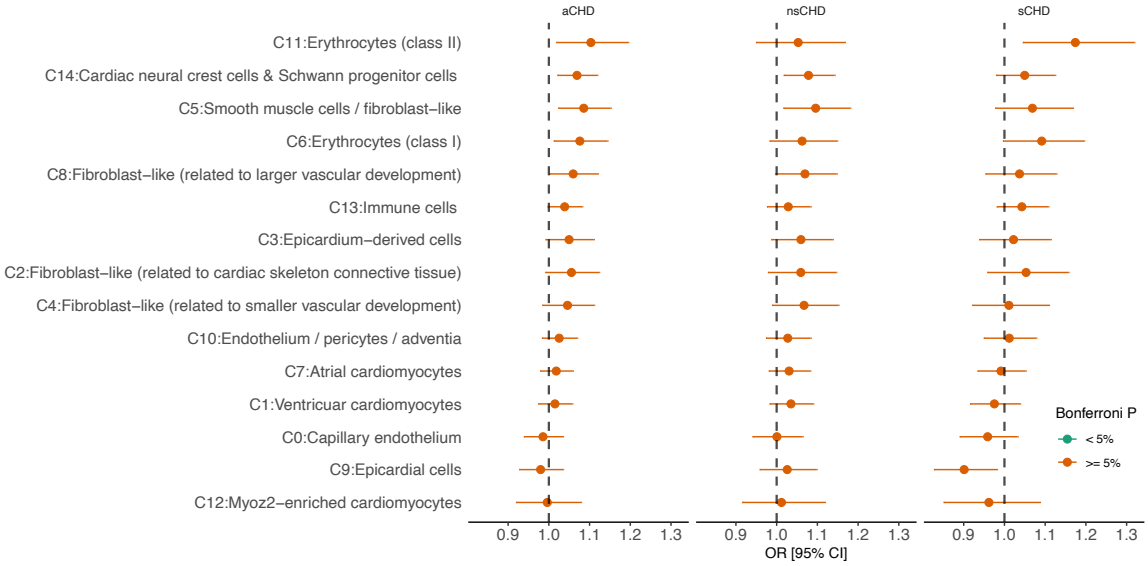
Supplemental Figure 2. ROC analysis of pathogenicity prediction scores. The analysis was performed on a balanced set of benign (true negative) and likely pathogenic (true positive) variants from the ClinVar database within known CHD genes. The top panels show the results for LOF constraint genes (LOUEF < 0.35). The bottom panels show the results for LOF non-constraint genes (LOUEF >= 0.35).



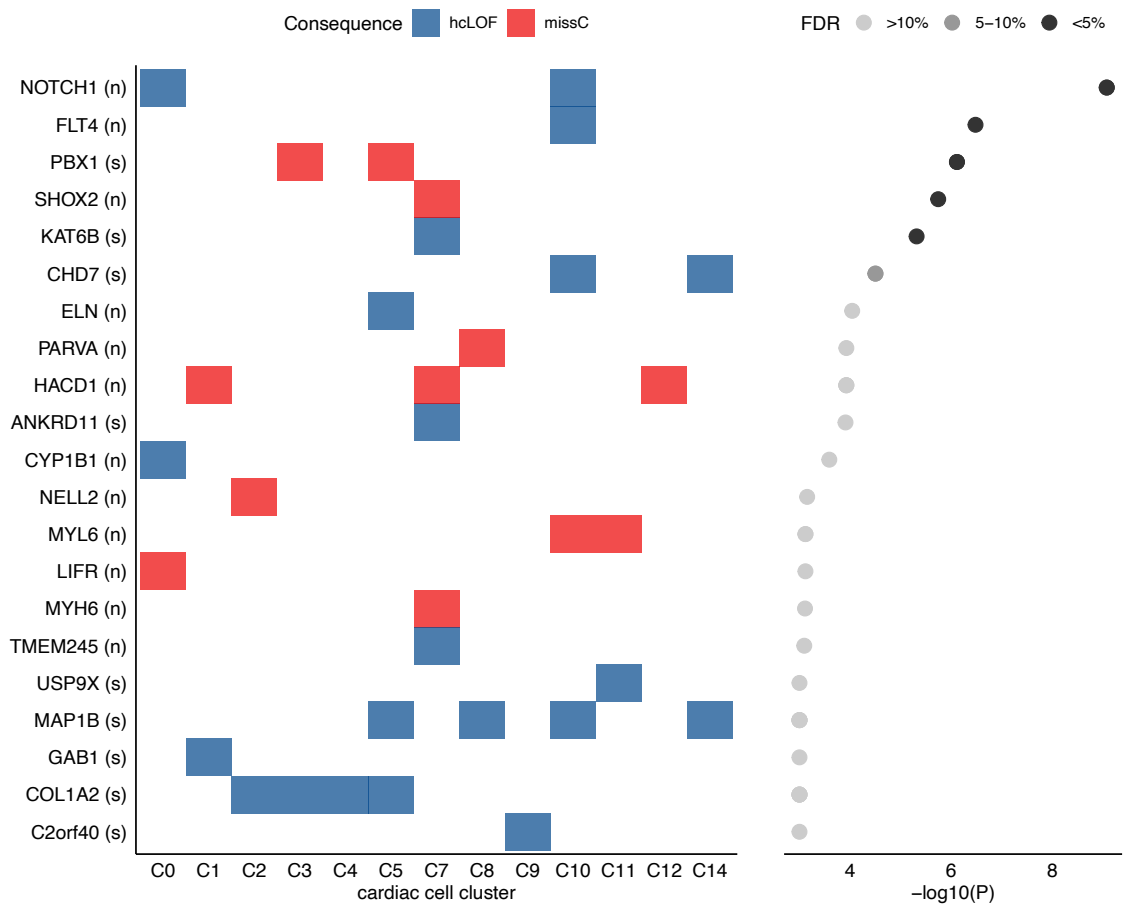
Supplemental Figure 3. Quantile-quantile plots. Expected vs observed p-values for synonymous variants stratified by syndromic status (MAF 0.1%). Q-Q plots for aCHD, sCHD and nsCHD vs controls are shown.



Supplemental Figure 4. Logistic regression-based enrichment analysis of differentially expressed genes (DEGs) in cardiac-specific cell clusters for missense constrained variants (missC). The analysis was stratified by syndromic status (aCHD, sCHD and nsCHD). The x-axis denotes the Odds Ratio (OR) and the 95% confidence interval. *P-values* were adjusted using the Bonferroni method (0.05 / 45 tests) to assess for significant enrichment.



Supplemental Figure 5. Logistic regression-based enrichment analysis of differentially expressed genes (DEGs) in cardiac-specific cell clusters for synonymous variants. The analysis was stratified by syndromic status (aCHD, sCHD and nsCHD). The x-axis denotes the Odds Ratio (OR) and the 95% confidence interval. *P-values* were adjusted using the Bonferroni method (0.05 / 45 tests) to assess for significant enrichment.



Supplemental Figure 6. Top enriched genes (unadjusted *P* < 0.001, case-control Fisher Exact test) found differentially expressed in at least one cardiac-specific cell cluster. The left plot shows the gene/cluster overlap and highlights the variant category with the highest enrichment (blue: hcLOF, red: missC). The x-axis denotes de cardiac clusters; the y-axis indicates the genes and the CHD category analysed (s: sCHD, n: nsCHD). The right plot shows the log-transformed *P* (x-axis) and the *FDR* significant level per gene. Six genes showed *FDR* < 10%: *NOTCH1, FLT4, PBX1, SHOX2, KAT6B* and *CHD7*. C0: Capillary endothelium, C1: Ventricular cardiomyocytes, C2: Fibroblast-like (related to cardiac skeleton connective tissue), C3: Epicardium-derived cells, C4: Fibroblast-like (related to smaller vascular development), C5: Smooth muscle cells, C7: Atrial cardiomyocytes, C8: Fibroblast-like (related to larger vascular development), C9: Epicardial cells, C10: Endothelium/pericytes/adventia, C11: Erythrocytes (class II), C12: Myoz2-enriched cardiomyocytes, C14: Cardiac neural crest cells & Schwann progenitor cells.

Supplemental Data (Sample and Variant QC)

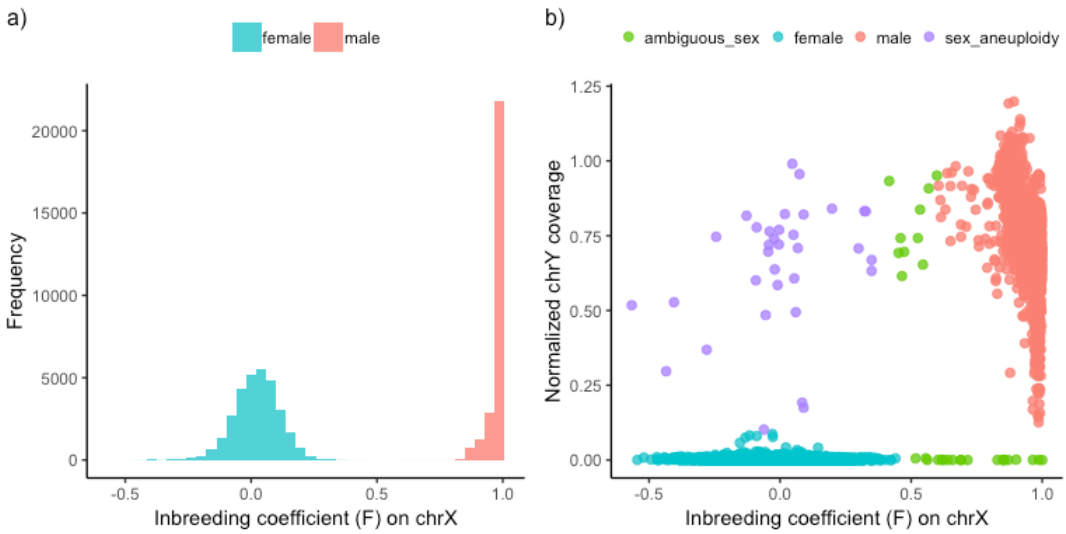


Figure S1. a) Inbreeding coefficient (F-stat) distribution computed over 57,628 samples. b) Inbreeding coefficient (x-axis) vs. normalized chromosome Y coverage (y-axis). Sample chromosomal sex was defined as follow, i) female: F < 0.4 and coverage chrY < 0.1, ii) male: F > 0.6 and coverage chrY > 0.1, iii) aneuploidy: F < 0.4 and coverage chrY >= 0.1, iv) samples failing any of these criteria were flagged as 'ambiguous sex'.

Table S1. The number of affected samples per hard filter.

Hard filters	N. of samples	Percent (%)
Low call rate	9	0.02
Low coverage	1	0.00
Ambiguous sex	30	0.05
Sex aneuploidy	34	0.06
Filters combined	72	0.12

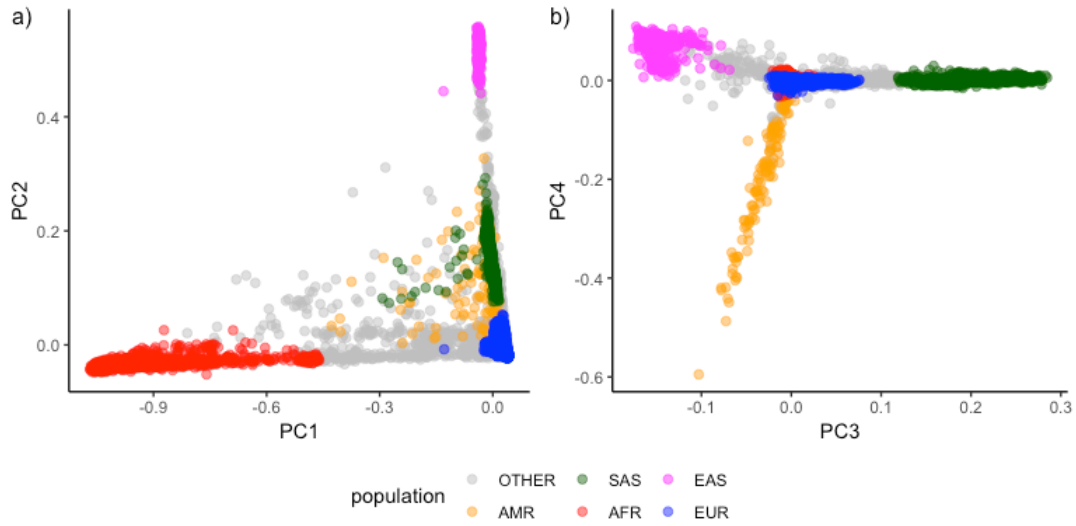


Figure S2. Samples projected onto the top four ancestry principal components (PCs) and their classification into five major ancestral populations. Samples were assigned to SAS, EAS, AMR, AFR or EUR if random forest probability (*p*) > 0.8. Samples failing this threshold were labelled as OTHER (grey). a) PC1 vs PC2 and b) PC3 vs PC4.

Table S2. The number of samples assigned per population. As expected, most samples were assigned to European ancestries (~91%). Approximately 3% of the samples were not assigned to a specific population (labelled as OTHER).

Population	N. of samples	Percent (%)
AFR	1,196	2.07
AMR	111	0.19
EAS	313	0.54
EUR	52,844	91.63
OTHER	1,772	3.07
SAS	1,437	2.49

Table S3. Confusion matrix with assigned population vs reported ethnicity for samples from the UK Biobank (UKBB).

Assigned population	Reported ethnicity	N. of samples per assigned population	N. of samples per reported ethnicity	Percent true classified (%)
AFR	African	319	332	96.08
EUR	British	42450	43184	98.30
EAS	Chinese	169	173	97.69
SAS	Indian	690	708	97.46
EUR	Irish	1495	1498	99.80
SAS	Pakistani	138	138	100.00

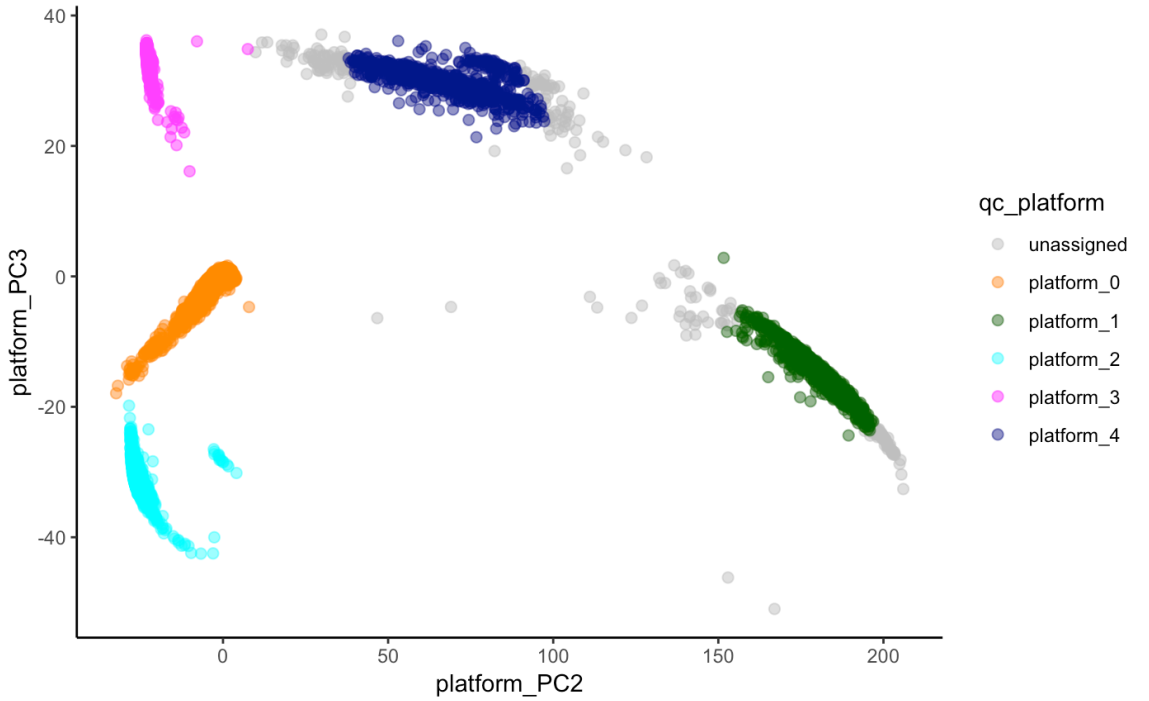


Figure S3. Samples projected onto the platform's principal components (PC) two and three. No generic platform (grey dots) was assigned for less than 0.5% of the samples (n=233). The proposed clustering approach accurately assigned 100% of the samples in the UK Biobank cohort (samples with known capture platform information, orange cluster). The exome capture platform intervals used in the analysis are described in the Methods section.

Table S4. The number of samples detected as outliers by evaluating different sample quality control (QC) metrics. Samples were grouped as per assigned population/platform, and QC metrics were computed per group. Multiple samples (n=104) were detected as outliers by two or more QC metrics.

QC metrics	N. of sample outliers	Percent (%)
Number of SNPs	134	0.23
Number of deletions	85	0.15
Number of insertions	85	0.15
Ratio transmission/transversion	89	0.15
Ratio insertion/deletion	14	0.02
Ratio heterozygous/homozygous	266	0.46

Table S5. Number of remaining samples after each filter stage. *Population filter refers here to samples with assigned European ancestries.

Filter stages	Remaining samples
Unfiltered	57,628
Hard filters	57,560
Hard filters, relatedness	53,862
Hard filters, relatedness, QC outliers	53,507
Hard filters, relatedness, QC outliers, *population	49,308

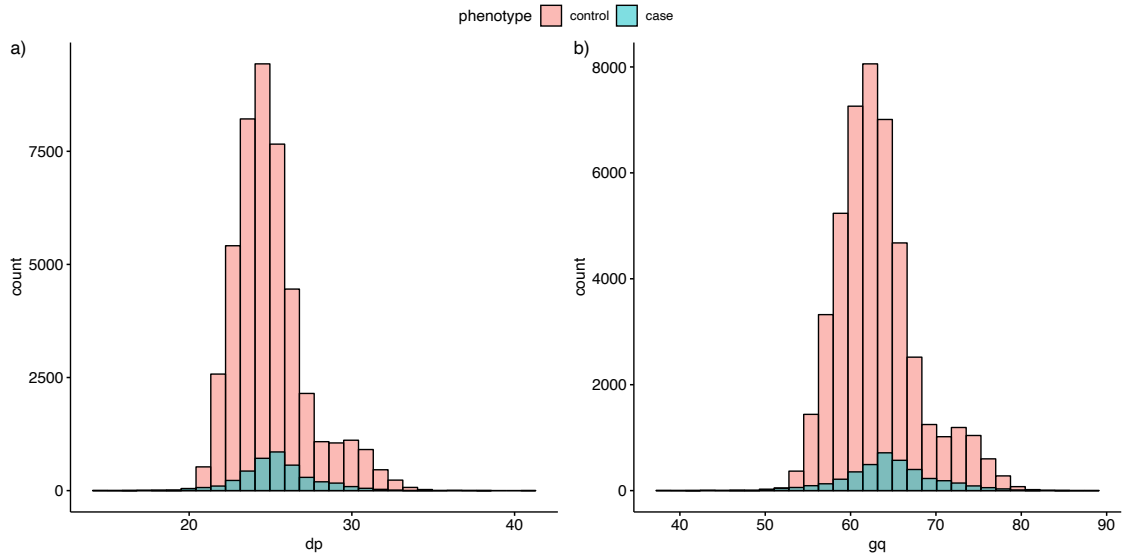


Figure S4. Distribution of per sample averaged QC metrics stratified by phenotype (case/control). a) Mean depth of coverage (DP) and b) Mean genotype quality (GQ). QC metrics were computed per sample across autosomal variants.

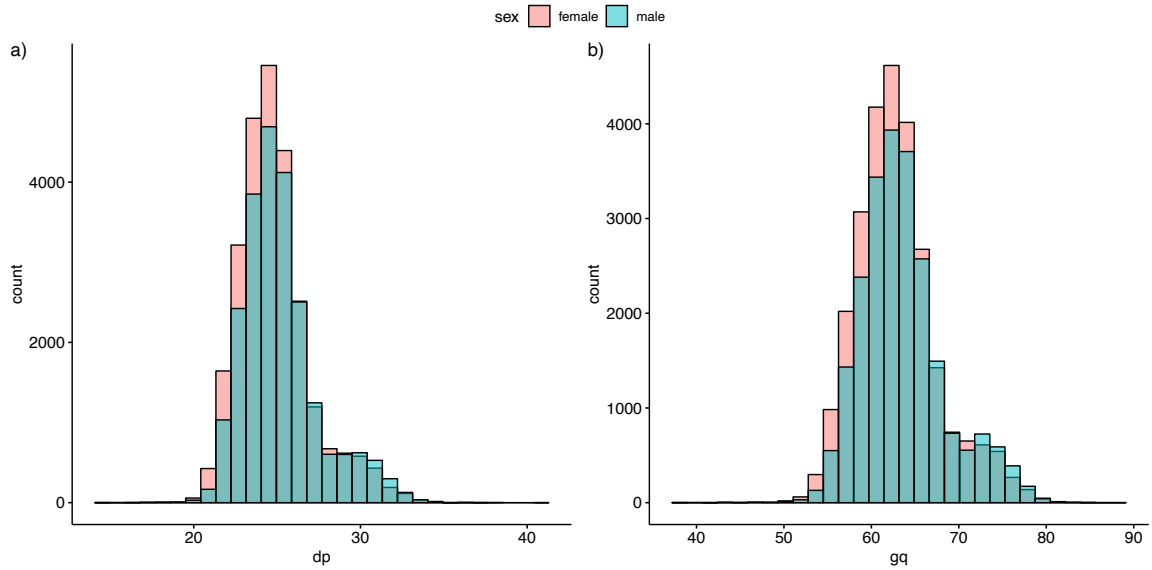


Figure S5. Distribution of per sample averaged QC metrics stratified by sex (female/male). a) Mean depth of coverage (DP) and b) Mean genotype quality (GQ). QC metrics were computed per sample across autosomal variants.

Table S6. Features used in the random forest model to predict the variant probability of being true positive or false positive.

RF features	Description	Importance
variant_type	SNV or indel	0.011
SOR	Symmetric Odds Ratio of 2x2 contingency table to detect strand bias	0.105
ReadPosRankSum	Z-score from Wilcoxon rank-sum test of Alt vs Ref read position bias	0.016
InbreedingCoeff	Inbreeding coefficient as estimated from the genotype likelihoods per-sample when compared against the Hardy-Weinberg expectation	0.103
FS	Phred-scaled p-value using Fisher's exact test to detect strand bias	0.041
DP	Approximate read depth	0.003
QD	Allele-specific Variant Confidence/Quality by Depth	0.704
was_mixed	True if both SNVs and indels are present at the site	0.001
n_alt_alleles	Number of alleles at the site	0.001
MQRankSum	Z-score From Wilcoxon rank-sum test of Alt vs Ref read mapping qualities	0.010

Table S7. The number of remaining variants per filter stage. RF: Random Forest filter, VQSR: Variant Quality Score Recalibration, Coverage: >10X in at least 90% of the samples in gnomAD genome cohort.

Filter stages	Remaining variants
Unfiltered	11,433,645
Hard filters	11,406,658
Hard filters, RF	9,490,151
Hard filters, RF, VQSR	9,191,448
Hard filters, RF, VQSR, Coverage	9,134,464

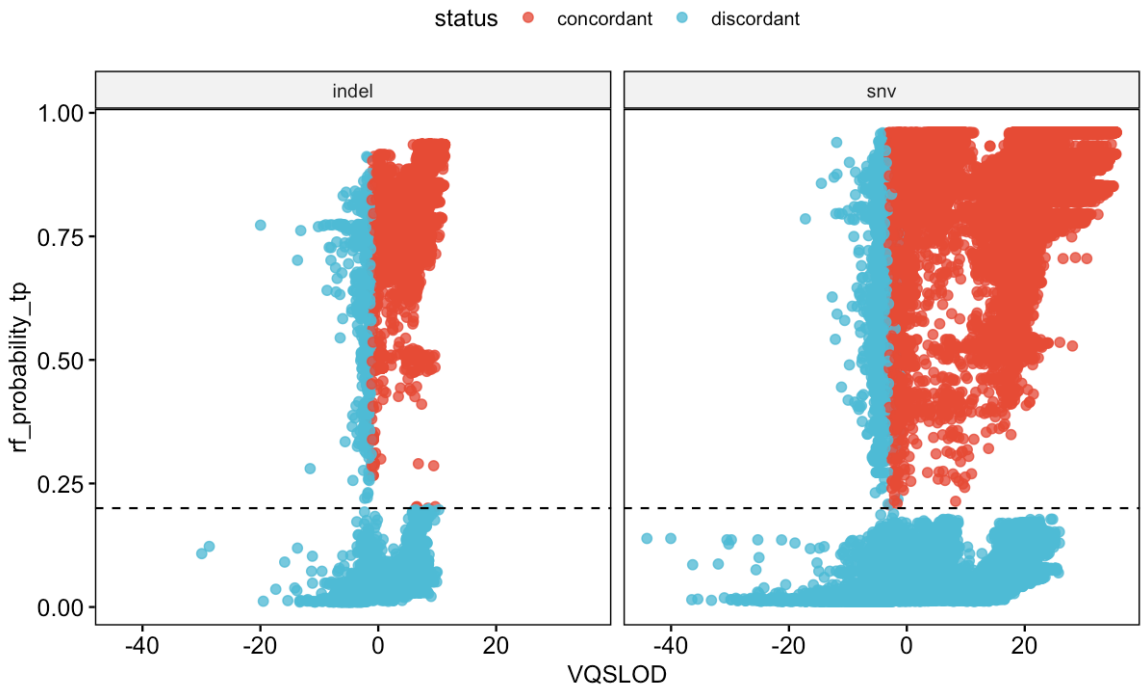


Figure S6. Variant quality score recalibration (x-axis) vs Radom Forest (RF) probability of being true positive (y-axis). Variants (SNVs and indels) are depicted for chromosome 20. The dashed line indicates the cut-off used for the RF probability (=0.2). Concordant (red dots): variants pass both the RF and VQSR filters; discordant (blue dots): variants fail at least one of the RF or VQSR filters.

Chapter VI. General discussion

Cardiovascular diseases (CVDs) remain the principal cause of mortality wordwide[1]. Within the wide spectrum of CVDs, Congenital Heart Disease (CHD) represents a significant burden, with a global incidence of ~7-9 cases per 1000 newborns[2]. CHD is a complex and multifactorial disorder, with genetics as a key component causing the disease[3]. Despite the advances in the last decades regarding our understanding of the causes of CHD, the aetiology of the disease, including its genetic risk factors, remains evasive[3]. Next-generation sequencing technologies (e.g., exome and genome sequencing) have become a crucial diagnostics tool to access genetic causes conferring risk of CHD. Moreover, due to dropping sequencing costs and its accompanying increasing number of patients to be sequenced and analysed independently, these results can be used in the context of large-scale sequencing to elucidate further the genetic causes associated with CHD[4,5].

Still, the interpretation of variants detected in exomes and genomes from CHD patients remains challenging due to the extreme heterogeneity underlying CHD[6]. Previous studies have suggested that likely over 500 CHD-associated genes remain undiscovered[7] and that reduced penetrance decreases the power to identify these genes through the analysis of case-control and parent-offspring cohorts. Identifying these genes, therefore, requires both improved analytical methods and larger sample sizes.

In this work, we have adopted an integrative approach to overcome some of these challenges by assuming that such methodology will increase the power to detect novel variants and genes associated with CHD, compared to previously described approaches[5,8]. Furthermore, we implemented novel statistical frameworks for the

meta-analysis of different genetic variants and developed efficient computational pipelines to process large datasets, usually in the order of hundreds of terabytes.

As shown in Chapter II, we proposed a set of feature selection workflows, including machine learning techniques with application in the context of classification and regression problems. These concepts were then applied to infer individual ancestries in a large exome CHD case-control cohort (~57,000 exomes, Chapter V). Defining matched-ancestry individuals within the cohort being analysed is a crucial step during the quality control process. It decreases the probability of finding spurious associations, which can arise from differences in the individual's ancestries, in case-control burden analysis[9]. The proposed methods can be applied to classification or regression tasks when analysing large multi-dimensional datasets (e.g., analysis of genes, transcripts, and protein expression profiles). Recently, the developed workflows and R-package (https://github.com/enriquea/feseR) have been used for analysing both proteomics[10] and transcriptomics datasets[11].

In Chapter III, we presented a community-driven cooperative effort involving research groups from seven countries (Belgium, Canada, Denmark, Germany, The Netherlands, United States of America, and United Kingdom). We accomplished one of the most extensive meta-analyses in the field of CHD so far, which allowed us to strengthen known CHD associations and discover novel candidate genes for CHD. Besides the gene-centred analysis and based on previous evidence showing that CHD is also influenced by oligo and possibly polygenic factors[12,13], we also investigated biological processes and protein complexes associated with CHD pathogenesis. Alterations in Notch-related pathways were confirmed as an essential mechanism in

the pathophysiology of CHD[14]. Our findings also identified disruption in the calcium signalling pathway as a novel cause conveying a higher risk to develop CHD (Chapter IV). Our work highlighted the advantage of exploiting the available protein-protein interactions (PPI) database for exploring the mutational burden of specific PPI modules (applied in Chapters III and IV). Besides, we demonstrated that such an approach, complemented by functional analysis in animal models and engineered cell lines (e.g., zebrafish, Chapter IV), constitutes a powerful tool to discover and validate new mechanisms underlying CHD. Therefore, functional characterisation of genomic variants identified from exome and genome sequencing data has been recently used to identify new genes associated with specific CHD sub-types (e.g., *KDR* associated with Tetralogy of Fallot)[15,16]. These works (Walree *et al.*[15] and Škorić-Milosavljević *et al.*[16]) are related to but not part of this thesis.

The analysis of a large exome cohort of patients (e.g., tens of thousands of samples) demonstrated being a powerful tool for discovering new genes and variants associated with CHD (see Chapter V). Nevertheless, such analysis adds further challenges regarding computation resources and tools to manage and analyse big data efficiently. Ultimately, the development of reliable distributed systems has allowed scaling such analysis within a reasonable amount of computational time and resources needed. Specifically, Spark and Hail (https://hail.is) are becoming a standard computational framework to analyse large-scale exome and genome cohorts[17]. Consequently, we took advantage of the potentialities of these frameworks to develop efficient bioinformatics pipelines, as shown in Chapter V.

When investigating the causality of variants and genes in CHD, an important aspect is considering the expression level within the organ of interest. The integration of bulk-RNA data[18], as well as transcript-aware expression at a single-cell resolution[19], with genetics data, have demonstrated to be a valuable source of complementary information in the interpretation of candidate genes (Chapter III and V). We showed how such combined analysis could add supporting evidence to strengthen the association of the identified candidate genes. Therefore, the interrogation of the existing heart transcriptomics atlas proves a valuable tool to further prioritise novel CHD causes. We anticipate that subsequence improvements of single-cell technologies will provide us with a complete expression map of the heart. Such maps would help to improve the accuracy of available tools for variants and genes prioritisation by integrating transcript-aware expression as a relevant feature.

Moving forward, we studied the genetic differences between non-syndromic CHD and syndromic forms of CHD. We observed a higher contribution of loss-of-function variants than constraint missense in non-syndromic and syndromic cases (Chapter V). Overall, the contribution of these two functional categories was observed higher in syndromic compared to non-syndromic CHD. Syndromic forms of CHD show a phenotype, which is not limited to the cardiovascular system but involves other non-cardiac organ systems. Therefore, the higher contribution of loss-of-function and constraint missense variants compared to non-syndromic forms of the disease is evident in our analysis and corroborates previous findings[8,20].

Nevertheless, our analysis has also shown the need to develop new tools to increase the power to detect variants and genes associated with non-syndromic forms of CHD.

Thus, future work will focus on improving the prioritisation of variants and genes associated with non-syndromic forms of CHD. Furthermore, as exposed during this research (Chapters III and V), the availability of detailed phenotypic descriptions using a standardised system will facilitate improved genotype-phenotype correlation in CHD subtypes.

The field of CHD continues to evolve, with still numerous challenges and research upfront to be accomplished. As well-reviewed earlier[3,21], CHD is influenced by both environmental and genetic factors. In this thesis, we focused mainly on the study of the genetics aspect of the disease. Nevertheless, such a multifactorial yet complex disease requires different angles to elucidate its causes. Multiple efforts have been initiated to improve our understanding of the disease. For instance, Hoff *et al.* studied epigenetics such as methylation patterns of different types of CHD[22]. Also, proteomics applied to CHD is becoming an active area of research that has helped reveal new aspects[23,24]. More recently, transcriptomics studies to explore gene expression patterns of cardiac-specific cells have been used to improve the interpretation of genes and variants associated with CHD[25,26]. We anticipate that future efforts will converge toward multi-omics approaches to enhance our capabilities to discover novel factors associated with CHD. Notably, such an integrative approach will require the development of novel statistical and computational methods.

References

1. GA, R. *et al.* Global, Regional, and National Burden of Cardiovascular Diseases for 10 Causes, 1990 to 2015. *J. Am. Coll. Cardiol.* **70**, 1–25 (2017).
2. van der Linde, D. *et al.* Birth Prevalence of Congenital Heart Disease Worldwide. *J. Am. Coll. Cardiol.* **58**, 2241–2247 (2011).
3. Zaidi, S. & Brueckner, M. Genetics and Genomics of Congenital Heart Disease. *Circ. Res.* **120**, 923–940 (2017).
4. Flannick, J. *et al.* Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. *Nature* **570**, 71–76 (2019).
5. Kaplanis, J. *et al.* Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature* **586**, 757–762 (2020).
6. Bean, L. J. H. & Hegde, M. R. Clinical implications and considerations for evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines. *Genome Med.* **9**, 111 (2017).
7. Andersen, T. A., Troelsen, K. D. L. L. & Larsen, L. A. Of mice and men: Molecular genetics of congenital heart disease. *Cell. Mol. Life Sci.* **71**, 1327–1352 (2014).
8. Jin, S. C. *et al.* Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. *Nat. Genet.* (2017). doi:10.1038/ng.3970
9. Singh, T., Neale, B. M. & Daly, M. J. Exome sequencing identifies rare coding variants in 10 genes which confer substantial risk for schizophrenia on behalf of the Schizophrenia Exome Meta-Analysis (SCHEMA) Consortium*. *medRxiv* 2020.09.18.20192815 (2020). doi:10.1101/2020.09.18.20192815
10. McGurk, K. A. *et al.* The use of missing values in proteomic data-independent acquisition mass spectrometry to enable disease activity discrimination. *Bioinformatics* **36**, 2217–2223 (2020).
11. Rychkov, D., Sur, S., Sirota, M. & Sarwal, M. M. Molecular Diversity of Clinically Stable Human Kidney Allografts. *JAMA Netw. open* **4**, e2035048 (2021).
12. Sevim Bayrak, C., Zhang, P., Tristani-Firouzi, M., Gelb, B. D. & Itan, Y. De novo variants in exomes of congenital heart disease patients identify risk genes and pathways. *Genome Med.* **12**, 9 (2020).
13. Chapman, G. *et al.* Functional genomics and gene-environment interaction highlight the complexity of congenital heart disease caused by Notch pathway variants. *Hum. Mol. Genet.* **29**, 566–579 (2020).
14. Luxán, G., D'Amato, G., MacGrogan, D. & de la Pompa, J. L. Endocardial Notch Signaling in Cardiac Development and Disease. *Circ. Res.* **118**, e1–e18 (2016).
15. van Walree, E. S. *et al.* Germline variants in HEY2 functional domains lead to congenital heart defects and thoracic aortic aneurysms. *Genet. Med.* **23**, 103–110 (2021).
16. Škorić-Milosavljević, D. *et al.* Rare variants in KDR, encoding VEGF Receptor 2, are associated with tetralogy of Fallot. *Genet. Med.* 1–9 (2021). doi:10.1038/s41436-021-01212-y
17. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
18. Cardoso-Moreira, M. *et al.* Gene expression across mammalian organ development. *Nature* **571**, 505–509 (2019).

19. Asp, M. *et al.* A Spatiotemporal Organ-Wide Gene Expression and Cell Atlas of the Developing Human Heart. *Cell* **179**, 1647-1660.e19 (2019).
20. Sifrim, A. *et al.* Distinct genetic architectures for syndromic and non-syndromic congenital heart defects identified by exome sequencing. *Nat. Genet.* **48**, 1060–5 (2016).
21. Edwards, J. J. & Gelb, B. D. Genetics of congenital heart disease. *Current Opinion in Cardiology* **31**, 235–241 (2016).
22. Hoff, K. *et al.* DNA methylation profiling allows for characterisation of atrial and ventricular cardiac tissues and hiPSC-CMs. *Clin. Epigenetics* **11**, 89 (2019).
23. AR, B. *et al.* The cardiac proteome in patients with congenital ventricular septal defect: A comparative study between right atria and right ventricles. *J. Proteomics* **191**, 107–113 (2018).
24. T, W. *et al.* Mechanisms of Congenital Heart Disease Caused by NAA15 Haploinsufficiency. *Circ. Res.* **128**, 1156–1169 (2021).
25. Kathiriya, I. S. *et al.* Modeling Human TBX5 Haploinsufficiency Predicts Regulatory Networks for Congenital Heart Disease. *Dev. Cell* **56**, 292-309.e9 (2021).
26. Liu, X. *et al.* Single-Cell RNA-Seq of the Developing Cardiac Outflow Tract Reveals Convergent Development of the Vascular Smooth Muscle Cells. *Cell Rep.* **28**, 1346-1361.e4 (2019).

List of abbreviations

Term	Definition
API	Application Programming Interface
ASD	Atrial septal defect
B-H	Benjamini-Hochberg (BH) false discovery rate (FDR) procedure
BAM	Binary Alignment Map
BAV	Bicuspid aortic valve
BMI	Body mass index
BP	Bipolar disorder
BWA	Burrows-Wheeler Aligner
CADD	Combined Annotation Dependent Depletion
CCR	Coding constraint region
CDG	Candidate disease gene
CHD	Congenital Heart Disease
CM	Correlation Matrix
cNCCs	Cardiac neural crest cells
CNV	Copy number variant
CPM	Count per million matrix
CRAM	Compressed sequence alignment map
CVD	Cardiovascular disease
DD	Developmental delay
DDD	Deciphering Developmental Disorder
DEG	Differentially expressed gene
DNA	Deoxyribonucleic Acid
DNV	De novo variant
DP	Depth of coverage
ENA	European Nucleotide Archive
ES	Exome sequencing
EVA	European Variation Archive
FASD	Foetal alcohol spectrum disorder
FDR	False discovery rate
FHF	First heart field
FS	Feature selection
FWER	Family-wise error rate
GATK	Genome Analysis Toolkit

GEO	Gene Expression Omnibus
GI	Gain Information
GO	Gene Ontology
GQ	Genotype quality
GS	Genome sequencing
GTPTS	Genitopatellar syndrome
GVCF	Genomic Variant Call Format
GWAS	Genome-wide association study
hcLOF	High-confidence loss-of-function variant
HLHS	Hypoplastic left heart syndrome
IHW	Independent hypothesis weighting
LD	Linkage disequilibrium
LOEUF	Loss of function observed/expected ratio upper fraction
LOF	Loss of function
LOFTEE	Loss-Of-Function Transcript Effect Estimator
LVOTO	Left ventricular outflow tract obstruction
MAC	Minor allele count
MAD	Median absolute deviation
MAF	Minor allele frequency
missC	Missense constraint variant
ML	Machine learning
MOEUF	Missense observed/expected ratio upper fraction
MPC	Missense badness, PolyPhen-2, and Constraint
MVP	Missense variant pathogenicity predictor
NCBI	National Center for Biotechnology Information
NGS	Next generation sequencing
NIPT	Non-invasive prenatal testing
NMD	Nonsense-mediated decay
OMIM	Online Mendelian Inheritance in Man
OR	Odds ratio
PAV	Protein altering variant
PCA	Principal component analysis
PCR	Polymerase chain reaction
PDA	Patent ductus arteriosus
PPI	Protein-protein interaction
PTV	Protein truncating variant

QC	Quality control
RefSeq	Reference Sequence
REVEL	Rare exome variant ensemble learner
RF	Random Forest
RFE	Recursive feature elimination
RMSE	Root-mean-square-error
RNA	Ribonucleic Acid
ROC	Receiver operating characteristic
RPKM	Reads per kilobase million
SAM	Sequence Alignment Map
SBBYSS	Say–Barber–Biesecker–Young–Simpson syndrome
SHF	Second heart field
sMVP	Severe mitral valve prolapse
SNV	Single nucleotide variant
SRA	Sequence Read Archive
SV	Structural variation
SVM	Support Vector Machine
TAA	Thoracic aortic aneurysm
TGA	Transposition of the great arteries
TNBC	Triple-negative breast cancer
TOF	Tetralogy of Fallot
UKBB	United Kingdom Biobank
VCF	Variant Call Format
VEP	Variant Effect Predictor
VQSR	Variant Quality Score Recalibration
VSD	Ventricular septal defect
WES	Whole exome sequencing

Curriculum Vitae

University Hospital of Schleswig- Holstein (UKSH)
Quincke-Forschungszentrum (QFZ)
Kardiovaskuläre Genetik AG Hitz
Forschungsbau 1, Haus U18
Rosalind-Franklin-Straße 9
24105 Kiel

enrique.audain@gmail.com
enrique.audain@uksh.de

ORCID: 0000-0002-9201-7840
GitHub: https://github.com/enriquea

Enrique Audain Martinez

Personal details

Last name: Audain Martinez

First name: Enrique

Nationality: Cuban

Date of birth: 07.08.1987

Gender: Male

Education

May 2017 – present **Kiel University**
PhD Student
Kiel, Germany

Jan 2012 – Feb 2014 **Center for Genetic Engineering and Biotechnology**
MSc, Biotechnology
Havana, Cuba

Sep 2007 – Jun 2011 **Technological University of Havana, CUJAE**
Engr, Biomedical Engineer
Havana, Cuba

Research Experience

May 2016 – present **Bioinformatician**
Universitätsklinikum Schleswig - Holstein, Kiel, Germany

Jun 2014 – Sep 2014 **Research Visitor**
University of Tübingen, Tübingen, Germany

Sep 2011 – Apr 2016 **Junior researcher**
Center of Molecular Immunology, Department of Proteomics
Havana, Cuba

Awards & Grants

Jun 2014 Grant: Boehringer Ingelheim Fonds (BIF)

Skills

Skills Proteomics/Genomics data analysis, R/Python programming (advanced), Java Programming (intermediate), Data Mining, Machine Learning, Apache Spark.

Languages Spanish, English, German

Meetings & Trainings

November 2021 CSHL Genome Informatics meeting (virtual), NY, USA

May 2020 Introductory course to HCP cluster (virtual), Sanger Institute, UK

March 2019 PROCEED Annual meeting, UMC Amsterdam, The Netherlands

April 2018 Hail 0.2 hands-on workshop, CCG, Cologne, Germany

November 2017 Variant and gene prioritization, KU Leuven, Belgium

Publications (Highlights)

Audain E, Wilsdon A, Breckpot J, […], Thienpont B, Larsen LA, Hitz MP. (2021) *Integrative analysis of genomic variants reveals new associations of candidate haploinsufficient genes with congenital heart disease*. PLoS Genet 17(7):e1009679. https://doi.org/10.1371/journal.pgen.1009679

Dai C, Füllgrabe A, Pfeuffer J, […], **Audain E**, […], Perez-Riverol Y. *A proteomics sample metadata representation for multiomics integration and big data analysis.* Nat Commun 12, 5854 (2021). https://doi.org/10.1038/s41467-021-26111-3

Izarzugaza JMG, Ellesøe SG, Doganli C, […], **Audain E**, […], Hitz MP, Larsen LA, Brunak S. *Systems genetics analysis identifies calcium-signaling defects as novel cause of congenital heart disease*. Genome Med 12, 76 (2020). https://doi.org/10.1186/s13073-020-00772-z

Perez-Riverol Y, Csordas A, Bai J, […], **Audain E**, […], Ternent T, Brazma T, Vizcaíno JA. *The PRIDE database and related tools and resources in 2019: improving support for quantification data*. Nucleic Acids Research, Volume 47, Issue D1, 08 January 2019, Pages D442–D450, https://doi.org/10.1093/nar/gky1106

Perez-Riverol Y, Kuhn M, Vizcaino JA, Hitz MP, **Audain E**. *Accurate and fast feature selection workflow for high-dimensional omics data*. PLoS ONE 12/2017; 12(12). https://doi.org/10.1371/journal.pone.0189875

Audain E, Uszkoreit J, Sachsenberg T, […], Tabb DL, Kohlbacher O, Perez-Riverol Y: *In-depth analysis of protein inference algorithms using multiple search engines and well-defined metrics*. Journal of Proteomics 08/2016; 150. https://doi.org/10.1016/j.jprot.2016.08.002

Audain E, Ramos Y, Hermjakob H, Flower DR, Perez-Riverol Y. *Accurate estimation of Isoelectric Point of Protein and Peptide based on Amino Acid Sequences*. Bioinformatics 11/2015; 32(6). https://doi.org/10.1093/bioinformatics/btv674

Audain E, Sanchez A, Vizcaíno JA, Perez-Riverol Y. *A Survey of Molecular Descriptors Used in Mass Spectrometry Based Proteomics*. Current topics in medicinal chemistry 12/2013; 14(3). https://doi.org/10.2174/15680266136661312041135 37

Publications (Others)

Spielmann N, Miller G, Oprea T, […], **Audain E**, […], Gailus-Durner V, Hrabe de Angelis M. *Identification of novel genes involved in congenital heart rhythm disorders, cardiomyopathy, and structural cardiac abnormalities by in-vivo and ex-situ screening of 3,894 knockout mouse genes.* (In revision, Nature Cardiovascular Research).

Umer H, **Audain E**, Zhu Y, […], Lehtiö J, Branca R, Perez-Riverol Y. *Generation of ENSEMBL-based proteogenomics databases boosts the identification of non-canonical peptides.* (In revision, Oxford Bioinformatics).

Škorić-Milosavljević D, Lahrouchi N, Bosada FM, […], **Audain E**, […], Christoffels VM, Postma AV, Bezzina CR. *Rare variants in KDR, encoding VEGF Receptor 2, are associated with tetralogy of Fallot*. Genet Med (2021). https://doi.org/10.1038/s41436-021-01212-y

van Walree ES, Dombrowsky G, Jansen IE, […], **Audain E**, […], Postma AV, Mathijssen IB. *Germline variants in HEY2 functional domains lead to congenital heart defects and thoracic aortic aneurysms*. Genet Med 23, 103–110 (2021). https://doi.org/10.1038/s41436-020-00939-4

Dass G, Vu M, Xu P, **Audain E**, Hitz MP, Grüning BA, Hermjakob H, Perez-Riverol Y. *The omics discovery REST interface*. Nucleic Acids Research, Volume 48, Issue W1, 02 July 2020, Pages W380–W384, https://doi.org/10.1093/nar/gkaa326

Wünnemann F, Ta-Shma A, Preuss C, […], **Audain E**, […], Hitz MP, Andelfinger G. *Loss of ADAMTS19 causes progressive non-syndromic heart valve disease.* Nat Genet 52, 40–47 (2020). https://doi.org/10.1038/s41588-019-0536-2

Hoff K, Lemme M, Kahlert AK, Runde K, **Audain E**, […], Hansen A, Ammerpohl O, Hitz MP. *DNA methylation profiling allows for characterization of atrial and ventricular cardiac tissues and hiPSC-CMs.* Clin Epigenetics. 2019 Jun 11;11(1):89. https://doi.org/10.1186/s13148-019-0679-0

Cabrales-Rico A, de la Torre BG, Garay HE, […], **Audain E**, […], Perea SE, Reyes O, González LJ. *Bio-analytical method based on MALDI-MS analysis for the quantification of CIGB-300 anti-tumor peptide in human plasma*. Journal of Pharmaceutical and Biomedical Analysis, Volume 105, 2015, Pages 107-114, ISSN 0731-7085, https://doi.org/10.1016/j.jpba.2014.11.043

Perez-Riverol Y, **Audain E**, Millan A, […], González LJ, Padrón G, Besada V. *Isoelectric point optimization using peptide descriptors and support vector machines*. Journal of Proteomics 02/2012; 75(7):2269-74. https://doi.org/10.1016/j.jprot.2012.01.029

Poster & Conference Proceedings

Hofmann P, **Audain E**, […], Hitz MP. *Non-coding Copy Number Variation in Congenital Heart Disease*. ESC Cardiovascular Development Meeting 2019, Malaga, Spain (Poster)

Audain E, […], Larsen LA, Hitz MP. *Integrative analysis of genomic variants reveals new protein-coding candidate genes associated with congenital heart diseases.* ESC Cardiovascular Development Meeting 2018, Marseille, France (Poster)

Perez-Riverol Y, Bernal-Linares M, **Audain E**, […], Vizcaino JA. *Systematic Integration of Millions of Peptideforms Evidence into ENSEMBL Genome Browser.* 66th Conference on Mass Spectrometry and Allied Topics 2018. (Poster)

Audain E, Carmenate T, de la Luz KR. *IL-2 mutein: Challenges in its characterization*. Biomanufacturing Challenges of Immunotherapy 2015, Havana, Cuba