Mosaicism and the genetic architecture of congenital heart disease

Alexander Hsieh

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2020

Abstract

Mosaicism and the genetic architecture of congenital heart disease

Alexander Hsieh

Congenital heart disease (CHD) is characterized by structural defects of the heart and great vessels. It is the most common birth defect, affecting an estimated 1% of live births, and is the leading cause of mortality among birth defects. Despite recent progress in genetic research, more than 50% of CHD cases remain unexplained. An estimated 23% are due to aneuploidies and copy number variants and up to 30% has been attributed to *de novo* variation, though that number ranges between 3-30% depending on CHD complexity.

The contribution of somatic mosaicism, or *de novo* genetic mutations arising after oocyte fertilization, to congenital heart disease (CHD) is not well understood due to limitations in sample size, detection method, and validation rate. Further, the relationship between mosaicism in blood and cardiovascular tissue has not been determined. We developed a computational method, Expectation-Maximization-based detection of Mosaicism (EM-mosaic), to analyze mosaicism in exome sequences of 2530 CHD proband-parent trios. EM-mosaic accurately detected 309 mosaic mutations in blood, with 85 of 94 (90%) candidates tested independently confirmed. We found twenty-five likely damaging mosaics in plausible CHD-risk genes, affecting 1% of our cohort. Variants in these genes predicted as damaging had higher variant

allele fraction than benign variants, suggesting a role in CHD. The frequency of protein-coding mosaic variants detectable in blood was 0.122 or roughly 1 in 8 individuals. Analysis of 66 individuals with matched cardiac tissue available revealed both tissue-specific and shared mosaicism, with shared mosaics generally having higher allele fraction.

CHD patients often present with comorbid cardiac and extracardiac anomalies that further their impact quality of life. Neurodevelopmental disorders (NDDs) are especially prevalent in CHD cases compared to the general population, yet the underlying genetic causes remain poorly explained. Further, patients with single ventricle defects undergoing surgery often later develop arrhythmias and experience worsening ventricular function. We used a statistical approach to dissect the association between *de novo* variation and these clinical outcomes and found that pleiotropic mutations contribute a large fraction of the risk of acquiring NDD and abnormal ventricular function phenotypes in CHD patients. We developed a proof-of-concept rare variant risk score that combines information from *de novo,* rare transmitted, and copy-number variants and show that prediction of outcomes such as NDD can be improved, especially in complex CHD cases.

Table of Contents

List of Figures iv

List of Tables vi

Acknowledgments vii

Dedication viii

Introduction 1

0.1 Introduction to congenital heart disease 1

0.2 Introduction to genetic mosaicism 3

0.3 Organization of this dissertation 8

Chapter 1: EM-mosaic: Expectation-Maximization-based detection of mosaicism 10

1.1 Introduction 11

1.2 Results 13

1.2.1 High-accuracy detection of mosaic mutations in WES data using EM-mosaic 13

1.2.2 Method evaluation via simulation experiment 19

1.2.3 Mosaic mutations found in blood derived DNA with MosaicHunter 21

1.2.4 Sequence confirmation of candidate mosaic variants and estimation of mosaicism in CHD 21

1.2.5 Mosaic variants occurred most frequently at CpG sequences. 26

1.2.6 Detection of mosaic mutations in CHD tissues 27

1.2.7 Blood and cardiac tissue mosaics likely to contribute to CHD 29

1.3 Discussion 35

1.4 Materials and Methods 40

1.4.1 Samples and sequencing data 40
1.4.2 *De novo* variant calling and annotation 41
1.4.3 Pre-processing and quality control 42
1.4.4 IGV visualization of low allele fraction *de novo* SNVs 44
1.4.5 Expectation-Maximization to estimate prior mosaic fraction and control FDR 44
1.4.6 Mosaic mutation detection model 45
1.4.7 Simulation experiment 46
1.4.8 Mutation confirmation by MiSeq amplicon sequencing 48
1.4.9 Investigating the relationship between VAF and pathogenicity 52
1.4.10 Estimated contribution of mosaicism to CHD 53
1.4.11 Union with validated *de novo* SNVs from Jin et al. *Nature Genetics* 2017 54
1.4.12 Mutation spectrum analysis 54
1.4.13 Mosaic detection power given sample average coverage 55
1.4.14 Filtering of MosaicHunter candidate variants 56
1.4.15 Filtering of MosaicHunter-detected cardiovascular tissue candidate variants 57
1.4.16 Clinical interpretation of mosaic variants – limitations 57
Chapter 2: Genetic factors associated with clinical outcomes in CHD 58
2.1 Introduction 59
2.2 Results 62
2.2.1 Complex CHD cases are more likely to acquire NDD than Isolated CHD cases 62
2.2.2 Damaging *de novo* variants are associated with NDD 63
2.2.3 Mutations with pleiotropic effects drive the acquisition of NDD in CHD 67

2.2.4 Damaging *de novo* variants are associated with abnormal ventricular function in patients with single ventricle defects 68

2.2.5 Rare variant risk score predicts NDD in CHD patients 72

2.3 Discussion 75

2.4 Materials and Methods 77

2.4.1 Sequencing data, variant calling, and quality control 77

2.4.2 Depth of coverage and D15 for NDD and non-NDD samples across batches 79

2.4.3 Annotations and gene sets 79

2.4.4 Association analysis 80

2.4.5 Prevalence analysis 82

2.4.6 NDD rare variant risk score 82

Conclusion 89

References 92

Appendix 108

List of Figures

Figure 1.1 EM-mosaic flowchart 15

Figure 1.2 Mosaic detection by Expectation-Maximization. 16

Figure 1.3 Blood variants with posterior odds between 1 and 10. 17

Figure 1.4 MosaicHunter workflow. 18

Figure 1.5 Simulation experiment results (60x dataset). 20

Figure 1.6 Targeted sequencing to validate candidate blood mosaic variants. 24

Figure 1.7 Estimated mosaic detection power using less stringent mosaic definitions. 25

Figure 1.8 Mutation spectrum of detected germline and mosaic variants. 26

Figure 1.9 Validated mosaics detected in probands with matched blood and cardiovascular tissue samples available. 28

Figure 1.10 Mosaic variants shared in blood and cardiovascular tissues have higher variant allele fraction. 28

Figure 1.11 Damaging mosaics in CHD-related genes have higher variant allele fraction than likely-benign mosaics. 31

Figure 1.12 Damaging CHD-related mosaics have higher VAF under less stringent definitions of mosaicism. 32

Figure 1.13 Mosaic rate by proband age. 34

Figure 1.14 Mosaic rate by parental age at birth. 34

Figure 1.15 FDR-based minimum N_{alt} threshold. 43

Figure 1.16 Overdispersion. 46

Figure 1.17 Simulation experiment results for 40x, 60x, 100x. 48

Figure 1.18 Confirmation rate across VAF bins. 51

Figure 1.19 Posterior odds comparison for tested vs. untested mosaics. 52

Figure 2.1 Damaging (LGD + Dmis) DNVs are enriched across gene sets. 65

Figure 2.2 Higher prevalence of NDD among cases carrying likely pathogenic DNVs. 66

Figure 2.3 Genes annotated as both HHE & NDD-risk contribute substantial PAR and suggest pleiotropic activity. 68

Figure 2.4 Higher abnormal phenotype prevalence among cases carrying likely pathogenic DNVs. 71

Figure 2.5 Risk score distribution in NDD vs. non-NDD cases, across CHD subtype groups. ... 73

Figure 2.6 Risk score performance, 10-fold cross validated Precision-Recall curves..... 74

Figure 2.7 Prevalence of NDD as a function of risk score percentile..... 74

Figure 2.8 Enrichment of NDD by risk score quartile, compared to bottom quartile 75

Figure 2.9 Comparison of DP and D15 in NDD and non-NDD samples. 79

Figure 2.10 Matthews Correlation Coefficient (MCC) for composite score thresholds, by percentile. 84

List of Tables

Table 1.1 Mosaic detection by EM-mosaic, MosaicHunter; validated by PCR product sequencing 17

Table 1.2 Mosaics detected in individuals with matched cardiovascular tissue and blood 29

Table 1.3 Damaging Mosaics in CHD-relevant genes..... 33

Table 2.1 Complete CHD cohort. 63

Table 2.2 Rates of LGD, Dmis DNVs across different gene groups. 64

Table 2.3 Prevalence of NDD in cases carrying LP DNVs vs. cases that do not carry LP DNVs. 66

Table 2.4 Rates of damaging DNVs among 114 patients with single ventricle defects with abnormal phenotypes. 71

Table 2.5 Prevalence of NDD in cases carrying LP DNVs vs. cases that do not. 72

Table 2.6 Risk score weights for *de novo* variants. 86

Table 2.7 Risk score weights for rare transmitted variants..... 87

Table 2.8 Risk score weights for copy number variants. 88

Acknowledgments

I'd first like to thank my thesis advisor Dr. Yufeng Shen for his support and guidance throughout my PhD. He has been an amazing mentor over these last few years and I'm grateful to have been a part of his lab. He showed me what it means to be a true scientist and his emphasis on analytic rigor is something I will carry with me in all future endeavors (especially the "sanity checks"). I've also been extremely lucky to have been mentored by Dr. Wendy Chung, who is one of the most brilliant and charismatic leaders I've ever met and a constant source of inspiration both professionally and personally. I'm thankful for the countless opportunities that she provided when it came to research, funding, and collaborations and I'm certain my PhD experience wouldn't have been the same without her. I would also like to thank my committee members: Dr. Aris Floratos, Dr. Nick Tatonetti, Dr. Bruce Gelb, and Dr. Shuang Wang for their support and guidance throughout my PhD career (as well as their patience with my doodle poll spam!). Special thanks to my collaborators at Harvard and Mount Sinai School of Medicine, and to the PCGC patients and families– this dissertation wouldn't have been possible without all of you. Also, thank you to all current and past lab mates for the helpful discussion and feedback and thank you the administrative staff members in the Departments of Biomedical Informatics and Systems Biology (especially Marina and Carlos!) – my life would have been infinitely harder without all of you. Finally, I want to thank my mother (Peyling) and my father (Ching-Tzong) for their continued support and love through this PhD and throughout my life. I love you both very much.

Dedication

Dedicated to my parents

Introduction

0.1 Introduction to congenital heart disease

Congenital heart disease (CHD) is characterized by structural defects of the heart and great vessels. It is the most common birth defect, affected an estimated 1% of live births, and is the leading cause of mortality among birth defects {van der Linde 2011; Yang 2006}. Incidence rate was found to increase between 1977 and 2005 but has since stabilized at 0.8%-1.1%, with minor differences attributable to race/ethnicity and methods of diagnosis {Oyen 2009; Bjonard 2013}. CHD severity is categorized according the complexity of the patient's anatomical and physiological abnormalities and approximately one third of patients have severe manifestations requiring surgical intervention shortly after birth {Zaidi 2017}. Furthermore, CHD patients often acquire cardiac and extracardiac abnormalities that impact quality of life, such as arrhythmias, myocardial dysfunctions, and neurodevelopmental disorders (NDDs) {Marino 2012; Calderon 2014; Miller 2005; Burnham 2010}. While a range of fetal developmental, surgical/post-operative, and genetic factors have been found to associate with these abnormalities {Marelli 2016}, thus far none have been identified as the primary contributor {Zaidi 2017}.

Environmental factors that affect risk of cardiovascular defects in the developing fetus (during the preconception stage through to the first trimester of pregnancy) have been studied extensively and reviews on the topic {Jenkins 2007; Mone 2004} summarizing a large body of work have grouped these factors into the following categories: maternal illness, maternal

nutrition, and maternal drug exposure. Maternal illnesses such as phenylketonuria, pregestational diabetes, rubella infection, influenza, and other febrile illnesses are most strongly associated with CHD, increasing risk by 2-fold in most cases and up to 18-fold for specific defects {Jenkins 2007}. Conditions such as obesity, HIV, systemic lupus erythematosus, and epilepsy have also been shown to be associated with CHD, though the quantifiable risk difference has yet to be fully understood. Poor maternal nutrition during pregnancy – specifically deficiency or excess of folic acid and retinoic acid (vitamin A) – also directly impacts fetal cardiovascular development {Mone 2004} and intake of multivitamin supplements have been shown to reduce the risk of CHD in offspring (OR 0.5-0.8) {Jenkins 2007}. Finally, maternal drug exposure has been associated with a 2- to 4-fold increase in risk of CHD {Jenkins 2007}. Examples of non-therapeutic drugs include alcohol, cocaine, marijuana, and other narcotics {Mone 2004}. In sum, these non-genetic environmental factors are estimated to explain roughly 10% of CHD cases {Zaidi 2017} – comprehensive reviews {Mone 2004; Jenkins 2007} are available that offer more detail than the summary presented here.

Given its impact on reproductive fitness and its sporadic occurrence in families with no prior history, CHD is believed to be driven largely by genetic variation, in particular by *de novo* events in more complex CHD presentations. In terms of etiology, CHD results from disruption of key biological pathways involved in normal cardiac development – including but not limited to chromatin remodeling {Zaidi 2013; Homsy 2015}, Notch and RAS signaling {Garg 2005; Preuss 2016; Gelb 2011; Weismann 2005}, and cilia and sarcomere genes {Li 2015; Kennedy 2007; Slough 2008}. Despite progress made by recent large-scale genetic studies, more than 50% of CHD cases remain unexplained. There is an established relationship between strength of genetic effect (odds ratio) and risk allele frequency {Manolio 2009} and genetic research in

CHD, consequently, has largely focused on rare variants with large effect sizes. Given the prevalence and etiology of CHD and our current model that describes many genomic loci each contributing weak effects in an additive manner, we are still severely underpowered to detect associations between common variation and CHD. While recent GWAS studies have uncovered evidence of potential CHD susceptibility loci {Cordell 2013; Hu 2013; Agopian 2017; Hanchard 2016}, the impact of common genetic variants has yet to be fully characterized, primarily due to sample size limitations. Larger patient cohorts will be necessary to investigate the complete spectrum of genetic variation in CHD; this dissertation will focus on rare genetic variation. In terms of rare genetic variation, roughly 1% of CHD cases are attributed to rare inherited variation {Schott 1998; Gebbia 1997; Dina 2015; Durst 2015; Garg 2005}, though this number is likely an underestimate due to insufficient sample sizes for detecting the smaller genetic effect of mutations that are inherited. An estimated 23% of CHD cases are due to aneuploidies {Hartman 2011} and copy number variants {Kim 2016} and up to 30% have been attributed to *de novo* variation {Zaidi 2013; Homsy 2015; Sifrim 2016}, though that number ranges between 3% in cases with isolated CHD to 30% in cases with complex CHD. The biological mechanisms governing this difference between isolated and complex CHD cases are not yet fully understood and we hypothesize that the mutations causing CHD have pleiotropic effects that also contribute to poor clinical outcomes in these same individuals.

0.2 Introduction to genetic mosaicism

Mosaicism is readily observable in nature – certain types of animal coat color variation, for example, have been understood to be manifestations of this biological phenomenon for over a century. However, its role in human disease is still an area of active research. Mosaicism (or

post-zygotic mutation) is largely driven by mutational processes of normal aging and development, such as errors in DNA replication and repair, retrotransposition, or gain/loss of chromosomes of ploidy, among others {De 2011}. They tend to occur in the early embryonic cells of the dividing zygote and result in two or more cell populations with distinct genotypes within the same individual {Biesecker 2013}. The developmental status of the early embryonic cell in which the mutation occurs determines the proportion of mutation-carrying cells and tissue distribution of these cells in the post-natal child {Acuna-Hidalgo 2015}.

Clinical manifestation and detection of mosaicism are not yet fully understood; results from previous studies suggest that both differ according to the size of the affected region. Large-scale chromosomal abnormalities can manifest as mosaic monosomies/trisomies {Hassold 1984; Daber 2011} or isochromosome-related disorders {Hook 1983; Raffel 1986} and are typically detected via cytogenetic analysis (e.g. cell-by-cell FISH). Mosaic copy number variants have been implicated in developmental disorders {Conlin 2010; King 2015}, aging {Forsberg 2012}, hematological malignancies {Jacobs 2012; Laurie 2012}, certain forms of cancer {Lonigro 2011; Amarasinghe 2014}, and congenital heart disease {Prabhu 2015} and are conventionally studied using array-based comparative genomic hybridization or SNP microarrays. Mosaic SNVs and indels can manifest as cutaneous/dermatological disorders {Happle 1986; Weinstein 1991; Konig 2000; Hafner 2006; Hafner 2007}, overgrowth disorders {Wiedemann 1983; Lindhurst 2012; Poduri 2012; Lee 2012; Riviere 2012; Kurek 2012}, clonal hematopoiesis {Jaiswal 2014; Genovese 2014; Xie 2014}, or cancer {Forbes 2008} and were detected in the past using Sanger sequencing but more recently have been detected via next-generation sequencing techniques.

The earliest investigations into mosaicism focused on cutaneous manifestations, as the outward manifestation of these phenotypes were more easily recognizable than manifestations in

the internal organs {Happle 1993}. Work by Alfred Blaschko in 1901 was among the first studies of mosaicism in humans. Invisible under normal conditions, the Lines of Blaschko describe the trajectory along which the cutaneous ectoderm and neural crest differentiate from the ectoderm and migrate radially from the dorsal neural tube. These lines become apparent in patients with disorders of the melanocytic system whereby mosaic mutations cause hyperpigmentation of the skin. The patterning – e.g. Type 1a "narrow bands" or Type 1b "broad bands" – is largely determined by embryological processes such as cell replication, migration, and apoptosis, as well as mutation timing and physiological effect {Happle 1993}. While benign in isolation, mosaic hyperpigmentation is often observed alongside more severe clinical features in patients with multisystem diseases such as McCune-Albright Syndrome {Happle 1986; Weinstein 1991}, CHILD Syndrome {Happle 1990; Konig 2000}, Segmental Neurofibromatosis Type 1 {Ruggieri 2011; Maertens 2007; Messiaen 2011}, or Sturge-Weber Syndrome {Shirley 2013}.

Overgrowth disorders are another class of phenotypically recognizable conditions with readily identifiable affected tissue that lend themselves to the study of mosaicism. Proteus Syndrome {Wiedemann 1983; Lindhurst 2012}, Megencephaly Syndromes {Lee 2012; Poduri 2012; Riviere 2012}, and CLOVES Syndrome {Kurek 2012} all involve mosaic activating mutations in the PIK3C-AKT pathway that result in asymmetric overgrowth of bones, skin, organs, or other tissue. While the studies of the cutaneous disorders described above used lower-throughput conventional experimental techniques (e.g. denaturing gradient gel electrophoresis {Weinstein 1991}, single-strand conformation analysis {Konig 2000}, etc.) available at the time, these more recent studies of overgrowth disorders took advantage of higher-throughput next-generation sequencing (NGS) technology.

In 2012, two large-scale exome-sequencing studies investigating the role of *de novo* variation in neurodevelopmental disorders uncovered pathogenic mosaic mutations in blood, prompting further investigation into the disease implications of mosaicism detectable in blood. Gilissen et al. and O'Roak et al. sequenced blood samples belonging to large cohorts of Intellectual Disability and Autism Spectrum Disorder patients, respectively, and found, in addition to pathogenic *de novo* variants, disease-relevant post-zygotic point mutations that were confirmed via Sanger sequencing and ultra-deep resequencing. In the following years, these same research groups and others would go on to publish a set of studies applying various NGS techniques to detect mosaicism in the blood of patients with Brain Malformations {Jamuar, Walsh 2014}, Intellectual Disability {Acuna-Hidalgo, Gilissen 2015}, Autism Spectrum Disorder {Pevsner 2016; Krupp, O'Roak 2017; Lim, Walsh 2017; Dou 2017}, Epilepsy {Stosser 2018}, Alzheimer's Disease {Sala Frigerio 2015}, CINCA/NOMID Syndrome {Tanaka 2011}.

Discrepancies in reported validation rates and fraction of apparent *de novo* variants arising post-zygotically brought to light several key considerations for studying mosaicism: (1) sequencing depth and coverage profile affect mosaic detection, (2) extensive filtering is necessary to control the false positive rate, and (3) validation should both confirm presence and resolve allele fraction.

There are two key technical challenges when it comes to mosaic detection: distinguishing low allele fraction mosaics mutations from technical artifacts and distinguishing high allele fraction mosaics from germline heterozygous mutations. Methods to detect mosaicism in high-throughput sequencing data tend to use one of two approaches: (1) paired-sample (tumor/normal) and (2) single-sample {Xu 2018}. Paired-sample approaches are most commonly used in cancer where DNA is extracted from tumor and benign (normal) tissue and compared to identify tumor-

specific mutations of clinical interest. Methods in this space generally focus on distinguishing low allele fraction mosaics from technical noise and can be broadly classified into heuristic approaches (e.g. VarScan2{Koboldt 2012}, VarDict {Lai 2016}), joint genotyping-based approaches (e.g. SomaticSniper {Larson 2012}, JointSNVMix2 {Roth 2012}), and allele frequency-based approaches (e.g. MuTect {Cibulskis 2013}, LoFreq {Wilm 2012}, Strelka {Saunders 2012}). Heuristic approaches identify candidate variants on the basis of alternate allele read support and variant allele fraction and use tests such as Fisher's Exact Test comparing the allelic depth (REF, ALT) between tumor and normal samples to test for nonrandom association {Koboldt 2012}. Joint genotyping-based approaches use Bayesian comparisons of the genotype likelihoods in tumor and normal samples to identify candidate mosaics {Larson 2012}. Allele-frequency-based approaches model joint allele fractions and formulate somatic variant calling as a 2-model (wild-type vs. mutant) comparison problem, with candidate mosaics variants identified on the basis of log likelihood score {Cibulskis 2013}. Single-sample methods (e.g. SomVarIUS {Smith 2016}, Somatic-Germline-Zygosity {Sun 2018}, MosaicHunter {Huang 2017}) have been used in cancer and other settings where paired normal tissues are not always available. These methods typically perform somatic-germline classification using a probabilistic framework involving estimating the sequencing error probability and the probability of being germline for each variant and identifying candidate mosaics using pre-defined thresholds {Smith 2016}. While existing approaches tend to perform well in resolving low allele fraction mosaics, distinguishing high allele fraction mosaics from germline variants remains a challenge.

Mosaicism in heart disease is still an emerging area of research. The earliest studies involved sequencing paired blood and tissue samples from a small number of patients with

Ventricular Tachycardia {Lerman 1998} or Atrial Fibrillation {Gollob, Bai 2015}. Priest et al. in 2016 implicated a mosaic *SCN5A* mutation in Long-QT Syndrome by applying a battery of sequencing techniques to blood, urine, and saliva samples from a single patient, followed by confirmatory RNA-seq of matched heart tissue. While these studies advanced our understanding of the genetics underlying various forms of cardiovascular disease, they were limited in sample size and throughput. The most recent large-scale study of mosaicism in congenital heart disease {Manheimer 2018} analyzed exome-sequencing data of blood samples belonging to 715 proband parent trios, followed by confirmation via digital-droplet PCR. While Manheimer et al. did not find that mosaics contributed significantly, limitations in terms of sample size, detection method, and validation rate suggest that future investigations into the role of mosaicism in CHD stand to benefit from a more systematic approach to mosaic detection involving larger study cohorts and different tissue types.

0.3 Organization of this dissertation

The remainder of this dissertation is organized into two chapters.

In Chapter 1, I will discuss a new method for the detection of mosaic single-nucleotide variants in exome-sequencing data of blood and the implications of mosaicism for congenital heart disease. Briefly, we developed a new computational method, EM-mosaic, that detected mosaicism CHD patients with 90% validation rate. We found that in genes related to CHD, mosaic variants predicted to be deleterious had higher allele fraction than those predicted to be benign, suggesting presence in a larger fraction of the cells in the individual, earlier occurrence in development, and a role in disease. Detected mosaics comprised 10.4% of apparent *de novo* SNVs and occurred at a frequency of 0.122/exome. Twenty-five patients in our cohort (1%)

carried a plausible disease-causing mosaic event, all of which were independently confirmed. Analysis of individuals with matched blood and cardiac tissue available supported the notion that mosaic mutations in blood samples with relatively high allele fraction were more likely to also be detected in the heart.

In Chapter 2, I discuss a statistical approach to investigating association between genetic variation and poor clinical outcomes in congenital heart disease patients. We examined the contribution of *de novo* variation to comorbid neurodevelopmental disorder (NDD) and abnormal ventricular function phenotypes by gene set and found that pleiotropic mutations contribute a large fraction of the risk. We also developed a proof-of-concept rare variant risk score combining information from *de novo,* inherited, and copy-number variants and demonstrate its utility in predicting NDD in CHD patients, particularly those with Complex or Unknown CHD presentations.

Chapter 1: EM-mosaic: Expectation-Maximization-based detection of mosaicism

In this section, I discuss the development of a computational method, EM-mosaic (Expectation-Maximization-based detection of Mosaicism), to detect mosaic single nucleotide variants (SNVs) using whole-exome sequencing data (WES) of proband and parent DNA. We evaluated our method using a simulation experiment to measure the accuracy of its mosaic fraction estimation and its posterior odds-based false discovery rate estimation. To optimize this method, we also measured mosaic detection power as a function of variant allele fraction and sequencing depth. We then compared EM-mosaic against an existing method, MosaicHunter {Huang 2014}, and applied both methods investigate mosaicism in 2530 CHD proband-parent trios from the Pediatric Cardiac Genomics Consortium (PCGC) {Jin 2017}, using exome sequences derived from blood-derived DNA. We detected predicted deleterious mosaic mutations in genes involved in known biological processes relevant to CHD or developmental disorders in 1% of probands. The accuracy of these mosaic variant detection algorithms was assessed using an independent re-sequencing method. We found that among high-confidence mosaic mutations in CHD-relevant genes, likely-damaging variants tended to have higher VAF

than likely-benign variants. In parallel, we assessed mosaicism by applying EM-mosaic and MosaicHunter to 70 discarded tissues from several heart regions obtained from 66 probands who underwent cardiac surgical repairs. While VAF varied significantly (>3 fold) between blood and cardiovascular tissue at about 60% of sites, in general mosaic variants with high (>15%) VAF were more likely shared between blood and cardiac tissue than variants with lower VAF.

1.1 Introduction

Mosaicism results from somatic mutations that arise post-zygotically in an early embryonic cell, resulting in two or more cell populations with distinct genotypes in the developing embryo {Biesecker 2013}. The developmental status of the early embryonic cell at the time of mutagenesis determines the proportion of variant-carrying cells and the tissue distribution of these cells in the post-natal child {Acuna-Hidalgo 2015}. While germline variants have a variant allele frequency (VAF) of 0.5, somatic mosaic variants have a significantly lower VAF.

Post-zygotic mosaic mutations have been implicated in several diseases including non-malignant developmental disorders such as overgrowth syndromes {Poduri 2013; Lindhurst 2012; Kurek 2016}, structural brain malformations {Poduri 2012; Jamuar 2014; Riviere 2012; Lee 2012}, epilepsy {Stosser 2018}, and autism spectrum disorder {Lim 2017; Krupp 2017; Freed 2016; Dou 2017}. Recent analyses also identified mosaic variants in a cohort of patients with congenital heart disease (CHD) {Manheimer 2018}, but the prevalence of these was far less than germline variants (CHD) {Zaidi 2013; Homsy 2015; Jin 2017; Zaidi 2017}.

Assessment of the frequency of mosaicism in human disease is confounded by technical issues, including differences in sequencing depth, DNA sources, and variant assessment pipelines. Low levels of mosaicism can escape the detection threshold of traditional sequencing

methods with standard read depths, while post-zygotic mutations with a higher percentage of affected cells are difficult to discriminate from germline de novo mutations {Acuna-Hidalgo 2015}. All of these issues can lead to substantially different conclusions. For example, analyses of mosaicism in autism spectrum disorder was recently assessed from whole exome sequence (WES) data from whole blood DNA from 2506 families (proband, parents and unaffected sibling; trios and quads) in the Simons Simplex Collection (SSC) {Fischbach 2010}. The primary sequence data were analyzed by three groups; one that identified a protein-coding somatic mosaic variant rate of 0.074 per individual {Freed 2016}, another that found a mosaic rate of 0.059 per individual {Lim 2017}, and a third group that reported a mosaic rate of 0.125 per individual {Krupp 2017}. This disparity suggests the need for more systematic mosaic mutation detection methods that account for dataset-specific confounding factors.

By contrast, analyses of affected tissues can improve the sensitivity and specificity of detection of somatic mosaicism. In cancer, methods to detect these events, such as MuTect {Cibulskis 2013}, compare tumor and benign tissues from the same patient. Mosaicism has also been demonstrated from the analyses of unpaired samples with cancer and other pathologies {Sun 2018; Huang 2017; Smith 2015} by the demonstration of variants in affected tissues that are absent from blood-derived DNA {Symoens 2017; McDonald 2018}. With access to cardiac tissues from patients with CHD obtained during surgical repair, we hypothesized that analyses of mosaicism in cardiac tissue might improve insights into the causes of this common congenital anomaly. As many cardiomyocyte lineages share a mesodermal origin with blood cells but exit the cell cycle during embryogenesis, we also sought to determine if mosaicism in the heart exhibited distinct patterns of mosaicism with regard to variant frequency and allele fractions.

1.2 Results

1.2.1 High-accuracy detection of mosaic mutations in WES data using EM-mosaic

We analyzed whole exome sequence (WES) data from 2530 CHD proband-parent trios {Homsy 2015; Jin 2017} **(>Table S1).** Among this cohort, 1205 probands had CHD with neurodevelopmental disorders (NDD) and/or extracardiac manifestations (EM), 788 had isolated CHD at the time of enrollment, 539 had undetermined NDD status due to young neonatal age at the time of enrollment, and 9 subjects had incomplete data **(>Table S2)**.

Previous WES analyses {Jin 2017} identified 1742 germline de novo SNVs among 838 cases with NDD and/or EM, 516 isolated cases, 644 cases of unknown NDD status, and 7 with incomplete data. These de novo variants were identified using the Genome Analysis Toolkit (GATK) pipeline {McKenna 2010; DePristo 2011} assuming a germline diploid model in which the expected VAF is 0.5. This model has limited sensitivity to detect mosaic mutations for which the fraction of alternative allele reads is significantly below 0.5, especially because de novo variants with VAF<0.2 were excluded to reduce false discovery.

To efficiently capture mosaic variants with VAF<0.4, we developed a new method (EM-mosaic) to detect mosaic variants in WES sequence of a proband and parents (trios). Potential mosaic variants were identified in WES sequence data using SAMtools mpileup {Li 2009} with settings designed to capture sites with VAF between 0.1-0.4 and merged with the variants found by the GATK pipeline {Jin 2017} **(>Fig. 1.1)** to create a union variant set. To reduce the elevated false positive rate inherent in low-VAF calls, we applied a set of empirical filters to remove likely technical artifacts due to sequencing errors associated with repetitive and/or low complexity sequences. We then manually inspected de novo SNVs with VAF<0.3 (n=582) using

IGV and filtered out an additional 188 likely false positives. After preprocessing and outlier removal, the remaining 2971 de novo SNVs were used as input to our mosaic detection model. Among the 2971 de novo SNVs, this pipeline identified 309 sites as candidate mosaics based on posterior odds score **(>Fig. 1.2A-B, Table S3)**, including 50 sites that were previously reported as germline de novo variants {Jin 2017}. An additional 86 sites were identified as having posterior odds below our threshold of 10 but greater than 1 **(>Fig. 1.3A-B)**, including a ZEB2 variant with posterior odds 4.7 that was previously confirmed via ddPCR {Manheimer 2018}. Among these 86 variants, 53 are likely mosaic and 33 are likely germline **(>Fig. 1.3B).** We chose not to include these sites since there was insufficient evidence to confidently resolve them individually as mosaic or germline.

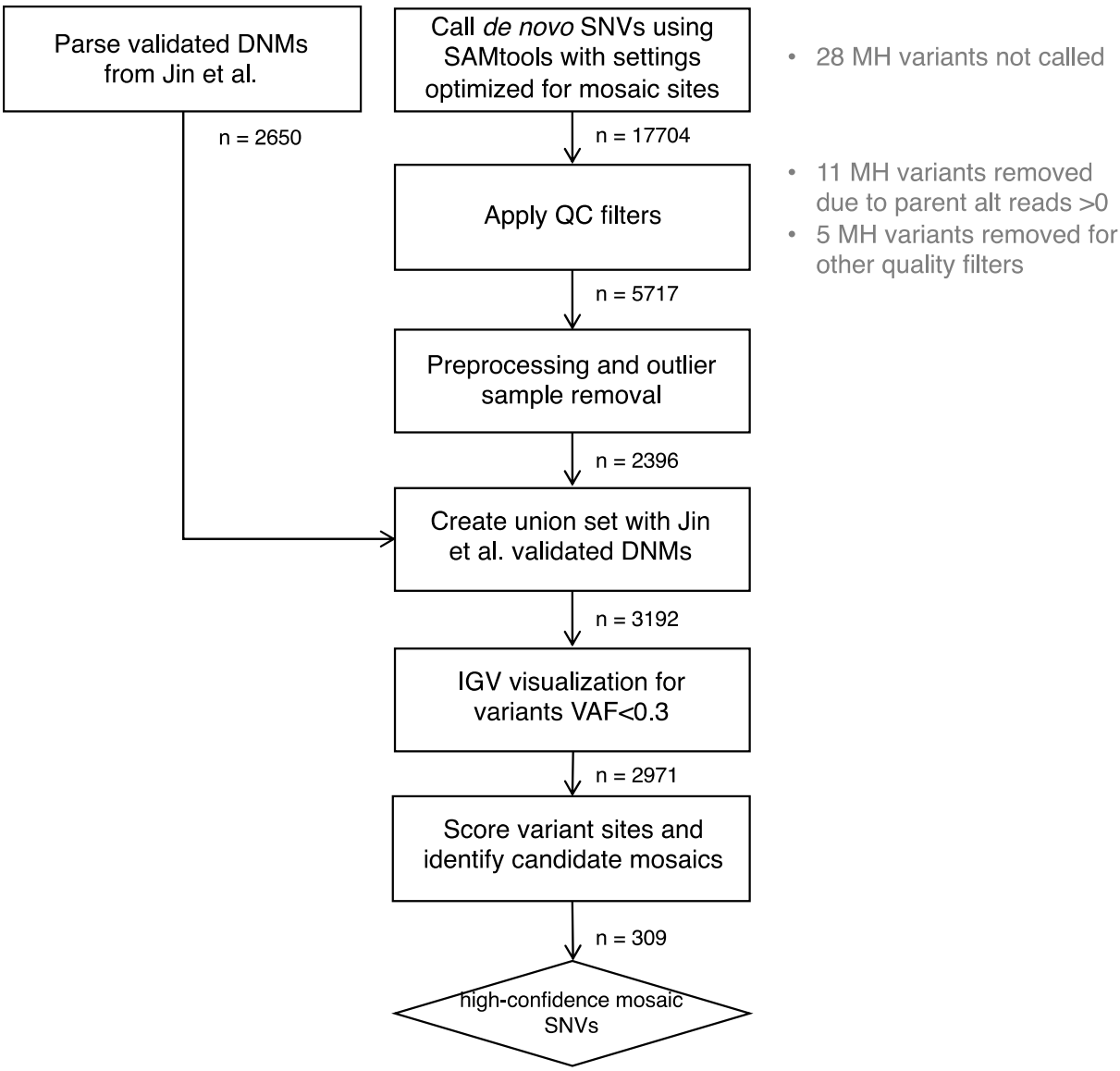

Figure 1.1. EM-mosaic flowchart

We first processed our SAMtools de novo calls using our upstream filters (n=2396 sites passing all filters). We then applied the same upstream filters to the published dnSNVs from Jin et al. (n=2650 sites passing all filters) before finally taking the union of these two call sets (n=3192). High-confidence mosaics (n=309) were defined as mosaics passing IGV inspection and having posterior odds > 10. Grey text indicates which filters removed candidate mosaic variants called by MosaicHunter but not by EM-mosaic.

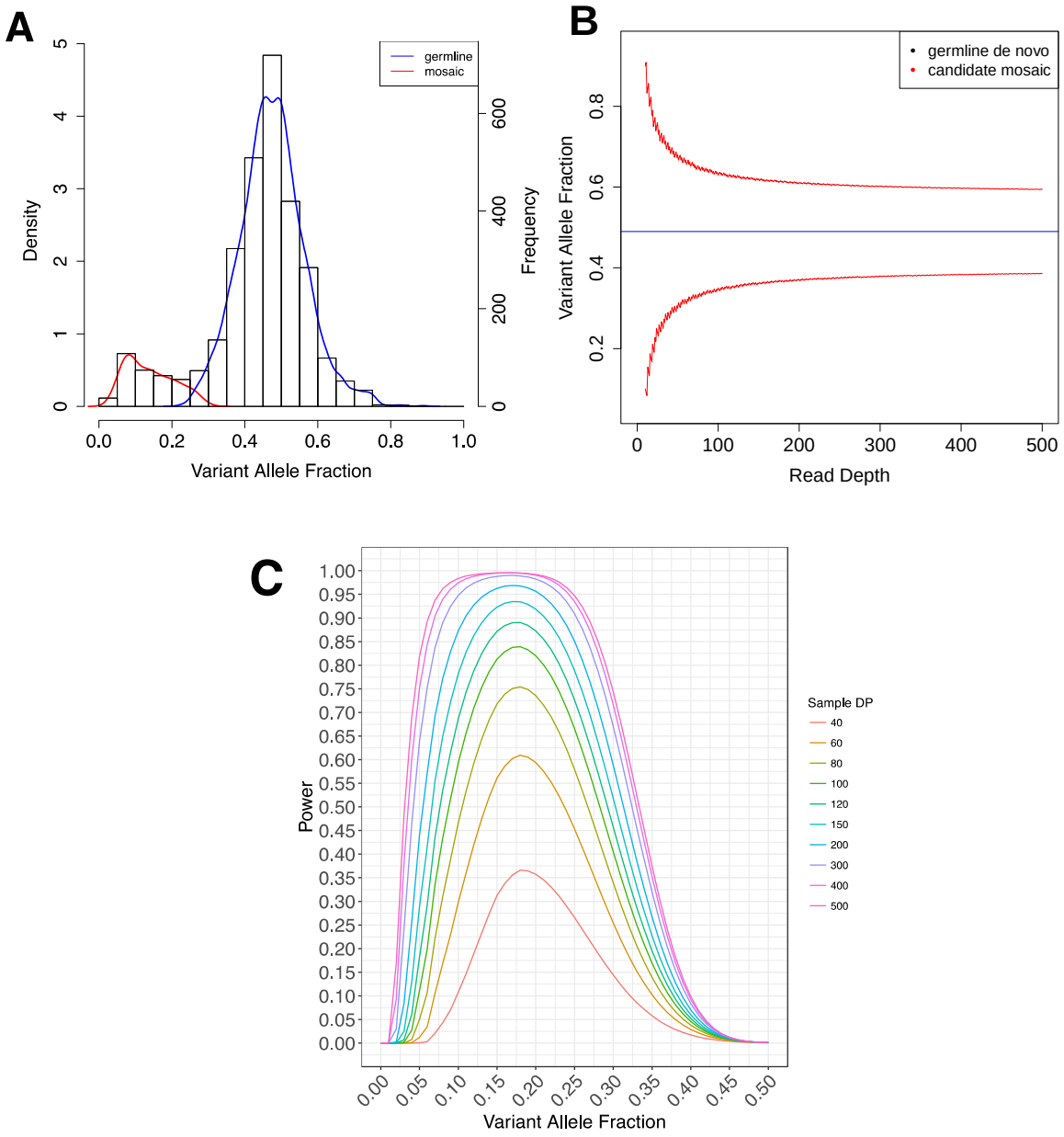

Figure 1.2. Mosaic detection by Expectation-Maximization.

(A) Expectation-Maximization (EM) Estimation to decompose the variant allele fraction (VAF) distribution of our input variants into mosaic and germline distributions. The EM-estimated prior mosaic fraction was 12.15% and the mean of the mosaic VAF distribution was 0.15. (B) Read depth vs. VAF distrubution of individual variants. The blue line denotes mean VAF (0.49) and the red lines denote the 95% confidence interval under our Beta-Binomial model. Mosaic variants are defined as sites with posterior odds > 10, corresponding to a False Discovery Rate of 9.1%. Germline variants are represented in black and mosaic variants are represented in red. (C) Estimated mosaic detection power as a function of average sample depth for values between 40x and 500x.

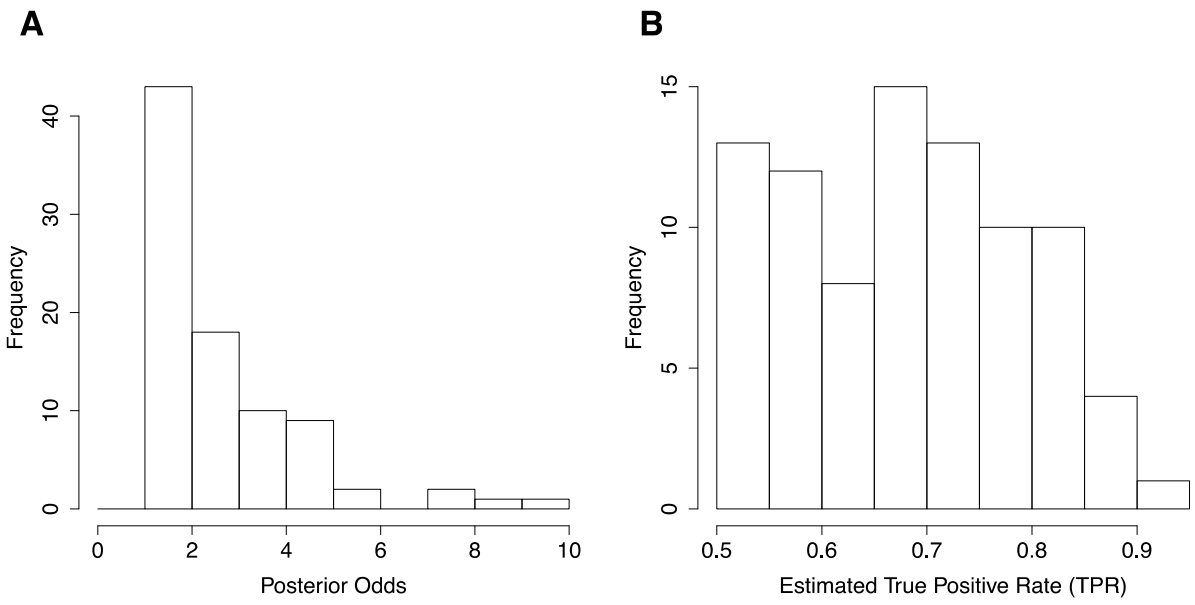

Figure 1.3. Blood variants with posterior odds between 1 and 10.

(A) Distribution of the 86 variants with posterior odds between 1 and 10. (B) Histogram of counts by bin. To estimate the number of potential mosaics missed by our threshold, counts of each bin were scaled by the estimated true positive rate (TPR; posterior odds / 1+posterior odds). By our estimate, 54/86 variants were likely mosaic and 32/86 were likely germline.

Table 1.1. Mosaic detection by EM-mosaic, MosaicHunter; validated by PCR product sequencing

		Union	**Shared**	**Unique**	
				EM-mosaic	**MosaicHunter**
Mosaic Variants (total)*		**315 (332)**	**56 (57)**	**218 (240)**	**29 (35)**
Mosaic Candidates		367	58	251	58
Mosaic Candidate VAF mean (SD)		0.13 (0.06)	0.12 (0.05)	0.13 (0.06)	0.10 (0.05)
MiSeq Confirmation	Total Tested	143	22	75	46
	Mosaic	108	21	64	23
	Germline	3	0	3	0
	No Variant	32	1	8	23
Validation Rate		76%	95%	85%	50%

*Estimated number of mosaic variants found among 2530 CHD probands (total number of mosaic variants detected by EM-mosaic and MosaicHunter).

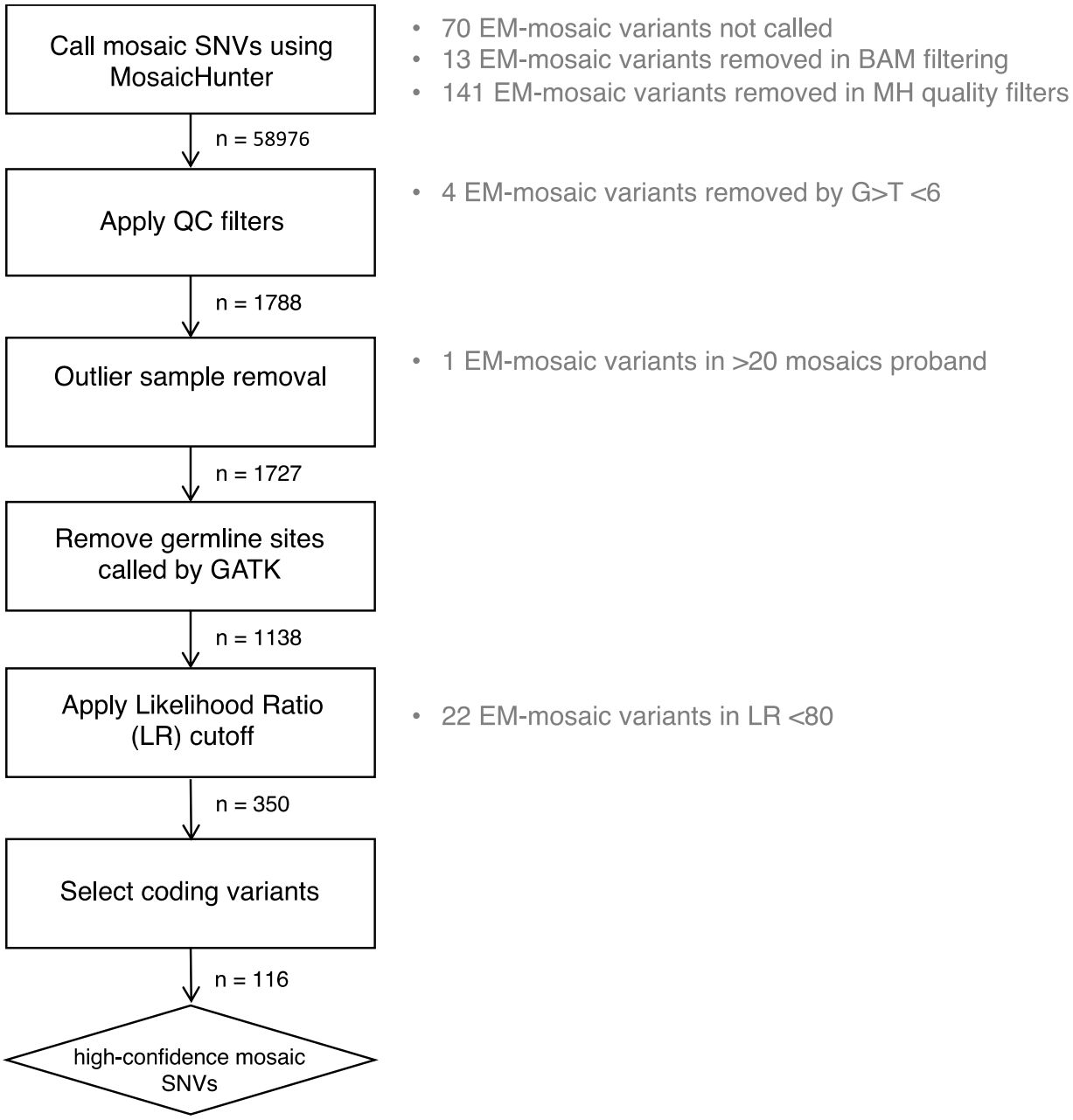

Figure 1.4. MosaicHunter workflow.

Quality Control filters excluded any sites that were (1) present in ExAC (2) G>T with Nalt<10 (3) parent Nalt>2. Outliers were defined as probands carrying more than 20 mosaics, or non-unique sites. We also removed sites called as germline by GATK Haplotype Caller. High-confidence mosaics (n=116) were defined as having Likelihood Ratio > 80 and affecting coding regions excluding *MUC*/*HLA* genes. Grey text indicates which filters removed variants called by EM-mosaic but not by MosaicHunter.

1.2.2 Method evaluation via simulation experiment

We evaluated EM-mosaic using a simulation experiment to measure the accuracy of its mosaic fraction estimation and its posterior odds-based false discovery rate (FDR) estimation. We simulated roughly 1 million variants (N_{alt}, N) for a range of sample average sequencing depth values (40x, 60x, 100x) and spiked in a known fraction of simulated mosaic variants. We then applied EM-mosaic to each dataset and compared the resulting mosaic and germline predictions for each variant against their ground truth labels. We found that our EM-estimated mosaic fraction was consistent with the true fraction across all datasets (± 0.3%), with slight overestimation at lower sequencing depth (40x) and slight underestimation at higher sequencing depth (100x). We next estimated the false discovery rate (FDR) for each variant as a function of posterior odds (1/(1+posterior odds)). Then, for FDR cutoffs j = {0, 0.01, …, 0.99, 1.0}, we calculated both the $qvalue_j = \frac{\sum_1^N fdr_i}{N}$ as well as the False Discovery Proportion (FDP_j ; the fraction of variants with a ground truth label of "germline") using the N variants with FDR < j. We found that our posterior odds-based FDR estimates were consistent with the true FDR values (>**Fig. 1.5C**). Simulation experiment results for the 60x dataset (representative of the sequencing depth used for the CHD patient cohort) are summarized in **Figure 1.5**; results from the other datasets (40x, 100x) can be found in **Figure 1.17**.

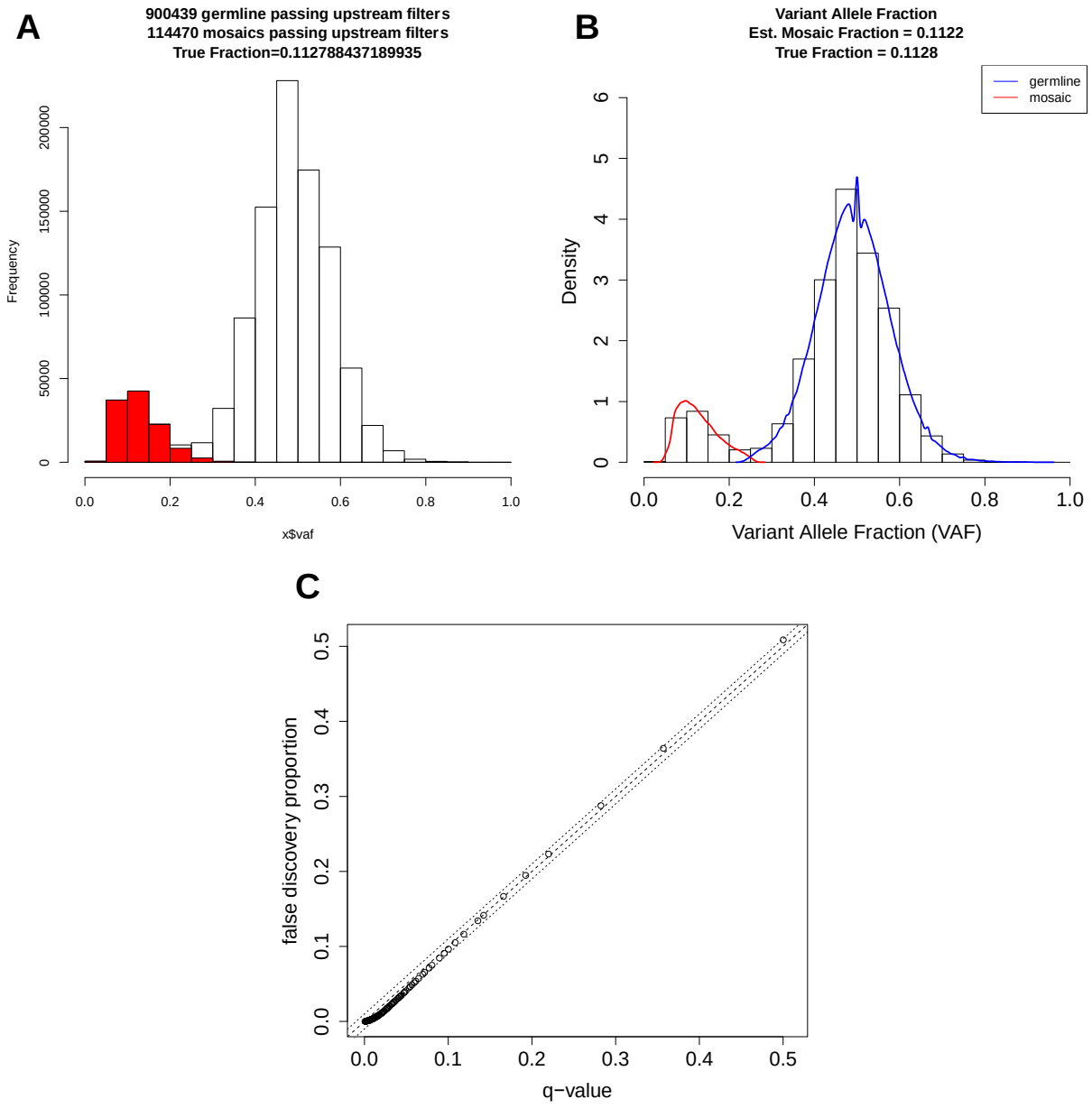

Figure 1.5. Simulation experiment results (60x dataset).

(A) We simulated n=1,015,017 variants at sample average sequencing depth 60x (n=900,439 germline, n=114,470 mosaic). (B) Our EM approach accurately estimated mosaic fraction (simulated = 11.28%, inferred = 11.23%). (C) Our posterior-odds-based FDR was consistent with the FDR as determined from the ground truth variant labels (# false positives / # predicted positives). The dashed lines denote perfect concordance with a margin of +/- 0.01.

1.2.3 Mosaic mutations found in blood derived DNA with MosaicHunter

We also employed MosaicHunter, which uses a Bayesian genotyping algorithm with a series of stringent filters (see Materials and Methods) for discovering mosaic variants using WGS genotype information from trios. {Huang 2017} Among the 2530 CHD trios, MosaicHunter identified an initial set of 58976 sites showing evidence of mosaicism, including 214 high-confidence variants located in coding regions. **(>Fig. 1.4).** After applying a minimum likelihood ratio (LR) cutoff of 80 for distinguishing mosaic from germline mutation, and additional heuristic filters (Materials and Methods), MosaicHunter identified 116 coding sites **(>Table S4)** or 0.05 mosaics /individual.

Of the mosaic candidates detected by MosaicHunter, 58/116 (50%) were also identified by EM-mosaic while 58/116 (50%) candidates were unique to MosaicHunter **(>Table 1.1)**. Of the 58 candidates unique to MosaicHunter, 35 were filtered out by EM-mosaic on the basis of insufficient alternate allele read support, 16 had a non-zero allelic depth in the parents, and 7 failed quality filters. The 251 candidates unique to EM-mosaic were discarded by the MosaicHunter pipeline during BAM reprocessing (n=13), quality filtering (n=146), application of LR cutoff (22), or were not called due to inadequate read depth (n=70) **(>Fig. 1.4).**

1.2.4 Sequence confirmation of candidate mosaic variants and estimation of mosaicism in CHD

From the 367 high-confidence EM-mosaic and/or MosaicHunter SNVs, we selected 143 candidates (97 identified by EM-mosaic; 68 identified by MosaicHunter) for experimental confirmation using MiSeq amplicon resequencing **(>Table S5; Tables S11 and S12; Methods)**. DNA fragments encompassing the putative mosaic variant were PCR-amplified from proband

and each parent DNA, sequenced on an Illumina MiSeq next generation sequencer and VAF was calculated for each individual. These candidate mosaics included SNVs on the extremes of the VAF spectrum, as well as mosaics that were flagged by MosaicHunter quality filters. Candidates mosaic variants were considered confirmed by MiSeq analyses if they demonstrated an amplicon VAF exceeding 0.01 but less than 0.45, so as to indicate a variant of post-zygotic origin. MiSeq VAF values closely correlated with those originally determined by exome sequencing (P= 2.2x10-16; >**Fig. 1.6**).

We confirmed 85/97 (88%) EM-mosaic candidate mosaic variants. Three candidate variants were likely germline de novo SNVs (VAF>0.45). Nine candidate variants were 'false positives' that were neither germline de novo SNVs or mosaic SNVs since either no variant reads were detected by MiSeq sequencing of the proband amplicon, or the same small fraction of variants were detected in proband amplicon and one parent's amplicon.

Parallel analyses with MosaicHunter confirmed 44/68 (65%) candidate mosaic variants. There were 23 sites for which no variant reads were detected by MiSeq amplicon sequencing (MiSeq VAF<0.001) or in which the same small fraction of variant reads was detected in the proband amplicon as in one parent's amplicon.

We considered whether estimates of mosaic variant frequency were sensitive to whole exome sequencing depth by calibrating estimates of mosaic detection power using properties of the sequence data (average read depth, prior mosaic fraction, and the value of our overdispersion parameter θ) (>**Fig. 1.16**; Materials and Methods). Our projected mosaic detection power curves demonstrated more than a doubling of power to detect mosaic variants with VAF 0.2 as sequencing depth increases from 40x to 80x **(>Fig. 1.2C).** Projected mosaic detection power

curves for less stringent mosaic cutoffs showed similar increases of power with increasing sequencing depth (>**Fig. 1.7**).

To estimate the 'true' frequency of mosaicism per blood DNA exome, independent of average coverage detection power constraints, we estimated the 'true' mosaic count in a VAF range by multiplying the number of mosaics by the inverse of the detection power for each VAF bin. Applying this method to the 184 of 309 high-confidence EM-mosaic variants with VAF>0.1, we estimated the adjusted number of mosaics with VAF>0.1 to be 361 **(>Fig. 1.7A)**. Thus, the true frequency of coding mosaics in the blood (0.4>VAF >0.1) is 0.14 variants per individual, representing a non-negligible class of mutations with potential contribution to genetic risk for congenital heart disease. The estimated true mosaic frequency does not change significantly when using less stringent mosaic definitions **(>Fig. 1.7B-C)**. In sum, we identified 315 blood mosaic variants in 2530 CHD probands or 0.13 mosaic variants per subject with a mean VAF of 0.13±0.06. We do not anticipate that doubling the sequencing depth would change significantly this estimate.

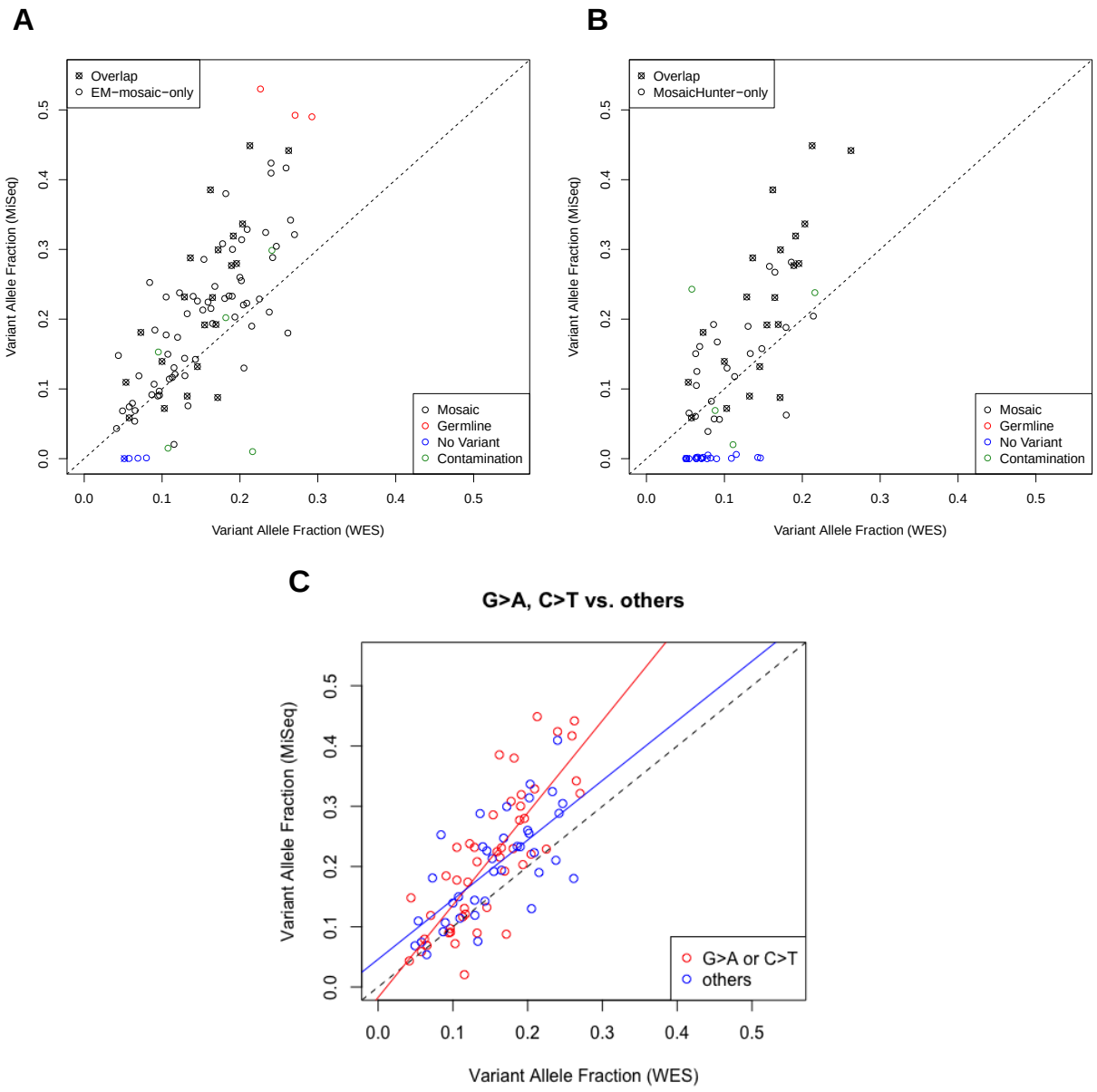

Figure 1.6. Targeted sequencing to validate candidate blood mosaic variants.

(A) EM-mosaic and (B) MosaicHunter variants were assayed using PCR followed by MiSeq for high-depth assessment of mosaicism. Variants with x symbols were shared by both pipelines. Mosaic variants that validated are black, while variants with VAF > 0.45 and therefore germline are red. Validation VAF values demonstrated significant correlation with the original WES-derived VAF for EM-mosaic (Pearson's correlation P=2.2x10-16) and MosaicHunter (P=8.2x10-11). (C) We observed an inflation of VAF when comparing MiSeq against WES data, which we attribute to a combination of reference bias and preferential amplification of G and C reference bases.

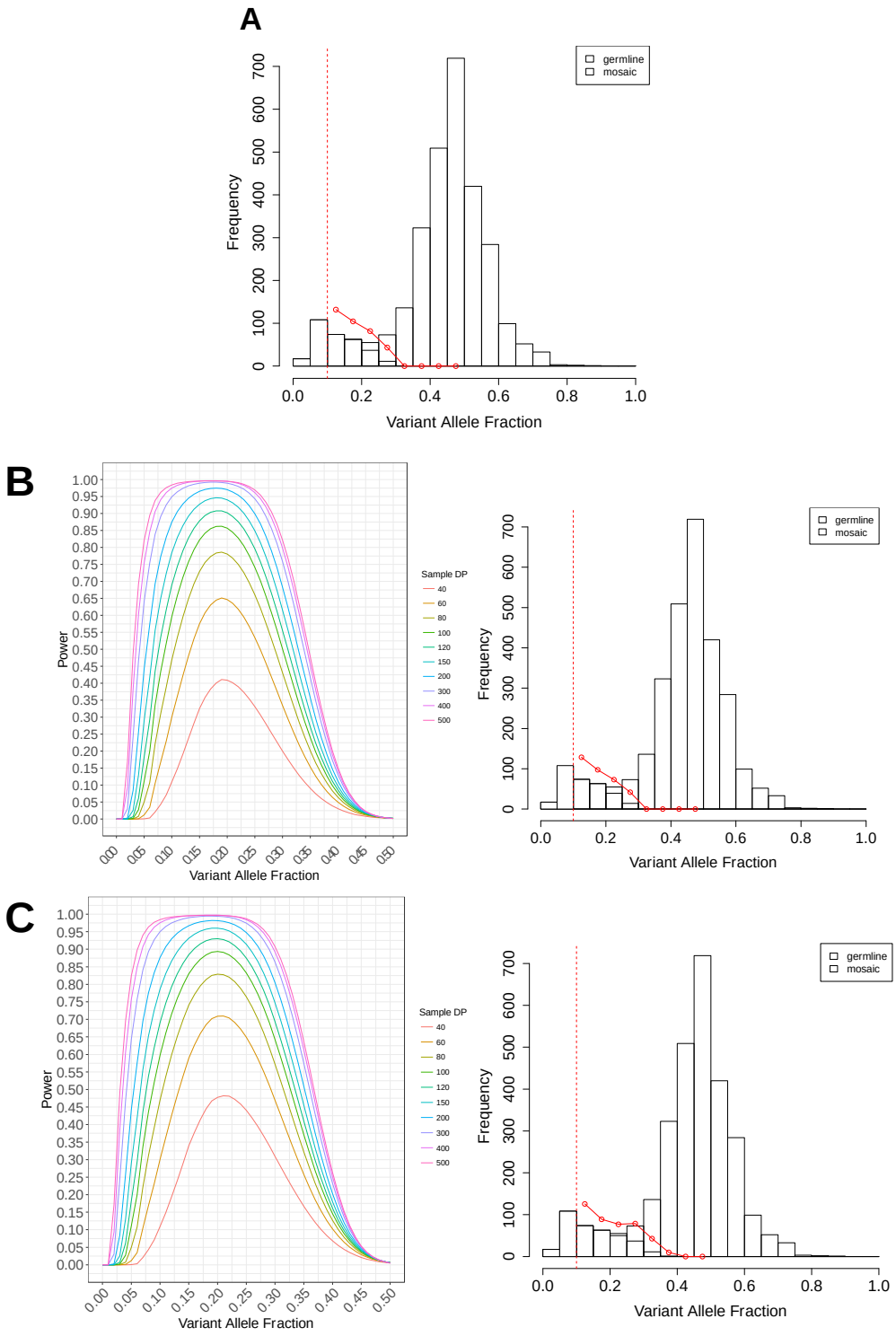

Figure 1.7. Estimated mosaic detection power using less stringent mosaic definitions.

(A) Estimated true frequency of detectable coding mosaics (0.4>VAF>0.1) adjusted by detection power (n=341; 0.135/exome) (B) Calibrated mosaic detection power and estimated true mosaic frequency of detectable coding mosaics, using posterior odds cutoff of 5 (n=361; 0.143/exome). (C) Calibrated mosaic detection power and

estimated true mosaic frequency of detectable coding mosaics, using posterior odds cutoff of 2 (0.4>VAF>0.1; n=424; 0.168/exome).

1.2.5 Mosaic variants occurred most frequently at CpG sequences.

Previous studies demonstrated a strong preference for de novo C>T mutations at CpG dinucleotides compared to other dinucleotides due to the spontaneous deamination of 5-methylcytosine {Fryxell 2005; Francioli 2015}. We asked whether the germline de novo variants observed in CHD probands and the 332 mosaic sites demonstrated a similar sequence preference **(>Fig. 1.8, Table 1.1, Tables S3 and S4)**. Of the 2662 germline de novo mutations identified in 2530 CHD probands, 979 variants (37% of all variants) involved mutation of the cytosine of a CpG dinucleotide (>**Fig. 1.8A**). By contrast, 99 (29% of all mosaic SNVs) of 332 mosaic SNVs altered the cytosine of a CpG dinucleotide more than expected by chance (2.2x above expectation; p=2.0E15). Ignoring the high CpG mutation frequency, cytosines and guanines were ~2-fold more likely to be mutated than adenines or thymidines both for germline mutations and for mosaic variants. Surprisingly, somatic mutations of A>C/T>G transversions in ApC dinucleotides were ~2-fold greater than the corresponding germline mutations (P=5x10-8; >**Fig. 1.8B**).

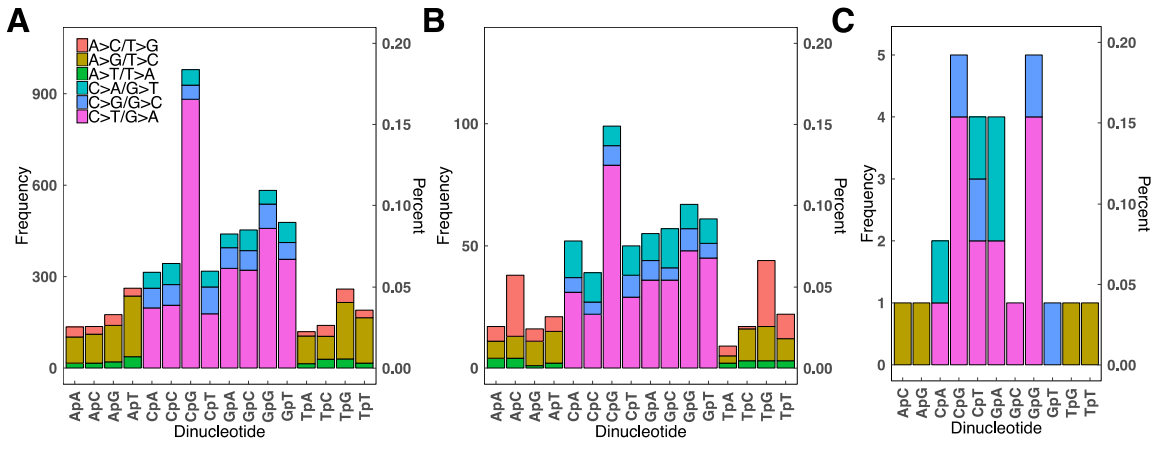

Figure 1.8. Mutation spectrum of detected germline and mosaic variants.

Rates of specific mutations for (A) germline, (B) blood mosaic, and (C) CHD tissue mosaic variants. Transitions predominated in both variant sets.

1.2.6 Detection of mosaic mutations in CHD tissues

Using EM-mosaic and MosaicHunter we analyzed exome sequences from 70 cardiac tissues derived from 66 subjects with CHD **(>Table S6)** and paired blood samples. Among 57 de novo variants (allele depth approximately 0.5) that were previously identified in blood-derived DNA, 54 were also found in CHD tissues. Of the 3 de novo variants not present in cardiac tissue, 1 was outside of the tissue WES capture region and 2 occurred in a single proband **(>Table 1.2)**. In addition, 23 distinct candidate mosaic variants were detected by EM-mosaic (n=13), MosaicHunter (n=6), or by both algorithms (n= 4). All 23 candidates were tested via MiSeq amplicon sequencing of blood and cardiac tissue DNAs; 15 of 23 unique candidate mosaics were confirmed **(>Table 1.2, S7)**, including a CCNC variant that was identified in two different CHD tissues from proband 1-01684. Ten (86%) confirmed mosaic variants were detected in blood and cardiac tissues (MAF>0.01), four were found only in cardiac tissue, and one was found only in blood. Of the 7 mosaics detected by blood WES analysis, 4 were confirmed in the corresponding cardiac tissue sample. Remarkably, five confirmed cardiac tissue mosaic variants occurred in one proband (1-07004), one of which was also present in blood DNA.

These analyses indicate a frequency of coding mosaics (0.4>VAF >0.1) in the cardiac tissues of 0.14 per individual (9 of 66 probands), which approximated our estimate of 0.14 blood mosaics per individual **(>Fig. 1.7A)**. Despite these similar frequencies, multiple distinct mosaic variants were identified in these tissues. Mosaics with highest VAF were more likely to be found in both tissues (Mann-Whitney U Test P=0.019), presumably indicating that the mutation occurred earlier in lineage development **(>Fig. 1.9, Fig. 1.10)**.

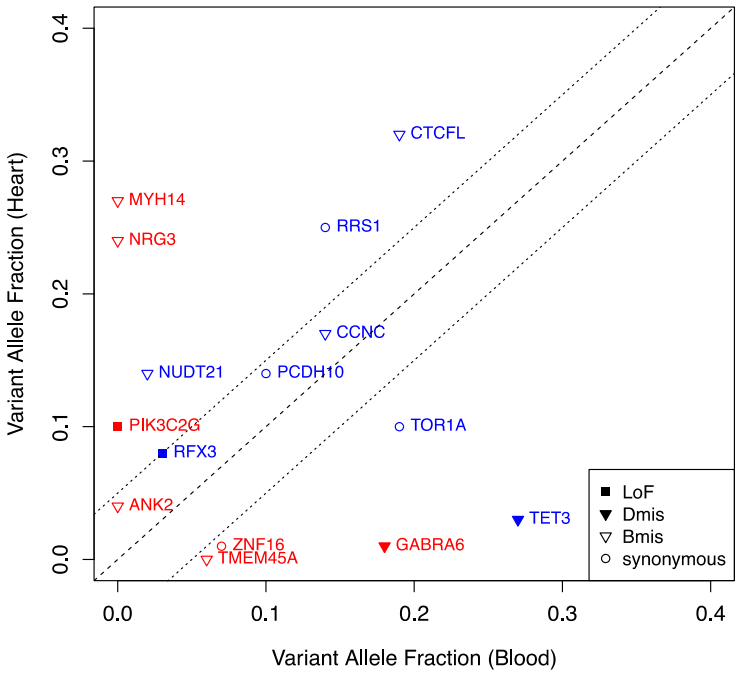

Figure 1.9. Validated mosaics detected in probands with matched blood and cardiovascular tissue samples available.

Validation VAF from blood compared to validation VAF from cardiovascular tissue demonstrated tissue-specific mosaicism (red) as well as shared mosaicism (blue). Predicted effect of mosaic variants corresponds to marker shape.

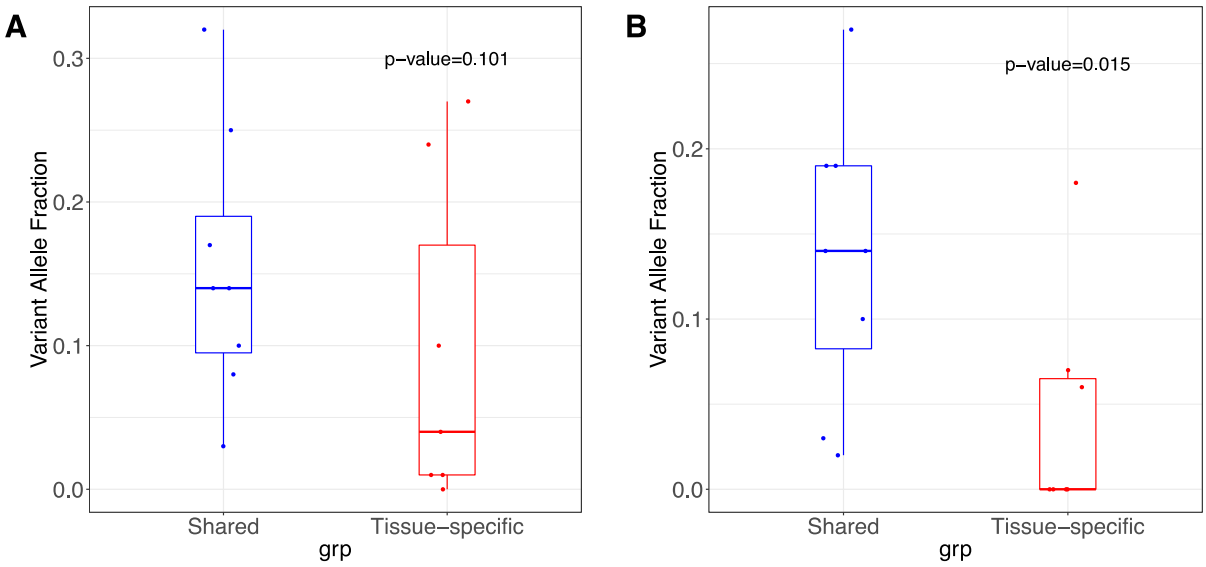

Figure 1.10. Mosaic variants shared in blood and cardiovascular tissues have higher variant allele fraction.

Validation VAF from (A) cardiovascular tissue and (B) blood had higher VAF for shared variants compared to tissue-specific variants (p=0.101 and 0.015, respectively).

Table 1.2. Mosaics detected in individuals with matched cardiovascular tissue and blood

ID	Gene	Variant Class	Pipeline	CHD Tissue				Blood WES VAF		
				Location	WES AD	WES VAF	MiSeq VAF	WES AD	WES VAF	MiSeq VAF
1-00543	*CTCFL*	Bmis	EM-mosaic	AO	138,36	0.21	0.32	29,8	0.22	0.19
1-00984	*ZNF16*	syn	EM-mosaic	LV	262,1	0.00	0.01	100,7	0.07	0.07
1-01282	*GABRA6*	Dmis	MosaicHunter	RV	104,1	0.01	0.01	55,12	0.18	0.18
1-01684	*CCNC*	Bmis	Both	AoValve, RV	36,7	0.16	0.17, 0.19	224,40	0.15	0.14
1-02672	*TOR1A*	syn	Both	AtrSpt	159,10	0.06	0.10	29,6	0.17	0.19
1-03512	*RFX3*	LoF	MosaicHunter	RV	156,15	0.09	0.08	39,0	0.00	0.03
1-04652	*PCDH10*	syn	Both	AtrSpt	154,19	0.11	0.14	15,1	0.06	0.10
1-07004	*ANK2*	Bmis	MosaicHunter	SubAoMembr	226,13	0.05	0.04	30,0	0.00	0.00
1-07004	*MYH14*	Bmis	Both	SubAoMembr	124,22	0.15	0.27	33,0	0.00	0.00
1-07004	*NRG3*	Bmis	EM-mosaic	SubAoMembr	152,30	0.16	0.24	43,0	0.00	0.00
1-07004	*NUDT21*	Bmis	Both	SubAoMembr	137,22	0.14	0.14	74,0	0.00	0.02
1-07004	*TET3*	Dmis	MosaicHunter	SubAoMembr	131,1	0.01	0.03	81,16	0.16	0.27
1-07299	*RRS1*	syn	Both	RV, UNK	160,25	0.14	0.25	22,2	0.08	0.14
1-09869	*PIK3C2G*	LoF	MosaicHunter	LV	126,9	0.07	0.10	31,0	0.00	0.00
1-11800	*TMEM45A*	Bmis	MosaicHunter	RV	213,0	0.00	0.00	32,7	0.18	0.06

Characteristics of mosaic variants predicted for individuals with blood and cardiovascular tissue WES data available. Among 15 mosaics, 5 were detected via analysis of blood WES, 8 were detected from cardiovascular tissue WES, and 2 were detected by both approaches. Six of 7 (86%) mosaics detected from analysis of blood were present in both DNA sources with MiSeq VAF≥0.01. Two additional variants previously identified as de novo germline variants in blood WES were absent from CHD tissue WES. Minimum 1023 MiSeq reads used to determine VAF. Abbreviations: AD, allelic depth (reference, alternate); AO, aorta; AtrSpt, atrial septum; Bmis, benign missense; Dmis, deleterious missense; LOF, Loss of function variant; LV, left ventricle; RV, right ventricle; VAF, variant allele fraction.

1.2.7 Blood and cardiac tissue mosaics likely to contribute to CHD

Our prior genetic studies of CHD studies showed that damaging de novo variants typically occurred in genes highly expressed in the top quartile of the developing E9.5 mouse heart (HHE) {Zaidi 2013; Homsy 2015} or that contribute to CHD in mouse models {Jin 2017}. Among the 342 mosaic variants identified from blood or cardiac tissue analyses that were not false by MiSeq, 65 altered these HHE and/or mouse CHD genes (n=4558, >**Table S8**). RefSeq functional annotation predicted 52 variants as likely-damaging variants (LOF, Dmis), and 46 as likely benign, missense (>**Tables S8, S9**). In total, we observed potentially CHD-causing mosaic mutations in 25 participants, representing 1% of the 2530 total participants in our CHD cohort.

Among these 25 mosaics, we confirmed 22/22 (100%) candidates tested via MiSeq. Notably, multiple likely-damaging mosaic variants altered genes (*ISL1, SETD2, NOVA2, SMAD9, LZTR1, KCTD10, KCTD20, FZD5, and QKI*) involved in key developmental pathways, which may account for the extra-cardiac phenotypes observed in these patients (>**Table 1.3, S10**). There was no difference in the proportion of individuals with extracardiac features among those with damaging mosaic variants compared to the overall cohort (11/25 vs 909/ 2521, P=0.68), and there was a wide range of CHD subtypes. Five subjects carried additional de novo LoF or Dmis variants (1-06216, *TYRP1*; 1-04046, *KRT13*; 1-06677, *TRIP4*; 1-05011, *KDM5B*; 1-00018, *SBF1*) and 4 genes harbored de novo LoF or Dmis variants other than those listed in **Table 1.3** (*FBN1; PKD1; LZTR1; PIK3C2G*). No CNVs were detected in these subjects, with the exception of 1-00192 (duplication at chr15:22062306-23062355; non-overlapping with the *GLYR1* mosaic).

If mosaic variants were unrelated to CHD, we would expect similar allelic fractions between mosaics with variants predicted as likely damaging or likely benign. However, we found that the allele fraction of likely damaging variants was significantly higher (Mann-Whitney U Test P=0.001, >**Fig. 1.11A**). Moreover, among mosaic variants in genes that are not included among HHE or mouse CHD genes, we found no significant difference of allele fraction (P=0.985, >**Fig. 1.11B**). We repeated these analyses using less stringent posterior odds cutoffs of 2 and 5 and found the same result (>**Fig. 1.12**). Together these data support our conclusion that at least some likely-damaging mosaic variants identified here contribute to CHD. These results were determined independently of MiSeq validation results.

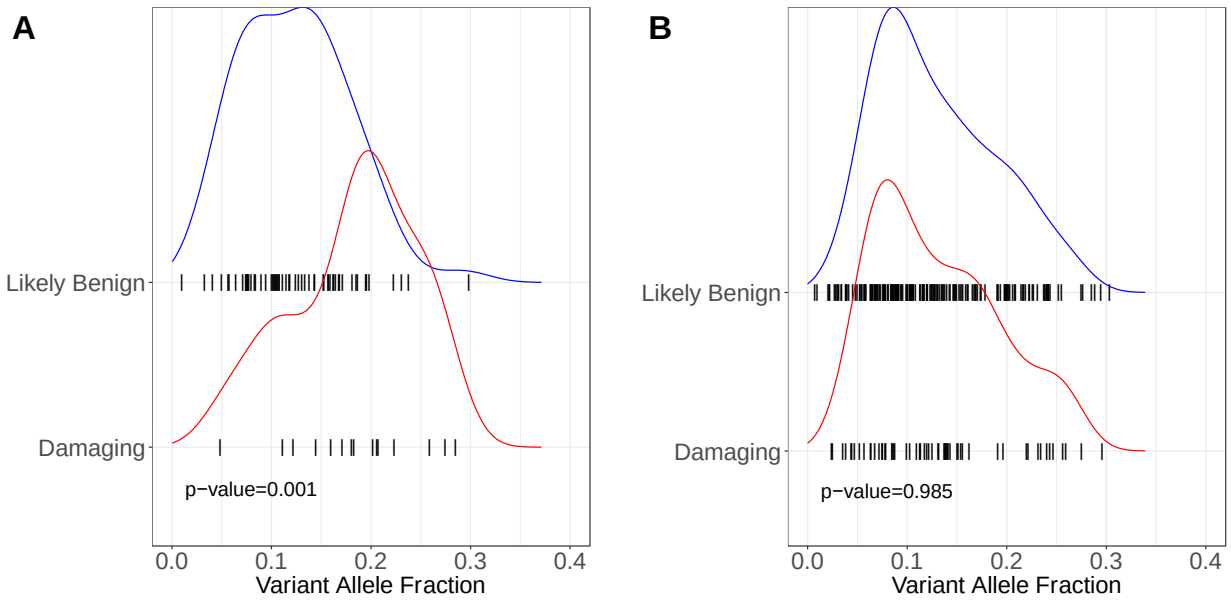

Figure 1.11. Damaging mosaics in CHD-related genes have higher variant allele fraction than likely-benign mosaics.

(A) Among the 76 mosaics in CHD-related genes, likely damaging variants have a higher VAF than likely benign (Mann-Whitney U p=0.001). (B) Among the 233 mosaics in Other (non-CHD-related) genes, there is no difference in VAF based on predicted effect (p=0.985).

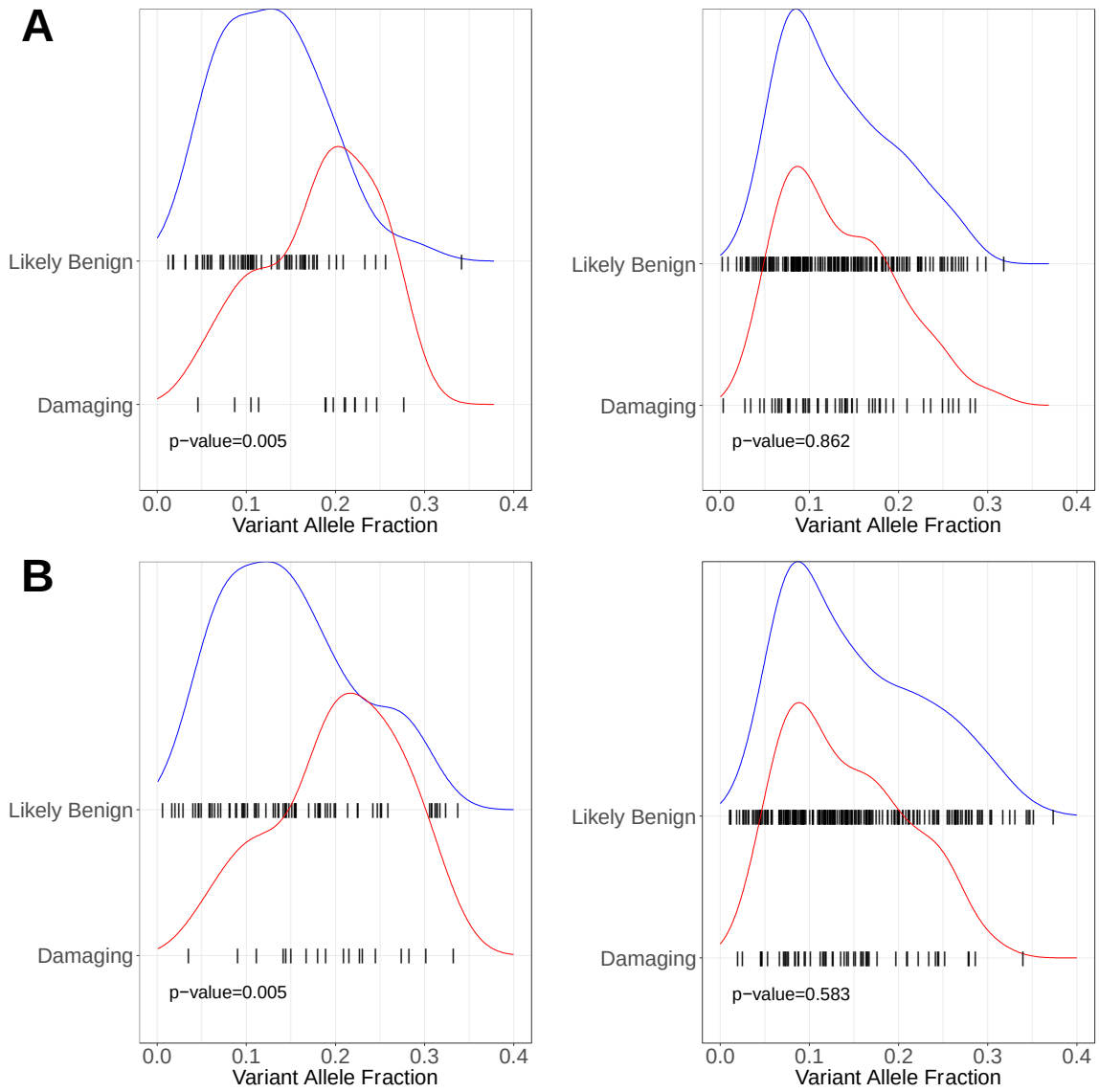

Figure 1.12. Damaging CHD-related mosaics have higher VAF under less stringent definitions of mosaicism.

(A) Using posterior odds cutoff of 5 (corresponding to 315 mosaics). Among 78 mosaics in CHD-related genes (left), there were 14 variants predicted as damaging, 63 variants predicted as likely-benign, and 1 variant of unknown functional consequence. Among 237 mosaics in non-CHD-related genes (right), there were 41 variants predicted as damaging, 184 variants predicted as likely-benign, and 2 variants of unknown functional consequence. (B) Using posterior odds cutoff of 2 (corresponding to 352 mosaics). Among 89 mosaics in CHD-related genes (left), there were 17 variants predicted as damaging, 71 variants predicted as likely-benign, and 1 variant of unknown functional consequence. Among 263 mosaics in non-CHD-related genes (right), there were 54 variants predicted as damaging, 206 variants predicted as likely-benign, and 3 variants of unknown functional consequence.

Table 1.3. Damaging Mosaics in CHD-relevant genes

*ASD, atrial septal defect; BAV, bicuspid aortic valve; Dmis, deleterious missense; episcore, haploinsufficiency score (percentile rank) {Han 2018}; Heart Exp, heart expression percentile rank; LoF, loss-of-function; pLI, probability of loss-of-function intolerance {gnomAD}; PCGC, Pediatric Cardiac Genomics Consortium; VAF, variant allele fraction; VSD, ventricular septal defect. *VAF refers to CHD tissue WES.*

ID	Gene	Gene Function	Variant Class	Blood VAF	gnomAD pLI	Episcore Percentile	Heart Expression Percentile	Clinical Phenotype			Additional WES denovos in Proband		PCGC denovo LoF/Dmis Variants in Mosaic Gene
								Age (y)	Cardiac Abnormalities	Extracardiac Abnormalities	LoF and Dmis	CNVs	
1-00761	FBN1	Extracellular matrix protein	Dmis	0.24	1.00	98	93	4.3	Mitral stenosis	dysmorphic features, subglottic stenosis, hypoplasic left mainstem bronchus, short stature	None	None	2
1-07004	TET3	DNA methylation	Dmis	0.16	1.00	7	87	10.3	Subaortic stenosis	None	None	None	0
1-05662	SETD2	Chromatin remodeling	LoF	0.13	1.00	99	85	0.8	Aortic coarctation, mitral valve hypoplasia	None	None	None	0
1-00344	UBR5	Ubiquitin ligase	splice-damaging	0.27	1.00	95	90	16	D-transposition of the great arteries, VSD, valvar and subvalvar pulmonary stenosis	None	None	None	0
1-03512	RFX3*	Transcription factor	LoF	0.09	1.00	100	46	0.4	Tetralogy of Fallot with pulmonary stenosis	None	None	None	0
1-06216	ITSN1	Endocytosis	Dmis	0.21	1.00	98	86	0.3	ASD	plagiocephaly, rib anomaly, single kidney, dysmorphic facial features	TYRP1	None	0
1-00363	QSER1	Endoplasmic reticulum stress response	Dmis	0.06	1.00	94	79	3.7	Tetralogy of Fallot with pulmonary stenosis, VSD	inguinal hernia	None	None	0
1-13185	PKD1	Cilia and transmembrane receptor	Dmis	0.10	1.00	87	84	0.8	VSD, partially anomalous pulmonary venous return	hemangioma	None	None	1
1-00192	GLYR1	Chromatin remodeling	Dmis	0.22	0.99	89	93	0.4	ASD, VSD, interrupted aortic arch, hypoplastic tricuspid valve, BAV	None	None	Dup 15:22062306 -23062355	0
1-04046	FZD5^	Receptor regulating Wnt-Frizzled pathway	Dmis	0.09	0.99	89	48	0.2	Tetralogy of Fallot with pulmonary stenosis, VSD	None	KRT13	None	0
1-06649	NOVA2	RNA processing	Dmis	0.15	0.95	75	56	0.6	Tetralogy of Fallot with pulmonary stenosis	None	None	None	0
1-05095	ISL1	Transcription factor	LoF	0.07	0.90	97	25	2.4	ASD	None	None	None	0
1-06677	KCTD10	Ion channel	Dmis	0.16	0.84	75	91	10.1	Aortic coarctation, pulmonary valve stenosis	dysmorphic facial features, hydrocephalus, pyloric stenosis, single kidney, imperforate/atretic anus	TRIP4	None	0
1-05447	HNRNPAB	RNA processing	Dmis	0.09	0.76	72	99	7.8	ASD, BAV, aortic coarctation	None	None	None	0
1-00021	QKI	RNA processing	LoF	0.13	0.76	94	97	0.5	Double outlet right ventricle, pulmonary stenosis, VSD	None	None	None	0
1-11871	FHOD3^	Actin regulation	Dmis	0.18	0.05	91	92	0.0	Tetralogy of Fallot with pulmonary atresia	hypocalcemia, thrombocytopenia, lymphopenia	None	None	0
1-01458	HK2	Glucose metabolism	Dmis	0.27	0.04	89	90	0.0	Hypoplastic left heart with aortic and mitral atresia, aortic coarctation	None	None	None	0
1-00669	PRKD3	Growth regulation	splice-damaging	0.19	0.02	77	82	0	D-transposition of the great arteries, conal VSD, bilateral conus, interrupted aortic arch	None	None	None	0
1-00524	RNF20^	Ubiquitin ligase	LoF	0.10	0.00	55	83	23.7	Left-dominant complete atrioventricular canal	Heterotaxy with situs inversus totalis, asplenia, duodenal atresia	None	None	0
1-01851	SUCLA2	Mitochondrial metabolism	LoF	0.11	0.00	72	89	15.5	Balanced complete atrioventricular canal, aortic coarctation	None	None	None	0
1-03885	LZTR1	Tumor suppressor	Dmis	0.20	0.00	31	84	14.7	Abnormal pulmonary vein draining into the right atrium	left sided/midline liver, asplenia, malrotation	None	None	2
1-05011	KCTD20	AKT regulation	Dmis	0.26	0.00	76	77	24.5	Transposition of the great arteries, Tricispid and Pulmonary valve atresia	left-sided/midline liver	KDM5B	None	0
1-00018	FIG4	Phosphoinositide regulation	Dmis	0.19	0.00	49	70	11.8	BAV, mitral atresia, aortic coarctation, VSD, total anomalous pulmonary venous return	nephritis	SBF1	None	0
1-05661	SMAD9	BMP signaling	Dmis	0.06	0.00	84	39	9.3	Common atrioventricular canal	None	None	None	0
1-09869	PIK3C2G*	Kinase	LoF	0.07*	0.00	73	28	6.3	Common atrioventricular canal, aortic stenosis, aortic arch hypoplasia, VSDs	dysmorphic facial features, low-set ears, campomelic dysplasia	None	None	1

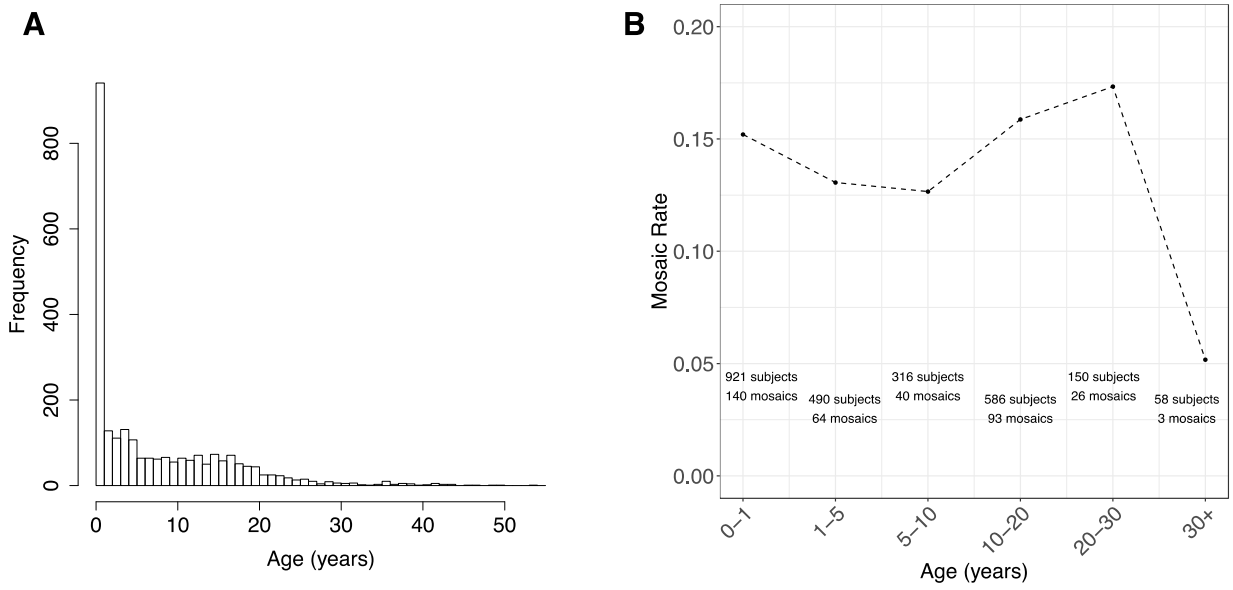

Figure 1.13. Mosaic rate by proband age.

(A) Age distribution for all 2530 probands in cohort. (B) Mosaic Rate across Age ranges. Rate = # mosaics/# probands in age bin. Note: 9/2530 probands missing Age information. 1/367 mosaic belong to a proband with missing Age.

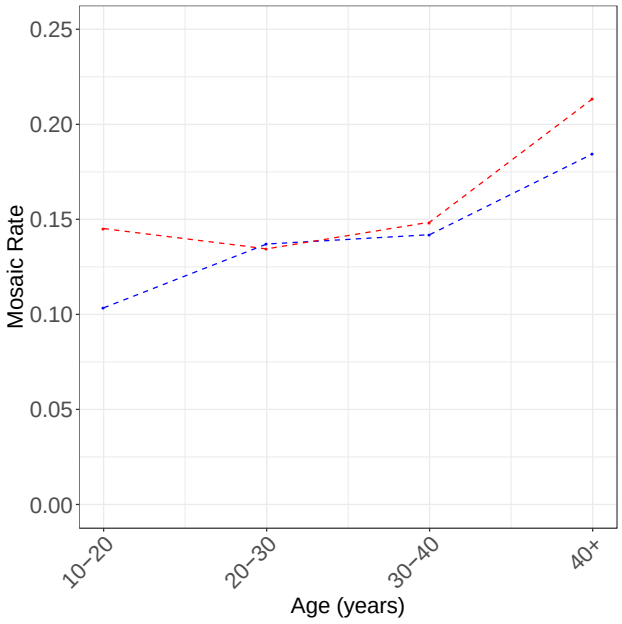

Figure 1.14. Mosaic rate by parental age at birth.

Mosaic rate by age of father (blue) and mother (red) at birth. Rate = # mosaics/# probands in each parental age bin. Note: 9/2530 probands missing age information. 1/367 mosaic belong to a proband with missing age.

1.3 Discussion

Distinguishing mosaic mutations from constitutional mutations has both clinical management and reproductive implications for proband and parents. Individuals with mosaic mutations are generally clinically less severely affected for conditions that affect multiple parts of the body {Happle 1986; Wallis 1990; Cohn 1990; Etheridge 2011; Donkervoort 2015; Weinstein 2016}. Mutations that occur post-zygotically should have no recurrence risk for the parents and could have a recurrence risk of less than 50% for the proband depending on gonadal involvement. This study is among the first investigations of the role of post-zygotic mosaic mutations in CHD. We developed a new computational method to robustly detect mosaic single nucleotide variants from blood WES data at standard read depth. Applying this method to a cohort of 2530 CHD patients, we detected 309 high-confidence mosaics (with a confirmation frequency of 88% in a subset of variants assessed) or 0.12 variants per proband. Sequencing of cardiac tissue to greater depth identified an additional 8 mosaic variants that had not been detected in blood WES, 6 of which are present in cardiac tissue but not blood. We found significantly more variants per proband in cardiac tissue DNA (0.23 variants per proband) than in blood DNA (0.12 variants per proband; p=0.02). While the increased numbers of mosaic variants in cardiac tissue DNA vs blood DNA may reflect technical differences such as sequencing read depth of cardiac tissue DNA vs blood DNA, it is possible that somatic variation occurs more frequently in cardiac tissue of CHD probands than in their blood. Whether or not there are more cardiac tissue mosaic variants in CHD probands than blood DNA variants, we found 10 mosaic variants among 66 CHD proband cardiac tissues with a higher VAF in tissue than in blood (**Table 1.2**) and 5 variants among these individuals with a higher VAF in blood than in tissue.

In total, we observed potentially CHD-causing mosaic mutations in 25 participants, representing 1% of the 2530 total participants in our CHD cohort. Among these 25 mosaics, we confirmed 22/22 (100%) candidates tested. We found that in CHD-related genes, likely-damaging mosaic mutations have significantly greater alternative allele fraction than likely-benign mosaics, suggesting that some of these variants contribute to CHD. Comparison of blood and cardiovascular tissues demonstrated tissue-specific mosaic variants, though those variants with a higher VAF were more likely to be shared between tissues. Due to limitations of conventional clinical interpretation for both mosaic and constitutional CHD variants (Materials and Methods), we cannot know with complete certainty which among these 25 variants is pathogenic and instead propose that, among our detected mosaics, the 23 detected from blood WES data provide an estimate of the disease-causing mosaics detectable in blood with standard exome-sequencing read depth. Nine of these variants affect genes known to have a role in cardiac development: *ISL1, SETD2, NOVA2, QKI, SMAD9, LZTR1, KCTD10, KCTD20, and FZD5*.

The mosaic LOF mutation in *ISL1* is likely to be the cause of CHD in participant 1-05095. ISL1 is a transcription factor essential to normal cardiac development that regulates expression of *NKX, GATA*, and *TBX* family genes {Golzio 2012; Colombo 2018} and controls secondary heart field differentiation and atrial septation {Colombo 2018; Briggs 2012}. *ISL1* deficiency has been shown to lead to severe CHD in mice {Cai 2003; Golzio 2012}. Participant 1-05095 has an isolated atrial septal defect consistent with a secondary heart field defect phenotype {Stevens 2010} and has no other previously reported damaging germline variants in CHD-related genes.

Damaging germline de novo variants in CHD subjects are enriched in genes related to chromatin modification and RNA processing {Homsy 2015; Jin 2017}. Three genes with damaging mosaic variants discovered here have related functions. *SETD2* is a histone methyltransferase required for embryonic vascular remodeling {Hu 2010}; it is both sensitive to haploinsufficiency and highly expressed in the heart during development. *NOVA2* is a key alternative-splicing regulator involved in angiogenesis that has been shown to disrupt vascular lumen formation when depleted {Giampietro 2015}. *QKI* encodes an RNA-binding protein that regulates splicing, RNA export from the nucleus, protein translation, and RNA stability {Lauriat 2008}. *QKI* is also highly expressed in the heart during development and has been shown to cause CHD and other blood vessel defects in mice when dysregulated {Noveroske 2002}. Other damaging mosaic variants affect processes known to be relevant to CHD. *SMAD9* is involved in the TGF-beta signaling pathway. TGF-beta signaling plays a critical role in cardiac development and cardiovascular physiology, leading to pulmonary arterial hypertension and cardiac abnormalities in mice when dysregulated {Drake 2015; Soubrier 2013}. *LZTR1* encodes a member of the BTB-Kelch superfamily that is highly expressed in the heart during development and has been associated with Noonan {Yamamoto 2015; Ghedira 2017} and DiGeorge Syndromes {Kurahashi 1995}, both of which are characterized by CHD. *KCTD10* binds to and represses the transcriptional activity of *TBX5* (T-box transcription factor), which plays a dose-dependent role in the formation of cardiac chambers {Tong 2014}. *KCTD10* is highly expressed in the heart during development and has been shown to produce CHD in mice when dysregulated {Ren 2014}. *KCTD20* is a positive regulator of Akt {Nawa 2013} also highly expressed in the heart during development. *FZD5* is haploinsufficient and encodes a

transmembrane receptor involved in Wnt, mTOR, and Hippo signaling pathways and has been shown to play a role in cardiac development {Dawson 2013}.

Finally, two mosaic variants found in cardiac tissue, genes encoding *RFX3* and *PIK3C2G*, may be disease-relevant. *PIK3C2G* is a signaling kinase involved in cell proliferation, survival, and migration, as well as oncogenic transformation and protein trafficking {OMIM: 609001; RefSeq}. The effects of *PIK3C2G* haploinsufficiency during cardiac development has not been characterized. *RFX3* is a highly-constrained ciliogenic transcription factor that leads to pronounced laterality defects {Rasmdell 2005} and disruption of *RFX3* leads to congenital heart malformations in mice {Lo 2011 MGI: 5560494}. Notably the *RFX3* LoF variant has a 4-fold higher VAF in cardiac tissue than in blood.

Several investigators who studied cancer and diseases with cutaneous manifestations proposed that the VAF correlates with time of mutation acquisition and disease burden {Belickova 2016; Sallman 2016; Happle 1986}. In this study, we used VAF as a proxy for cellular percentage and mutational timing, with increasing VAF corresponding to events occurring earlier in development. Thus, we assume that CHD-causal mosaic events identified in blood-derived DNA occurred during or shortly after the gastrulation process (3rd week of development) {Moorman 2003} in the mesodermal progenitor cells that differentiate into both heart precursor cells (cardiogenic mesoderm) and blood precursor cells (hemangioblasts). We found that in CHD-relevant genes, mosaic sites predicted to be damaging tended to have higher VAF than sites predicted to be likely benign, consistent with the hypothesis that these mutations arose early in fetal development and play significant roles in CHD. However, additional functional studies are necessary to fully assess causality.

Finally, we recognize that while our method is able to detect a large fraction of mosaic variants in blood, our calibrated estimates for the true number of mosaics suggest there are a non-negligible number of additional mutations that were not identified by our method. At our current average sequencing depth of 60x, we have limited sensitivity in the low VAF (<0.05) range. To reliably identify these low allelic fraction sites, ultra-deep sequencing will be critical to distinguishing true variants from noise. At 500x, we estimate detection sensitivity for mosaic events at VAF 0.05 to be above 80%. We also recognize age-related clonal hematopoiesis {Jaiswal 2014; Genovese 2014} as a potential confounding factor in somatic mutation detection; however, our study cohort includes mostly pediatric cases and we did not observe mosaic mutations in genes related to clonal expansion (e.g. *ASXL1*, *DNMT3A*, *TET2*, *JAK2*) nor did we observe a relationship between proband age and mosaic rate **(>Fig. 1.13, Fig. 1.14)**, suggesting minimal impact from this process.

This study is among the first investigations of the role of post-zygotic mosaic mutations in CHD. Despite limitations in sequencing depth and sample type, EM-mosaic was able to detect 309 high-confidence mosaics with resequencing confirmation in 88% of cases assessed. Using MosaicHunter, an additional 64 candidate mosaic sites were identified, of which 23/46 (50%) candidates from blood DNA and 4/6 (67%) from CHD tissue DNA validated. In total, we observed potentially CHD-causing mosaic mutations in 25 participants, representing 1% of our CHD cohort, and propose that these 25 cases provide an estimate of the disease-causing mosaics detectable in blood with standard exome-sequencing read depth. Additionally, we found that in CHD-related genes, likely-damaging mosaics have significantly greater alternative allele fraction than likely benign mosaics, suggesting that many of these variants cause CHD and occurred early in development. In the subset of our cohort for which cardiovascular tissue samples were

available, we show that mosaics detected in blood can also be found in the disease-relevant tissue and that, while the VAF for mosaic variants often differed between blood and cardiovascular tissue DNA, variants with higher VAF were more likely to be shared between tissues. Given current limitations in sequencing depth and on the availability of relevant tissues, particularly for conditions impacting internal organs like the heart, the full extent of the role of mosaicism in many diseases remains to be explored. However, as datasets containing larger numbers of blood and other tissue samples sequenced at higher depths become increasingly available, we will be able to more fully characterize the biological processes underlying post-zygotic mutation and, by extension, the contribution of mosaicism to disease using the methods presented here.

1.4 Materials and Methods

1.4.1 Samples and sequencing data

We analyzed WES data from 2530 Congenital Heart Disease (CHD) proband-parents trio families who were recruited as part of the Pediatric Cardiac Genomics Consortium (PCGC) study {Homsy 2015; Jin 2017}. Genomic DNA from venous blood or saliva was captured using Nimblegen v.2 exome capture reagent (Roche) or Nimblegen SeqCap EZ MedExome Target Enrichment Kit (Roche) followed by Illumina DNA sequencing (paired-end, 2x75bp) {Jin 2017, Zaidi 2013}. Genomic DNA from 70 surgically-discarded cardiovascular tissue samples (2-10mg) was isolated using DNeasy Blood & Tissue Kit (QIAgen), then captured using xGen Exome Research Panel v1.0 reagent (IDT) followed by Illumina DNA sequencing (paired-end, 2x75bp). Sequence reads were mapped to the hg19 human reference genome with BWA-MEM and BAM files were further processed following GATK Best Practices, which included duplication marking, indel realignment, and base quality recalibration steps. Blood and saliva

samples had sample average depth 60x and cardiovascular tissue samples had sample average depth 160x.

1.4.2 *De novo* variant calling and annotation

We processed our sample BAMs and called variants on a per-trio basis using SAMtools (v1.3.1-42) and BCFtools (v1.3.1-174). Pileups were generated using samtools 'mpileup' command with mapQ 20 and baseQ 13 to minimize the effect of poorly mapped reads on variant allele fraction, followed by bcftools 'call' using a cutoff of 1.1 for the posterior probability of the homozygous reference genotype parameter (-p) to capture additional sites with variant allele fraction suggestive of post-zygotic origin that would otherwise be excluded under the default threshold of 0.01. To identify de novo mutations from trio VCF files, we selected sites with (i) a minimum of 6 reads supporting the alternate allele in the proband and (ii) for both parents, a minimum depth of 10 reads and 0 alternate allele read support. Variants were then annotated using ANNOVAR (v2017-07-17) to include information from refGene, gnomAD (March 2017), 1000 Genomes (August 2015), ExAC, genomicSuperDups, CADD (v1.3) COSMIC (v70), and dbSNP (v147) databases, as well as pathogenicity predictions from a variety of established methods included as part of the dbNSFP (v3.0a) database or generated in-house (MCAP, REVEL, MVP, MPC). We used REVEL {Ionnidis 2016} to evaluate missense variant functional consequence, using the recommended threshold of 0.5 corresponding to sensitivity of 0.754 and specificity of 0.891. We used spliceAI {Jaganathan 2019} to predict the variant functional impact on splicing using the delta score thresholds of 0.2 for likely pathogenic (high recall), 0.5 for pathogenic (recommended), and 0.8 for pathogenic (high precision). We considered sites predicted to be Likely Gene-Disrupting (LOF) (stopgain, stoploss, frameshift indels, splice-site),

Deleterious Missense (Dmis; nonsynonymous SNV with REVEL>0.5), or splice-damaging (Benign Missense or synonymous SNV with delta score > 0.5) to be damaging and likely disease causing. We considered sites predicted to be Synonymous (delta score ≤ 0.5) or Benign missense (Bmis; nonsynonymous SNV with REVEL ≤ 0.5 and delta score ≤ 0.5) to be non-damaging.

1.4.3 Pre-processing and quality control

To reduce the number of low VAF technical artifacts introduced by our variant calling approach, we pre-processed our variants using a variety of filters. We first excluded indels from further analysis, as their downstream model parameter estimates were less stable than those of SNVs. We then filtered our variant call set for rare heterozygous coding mutations (Minor Allele Frequency (MAF) ≤ 10-4 across all populations represented in gnomAD and ExAC databases). To account for regions in the reference genome that are likely to affect read-depth estimates, we removed variant sites found in regions of non-unique mappability (score<1; 300bp), likely segmental duplication (score>0.95), and known low-complexity {Li 2014}. We then excluded sites located in MUC and HLA genes and imposed a maximum variant read depth threshold of 500. We used SAMtools PV4 to exclude sites with evidence of technical issues using a cutoff of 1e-3 for baseQ Bias and Tail Distance Bias and a cutoff of 1e-6 for mapQ Bias. To account for potential strand bias, we used an in-house script to flag sites that have either (1) 0 alternate allele read support on either the forward or reverse strand or (2) p<1e-3 and (Odds Ratio (OR)<0.33 or OR>3) when applying a two-sided Fisher's Exact Test to compare proportions of reference and alternate allele read counts on the forward and reverse strands. We also excluded sites with cohort frequency>1%, as well as sites belonging to outlier samples (with abnormally high de novo SNV (dnSNV) counts, cutoff = 8) and variant clusters (defined as sites with neighboring

SNVs within 10bp). Finally, we applied an FDR-based minimum N_{alt} filtering step **(>Fig 1.15)** to control for false positives caused purely by sequencing errors.

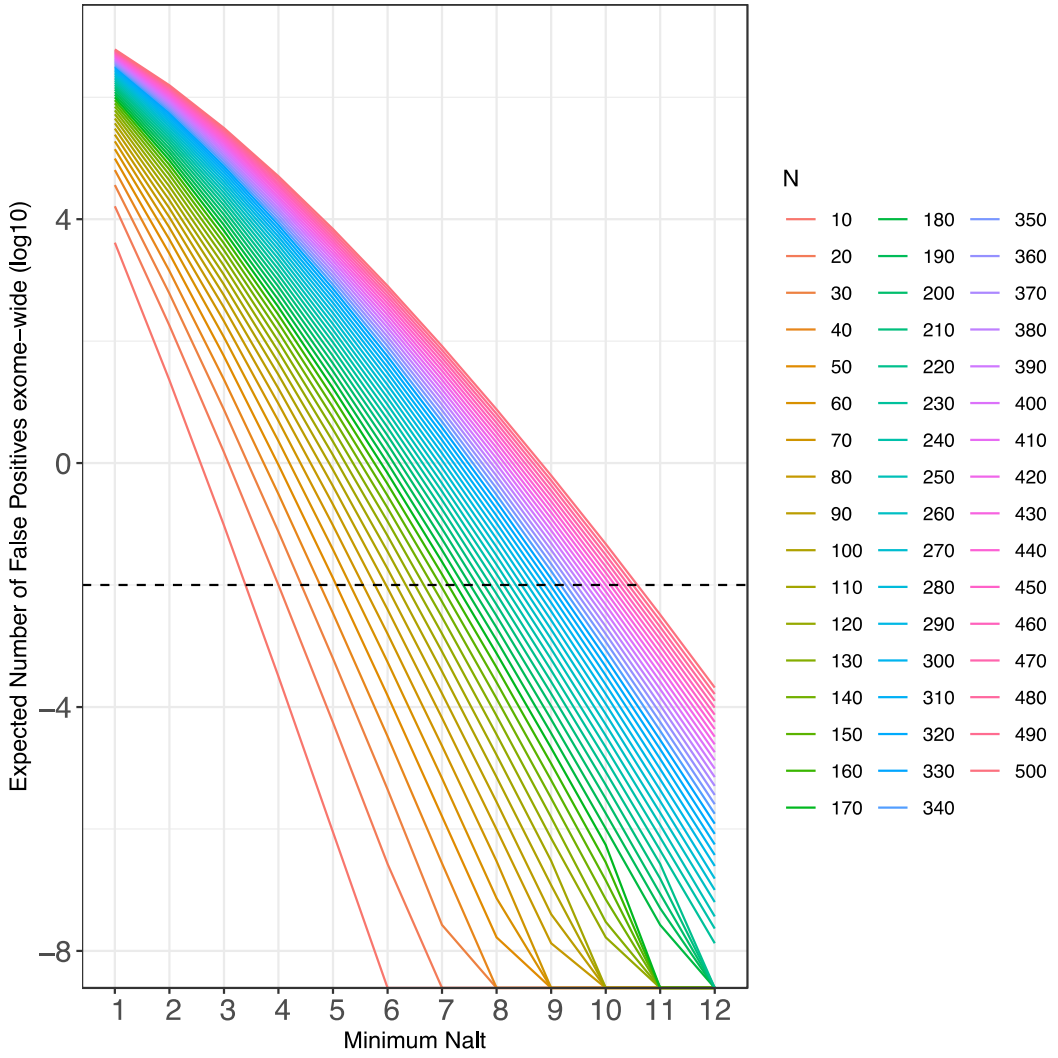

Figure 1.15. FDR-based minimum N_{alt} threshold.

An FDR-based approach was used to determine a threshold for the minimum number of reads supporting the alternate allele for each site to avoid false positives caused purely by sequencing errors. Assuming that sequencing errors are independent and that errors occur with probability 0.005, with the probability of an allele-specific error being 0.005/3=0.00167, and given the total number of reads (N) supporting a variant site, we iterated over a range of possible N_{alt} values between 1 and 0.5*N and estimated the expected number of false-positives due to sequencing error, exome-wide ((1- $f_{Poisson}(x=n, \lambda=N*0.005/3)$)*3x107 ; where $f_{Poisson}$ is the probability of x events in a Poisson process with mean λ). Assuming one coding de novo SNV per exome {Acuna-Hidalgo 2016} and that roughly 10% of de novo SNVs arise post-zygotically {Lim 2017; Krupp 2017; Freed 2016}, we used a conservative assumption of 0.1 mosaic mutation per exome. To constrain theoretical FDR to 10% we allowed a maximum of 0.01 false positives per exome and used the corresponding N_{alt} value to define an FDR-based minimum N_{alt} threshold for each variant. We then excluded variants with alternate allele read counts below this threshold.

1.4.4 IGV visualization of low allele fraction *de novo* SNVs

To reduce the impact of technical artifacts on model parameter estimation, we manually inspected de novo SNVs with VAF<0.3 (n=558) using Integrative Genomics Viewer (v2.3.97) to visualize the local read pileup at each variant across all members of a given trio family. We focused on the allele fraction range 0.0-0.3 since this range is enriched for technical artifacts that could potentially impact downstream parameter estimation. Sites were filtered out if (1) there are inconsistent mismatches in the reads supporting the mosaic allele, (2) the site overlaps or is adjacent to an indel, (3) the site has low MAPQ or is not Primary alignment, (4) there is evidence of technical bias (strand, read position, tail distance), or (5) the site is mainly supported by soft-clipped reads.

1.4.5 Expectation-Maximization to estimate prior mosaic fraction and control FDR

Current estimates for the fraction of de novo events occurring post-zygotically are unstable due to differences in study factors such as variant calling methods, average sequencing depth, and paternal ages. In order to use this fraction as a prior probability in our posterior odds and false discovery calculations, we reason that this value must be estimated from the data itself. We used an Expectation-Maximization algorithm to jointly estimate the fraction of mosaics among apparent de novo mutations and to calculate a per-site likelihood ratio score. This initial mosaic fraction estimate gives us a prior probability of mosaicism, independent of sequencing depth or variant caller, and allows us to calculate for each variant the posterior odds that a given site is mosaic rather than germline. To control for false discovery among our predicted mosaic candidates, we chose a posterior odds threshold of 10 to restrict FDR to 9.1%.

1.4.6 Mosaic mutation detection model

To distinguish variant sites that show evidence of mosaicism from germline heterozygous sites, we modeled the number of reads supporting the variant allele (N_{alt}) as a function of the total variant position read depth (N). In the typical case, N_{alt} follows a Binomial distribution with parameters N (site depth) and p (mean VAF). However, we observed notable overdispersion in the distribution of variant allele fraction compared to the expectations under this Binomial model **(>Fig. 1.16)**. To account for this overdispersion, we instead modeled N_{alt} using a Beta-Binomial distribution {Heinrich 2012; Ramu 2013}. We estimated an overdispersion parameter θ for our model as follows: for site depth values N in the range 1 to 500, we (1) bin variants by identifying all sites with depth N, (2) calculate a maximum-likelihood estimate θ value using N and all N_{alt} values observed for variants in a given bin, and (3) estimate a global θ value by taking the average of θ values across all bins, weighted by the number of variants in each bin. We then used θ in our Expectation-Maximization approach to jointly estimate prior mosaic fraction and to calculate per-site likelihood ratios.

To calculate the posterior odds that a given variant arose post-zygotically, we first calculated a likelihood ratio (LR) of two models: M_0: germline heterozygous variant, and M_1: mosaic variant. Under our null model M_0, we calculated the probability of observing N_{alt} from a Beta-Binomial distribution with site depth N, observed mean germline VAF p, and overdispersion parameter θ. Under our alternate model M_1, we calculated the probability of observing N_{alt} from a Beta-Binomial distribution with site depth N, observed site VAF $p=N_{alt}/N$, and overdispersion parameter θ. Finally, for each variant, we calculated LR by using the ratio of probabilities under each model and posterior odds by multiplying LR by our E-M estimated prior mosaic fraction estimate. We defined mosaic sites as those with posterior odds greater than 10

(corresponding to 9.1% FDR). We used posterior odds in this context to be able to control for false discovery, but we output similarly valid p-value and likelihood ratio scores for each de novo SNV.

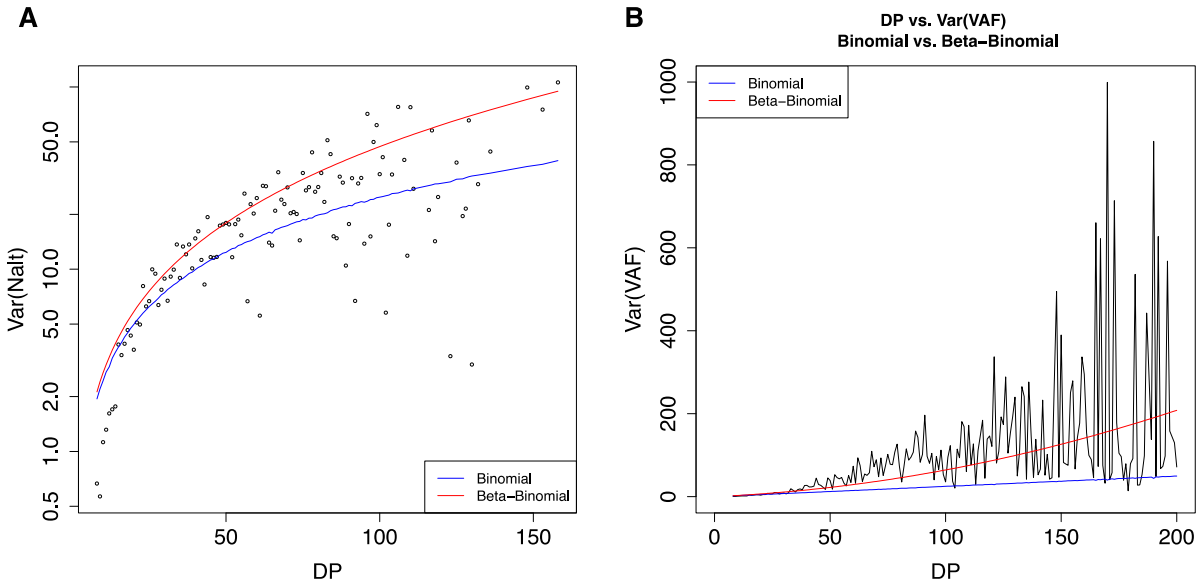

Figure 1.16. Overdispersion.

Overdispersion is commonly seen in WES data {Heinrich 2012; Ramu 2013} and is defined as observing variance (in terms of (A) N_{alt} or (B) VAF of variants with a given DP value) higher than expected across DP values, under a given statistical model. The blue line denotes the expectation under a Binomial model and the red line denotes the expectation under a Beta-Binomial model.

1.4.7 Simulation experiment

Variant datasets used in the simulation experiment were generated as follows:

For a given sample average sequencing depth value *S*,

1) Generate $n > 1{,}000{,}000$ VAF values where VAF ~ Beta(α=0.8, β=7) / 2
2) Generate *n* variant position read depth values (*N*) where *N* ~ NegativeBinomial(θ=4, mean=*S*)
3) Generate *n* variant alternate allele read depth values (N_{alt}) where N_{alt} = VAF * *N* and recalculate VAF = N_{alt} / *N*

4) Apply FDR-based minimum N_{alt} threshold (used to control false positives during variant calling), removing ~90% of variants and leaving ~100,000 mosaics
5) Apply the same procedure to generate 10 * *n* germline variants
6) Combine mosaic variants with germline variants to produce final dataset
7) Calculate true mosaic fraction

To evaluate method performance on each dataset, we first estimated the false discovery rate (FDR) for each variant as a function of posterior odds (1/(1+posterior odds)). Then, for FDR cutoffs $j = \{0, 0.01, \ldots, 0.99, 1.0\}$, we calculated both the $qvalue_j = \frac{\sum_1^N fdr_i}{N}$ as well as the False Discovery Proportion (FDP_j ; the fraction of variants with a ground truth label of "germline") using the *N* variants with FDR < *j* before comparing the results. Results are shown in **Fig. 1.17.**

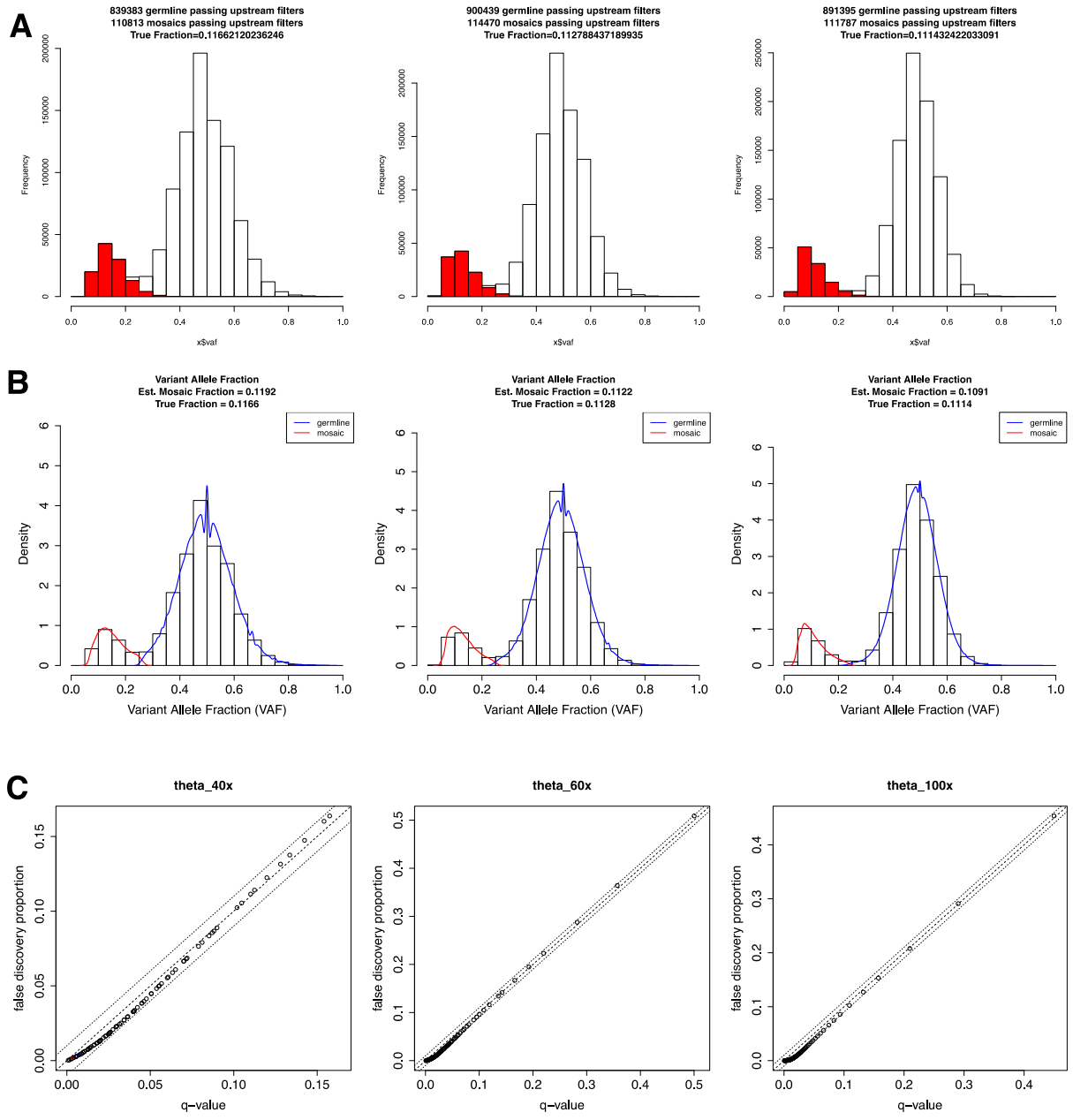

Figure 1.17. Simulation experiment results for 40x, 60x, 100x.

Panel order from left to right correspond to 40x, 60x, 100x, respectively. (A) Simulated variant datasets with a known true fraction of spiked in mosaic variants. (B) EM-estimated mosaic fraction compared to true mosaic fraction. (C) q-value vs. FDP_{truth} plots.

1.4.8 Mutation confirmation by MiSeq amplicon sequencing

Chromosome coordinates were expanded 500 bp upstream and downstream of the candidate mosaic variants in the UCSC Genome Browser. Primer 3 Plus software was used to

design forward and reverse primers to generate 150-300 bp amplimers containing the candidate site. PCR reactions consisting of genomic DNA, primers, and Phusion polymerase were amplified by thermal cycling and purified with AMPure XP beads. The purified PCR product was quantified, and 0.5-1.0 ng of product was used to construct Nextera XT libraries according to the protocol published by Illumina. Libraries were purified using AMPure XP beads, and final libraries were quantified and pooled to undergo sequencing through Illumina MiSeq.

We experimentally tested for the presence our predicted post-zygotic sites in the original blood DNA and cardiovascular tissue DNA samples using Illumina MiSeq Amplicon sequencing. The Amplicon Deep Sequencing workflow, optimized for the detection of somatic mutations in tumor samples, offers ultra-high sequencing depth (>1000x) that gives us the resolution to confirm low VAF variants, accurately estimate site VAF, and to distinguish true variant calls from technical artifacts. Mosaic candidates were considered validated if the variant allele matched the MiSeq call and both the mosaic VAF and MiSeq VAF indicated post-zygotic origin (VAF<0.45).

Mosaic candidates were selected for confirmation on the basis of VAF, plausible involvement in CHD (based on predicted pathogenicity and HHE status), and detection method (Table S11; Table S12). We sampled mosaics from both ends of the VAF spectrum to evaluate our ability to distinguish high VAF mosaics (VAF>0.2; n=29) from germline variants and to distinguish low VAF mosaics (VAF<=0.1; n=52) from technical artifacts. Confirmation rate across different VAF bins is shown in >**Figure 1.18**. We also selected for confirmation mosaics detected uniquely by either EM-mosaic or MosaicHunter, for the sake of method comparison (Table 1).

To examine a potential source of bias in our candidate selection process, we compared the posterior odds distribution of selected candidate mosaics (n=97) against those not chosen (n=212). We found that our tested candidates had lower posterior odds than untested mosaics ($mean_{tested}$=5.382, $mean_{untested}$=7.050, log10-scale; Mann Whitney U P=0.002) **(>Fig 1.19),** suggesting that our validation rate is not buoyed by testing variants with the strongest evidence of mosaicism. For method development purposes, we intentionally focused on mosaics with lower posterior odds as these variants fall in the VAF range for which it is most difficult to distinguish germline from mosaic.

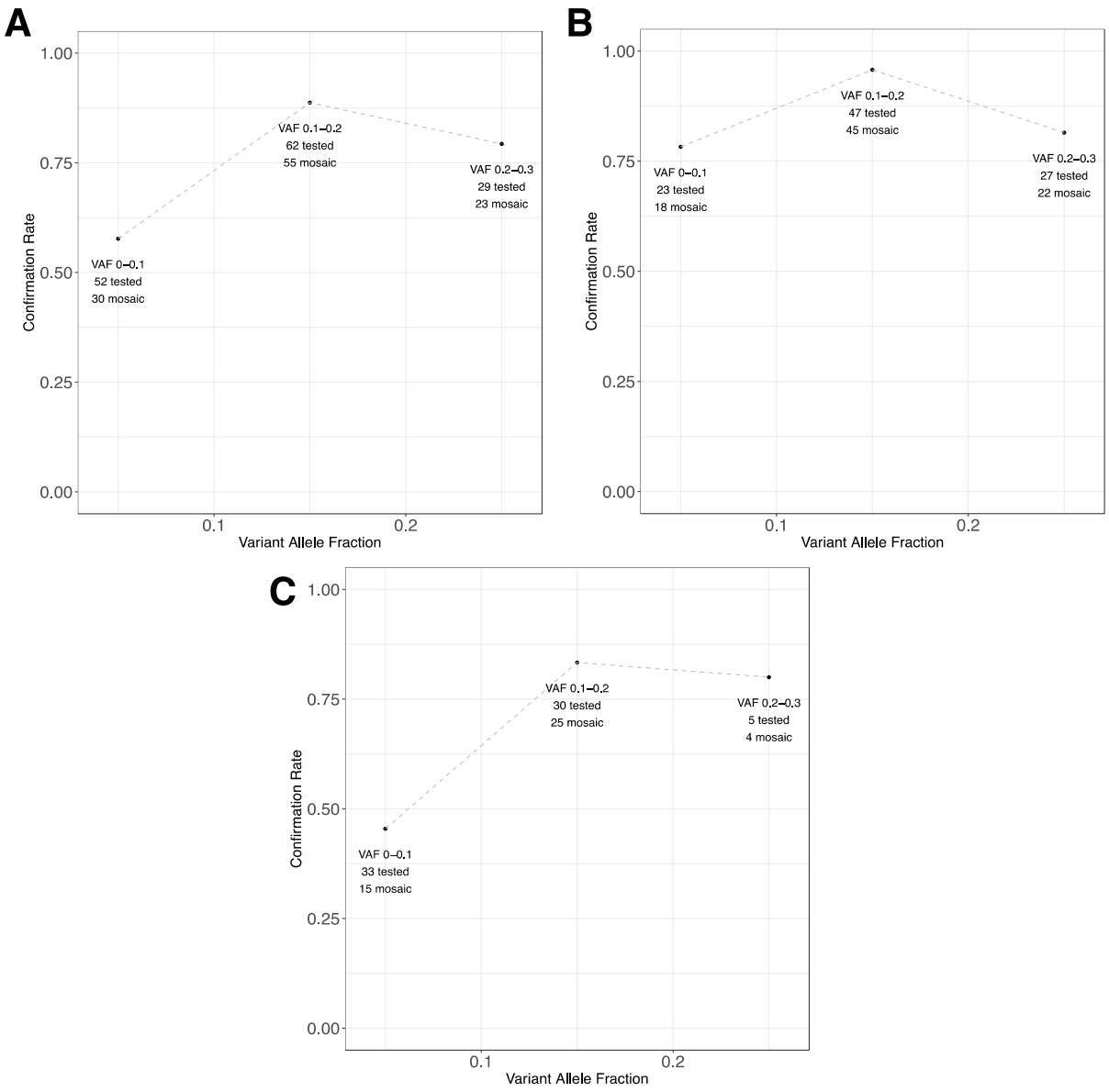

Figure 1.18. Confirmation rate across VAF bins.

The number of candidates for which we performed MiSeq resequencing among (A) the union set (n=143 tested) (B) all EM-mosaic calls (n=97) and (C) all MosaicHunter (n=68) calls vs. the number confirmed as mosaic for VAF ranges [0, 0.1), [0.1, 0.2), and [0.2, 0.3).

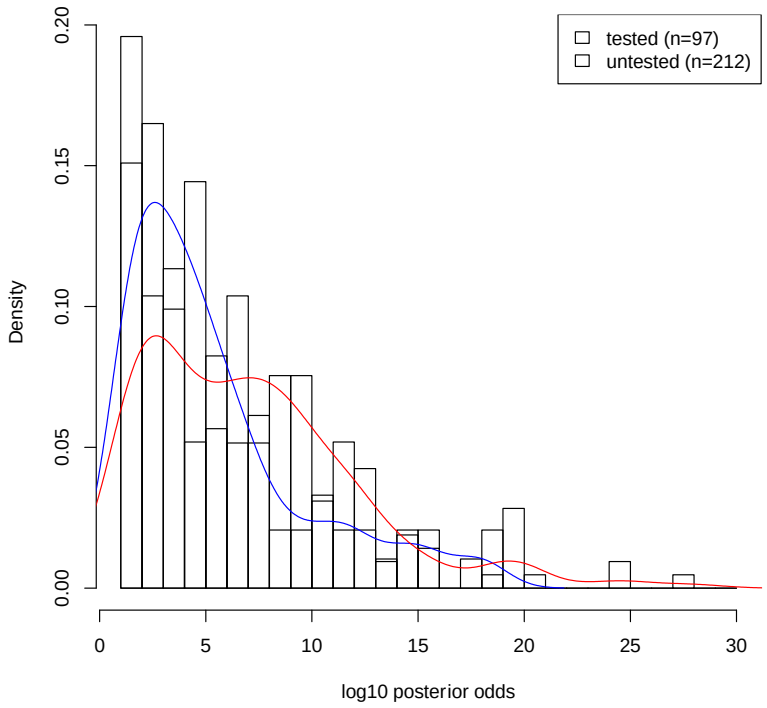

Figure 1.19. Posterior odds comparison for tested vs. untested mosaics.

Among 309 candidates with EM-mosaic posterior odds scores available, we compared the distribution of tested (n=97) vs. untested (n=212) mosaics. The log10-scaled posterior odds distribution for the tested group is shown in blue (mean=5.382). The log10-scaled mean posterior odds for the untested group is shown in red

1.4.9 Investigating the relationship between VAF and pathogenicity

We hypothesized that mosaic contribution to disease is positively correlated with cellular percentage and by extension mutational timing. Here, we used variant allele fraction as a proxy for cellular percentage. We grouped mosaics into likely-damaging and likely-benign and compared the distribution of allele fraction in CHD-related genes. We defined likely-damaging variants as: (a) likely gene-disrupting (LOF) variants (including premature stop-gain, frameshifting, and variants located in canonical splice sites), (b) missense variants predicted to be damaging by REVEL {Ioannidis 2016} (with score ≥ 0.5) or (c) missense variants and synonymous predicted to be splice-damaging by spliceAI (with score > 0.5). One of the main

findings from previous CHD studies is that damaging de novo variants in genes highly expressed in the developing heart ("HHE", ranked in the top 25% by cardiac expression data in mouse at E14.5 {Zaidi 2013; Homsy 2015}) contribute to non-isolated CHD cases that have additional congenital anomalies or neurodevelopmental disorders. Therefore, we considered the union of HHE genes and known candidate CHD genes {Jin 2017} as CHD-related genes (n=4558). For mosaics in CHD-related genes and for mosaics in other genes, we used a Mann-Whitney U Test to compare the VAF distributions of likely-damaging and likely-benign groups.

1.4.10 Estimated contribution of mosaicism to CHD

We identified likely disease-causing mosaic mutations on the basis of predicted pathogenicity and presence in genes involved in biological processes relevant to CHD or developmental disorders. Each mosaic mutation was annotated with gene-specific information, including heart expression percentile, probability of loss-of-function intolerance (pLI) score {Lek 2016}, whether dysregulation causes CHD in mice {Smith 2018; Finger 2017}, and gene function {NCBI RefSeq}. We focused on HHE genes, genes with high pLI (pLI>0.9), genes that cause CHD phenotypes in mice, and genes involved in key developmental processes such as Wnt, mTOR, and TGF-beta signaling pathways. Then, for each patient, we used the clinical phenotype to further prioritize mosaic mutations most likely contributing to that individual's clinical features. Detailed mutation annotation and clinical phenotypes for the mosaic carriers described above can be found in **Table S10**. We estimate the contribution of mosaicism to CHD as the percentage of individuals carrying likely disease-causing mosaic mutations among all individuals in our CHD cohort.

1.4.11 Union with validated *de novo* SNVs from Jin et al. *Nature Genetics* 2017

As part of the PCGC program, Jin et al. previously sequenced and processed a cohort of 2871 CHD probands – including 2530 parent-offspring trios used in this study – to investigate the contribution of rare inherited and de novo variants to CHD. They called a total of 2992 proband de novo variants, including 2872 SNVs and 118 indels, and Sanger confirmed a subset of the most likely-disease causing variants. Since we processed the same proband-parent trios using different variant calling pipelines, we combined the results of our two approaches to provide a more complete input de novo call set for mosaic variant detection.

We first processed our SAMtools de novo calls using our upstream filters (n=2396 sites passing all filters). We then applied the same upstream filters to the published dnSNVs from Jin et al. (n=2650 sites passing all filters) before finally taking the union of these two call sets (n=3192). There were 1814 sites in the intersection, with 836 sites unique to the Jin et al. calls and 542 sites unique to our SAMtools calls. After preprocessing, outlier removal, and FDR-based minimum Nalt filtering, the remaining 2971 dnSNVs were used as input to our mosaic detection model.

1.4.12 Mutation spectrum analysis

We compared the mutation spectrum – the frequencies of all possible base changes – of our predicted mosaic candidates against the spectrum of our predicted germline heterozygous variants. Under the assumption that that post-zygotic events occur randomly (i.e. due to errors in DNA replication rather than a specific biological process), the mosaic mutation spectrum should not differ significantly from the germline mutation spectrum. We used Pearson's Chi-square Test to test for a difference in frequencies across all base changes between our predicted sets of

variants. We interpreted large qualitative differences in base change frequencies as evidence of technical artifacts and rejection of the Chi-square null as evidence of systemic issues in our pipeline.

1.4.13 Mosaic detection power given sample average coverage

To model statistical power in the context of mosaic variant detection, we considered two conditional probabilities: (i) the probability of detecting a mosaic event (i.e. the probability of a variant's posterior odds exceeding a threshold) given site depth N, VAF, and overdispersion parameter θ and (ii) the probability of observing site depth N, given sample-wide average coverage DP_{sample}.

(i) Pr(detect mosaic | N, VAF, θ) was calculated by first identifying the VAF range (and by extension, the range of N_{alt}) over which posterior odds > cutoff, then by integrating the Beta-Binomial probability mass function over this range, with considerations for the probability of strand bias (P(strand bias | N) ~ Binomial(N_{alt}, N, p=0.5)).

(ii) Pr(observe N | DP_{sample}) follows an overdispersed Poisson distribution that we approximated using a negative binomial model with overdispersion parameter θ {Sampson 2011}. For each N value, we calculated a vector of weights corresponding to Pr(N | DP_{sample}) for N values in the range (1, 1500).

Finally, we took the sum of the detection probabilities described in (i) multiplied by the weights described in (ii) to determine the probability of detecting a mosaic variant given a sample average coverage value – Pr(detect mosaic | N). Our estimated detection power curves for a range of sample average coverage values typical of exome-sequencing studies are shown in

(>Fig. 1.2C). Our CHD cohort was sequenced to sample average depth of 60x, with prior mosaic fraction=0.121 and estimated θ=116.

To estimate the true rate of mosaicism per exome given sample average coverage, we first split our set of predicted mosaics into VAF bins of size 0.05. For each bin above VAF 0.1, we multiplied the number of mosaics by the inverse of the detection power for that given VAF bin to estimate the true count of mosaic variants in that VAF range, assuming full detection power. Since EM-mosaic is underpowered to detect mosaics with VAF < 0.1 in the blood and since this range is enriched for technical artifacts that potentially affect our counts, we did not apply this scaling procedure to these bins to avoid over-inflating our adjusted mosaic rate estimate **(>Fig. 1.7A).**

1.4.14 Filtering of MosaicHunter candidate variants

MosaicHunter was used to identify candidate mosaic variants from blood exome-sequencing trio data using default settings {Huang 2014}. Filtering of original MosaicHunter candidate variants excluded, in order, any variant present in ExAC (n=46634), G to T mutations with fewer than N_{alt}<10 oxidative indicating DNA damage {Costello 2013} (n=3995), non-uniquely called sites (n=4719), germline SNVs previously called by GATK HaplotypeCaller (n=591), probands with >20 mosaic variants (n=1490 in 10 probands), mosaic log posterior likelihood ratio <10 (n=940), variants with >2 parental alternative allele reads (n=244), variants with gnomAD population frequency > 1e-4 or located in MUC or HLA genes (n=40).

1.4.15 Filtering of MosaicHunter-detected cardiovascular tissue candidate variants

We used the MosaicHunter pipeline in trio mode to identify candidate variants in WES data from 70 cardiovascular tissue samples (belonging to 66 unique probands). From the list of variants initially reported by the pipeline using default settings, we applied the same filtration steps listed for MosaicHunter candidate variants in blood samples with the exception of the removal of G to T mutations with fewer than 10 alternative allele reads and the mosaic log posterior likelihood ratio <10. Finally, we removed variants that were identified in either parent or had a total read depth <10 in either parent.

1.4.16 Clinical interpretation of mosaic variants – limitations

We note that conventional clinical interpretation of mosaic mutations is challenging for several reasons: (i) it is unclear in which tissues each mosaic mutation is expressed (ii) several study participants were very young at time of clinical assessment and many classical disease features may not yet have developed or been noted, and (iii) the absence of additional clinical features does not necessarily rule out a mosaic mutation as being for the cause of the CHD. For the purposes of this study, we selected these mosaic mutations on the basis of predicted pathogenicity and detection in genes involved in biological processes relevant to CHD or developmental disorders

Chapter 2: Genetic factors associated with clinical outcomes in CHD

In this section, I discuss the development of an analytical framework to investigate the association between rare genetic variation and clinical outcomes in congenital heart disease patients (CHD). Rare *de novo*, transmitted, and copy number variants were called from a cohort of 3966 CHD proband-parent trios. We show that damaging *de novo* variants (DNVs) are associated with neurodevelopmental disorders (NDD) in CHD patients and that the enrichment is stronger when focusing on variants in genes highly expressed in developing hearts (HHE), known NDD risk genes, and genes that are both HHE and NDD-risk. The prevalence of NDD is higher in CHD patients carrying likely pathogenic (LP) variants than in cases that do not carrying LP variants and the difference increases when focusing on variants in the gene sets of interest above. Despite comprising roughly half of NDD-risk genes and only 5% of HHE genes, the genes that are annotated as both HHE and NDD-risk appear to drive the majority of the association and suggest that disruptive mutations in these genes have pleiotropic effects that likely play a role in the acquisition of NDD in our CHD patients. We next focused on CHD patients diagnosed with single ventricle defects and found that damaging DNVs are enriched in patients with abnormal ventricular function phenotypes (decreased systemic ventricular function, worsening ventricular function, arrhythmia). The enrichment is increases when considering variants in HHE genes, constrained genes (pLI>0.5), and genes that are both HHE and

constrained. The prevalence of abnormal phenotypes is higher in CHD patients carrying likely pathogenic (LP) variants than in cases that do not carrying LP variants and the difference again increases when focusing on variants in the gene sets of interest above. Genes that are annotated as both HHE and constrained comprise 57% of HHE genes and 42% of constrained genes and drive the majority of the association, suggesting pleiotropic effects of disruptive mutations in these genes. Finally, we created a proof-of-concept rare variant risk score model to predict NDD on a per-patient basis by combining counts of rare *de novo,* transmitted, and copy number variants with weights defined by the strength of association with NDD for each particular gene set. Our risk score achieved a 10-fold cross validated AUPRC of 0.44 when applied to all cases and AUPRCs of 0.32, 0.53, 0.46 when applied to cases with Isolated, Complex, and Unknown CHD subtypes, respectively. We found that prevalence of NDD increased as function of risk score percentile and that patients with scores in the top 25% were 3.71-fold as likely to have NDD than patients in the bottom 25%, demonstrating that our score is able to stratify patients in a clinically meaningful way and identify patients at increased risk of NDD.

2.1 Introduction

Congenital heart disease (CHD) patients often acquire cardiac and non-cardiac comorbidities that impact quality of life, such as arrhythmias, myocardial dysfunctions, and neurodevelopmental disorders (NDDs) {Marino 2012; Calderon 2014; Miller 2005; Burnham 2010}. While these clinical outcomes have been associated with a variety of fetal developmental, surgical/post-operative, and genetic factors {Marelli 2016}, thus far none have been identified as the primary contributor {Zaidi 2017} and early identification of patients at risk for these poor outcomes remains a challenge.

CHD patients, especially those with single ventricle defects such as hypoplastic left heart syndrome or tricuspid atresia, often experience a range of poor outcomes following surgery that last into adulthood, including impaired systemic ventricular function, arrhythmias, and neurodevelopmental disorders {Feinstein 2012}. Arrhythmias in particular may surface later in life and in conjunction with co-existing hemodynamic alterations are a common cause of mortality in adult congenital heart disease patients {Kairy 2006; Kanter 1997}. Identification of CHD patients most at risk of developing these poor cardiac outcomes creates opportunities for improved care strategies and earlier therapeutic intervention.

Among non-cardiac comorbidities, neurodevelopmental disorders (NDD) affect a disproportionately high number of CHD patients and have the largest impact on quality of life {Zaidi 2017}. Neurodevelopmental disorders describe a spectrum of conditions including but not limited to intellectual disability, autism spectrum, and other cognitive, motor, social, and language deficits {Homsy 2015}. Risk of acquiring NDD in CHD patients is a function of CHD severity/complexity and prevalence estimates range from 10% to over 50%, compared to 4-6% in the general population {Zaidi 2017; Homsy 2015; Marino 2012; Dixon-Salazar 2012}. Both CHD and NDD impair reproductive fitness and tend to occur sporadically in individuals with no prior family history, pointing to strong contribution from *de novo* genetic variation. Recent large-scale genetic studies in CHD {Zaidi 2013; Homsy 2015} and NDD cohorts {De Rubeis 2014} have implicated disruption of chromatin modifying genes in both conditions, suggesting shared genetic etiology. Further, CHD patients presenting with extracardiac anomalies, NDD, or both were found to be more likely to carry damaging *de novo* variants {Homsy 2015}. Given disease heterogeneity and difficulty in resolving diagnosis criteria in infants, predicting risk of

acquiring NDD in CHD patients using genetic information presents an exciting clinical opportunity.

Genetic disease risk is a combination of rare variants of large effect size and common variants of small effect sizes (and to varying degrees, environmental factors). The Common Disease/Common Variants hypothesis {Reich 2001} posits that common variants drive risk of common disease and, by extension, that rare variants drive risk of rare disease. Identification of individuals at high risk of acquiring specific diseases enables earlier therapeutic intervention and changes in patient management {Khera 2018}. Polygenic risk scores (PRS) aim to stratify patients and identify individuals at clinically significant increased risk by integrating the contribution of a large number of loci genome-wide. Introduced in 2010 {Ripatti 2010}, PRS have begun to gain traction as patient study cohorts have dramatically increased in size. Using data from ~500,000 participants recruited as part of the recently released UK Biobank database, Khera and Kathiresan et al. developed a PRS that was able to identify 57,115 (19.8%) of participants in their testing dataset (n=288,978) at >3-fold risk of coronary artery disease, atrial fibrillation, type 2 diabetes, inflammatory bowel disease, and breast cancer {Khera 2018}. Scores have also been developed for a range of other common traits including but not limited to height, body mass index, and total cholesterol {Chatterjee 2013}, though the authors note that clinical utility of PRS depends on factors such as association study sample size and genetic architecture. Given that rare genetic variation typically contributes larger effect size, the field has also seen the development of a genome-wide rare variant risk score for schizophrenia {Purcell 2014} and a genome-wide de novo risk score for autism spectrum disorder {An 2018}, highlighting considerable interest in patient risk stratification on the basis of genetic variation.

Currently no risk score exists for predicting NDD risk in CHD, which we believe is an unmet need with significant clinical utility.

2.2 Results

2.2.1 Complex CHD cases are more likely to acquire NDD than Isolated CHD cases

Our cohort of 3966 CHD cases was annotated with information about extracardiac anomalies and NDD diagnosis and stratified along these two axes (>**Table 2.1**). Here, Complex cases were defined as having at least one other extracardiac anomaly (e.g. Skeletal, Craniofacial, Genitourinary, etc), Isolated cases were defined as having no extracardiac anomalies, with Unknown indicating cases where this information was not available. NDD cases were defined as having received services for cognitive, motor, social, or language impairments, non-NDD defined as not having received the services above, with Unknown NDD describing patients with unclear diagnosis (typically patients <1 year old at evaluation). Recent genetic and clinical studies have established a relationship between increased CHD complexity and increased prevalence of NDD {Homsy 2015; Marino 2012}. As a sanity check, we compared the relative numbers of NDD and non-NDD cases among our CHD patients with Complex and Isolated presentations to see if we could reproduce this finding. Among 565 patients with Complex CHD and a definitive NDD diagnosis, 240 were annotated as having NDD and 325 were non-NDD cases. Among 1175 patients with Isolated CHD, 305 were annotated as having NDD and 870 were non-NDD cases. We found that cases with Complex CHD were more likely to acquire NDD than those with Isolated CHD (OR=2.1, p=6e-12, Fisher's Exact Test).

Table 2.1. Complete CHD cohort.
NDD=neurodevelopmental disorder.

PCGC cases	All	NDD*	Non-NDD	Unknown NDD**
All	3966	652	1588	1726
Isolated	1803	305	870	628
Complex	996	240	325	431
Unknown	1167	107	393	667

2.2.2 Damaging *de novo* variants are associated with NDD

We called a total of 5271 *de novo* variants (DNVs) from our cohort of 3966 CHD cases. We then compared counts of variants per individual across different classes of functional consequence (>**Table 2.2**). Excluding cases with an unknown NDD diagnosis, we found that likely gene-disrupting (LGD) DNVs are enriched in cases with NDD (Relative Risk (RR)=1.59, p=3e-05, Binomial Test). This enrichment is stronger when considering only LGD DNVs located in genes highly expressed in developing heart (HHE; RR=2.38, p=8e-08) and genes that are known NDD risk genes (RR=8.43, p=3e-14). Interestingly, the signal further increases when focusing on the variants located in genes that are at the intersection of HHE and NDD-risk genes (RR=9.26, p=1e-12). We see a similar trend when grouping LGD and Dmis variants **(>Fig. 2.1**). We observed a depletion of synonymous DNVs in NDD cases (RR=0.87, p=0.12) and we believe this to be due to technical differences (average depth and uniformity between batches of

patients sequenced at different times with different capture kits) rather than biological differences (**>Fig. 2.7**).

We next compared the prevalence of cases with NDD and the prevalence of cases without NDD among patients carrying likely pathogenic (LP; LGD + Dmis) DNVs and those that do not. We found that the prevalence of NDD is higher among patients carrying LP DNVs (21% NDD_{LP} vs. 14% NDD_{nonLP}) and that the NDD prevalence increases when considering only LP variants in HHE genes (26% NDD), NDD risk genes (41% NDD), and genes that are both HHE and NDD-risk (45%) **(>Fig. 2.2, Table 2.3**). The difference in prevalence (NDD_{LP} – NDD_{nonLP}) also follows the same trend, with difference values of 7%, 11%, 26%, and 29% for All, HHE, NDD-risk, and HHE&NDD-risk gene sets, respectively.

Table 2.2. Rates of LGD, Dmis DNVs across different gene groups.

De novo	Gene Set	Variant class	Rate in NDD	Rate in non-NDD	PAR (Δ_{rate})	Relative Risk	P-value
All Cases **#NDD = 652** **#non = 1588** **#unk = 1726**	All	LGD	0.21	0.14	0.08	1.59	3.1E-05
		Dmis	0.23	0.17	0.06	1.35	0.004
	HHE	LGD	0.12	0.05	0.07	2.38	7.9E-08
		Dmis	0.11	0.06	0.05	1.73	5.0E-04
	NDD-risk	LGD	0.07	0.01	0.06	8.43	3.1E-14
		Dmis	0.04	0.02	0.02	2.34	0.003
	NDD-risk & HHE	LGD	0.06	0.01	0.05	9.26	1.0E-12
		Dmis	0.03	0.01	0.02	3.37	0.001

LGD=likely-gene-disrupting, Dmis=deleterious missense, NDD=neurodevelopmental disorder, non=non-NDD, unk=unknown NDD, HHE=high heart expression genes, NDD-risk=known NDD risk genes.

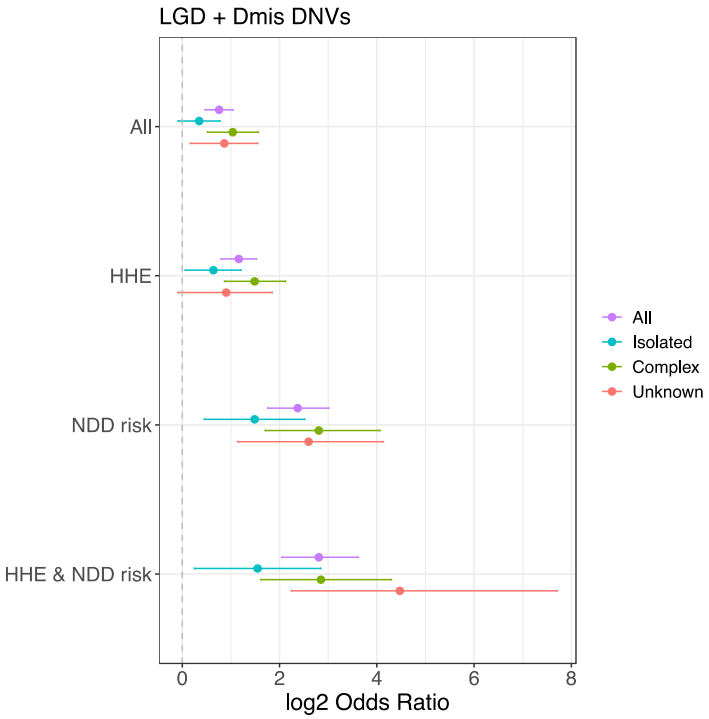

Figure 20. Damaging (LGD + Dmis) DNVs are enriched across gene sets.

We observe an enrichment across All genes, HHE, and NDD-risk genes across all CHD subtypes (Isolated, Complex, Unknown). The strongest enrichment is observed in genes that are annotated as both HHE and NDD-risk. HHE=high heart expression, NDD=neurodevelopmental disorder, NDD risk=known NDD risk genes, CHD=congenital heart disease.

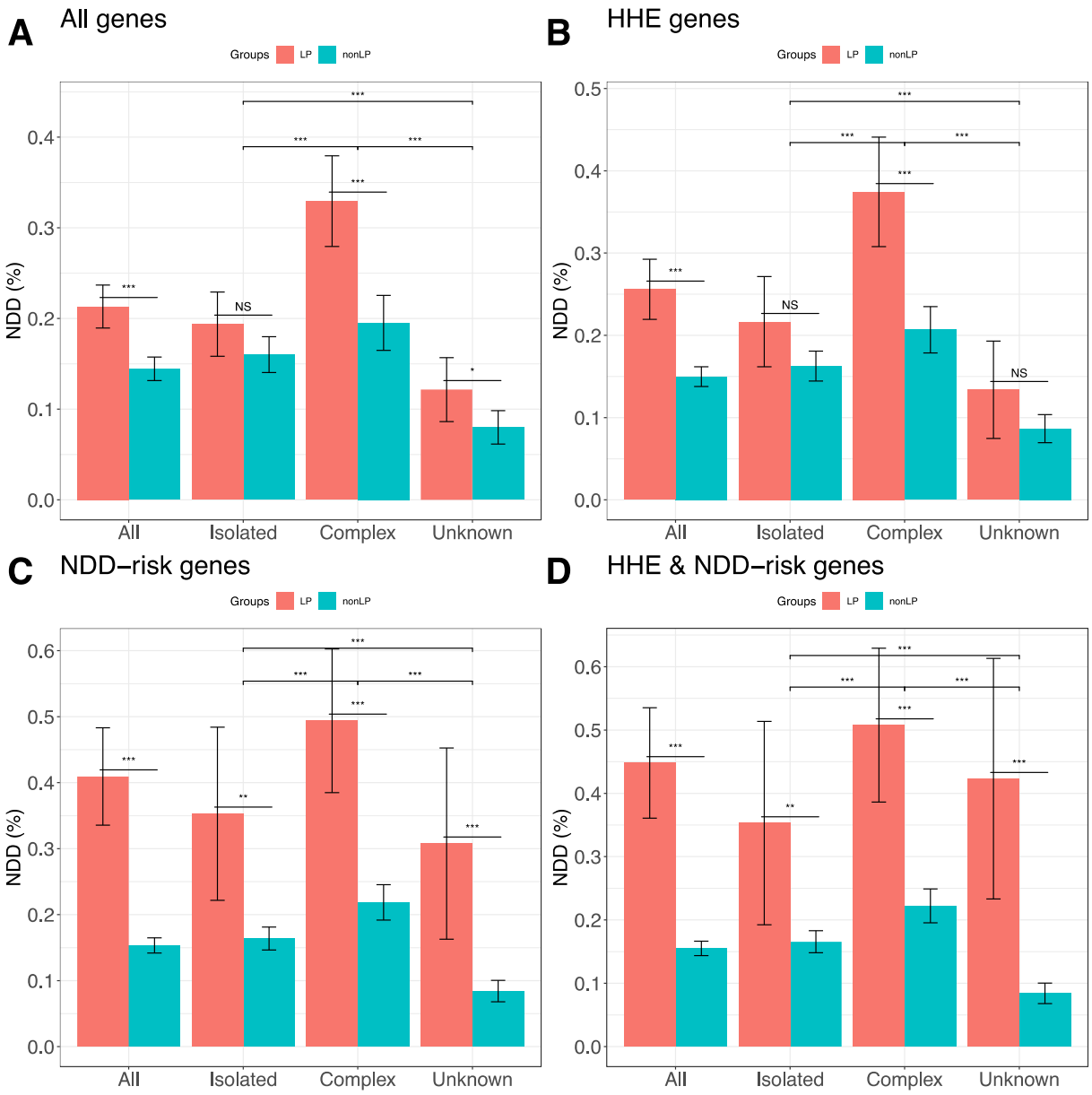

Figure 21. Higher prevalence of NDD among cases carrying likely pathogenic DNVs.

(A) DNVs in all genes. (B) DNVs in HHE genes. (C) DNVs in NDD-risk genes. (D) DNVs in genes annotated as both HHE and NDD-risk genes. The difference in NDD prevalence between LP and nonLP cases increases when considering DNVs in either HHE or NDD-risk genes, with the largest difference observed when considering DNVs in genes at the intersection of HHE and NDD-risk gene sets. The largest difference is observed in Complex CHD cases and is consistent across gene groups. For Isolated CHD cases, however, the difference is most noticeable in NDD-risk genes and genes at the intersection of HHE and NDD-risk. LP=cases carrying likely pathogenic DNVs, nonLP=cases that do not carry likely pathogenic DNVs, HHE=high heart expression, NDD=neurodevelopmental disorder. Stars indicate statistical significance: NS=non-significant, '*'=p<0.05, '**'=p<0.01, '***'=p<0.001, Fisher's Exact Test.

Table 2.3. Prevalence of NDD in cases carrying LP DNVs vs. cases that do not carry LP DNVs.

Gene set	CHD subtype	# LP, NDD	# LP, noNDD	# nonLP, NDD	# nonLP, nonNDD	LP NDD%	nonLP NDD%	Delta %	OR	P-value	sig
All	All	245	904	407	2410	0.21	0.14	0.07	1.60	2.4E-07	***
All	Isolated	93	387	212	1111	0.19	0.16	0.03	1.26	0.10	NS
All	Complex	112	228	128	528	0.33	0.20	0.13	2.02	3.8E-06	***
All	Unknown	40	289	67	771	0.12	0.08	0.04	1.59	0.03	*
HHE	All	140	407	512	2907	0.26	0.15	0.11	1.95	3.9E-09	***
HHE	Isolated	47	170	258	1328	0.22	0.16	0.05	1.42	0.05	NS
HHE	Complex	76	127	164	629	0.37	0.21	0.17	2.29	1.5E-06	***
HHE	Unknown	17	110	90	950	0.13	0.09	0.05	1.63	0.10	NS
NDD-risk	All	70	101	582	3213	0.41	0.15	0.26	3.82	4.8E-15	***
NDD-risk	Isolated	18	33	287	1465	0.35	0.16	0.19	2.78	0.001	**
NDD-risk	Complex	40	41	200	715	0.49	0.22	0.28	3.48	2.3E-07	***
NDD-risk	Unknown	12	27	95	1033	0.31	0.08	0.22	4.82	9.1E-05	***
HHE & NDD-risk	All	56	69	596	3245	0.45	0.16	0.29	4.42	3.2E-14	***
HHE & NDD-risk	Isolated	12	22	293	1476	0.35	0.17	0.19	2.75	0.01	**
HHE & NDD-risk	Complex	33	32	207	724	0.51	0.22	0.29	3.60	1.5E-06	***
HHE & NDD-risk	Unknown	11	15	96	1045	0.42	0.08	0.34	7.95	5.6E-06	***

LP=cases carrying likely pathogenic DNVs, nonLP=cases that do not carry likely pathogenic DNVs, HHE=high heart expression, NDD=neurodevelopmental disorder, NDD-risk=known NDD risk genes.

2.2.3 Mutations with pleiotropic effects drive the acquisition of NDD in CHD

There are 4420 genes that are highly expressed in the developing mouse heart (HHE) and 539 known NDD risk genes. The 261 genes annotated as both HHE and NDD-risk comprise about half of NDD-risk genes and only about 5% of HHE genes (>**Fig. 2.3A**), yet the LGD DNVs located in this subset of genes show the strongest association with NDD in our CHD cohort (RR=9.26, p=1e-12) (>**Table 2.2**). If we use the difference in rates of DNVs in NDD cases and non-NDD cases as a proxy for population attributable risk (PAR), we observe a PAR of 8% when we consider LGD DNVs in all genes, a PAR of 7% for LGD DNVs in HHE genes, and a PAR of 6% for DNVs in NDD risk genes. The PAR of 5% for LGD DNVs in genes that are both HHE and NDD-risk represents a substantial fraction of the overall PAR and suggests again that these genes are most strongly associated with NDD (>**Fig. 2.3B**). Though less

striking, we observe a similar trend when considering Dmis DNVs (>**Fig. 2.3C**). Given the gene set size and relative PAR, we hypothesize that mutations disrupting these specific genes have pleiotropic effects that play a role in the acquisition of NDD in CHD.

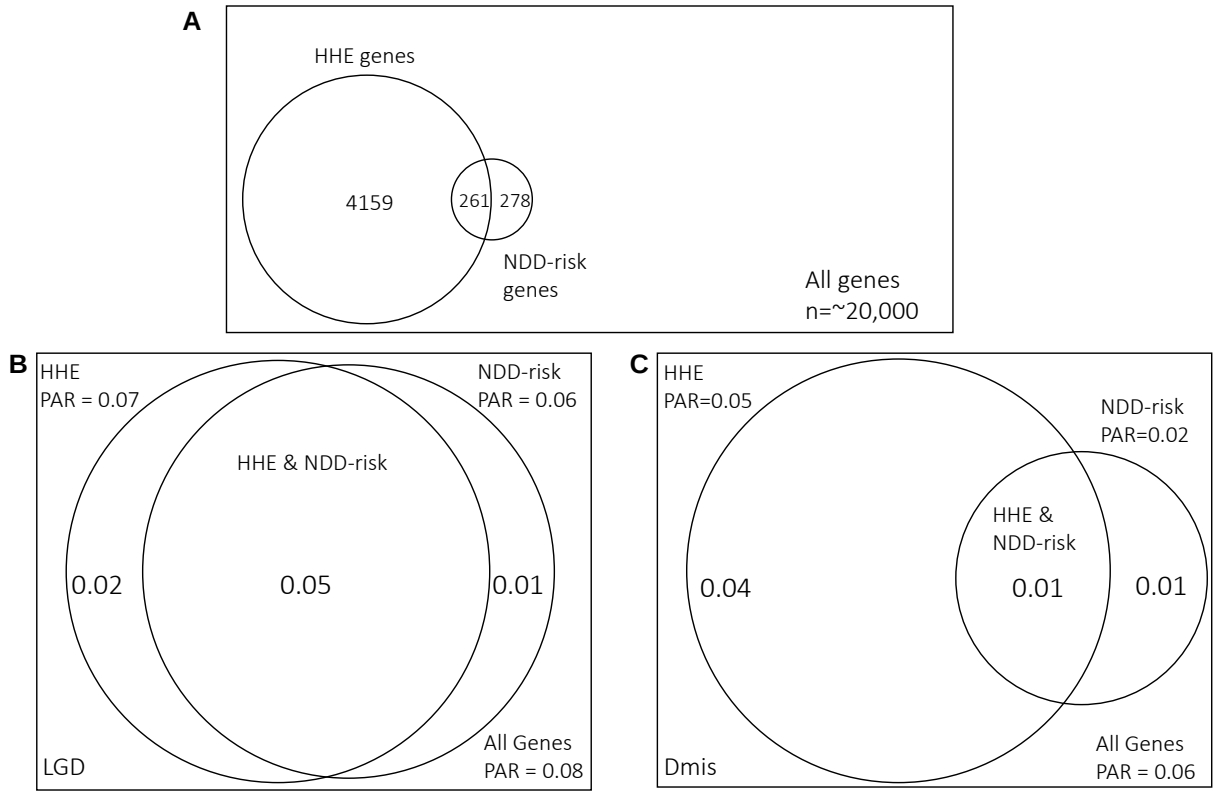

Figure 22. Genes annotated as both HHE & NDD-risk contribute substantial PAR and suggest pleiotropic activity.

Despite comprising roughly half of NDD-risk genes and ~5% of HHE genes, the genes at the intersection show the strongest association with NDD and contribute a substantial fraction of the PAR. The trend is most noticeable when considering LGD DNVs but also seen to a lesser degree in Dmis DNVs. PAR=population attributable risk, HHE=high heart expression, NDD=neurodevelopmental disorder, NDD-risk=known NDD risk genes, LGD=likely-gene-disrupting, Dmis=deleterious missense.

2.2.4 Damaging *de novo* variants are associated with abnormal ventricular function in patients with single ventricle defects

We next focused on the 114 CHD cases seen at Columbia University Medical Center who were diagnosed with single ventricle defects and analyzed the 654 *de novo* variants belonging to this patient subset. Instead of NDD, we used Decreased Systemic Ventricular Function,

Worsening Ventricular Function, and Arrhythmia as our clinical outcome variable and repeated the analyses above (>**Table 2.4**). We found that damaging DNVs were associated with Decreased Systemic Ventricular Function (RR=2.72, p=0.01) and that this association increased when focusing on constrained genes (pLI) (RR=3.63, p=0.02), HHE genes (RR=4.49, p=0.01), and genes that are annotated as both HHE and pLI (RR=5.18, p=0.02). We observed a similar trend in the association between damaging DNVs and Worsening Ventricular function between all genes (RR=2.11, p=0.05), HHE genes (RR=3.78, p=0.02), pLI (RR=4.47, p=0.003), and genes that are both HHE and pLI (RR=8.59, p=0.001). We also see a similar trend in Arrhythmia for damaging DNVs in all genes (RR=2.31, p=0.08), HHE genes (RR=2.52, p=0.11), pLI genes (RR=3.46, p=0.05), and HHE&pLI genes (RR=3.46, p=0.05).

We next compared the prevalence of cases with the phenotypes described above among patients carrying likely pathogenic DNVs and those that do not (**>Fig. 2.4; Table 2.5**). We did not stratify CHD cases into Isolated and Complex due to sample size constraints. We found that the Decreased Systemic Ventricular Function phenotype prevalence is higher among patients carrying LP DNVs (73% LP vs. 42% nonLP; OR=3.72, p=0.007, Fisher's Exact Test) and that the phenotype prevalence increases when considering only LP variants in constrained (pLI) genes (76%), HHE genes (80%), and genes that are both HHE and pLI (82%). The difference in prevalence (% in LP – % in nonLP) also follows the same trend, with difference values of 31%, 32%, 36%, and 46% for All, pLI (OR=4.05, p=0.02), HHE (OR=4.96, p=0.01), and HHE&pLI (OR=5.31, p=0.03) gene sets, respectively. The findings are similar for the Worsening Ventricular Function phenotype – 58% in LP vs. 30% in nonLP cases (OR=3.14, p=0.02) – though the trend shows both a higher phenotype prevalence and a larger difference in phenotype prevalence between LP and nonLP carriers in the constrained genes (71% LP vs. 30% nonLP;

OR=5.41, p=0.002) compared HHE genes (67% LP vs. 32% nonLP; OR=4.2, p=0.02) suggesting that the relative contributions of genes in these two gene sets differs by phenotype. The HHE&pLI gene set again shows the strongest signal (82% LP vs. 32% nonLP; OR=9.53, p=0.002). The Arrhythmia phenotype did not show as strong evidence of a trend as the others; the main difference in phenotype prevalence between LP carriers and nonLP carriers was limited to constrained genes (29% LP vs. 10% nonLP; OR=3.87, p=0.04).

There are 4420 genes that are highly expressed in the developing mouse heart (HHE) and 6050 constrained genes (pLI). The 2520 genes annotated as both HHE and pLI comprise 57% of HHE genes and 42% of pLI genes (**>Fig. 2.5**), yet the damaging DNVs located in this subset of genes show the strongest association with the three clinical outcomes described above. Given the gene set size and relative PAR across the different phenotypes (>**Table 2.4**), the mutations disrupting these specific genes again appear to have pleiotropic effects that play a role in the acquisition of abnormal ventricular function phenotypes in CHD patients diagnosed with single ventricle defects. While not reaching statistical significance, the results of this results of this section provide additional evidence for the pleiotropy hypothesis proposed in the previous sections; however, larger sample sizes will be necessary to draw strong conclusions about the association between damaging DNVs and abnormal ventricular function outcomes in single ventricle patients.

Table 2.4. Rates of damaging DNVs among 114 patients with single ventricle defects with abnormal phenotypes.

Damaging DNVs	Gene Set	Rate Yes	Rate No	PAR (Δ_{rate})	Relative Risk	P-value
Decreased Systemic Ventricular Function (n = 112) (55 yes, 57 no)	All	0.38	0.14	0.24	2.72	0.01
	HHE	0.24	0.05	0.18	4.49	0.01
	pLI	0.25	0.07	0.18	3.63	0.02
	HHE & pLI	0.18	0.04	0.15	5.18	0.02
Worsening Ventricular Function (n = 106) (39 yes, 67 no)	All	0.41	0.19	0.22	2.11	0.05
	HHE	0.28	0.07	0.21	3.78	0.02
	pLI	0.33	0.07	0.26	4.47	0.003
	HHE & pLI	0.26	0.03	0.23	8.59	0.001
Arrhythmia (n = 111) (14 yes, 97 no)	All	0.50	0.22	0.28	2.31	0.08
	HHE	0.29	0.11	0.17	2.52	0.11
	pLI	0.43	0.12	0.30	3.46	0.02
	HHE & pLI	0.29	0.08	0.20	3.46	0.05

DNV=de novo variant, damaging=LGD and Dmis DNVs, PAR=population attributable risk, HHE=high heart expression, pLI=constrained (pLI>0.5).

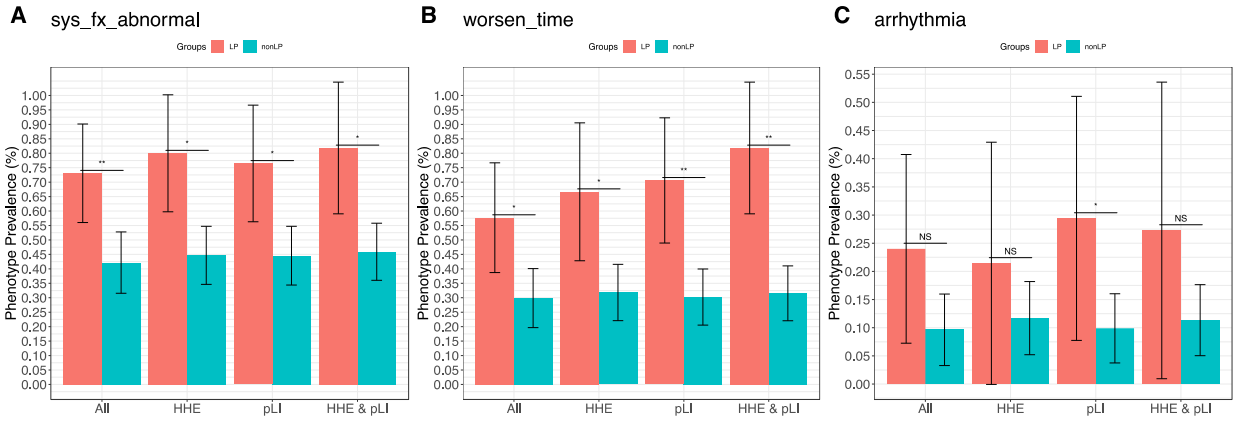

Figure 23. Higher abnormal phenotype prevalence among cases carrying likely pathogenic DNVs.

(A) Prevalence of the Decreased Systemic Ventricular Function phenotype. The phenotype prevalence increases when considering constrained or HHE genes and is the largest in genes annotated as both. (B) Prevalence of the Worsening Ventricular Function phenotype. The phenotype prevalence shows a similar trend, except that constrained genes have higher prevalence than HHE genes. Again, genes that are HHE&pLI show the highest phenotype prevalence(C) Prevalence of the Arrhythmia phenotype. LP=cases carrying likely pathogenic DNVs, nonLP=cases that do not carry likely pathogenic DNVs, HHE=high heart expression, pLI=constrained (pLI>0.5), sys_fx_abnormal=decreased systemic ventricular function, worsen_time=worsening ventricular function. Stars indicate statistical significance: NS=non-significant, '*'=p<0.05, '**'=p<0.01, '***'=p<0.001, Fisher's Exact Test.

Table 2.5. Prevalence of NDD in cases carrying LP DNVs vs. cases that do not.

Phenotype	Gene Set	# LP, yesPheno	# LP, noPheno	# nonLP, yesPheno	# nonLP, noPheno	LP Pheno %	nonLP Pheno %	Delta %	OR	P-value	sig
Decreased Systemic Ventricular Function	All	19	7	36	50	0.73	0.42	0.31	3.72	0.01	**
	HHE	12	3	43	54	0.80	0.44	0.36	4.96	0.01	*
	pLI	13	4	42	53	0.76	0.44	0.32	4.05	0.02	*
	HHE&pLI	9	2	46	55	0.82	0.46	0.36	5.31	0.03	*
Worsening Ventricular Function	All	15	11	24	56	0.58	0.30	0.28	3.14	0.02	*
	HHE	10	5	29	62	0.67	0.32	0.35	4.21	0.02	*
	pLI	12	5	27	62	0.71	0.30	0.40	5.41	0.002	**
	HHE&pLI	9	2	30	65	0.82	0.32	0.50	9.53	0.002	**
Arrhythmia	All	6	19	8	78	0.24	0.09	0.15	3.04	0.08	NS
	HHE	3	11	11	86	0.21	0.11	0.10	2.11	0.38	NS
	pLI	5	12	9	85	0.29	0.10	0.20	3.87	0.04	*
	HHE&pLI	3	8	11	89	0.27	0.11	0.16	2.99	0.14	NS

LP=cases carrying likely pathogenic DNVs, nonLP=cases that do not carry likely pathogenic DNVs, HHE=high heart expression genes, pLI=constrained (pLI>0.5) genes.

2.2.5 Rare variant risk score predicts NDD in CHD patients

We developed a proof-of-concept rare variant risk score model to predict NDD on a per-patient basis by combining counts of rare *de novo,* transmitted, and copy number variants with weights defined by the strength of association with NDD for each particular gene set. For each variant type (*de novo*, transmitted, CNV), weights were estimated using the enrichment (RR) in NDD cases of each combination of functional class (LGD, Dmis, DEL, DUP) and gene set (genes annotated as both HHE and NDD-risk genes, CHD risk genes, genes annotated as HHE and/or constrained (pLI), and in all other genes) (>**Tables 2.6, 2.7, 2.8**). The per-patient risk score was calculated as the sum of the log2-scaled weights across all variant-geneset combinations observed in the individual, given all rare variants detected for that individual.

We compared the score distribution between NDD cases and non-NDD cases and found that while many cases with lower scores overlapped between the two groups, cases with NDD tended to have higher risk scores and the mean risk score for NDD cases was higher than the mean risk score for non-NDD cases (NDD mean=1.38, non-NDD mean=0.75; p=1.1e-19,

Wilcoxon Rank-Sum Test) (>**Fig. 2.5**). We observed the greatest separation in Unknown CHD cases (NDD mean=1.65, non-NDD mean=0.65; p=2.9e-06) and Complex CHD cases (NDD mean=2.29, non-NDD mean=1.68; p=1e-05) whereas we observed the least separation in Isolated CHD cases (NDD mean=0.63, non-NDD mean=0.47; p=0.001). We next evaluated the ability to discriminate between NDD and non-NDD cases using 10-fold cross validation. Our risk score achieved a mean AUPRC of 0.44 (across folds) when applied to all cases and mean AUPRCs of 0.32, 0.53, 0.46 when applied to cases with Isolated, Complex, and Unknown CHD subtypes, respectively (>**Fig. 2.6**). We believe the observed trends in score performance per group reflect the established relationship between CHD complexity and NDD and that our Unknown CHD group represent a mixture of Isolated and Complex cases.

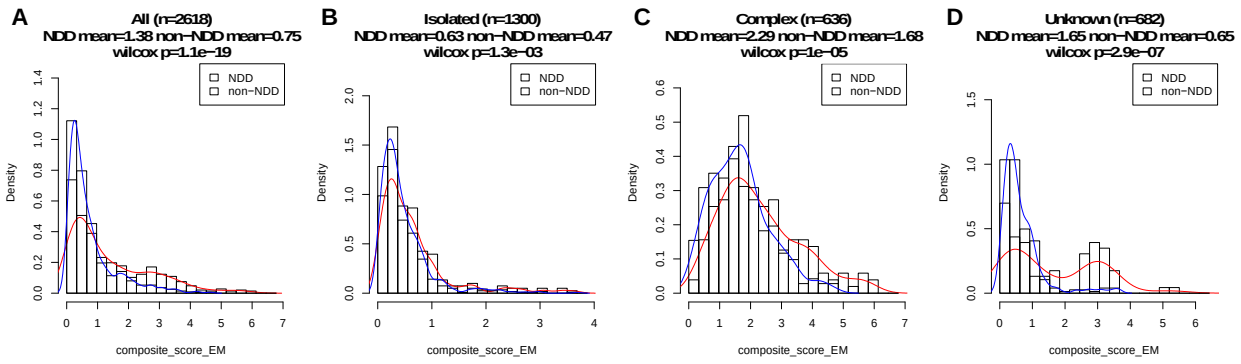

Figure 24. Risk score distribution in NDD vs. non-NDD cases, across CHD subtype groups.

(A) All CHD cases. (B) Isolated cases. (C) Complex cases. (D) Unknown cases. Higher risk score values showed the largest separation between NDD and non-NDD cases. The score distributions showed a larger difference in Complex and Unknown cases than in Isolated cases. NDD=neurodevelopmental disorder, wilcox.p=Wilcoxon Rank-Sum Test p-value

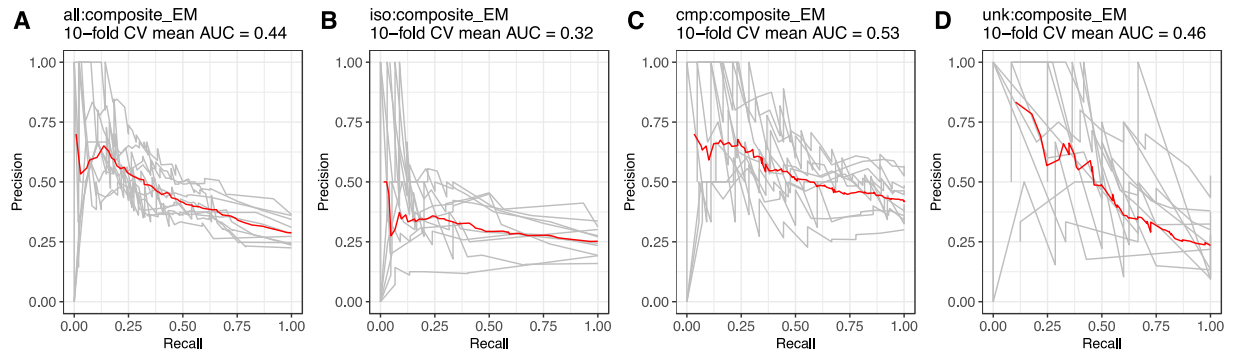

Figure 25. Risk score performance, 10-fold cross validated Precision-Recall curves.

(A) All CHD cases. (B) Isolated cases. (C) Complex cases. (D) Unknown cases. The score showed the strongest performance in discriminating NDD from non-NDD cases in Complex cases and showed the weakest performance in Isolated cases. The performance in Unknown cases fell in between that in Complex and Isolated groups, likely due to the Unknown group containing a mixture of true Complex and Isolated presentations. Iso=isolated, cmp=complex, unk=unknown, AUC=area under (precision-recall) curve.

We next considered the prevalence and enrichment of NDD as a function of patient risk score percentile. We found that the prevalence of NDD increased among patients with higher risk score percentiles (>**Fig 2.7**). The increase in prevalence was greatest in cases with Complex CHD and least in cases with Isolated CHD, with Unknown CHD in the middle. We also found that patients with risk scores in the top 25% were >3-fold as likely to have NDD (OR=3.71, p=8.1E-17, Fisher's Exact Test) compared to patients in the bottom 25% (>**Fig 2.8**). Again, the enrichment was greatest in Complex (OR=3.43, p=0.00003) and Unknown (OR=3.43, p=0.0005) CHD cases and least in Isolated CHD (OR=1.90, p=0.006).

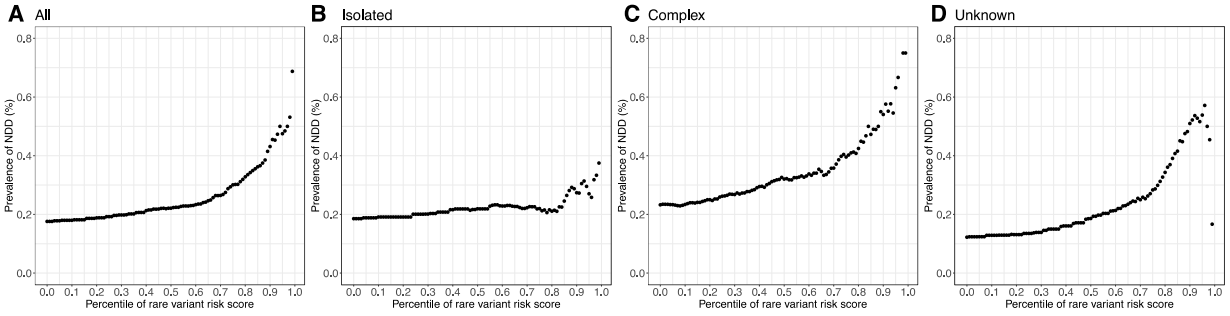

Figure 26. Prevalence of NDD as a function of risk score percentile

(A) All CHD cases. (B) Isolated cases. (C) Complex cases. (D) Unknown cases. The score increases most dramatically for Complex CHD cases, least dramatically for Isolated CHD cases. Unknown CHD cases show an increase in between Complex and Isolated cases.

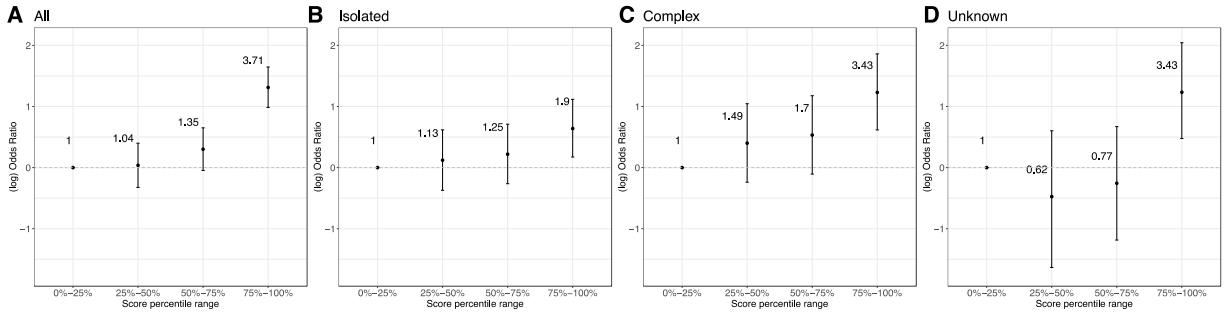

Figure 27. Enrichment of NDD by risk score quartile, compared to bottom quartile

(B) All CHD cases. (B) Isolated cases. (C) Complex cases. (D) Unknown cases. Patients with risk scores in the top 25% were >3-fold as likely to have NDD (OR=3.71, p=8.1E-17) compared to patients in the bottom 25%. Complex (OR=3.43, p=0.00003) and Unknown (OR=3.43, p=0.0005) CHD cases showed the largest enrichment while Isolated CHD (OR=1.90, p=0.006) cases showed the least.

2.3 Discussion

In this study, we conducted an association analysis between rare variants and clinical outcomes in 3966 CHD patients. We found that damaging *de novo* variants are associated with NDD cases and that the association is stronger for variants in HHE genes or in NDD-risk genes, and strongest for variants in genes annotated as both HHE&NDD-risk. We see a similar trend when comparing the relative prevalence of NDD in cases carrying likely pathogenic DNVs in each of the gene sets above. The genes annotated as HHE&NDD-risk comprise roughly half of NDD-risk genes and ~5% of HHE genes yet appear to drive the association with NDD. We believe this to suggests that disruptive mutations in these critical genes have pleiotropic effects that play a role in the acquisition of NDD in CHD patients. While we did not observe the same strength or significance of association with rare transmitted variants, we found that transmitted LGD variants in HHE genes had a PAR of 8% -- comparable with that observed in *de novo* variants and suggestive of an underlying association missed by the limited sample size in this analysis (>**Table 2.7**). CNVs overall were enriched in NDD cases (RR=2.09, p=7e-06), with deletion events appearing to drive the association (RR=3.00, p=1e-07) **(>Table 2.8**). However,

given the small number of CNVs detected, the generalizability of this association remains to be determined.

Focusing on the 114 CHD patients diagnosed with single ventricle defects, we found a similar association between damaging DNVs and abnormal ventricular function phenotypes (Decreased Systemic Ventricular Function, Worsening Ventricular Function, Arrhythmia). The association is stronger when considering damaging DNVs in HHE genes or in constrained genes and is strongest for damaging DNVs in genes annotated as both HHE and constrained, providing additional support for the pleiotropy hypothesis. Damaging DNVs in HHE genes appear to be more strongly associated with the Decreased Systemic Ventricular Function phenotype whereas damaging DNVs in constrained genes show stronger association with Worsening Ventricular Function, potentially hinting at different mechanisms driving these respective phenotypes.

Finally, we combined information from rare *de novo,* transmitted, and copy number variants into a proof-of-concept per-patient rare variant risk score. The score distributions between NDD and non-NDD cases were more distinct for higher score values and for cases with Complex and Unknown CHD presentations. Using 10-fold cross validation to evaluate the performance of our score in distinguishing NDD from non-NDD cases, we achieved an AUPRC of 0.44 for all cases and auPRC values of 0.32, 0.53, and 0.46 for Isolated, Complex, and Unknown cases, respectively. Here, weights were derived by comparing NDD vs. non-NDD cases within our CHD cohort; estimates from a comparison of CHD-NDD cases vs. non-CHD non-NDD age and sex matched controls would likely provide more information and improve the overall discriminatory performance of our method, particularly in Isolated CHD cases.

In conclusion, we found that genes annotated as both HHE and NDD-risk are most strongly associated with NDD in CHD and that disruptive *de novo* variants in these genes likely

have pleiotropic effects. We also found that genes annotated as both HHE and constrained are most strongly associated with abnormal ventricular function phenotypes in CHD patients with single ventricle defects and that the disruptive *de novo* variants in these genes provide additional evidence of pleiotropy. Our rare variant risk score shows potential in distinguishing CHD cases with NDD from non-NDD cases and represents a proof-of-concept application of genomic information in predicting clinical outcomes. As study cohorts increase in size, we will soon be able to more accurately and robustly quantify the association between different classes of genetic variation and phenotypes of interest using the methods described here. With improved measures and further development, we believe that genetic risk scores have the potential to provide clinically actionable information and guide the refinement of existing disease diagnosis and management strategies.

2.4 Materials and Methods

2.4.1 Sequencing data, variant calling, and quality control

We analyzed data from 3966 congenital heart disease (CHD) proband-parent trios recruited as part of the Pediatric Cardiac Genomics Consortium (PCGC) study {Homsy 2015; Jin 2017}. Genomic DNA from venous blood or saliva was captured using Nimblegen v.2 exome capture reagent (Roche) or Nimblegen SeqCap EZ MedExome Target Enrichment Kit (Roche) (for whole-exome sequencing datasets) followed by Illumina DNA sequencing (paired-end, 2x75bp) {Jin 2017, Zaidi 2013}. Sequence reads were mapped to the hg19 human reference genome with BWA-MEM and BAM files were further processed following GATK Best Practices, which included duplication marking, indel realignment, and base quality recalibration steps.

Candidate *de novo* variants were defined as sites present in the offspring with homozygous reference genotypes in both parents. Candidates satisfying any of the following criteria were filtered out and excluded from subsequent analysis: (1) failing VQSR filter (2) Fisher Strand (FS)>25 (2) Quality by Depth (QD)<2 (3) <5 reads supporting the alternate allele in proband (4) <20% alternate allele fraction in proband (5) Phred-scaled genotype likelihood (GQ) <60 (6) ExAC population allele frequency >0.1% (7) <10 reference reads in either parent (8) >5% alternate allele fraction in either parent or (9) GQ<30 in either parent. There were in total 5271 *de novo* variants passing filters belonging to 3966 patients.

Transmitted variants were extracted from the joint-genotype VCFs and defined as sites present in the offspring with at least one parent having a non-homozygous reference genotype. Candidates satisfying any of the following criteria were filtered out and excluded from subsequent analysis: (1) GQ<30 (2) average depth across interval (IDP) <=9 (3) <25% alternate allele fraction (4) gnomAD exome/genome population allele frequency >0.001% (5) Phred-scaled p-value for exact test of excess heterozygosity (ExcessHet) >55 (6) DP<10 in proband (7) DP<10 in either parent or (8) non-European ethnicity. There were in total 166100 rare transmitted variants passing filters belonging to 2618 patients of European (EUR) ethnicity.

Copy number variants were called using PennCNV. Samples with total # CNV calls >4 standard deviations (SD) from the cohort mean were considered outliers and removed from subsequent analysis. Candidates satisfying any of the following criteria were filtered out and excluded: (1) Log R Ratio (LRR) >0.35 SD from sample mean (2) B Allele Frequency (BAF) >4 SD from sample mean (3) BAF drift >4 SD from mean (4) Wave factor <-0.03 or >0.03 (5) # SNPs <10 (6) CNV size <100kb (7) Confidence score <30 (8) >80% overlap with repetitive regions (9) AC in parents >2 (10) gnomAD population frequency >0.1% (11) >5 *de novo* large

CNVs per individual (13) not annotated as overlapping genes. There were in total 237 CNVs passing filters belonging to 1794 patients for which array data was available.

2.4.2 Depth of coverage and D15 for NDD and non-NDD samples across batches

CHD samples were collected in 9 batches. For quality control purposes, we calculated sample average depth (DP) and D15 values for each sample BAM file using the GATK DepthOfCoverage tool. We then compared NDD cases and non-NDD cases to identify potential technical sources of bias in our downstream analysis. We found that overall mean DP and D15 were comparable between NDD and non-NDD cases (DP_{NDD}=58.08, DP_{nonNDD}=58.5, $D15_{NDD}$=87.47, $D15_{nonNDD}$=88.39). However, we note that were fewer NDD cases with high D15 than non-NDD cases (>Fig. 2.9).

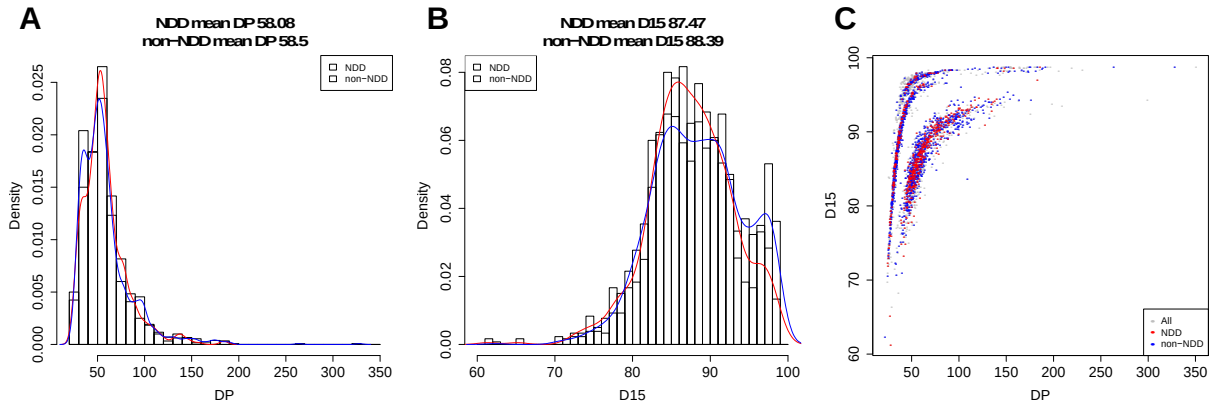

Figure 28. Comparison of DP and D15 in NDD and non-NDD samples.

Per-sample summary statistics were generated using the GATK DepthOfCoverage tool. (A) Sample average depth. (B) D15 (% of bases covered by >15 reads). (C) Sample average depth vs. D15.

2.4.3 Annotations and gene sets

Variants were annotated using both Ensembl Variant Effect Predictor (VEP) (release 96) and ANNOVAR (v2017-07-17) to include information from the dbNSFP version 4.0a database, as well as pathogenicity predictions from a variety of established methods (CADD, MCAP,

REVEL, MPC, MVP, MVP2, spliceAI). We used defined likely-gene-disrupting (LGD) variants as stopgain, stoploss, frameshift, startloss, or spliceAI >0.5. We defined deleterious missense (Dmis) variants as nonsynonymous sites with REVEL score >=0.5 and probably-deleterious-missense (PDmis) variants as nonsynonymous sites with REVEL score <0.5 and CADD >=20. Splice-likely-pathogenic (spliceLP) variants were defined as variants with a spliceAI score between 0.2 and 0.5. Benign missense (Bmis) variants were defined as nonsynonymous sites not in the groups above. For synonymous variants and inframe-indels, we excluded sites with spliceAI score >0.2. We considered sites predicted to be LGD or Dmis as damaging (likely pathogenic). Non-likely pathogenic sites include all variants not in the damaging group.

Gene sets were defined as follows. High Heart Expression (HHE) genes (n=4420) include those ranked in the top 25% by cardiac expression data in mouse at E14.5 {Zaidi 2013; Homsy 2015}. CHD-risk genes (n=156) were defined as known candidate CHD genes with autosomal dominant mode of inheritance {Jin 2017}. NDD-risk genes (n=539) were defined as the union of genes with SFARI score 1 or 2 (n=86), genes discovered by the Autism Sequencing Consortium (ASC) with FDR<0.1 (n=102) {Satterstrom 2019}, and genes in the Developmental Disorders Genotype-Phenotype (DDG2P) database with indicated organ 'brain' and with human phenotype ontology (HPO) terms 'abnormal brain morphology' (HP: 0012443) or 'cognitive impairment' (HP:0100543) (n=454). Constrained/pLI genes (n=6050) were defined as genes with gnomAD pLI>0.5 {Lek 2016}.

2.4.4 Association analysis

We used a Binomial Test and a Fisher's Exact Test to investigate the association between genetic variation and clinical outcomes of interest. Patients were first divided into groups based

on phenotype (e.g. NDD, non-NDD, Decreased Systemic Ventricular Function, no Decreased Systemic Ventricular Function, etc). Then, for each variant functional class (e.g. LGD, Dmis, Bmis, DEL, DUP, etc), we counted the number of variants detected in patients belonging to each phenotype group and calculated a per-group rate. The enrichment (relative risk) was calculated as the ratio of the rates in the positive phenotype and corresponding negative phenotype groups. We also calculated the difference in rates between positive and negative phenotype groups as a proxy for population attributable risk (PAR). To assess significance, we used a Binomial Test with the total number of variants detected across both groups as our number of trials (N), the proportion of cases with the positive phenotype among all patients in both groups as our null probability of success (p), and the number of variants detected in cases with the positive as our number of successes (x). Since the number of variant counts varied between functional classes, we also used a Fisher's Exact Test to improve association accuracy for classes with low counts. Patients were further divided into 4 subsets – cases with positive phenotype and carrying the variant, cases with negative phenotype and carrying the variant, cases with positive phenotype and not carrying the variant, and cases with negative phenotype and not carrying the variant. We then tested for nonrandom association between phenotype and variant functional class variables. This analysis was repeated for each CHD subtype (All, Isolated, Complex, Unknown CHD), CHD category (conotruncal defect (CTD), heterotaxy (HTX), hypoplastic left heart syndrome (HLHS), left ventricular outflow (LVO), other), and gene set (All genes, HHE, constrained, CHD-risk, NDD-risk, and combinations of these).

2.4.5 Prevalence analysis

We used a Fisher's Exact Test to investigate whether patients carrying likely pathogenic (LP; LGD + Dmis) variants were more likely to also have the clinical outcome of interest. Patients were first divided into two groups – cases carrying LP variants and cases that do not carry an LP variant. We then counted the number of patients with the positive phenotype and the negative phenotype within each LP group and calculated a per-group phenotype prevalence and a prevalence difference value. Using the number of cases carrying LP variants with the positive phenotype, cases carrying LP Variants with the negative phenotype, cases without LP variants with the positive phenotype, and cases without LP variants with the negative phenotype, we used a Fisher's Exact Test to test for significance and strength of nonrandom association between phenotype and LP variant variables. This analysis was repeated for each CHD subtype (All, Isolated, Complex, Unknown CHD), CHD category (conotruncal defect (CTD), heterotaxy (HTX), hypoplastic left heart syndrome (HLHS), left ventricular outflow (LVO), other), and gene set (All genes, HHE, constrained, CHD-risk, NDD-risk, and combinations of these).

2.4.6 NDD rare variant risk score

We developed a simple framework for predicting clinical outcomes of interest (e.g. NDD) by combining information from rare *de novo*, transmitted, and copy number variants into a per-patient rare variant risk score. Each patient was represented as a vector of *de novo*, transmitted, and copy number variant counts for all combinations of relevant variant functional classes (LGD, Dmis, DEL, DUP) and gene sets (HHE&NDD-risk, CHD-risk, HHE&pLI, HHE or pLI, other). We used the association analysis described above to identify the gene groups most relevant to NDD and used the Fisher's Exact Test enrichment values (odds ratio) as the

weights in our score. Weights from the association analysis in specific CHD subtypes (Isolated, Complex, Unknown) were used where available. Final weights can be found in **Tables 2.6, 2.7, 2.8**. For each patient, a rare variant risk score was calculated by taking the sum of the log2-scaled product of the variants vector and the weights vector ($score = \sum_i^n log_2 RR_{i,v,g} * 1_{i,v,g}$ for each variant type *v* in {LGD, Dmis, DEL, DUP} and gene set *g* in {HHE&NDD-risk, CHD-risk, HHE&pLI, HHE or pLI, other}). We generated per-patient risk scores within each variant class (*de novo,* transmitted, CNV) as well as a composite score combining information across the variant classes. Given that transmitted variants were only available for the 2618 EUR cases, we limited subsequent analysis involving this composite rare variant risk score to these 2618 cases.

We next evaluated the utility of our score in discriminating between patients with and without the clinical outcome of interest (NDD). We first compared the mean score values in NDD and non-NDD to assess the magnitude of difference. We used a two-sided Wilcoxon Rank-Sum Test to test the null hypothesis that it is equally likely that a randomly selected score value from our NDD group will be greater/less than a randomly selected score from our non-NDD group. To identify optimal score thresholds for our rare variant risk score, we iterated over score values by percentile (0 to 100) and calculated the Matthews correlation coefficient (MCC = [TPxTN–FPxFN]/sqrt[(TP+FP)(TP+FN)(TN+FP)(TN+FN)]) to quantify the correlation between that particular score value threshold and NDD status while accounting for imbalanced class sizes (>**Fig. 2.10**). We next used a 10-fold cross validation approach to evaluate our score's predictive ability. Patients were first randomly divided into 10 folds. For each fold *k*, we split our dataset into testing (fold *k*) and training (remaining 9 folds) sets. We then estimated weights using the cases in the training set and calculated risk scores for each patient in the held-out testing set. To calculate per-fold area under precision-recall curve (AUPRC), we iterated over score values by

percentile (0 to 100) and calculated Precision (TP/(TP+FP)) and Recall (TP/(TP+FN)) for each score threshold. We used R package *DescTools* (v.0.99.30) to estimate AUPRC. We then calculated mean AUPRC across folds as our final risk score performance metric. This analysis was repeated for each CHD subtype (Isolated, Complex, Unknown).

To calculate prevalence of NDD as a function of risk score percentile, we iterated over score values by percentile (0 to 100) and calculated the fraction of NDD patients with scores above the percentile cutoff (prevalence = # NDD patients / # total patients). To calculate enrichment by quartile, we compared the number of NDD and non-NDD cases in each quartile against the number of NDD and non-NDD cases in the bottom quartile and performed a Fisher's Exact Test.

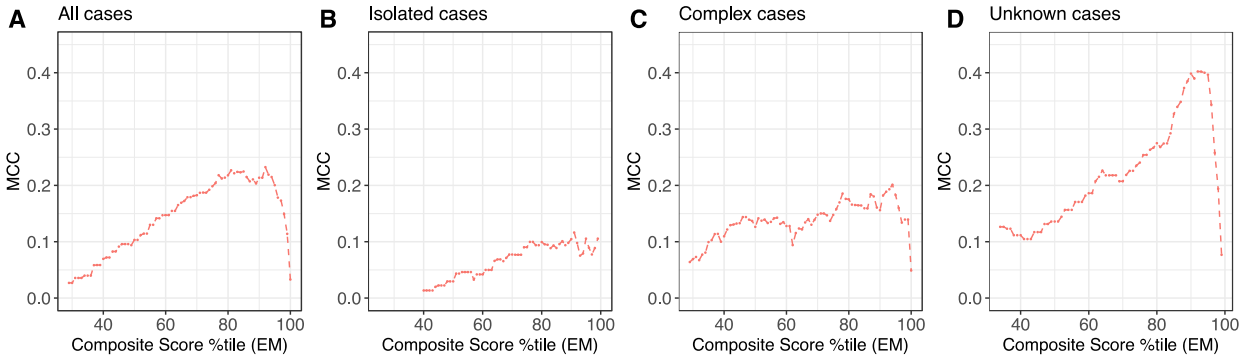

Figure 29. Matthews Correlation Coefficient (MCC) for composite score thresholds, by percentile.

(A) All cases. (B) Isolated cases. (C) Complex cases. (D) Unknown cases. High risk score values have the strongest correlation in Unknown CHD cases.

Table 2.6. Risk score weights for *de novo* variants.

Weights *de novos*		Variant class	# DNV in NDD cases	Rate in NDD	# DNV in non-NDD cases	Rate in non-NDD	Delta	RR	P-value	OR	P-value
All cases N_NDD = 652 N_non = 1588	NDDrisk & HHE	LGD	38	0.06	10	0.01	0.05	9.26	1.0E-12	9.48	1.5E-12
		Dmis	18	0.03	13	0.01	0.02	3.37	0.001	3.44	0.001
	CHD Risk Genes	LGD	6	0.01	15	0.01	0.00	0.97	1	0.97	1
		Dmis	17	0.03	24	0.02	0.01	1.73	0.09	1.74	0.08
	HHE & pLI	LGD	25	0.04	32	0.02	0.02	1.90	0.02	2.00	0.02
		Dmis	18	0.03	38	0.02	0.00	1.15	0.66	1.16	0.66
	HHE or pLI	LGD	26	0.04	63	0.04	0.00	1.01	1	1.02	0.91
		Dmis	58	0.09	83	0.05	0.04	1.70	0.002	1.74	0.003
	Other Genes	LGD	45	0.07	95	0.06	0.01	1.15	0.46	1.19	0.38
		Dmis	38	0.06	111	0.07	-0.01	0.83	0.37	0.87	0.51
Isolated N_NDD = 412 N_non = 1263	NDDrisk & HHE	LGD	14	0.03	8	0.01	0.03	5.36	0.0001	5.51	0.0001
		Dmis	8	0.02	8	0.01	0.01	3.07	0.04	3.10	0.04
	CHD Risk Genes	LGD	5	0.01	10	0.01	0.00	1.53	0.39	1.54	0.38
		Dmis	7	0.02	16	0.01	0.00	1.34	0.48	1.35	0.47
	HHE & pLI	LGD	10	0.02	24	0.02	0.01	1.28	0.55	1.34	0.42
		Dmis	9	0.02	28	0.02	0.00	0.99	1	0.99	1
	HHE or pLI	LGD	14	0.03	53	0.04	-0.01	0.81	0.57	0.82	0.56
		Dmis	33	0.08	72	0.06	0.02	1.41	0.11	1.46	0.10
	Other Genes	LGD	31	0.08	67	0.05	0.02	1.42	0.13	1.45	0.11
		Dmis	24	0.06	84	0.07	-0.01	0.88	0.66	0.91	0.81
Complex N_NDD = 240 N_non = 325	NDDrisk & HHE	LGD	24	0.10	2	0.01	0.09	16.25	1.4E-07	17.05	1.7E-07
		Dmis	10	0.04	5	0.02	0.03	2.71	0.07	2.78	0.07
	CHD Risk Genes	LGD	1	0.00	5	0.02	-0.01	0.27	0.25	0.27	0.25
		Dmis	10	0.04	8	0.02	0.02	1.69	0.34	1.72	0.33
	HHE & pLI	LGD	15	0.06	8	0.02	0.04	2.54	0.03	2.64	0.03
		Dmis	9	0.04	10	0.03	0.01	1.22	0.66	1.23	0.81
	HHE or pLI	LGD	12	0.05	10	0.03	0.02	1.63	0.28	1.66	0.28
		Dmis	25	0.10	11	0.03	0.07	3.08	0.001	3.02	0.004
	Other Genes	LGD	14	0.06	28	0.09	-0.03	0.68	0.28	0.71	0.41
		Dmis	14	0.06	27	0.08	-0.02	0.70	0.34	0.71	0.41

N_NDD=number of NDD cases, N_non=number of non-NDD cases, DNV=de novo variant, Delta=difference in rates, RR=relative risk (Binomial Test), OR=Odds Ratio (Fisher's Exact Test), HHE=high heart expression genes, pLI=constrained genes, NDDrisk=known NDD risk genes, Other Genes=genes not in above categories

Table 2.7. Risk score weights for rare transmitted variants.

Weights Transmitted (1e-5)		Variant class	# var in NDD cases	Rate in NDD	# var in non-NDD cases	Rate in non-NDD	Delta	RR	P-value	OR	P-value
All cases N_NDD = 454 N_non = 1122	HHE	LGD	394	0.87	887	0.79	0.08	1.10	0.12	1.14	0.24
		Dmis	743	1.64	2016	1.80	-0.16	0.91	0.03	0.87	0.33
	Other Genes	LGD	1482	3.26	3869	3.45	-0.18	0.95	0.07	0.88	0.58
		Dmis	1994	4.39	5218	4.65	-0.26	0.94	0.03	0.87	0.60
Isolated N_NDD = 295 N_non = 901	HHE	LGD	246	0.83	708	0.79	0.05	1.06	0.43	1.07	0.64
		Dmis	483	1.64	1626	1.80	-0.17	0.91	0.06	0.85	0.35
	Other Genes	LGD	967	3.28	3102	3.44	-0.16	0.95	0.18	0.75	0.25
		Dmis	1254	4.25	4227	4.69	-0.44	0.91	0.002	0.80	0.51
Complex N_NDD = 159 N_non = 221	HHE	LGD	148	0.93	179	0.81	0.12	1.15	0.22	1.21	0.40
		Dmis	260	1.64	390	1.76	-0.13	0.93	0.36	0.98	1
	Other Genes	LGD	515	3.24	767	3.47	-0.23	0.93	0.23	1.28	0.66
		Dmis	740	4.65	991	4.48	0.17	1.04	0.45	1.18	0.82

N_NDD=number of NDD cases, N_non=number of non-NDD cases, DNV=de novo variant, Delta=difference in rates, RR=relative risk (Binomial Test), OR=Odds Ratio (Fisher's Exact Test), HHE=high heart expression genes, Other Genes=genes not in above categories

Table 2.8. Risk score weights for copy number variants.

Weights CNV		Variant class	#var in NDD cases	Rate in NDD	# var in non-NDD cases	Rate in non-NDD	Delta	RR	P-value	OR	P-value
All cases N_NDD = 315 N_non = 748	NDDrisk & HHE	DEL	31	0.10	11	0.01	0.08	6.69	4.4E-09	7.04	5.3E-09
		DUP	7	0.02	3	0.00	0.02	5.54	0.01	5.63	0.01
	CHD Risk Genes	DEL	5	0.02	4	0.01	0.01	2.97	0.14	3.00	0.14
		DUP	1	0.00	1	0.00	0.00	2.37	0.50	2.38	0.51
	HHE & pLI	DEL	3	0.01	3	0.00	0.01	2.37	0.37	2.39	0.37
		DUP	3	0.01	11	0.01	-0.01	0.65	0.77	0.64	0.77
	HHE or pLI	DEL	4	0.01	10	0.01	0.00	0.95	1	0.95	1
		DUP	4	0.01	10	0.01	0.00	0.95	1	1.06	1
	Other Genes	DEL	10	0.03	14	0.02	0.01	1.70	0.26	1.72	0.26
		DUP	6	0.02	17	0.02	0.00	0.84	0.82	0.69	0.64
Isolated cases N_NDD = 220 N_non = 680	NDDrisk & HHE	DEL	27	0.12	9	0.01	0.11	8.29	1.9E-09	8.89	2.2E-09
		DUP	4	0.02	3	0.00	0.01	3.68	0.09	3.73	0.09
	CHD Risk Genes	DEL	5	0.02	4	0.01	0.02	3.45	0.06	3.51	0.06
		DUP	1	0.00	1	0.00	0.00	2.76	0.46	2.77	0.46
	HHE & pLI	DEL	1	0.00	2	0.00	0.00	1.38	1	1.38	1
		DUP	2	0.01	9	0.01	-0.01	0.61	0.74	0.61	0.74
	HHE or pLI	DEL	3	0.01	9	0.01	0.00	0.92	1	0.92	1
		DUP	3	0.01	9	0.01	0.00	0.92	1	1.04	1
	Other Genes	DEL	7	0.03	13	0.02	0.01	1.49	0.45	1.50	0.44
		DUP	5	0.02	14	0.02	0.00	0.99	1	0.79	0.79
Complex cases N_NDD = 95 N_non = 140	NDDrisk & HHE	DEL	4	0.04	2	0.01	0.03	2.95	0.23	3.02	0.22
		DUP	3	0.03	0	0.00	0.03	Inf	0.07	Inf	0.06
	CHD Risk Genes	DEL	0	0.00	0	0.00	0.00	NA	0	0	1
		DUP	0	0.00	0	0.00	0.00	NA	0	0	1
	HHE & pLI	DEL	2	0.02	1	0.01	0.01	2.95	0.57	2.98	0.57
		DUP	1	0.01	2	0.01	0.00	0.74	1	0.73	1
	HHE or pLI	DEL	1	0.01	1	0.01	0.00	1.47	1	1.48	1
		DUP	1	0.01	1	0.01	0.00	1.47	1	1.48	1
	Other Genes	DEL	3	0.03	1	0.01	0.02	4.42	0.31	4.50	0.31
		DUP	1	0.01	3	0.02	-0.01	0.49	0.65	0.49	0.65

N_NDD=number of NDD cases, N_non=number of non-NDD cases, var=variant, DEL=deletion, DUP=duplication, Delta=difference in rates, RR=relative risk (Binomial Test), OR=Odds Ratio (Fisher's Exact Test), HHE=high heart expression genes, pLI=constrained genes, NDDrisk=known NDD risk genes, Other Genes=genes not in above categories

Conclusion

In this dissertation, I have discussed the contribution of mosaicism and other types of variation to the genetic architecture of congenital heart disease. In the first chapter, I presented the development of a novel computational method for detecting mosaic single-nucleotide variants in exome-sequencing data, EM-mosaic. Recent publications have reported discordant validation rates and mosaic fraction/rate estimates due to differences in sequencing depth, variant calling, and mosaic detection approach. Further, distinguishing mosaic from germline heterozygous mutations remains a challenge for current methods. We addressed these gaps by developing an approach that combines heuristic variant filters, error modeling, and data-driven parameter estimation. EM-mosaic achieved a 90% validation rate, among the highest in recent publications. Simulation experiments demonstrated that our estimated prior mosaic fraction and posterior-odds based false discovery rate (FDR) estimate were consistent with the truth. We found that 1% of CHD patients carries a mosaic likely contributing to their heart malformation and that roughly 1 in 8 individuals carries a mosaic event detectable in blood exome sequencing data. Analysis of subjects with matched blood and heart tissue demonstrated that mutations in blood with relatively high allele fraction were more likely to also be found in the heart, supporting the notion of allele fraction as a proxy of cellular percentage and that mutations occurring earlier in development are more likely to be found across multiple tissues. In the second chapter, to disentangle the biological mechanisms governing differences in genetic etiology across CHD complexities (Isolated, Complex, Unknown), I presented a statistical

approach to characterizing the association between genetic variation and clinical outcomes in CHD patients. I found that damaging *de novo* variants are enriched in CHD patients with neurodevelopmental disorders (NDD) or with ventricular dysfunction phenotypes and that variants in high heart expression (HHE) genes, known NDD-risk genes, and constrained genes are most strongly associated. I then showed that pleiotropic *de novo* variants in HHE&NDD-risk genes and HHE&constrained genes contribute a disproportionately large fraction of the risk of acquiring comorbid neurodevelopmental disorder or ventricular dysfunction, respectively. Finally, using the association analysis results, I developed a proof-of-concept rare variant risk score to predict NDD in CHD patients on the basis of their genetic profile (detected *de novo*, rare inherited, and copy number variants) and the relative contributions of these variants across a variety of gene sets. I show that this risk score can stratify patients in our CHD cohort in a clinically meaningful way and identify patients at increased risk of NDD.

Future directions for this work including expanding EM-mosaic to detect post-zygotic small insertions/deletion (indels) and developing a method of *in silico* variant validation at scale. Modeling indels is challenging since indel calling in general is less refined than SNV calling and many additional factors influence their deviation in allele fraction from expectation under germline conditions. These factors include the type of event (insertion vs deletion), the size of the event, and the local sequence content (particularly GC%), all of which would need to be considered in, for example, a regression-based approach. Currently, *in silico* variant validation remains a bottleneck for large-scale genetic studies. In this work, we manually reviewed variant read pileup screenshots generated in IGV, which would be intractable for larger datasets containing an order of magnitude more variants. Automated orthogonal validation approaches present an attractive alternative to manual review. DeepVariant {Poplin 2018}, for example, is

the current state-of-the-art for germline variant quality control. However, extending its framework to mosaic variant validation would require training data that does not currently exist. However, it would be feasible to develop a synthetic training dataset by, for example, using transmitted germline variants with subsampled alternate allele read depth as “positives” and using mendelian error events with comparable variant allele fractions as “negatives”.

As study cohorts increase in size and our ability to detect different classes of variants improve, we will soon be able to accurately determine the association between the full spectrum of genetic variation and clinical phenotypes of interest. With improved measures and further development, the genetic risk score and other methods discussed here have the potential to provide clinically actionable information and to improve current disease diagnosis and management strategies, both for CHD and for other rare developmental disorders.

References

Acuna-Hidalgo, R., Bo, T., Kwint, M. P., van de Vorst, M., Pinelli, M., Veltman, J. A., … Gilissen, C. (2015). Post-zygotic Point Mutations Are an Underrecognized Source of De Novo Genomic Variation. *The American Journal of Human Genetics*, *97*(1), 67–74. https://doi.org/10.1016/J.AJHG.2015.05.008

Agopian, A. J., Goldmuntz, E., Hakonarson, H., Sewda, A., Taylor, D., Mitchell, L. E., & Pediatric Cardiac Genomics Consortium*. (2017). Genome-Wide Association Studies and Meta-Analyses for Congenital Heart Defects. *Circulation: Cardiovascular Genetics*, *10*(3), e001449. https://doi.org/10.1161/CIRCGENETICS.116.001449

Amarasinghe, K. C., Li, J., Hunter, S. M., Ryland, G. L., Cowin, P. A., Campbell, I. G., & Halgamuge, S. K. (2014). Inferring copy number and genotype in tumour exome data. *BMC Genomics*, *15*(1), 732. https://doi.org/10.1186/1471-2164-15-732

An, J.-Y., Lin, K., Zhu, L., Werling, D. M., Dong, S., Brand, H., … Sanders, S. J. (2018). Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. *Science*, *362*(6420), eaat6576. https://doi.org/10.1126/science.aat6576

Belickova, M., Vesela, J., Jonasova, A., Pejsova, B., Votavova, H., Merkerova, M. D., … Cermak, J. (2016). TP53 mutation variant allele frequency is a potential predictor for clinical outcome of patients with lower-risk myelodysplastic syndromes. *Oncotarget*, *7*(24), 36266–36279. https://doi.org/10.18632/oncotarget.9200

Biesecker, L. G., & Spinner, N. B. (2013). A genomic view of mosaicism and human disease. *Nature Reviews Genetics*, *14*(5), 307–320. https://doi.org/10.1038/nrg3424

Bjornard, K., Riehle-Colarusso, T., Gilboa, S. M., & Correa, A. (2013). Patterns in the prevalence of congenital heart defects, metropolitan Atlanta, 1978 to 2005. *Birth Defects Research Part A: Clinical and Molecular Teratology*, *97*(2), 87–94. https://doi.org/10.1002/bdra.23111

Blue, G. M., Kirk, E. P., Sholler, G. F., Harvey, R. P., & Winlaw, D. S. (2012). Congenital heart disease: current knowledge about causes and inheritance. *Medical Journal of Australia*, *197*(3), 155–159. https://doi.org/10.5694/mja12.10811

Briggs, L. E., Kakarla, J., & Wessels, A. (2012). The pathogenesis of atrial and atrioventricular septal defects with special emphasis on the role of the dorsal mesenchymal protrusion. *Differentiation*, *84*(1), 117–130. https://doi.org/10.1016/j.diff.2012.05.006

Burnham, N., Ittenbach, R. F., Stallings, V. A., Gerdes, M., Zackai, E., Bernbaum, J., … Gaynor, J. W. (2010). Genetic factors are important determinants of impaired growth after infant cardiac surgery. *The Journal of Thoracic and Cardiovascular Surgery*, *140*(1), 144–149. https://doi.org/10.1016/j.jtcvs.2010.01.003

Cai, C.-L., Liang, X., Shi, Y., Chu, P.-H., Pfaff, S. L., Chen, J., & Evans, S. (2003). Isl1 identifies a cardiac progenitor population that proliferates prior to differentiation and contributes a majority of cells to the heart. *Developmental Cell*, *5*(6), 877–889. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/14667410

Calderon, J., & Bellinger, D. C. (2015). Executive function deficits in congenital heart disease: why is intervention important? *Cardiology in the Young*, *25*(7), 1238–1246. https://doi.org/10.1017/S1047951115001134

Chatterjee, N., Wheeler, B., Sampson, J., Hartge, P., Chanock, S. J., & Park, J.-H. (2013). Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nature Genetics*, *45*(4), 400–405. https://doi.org/10.1038/ng.2579

Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., … Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, *31*(3), 213–219. https://doi.org/10.1038/nbt.2514

Cohn, D. H., Starman, B. J., Blumberg, B., & Byers, P. H. (1990). Recurrence of lethal osteogenesis imperfecta due to parental mosaicism for a dominant mutation in a human type I collagen gene (COL1A1). *American Journal of Human Genetics*, *46*(3), 591–601. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2309707

Colombo, S., de Sena-Tomás, C., George, V., Werdich, A. A., Kapur, S., MacRae, C. A., & Targoff, K. L. (2017). *nkx* genes establish SHF cardiomyocyte progenitors at the arterial pole and pattern the venous pole through Isl1 repression. *Development*, dev.161497. https://doi.org/10.1242/dev.161497

Conlin, L. K., Thiel, B. D., Bonnemann, C. G., Medne, L., Ernst, L. M., Zackai, E. H., … Spinner, N. B. (2010). Mechanisms of mosaicism, chimerism and uniparental disomy identified by single nucleotide polymorphism array analysis. *Human Molecular Genetics*, *19*(7), 1263–1275. https://doi.org/10.1093/hmg/ddq003

Cordell, H. J., Bentham, J., Topf, A., Zelenika, D., Heath, S., Mamasoula, C., … Keavney, B. D. (2013). Genome-wide association study of multiple congenital heart disease phenotypes identifies a susceptibility locus for atrial septal defect at chromosome 4p16. *Nature Genetics*, *45*(7), 822–824. https://doi.org/10.1038/ng.2637

Costello, M., Pugh, T. J., Fennell, T. J., Stewart, C., Lichtenstein, L., Meldrim, J. C., … Getz, G. (2013). Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Research*, *41*(6), e67–e67. https://doi.org/10.1093/nar/gks1443

Daber, R., Chapman, K. A., Ruchelli, E., Kasperski, S., Mulchandani, S., Thiel, B. D., … Spinner, N. B. (2011). Mosaic trisomy 17: Variable clinical and cytogenetic presentation. *American Journal of Medical Genetics Part A, 155*(10), 2489–2495. https://doi.org/10.1002/ajmg.a.34172

Dawson, K., Aflaki, M., & Nattel, S. (2013). Role of the Wnt-Frizzled system in cardiac pathophysiology: A rapidly developing, poorly understood area with enormous potential. *Journal of Physiology, 591*(6), 1409–1432. https://doi.org/10.1113/jphysiol.2012.235382

De Rubeis, S., He, X., Goldberg, A. P., Poultney, C. S., Samocha, K., Ercument Cicek, a., … Buxbaum, J. D. (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature*, *515*(7526), 209–215. https://doi.org/10.1038/nature13772

De, S. (2011). Somatic mosaicism in healthy human tissues. *Trends in Genetics*, *27*(6), 217–223. https://doi.org/10.1016/j.tig.2011.03.002

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V, Maguire, J. R., Hartl, C., … Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–498. https://doi.org/10.1038/ng.806

Dina, C., Bouatia-Naji, N., Tucker, N., Delling, F. N., Toomer, K., Durst, R., … Leducq Transatlantic MITRAL Network. (2015). Genetic association analyses highlight biological pathways underlying mitral valve prolapse. *Nature Genetics*, *47*(10), 1206–1211. https://doi.org/10.1038/ng.3383

Dixon-Salazar, T. J., Silhavy, J. L., Udpa, N., Schroth, J., Bielas, S., Schaffer, A. E., … Gleeson, J. G. (2012). Exome Sequencing Can Improve Diagnosis and Alter Patient Management. *Science Translational Medicine*, *4*(138), 138ra78-138ra78. https://doi.org/10.1126/scitranslmed.3003544

Donkervoort, S., Hu, Y., Stojkovic, T., Voermans, N. C., Foley, A. R., Leach, M. E., … Bönnemann, C. G. (2015). Mosaicism for Dominant Collagen 6 Mutations as a Cause for Intrafamilial Phenotypic Variability. *Human Mutation*, *36*(1), 48–56. https://doi.org/10.1002/humu.22691

Dou, Y., Yang, X., Li, Z., Wang, S., Zhang, Z., Ye, A. Y., … Wei, L. (2017). Postzygotic single-nucleotide mosaicisms contribute to the etiology of autism spectrum disorder and autistic traits and the origin of mutations. *Human Mutation*, *38*(8), 1002–1013. https://doi.org/10.1002/humu.23255

Drake, K. M., Comhair, S. A., Erzurum, S. C., Tuder, R. M., & Aldred, M. A. (2015). Endothelial Chromosome 13 Deletion in Congenital Heart Disease–associated Pulmonary Arterial Hypertension Dysregulates SMAD9 Signaling. *American Journal of Respiratory and Critical Care Medicine*, *191*(7), 850–854.

Durst, R., Sauls, K., Peal, D. S., deVlaming, A., Toomer, K., Leyne, M., … Slaugenhaupt, S. A. (2015). Mutations in DCHS1 cause mitral valve prolapse. *Nature*, *525*(7567), 109–113. https://doi.org/10.1038/nature14670

Erickson, R. P. (2010). Somatic gene mutation and human disease other than cancer: An update. *Mutation Research - Reviews in Mutation Research*, *705*(2), 96–106. https://doi.org/10.1016/j.mrrev.2010.04.002

Etheridge, S. P., Bowles, N. E., Arrington, C. B., Pilcher, T., Rope, A., Wilde, A. A. M., … Tristani-Firouzi, M. (2011). Somatic mosaicism contributes to phenotypic variation in Timothy syndrome. *American Journal of Medical Genetics Part A*, *155*(10), 2578–2583. https://doi.org/10.1002/ajmg.a.34223

Feinstein, J. A., Benson, D. W., Dubin, A. M., Cohen, M. S., Maxey, D. M., Mahle, W. T., … Martin, G. R. (2012). Hypoplastic Left Heart Syndrome. *Journal of the American College of Cardiology*, *59*(1), S1–S42. https://doi.org/10.1016/j.jacc.2011.09.022

Finger, J. H., Smith, C. M., Hayamizu, T. F., McCright, I. J., Xu, J., Law, M., … Ringwald, M. (2017). The mouse Gene Expression Database (GXD): 2017 update. *Nucleic Acids Research*, *45*(D1), D730–D736. https://doi.org/10.1093/nar/gkw1073

Fischbach, G. D., & Lord, C. (2010). The Simons Simplex Collection: A Resource for Identification of Autism Genetic Risk Factors. *Neuron*, *68*(2), 192–195. https://doi.org/10.1016/j.neuron.2010.10.006

Forbes, S. A., Bhamra, G., Bamford, S., Dawson, E., Kok, C., Clements, J., … Stratton, M. R. (2008). The Catalogue of Somatic Mutations in Cancer (COSMIC). In *Current Protocols in Human Genetics* (Vol. Chapter 10, p. Unit 10.11). Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/0471142905.hg1011s57

Forsberg, L. A., Rasi, C., Razzaghian, H. R., Pakalapati, G., Waite, L., Thilbeault, K. S., … Dumanski, J. P. (2012). Age-Related Somatic Structural Changes in the Nuclear Genome of Human Blood Cells. *The American Journal of Human Genetics*, *90*(2), 217–228. https://doi.org/10.1016/J.AJHG.2011.12.009

Francioli, L. C., Polak, P. P., Koren, A., Menelaou, A., Chun, S., Renkens, I., … Sunyaev, S. R. (2015). Genome-wide patterns and properties of de novo mutations in humans. *Nature Genetics*, *47*(7), 822–826. https://doi.org/10.1038/ng.3292

Freed, D., & Pevsner, J. (2016). The Contribution of Mosaic Variants to Autism Spectrum Disorder. *PLOS Genetics*, *12*(9), e1006245. https://doi.org/10.1371/journal.pgen.1006245

Fryxell, K. J., & Moon, W.-J. (2005). CpG Mutation Rates in the Human Genome Are Highly Dependent on Local GC Content. *Molecular Biology and Evolution*, *22*(3), 650–658. https://doi.org/10.1093/molbev/msi043

Garg, V., Muth, A. N., Ransom, J. F., Schluterman, M. K., Barnes, R., King, I. N., … Srivastava, D. (2005). Mutations in NOTCH1 cause aortic valve disease. *Nature*, *437*(7056), 270–274. https://doi.org/10.1038/nature03940

Gebbia, M., Ferrero, G. B., Pilia, G., Bassi, M. T., Aylsworth, A. S., Penman-Splitt, M., … Casey, B. (1997). X-linked situs abnormalities result from mutations in ZIC3. *Nature Genetics*, *17*(3), 305–308. https://doi.org/10.1038/ng1197-305

Gelb, B. D., & Tartaglia, M. (2011). RAS signaling pathway mutations and hypertrophic cardiomyopathy: getting into and out of the thick of it. *Journal of Clinical Investigation*, *121*(3), 844–847. https://doi.org/10.1172/JCI46399

Genovese, G., Kähler, A. K., Handsaker, R. E., Lindberg, J., Rose, S. A., Bakhoum, S. F., … McCarroll, S. A. (2014). Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood DNA Sequence. *New England Journal of Medicine*, *371*(26), 2477–2487. https://doi.org/10.1056/NEJMoa1409405

Ghedira, N., Kraoua, L., Lagarde, A., Abdelaziz, R. Ben, Olschwang, S., Desvignes, J. P., … Mrad, R. (2017). Further Evidence for the Implication of LZTR1, a Gene not Associated with the Ras-Mapk Pathway, in the Pathogenesis of Noonan Syndrome. *Biology and Medicine*, *09*(06), 4–7. https://doi.org/10.4172/0974-8369.1000414

Giampietro, C., Deflorian, G., Gallo, S., Di Matteo, A., Pradella, D., Bonomi, S., … Ghigna, C. (2015). The alternative splicing factor Nova2 regulates vascular development and lumen formation. *Nature Communications*, *6*, 1–15. https://doi.org/10.1038/ncomms9479

Gilissen, C., Hehir-Kwa, J. Y., Thung, D. T., van de Vorst, M., van Bon, B. W. M., Willemsen, M. H., … Veltman, J. A. (2014). Genome sequencing identifies major causes of severe intellectual disability. *Nature*, *511*(7509), 344–347. https://doi.org/10.1038/nature13394

Gollob, M. H., Jones, D. L., Krahn, A. D., Danis, L., Gong, X.-Q., Shao, Q., … Bai, D. (2006). Somatic Mutations in the Connexin 40 Gene (*GJA5*) in Atrial Fibrillation. *New England Journal of Medicine*, *354*(25), 2677–2688. https://doi.org/10.1056/NEJMoa052800

Golzio, C., Havis, E., Daubas, P., Nuel, G., Babarit, C., Munnich, A., … Etchevers, H. C. (2012). ISL1 Directly Regulates FGF10 Transcription during Human Cardiac Outflow Formation. *PLoS ONE*, *7*(1), e30677. https://doi.org/10.1371/journal.pone.0030677

Hafner, C., Lopez-Knowles, E., Luis, N. M., Toll, A., Baselga, E., Fernandez-Casado, A., … Real, F. X. (2007). Oncogenic PIK3CA mutations occur in epidermal nevi and seborrheic keratoses with a characteristic mutation pattern. *Proceedings of the National Academy of Sciences*, *104*(33), 13450–13454. https://doi.org/10.1073/pnas.0705218104

Hafner, C., van Oers, J. M. M., Vogt, T., Landthaler, M., Stoehr, R., Blaszyk, H., … Hartmann, A. (2006). Mosaicism of activating FGFR3 mutations in human skin causes epidermal nevi. *The Journal of Clinical Investigation*, *116*(8), 2201–2207. https://doi.org/10.1172/JCI28163

Han, X., Chen, S., Flynn, E., Wu, S., Wintner, D., & Shen, Y. (2018). Distinct epigenomic patterns are associated with haploinsufficiency and predict risk genes of developmental disorders. *Nature Communications*, *9*(1), 2138. https://doi.org/10.1038/s41467-018-04552-7

Hanchard, N. A., Swaminathan, S., Bucasas, K., Furthner, D., Fernbach, S., Azamian, M. S., … McBride, K. L. (2016). A genome-wide association study of congenital cardiovascular left-sided lesions shows association with a locus on chromosome 20. *Human Molecular Genetics*, *25*(11), 2331–2341. https://doi.org/10.1093/hmg/ddw071

Happle, R. (1986). The McCune-Albright syndrome: a lethal gene surviving by mosaicism. *Clinical Genetics*, *29*(4), 321–324. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/3720010

Happle, R. (1993). Mosaicism in human skin. Understanding the patterns and mechanisms. *Archives of Dermatology*, *129*(11), 1460–1470. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8239703

Hartman, R. J., Rasmussen, S. A., Botto, L. D., Riehle-Colarusso, T., Martin, C. L., Cragan, J. D., … Correa, A. (2011). The Contribution of Chromosomal Abnormalities to Congenital Heart Defects: A Population-Based Study. *Pediatric Cardiology*, *32*(8), 1147–1157. https://doi.org/10.1007/s00246-011-0034-5

Hassold, T. J., & Jacobs, P. A. (1984). Trisomy in Man. *Annual Review of Genetics*, *18*(1), 69–97. https://doi.org/10.1146/annurev.ge.18.120184.000441

Heinrich, V., Stange, J., Dickhaus, T., Imkeller, P., Krüger, U., Bauer, S., … Krawitz, P. M. (2012). The allele distribution in next-generation sequencing data sets is accurately described as the result of a stochastic branching process. *Nucleic Acids Research*, *40*(6), 2426–2431. https://doi.org/10.1093/nar/gkr1073

Homsy, J., Zaidi, S., Shen, Y., Ware, J. S., Samocha, K. E., Karczewski, K. J., … Chung, W. K. (2015). De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. *Science*, *350*(6265), 1262–1266. https://doi.org/10.1126/science.aac9396

Hook, E. B., & Warburton, D. (1983). The distribution of chromosomal genotypes associated with Turner's syndrome: livebirth prevalence rates and evidence for diminished fetal mortality and severity in genotypes associated with structural X abnormalities or mosaicism. *Human Genetics*, *64*(1), 24–27. https://doi.org/10.1007/BF00289473

Hu, M., Sun, X.-J., Zhang, Y.-L., Kuang, Y., Hu, C.-Q., Wu, W.-L., … Chen, Z. (2010). Histone H3 lysine 36 methyltransferase Hypb/Setd2 is required for embryonic vascular remodeling. *Proceedings of the National Academy of Sciences*, *107*(7), 2956–2961. https://doi.org/10.1073/pnas.0915033107

Hu, Z., Shi, Y., Mo, X., Xu, J., Zhao, B., Lin, Y., … Shen, H. (2013). A genome-wide association study identifies two risk loci for congenital heart malformations in Han Chinese populations. *Nature Genetics*, *45*(7), 818–821. https://doi.org/10.1038/ng.2636

Huang, A. Y., Zhang, Z., Ye, A. Y., Dou, Y., Yan, L., Yang, X., … Wei, L. (2017). MosaicHunter: accurate detection of postzygotic single-nucleotide mosaicism through next-generation sequencing of unpaired, trio, and paired samples. *Nucleic Acids Research*, *45*(10), e76–e76. https://doi.org/10.1093/nar/gkx024

Ioannidis, N. M., Rothstein, J. H., Pejaver, V., Middha, S., McDonnell, S. K., Baheti, S., … Sieh, W. (2016). REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *The American Journal of Human Genetics*, *99*(4), 877–885. https://doi.org/10.1016/j.ajhg.2016.08.016

Jacobs, K. B., Yeager, M., Zhou, W., Wacholder, S., Wang, Z., Rodriguez-Santiago, B., … Chanock, S. J. (2012). Detectable clonal mosaicism and its relationship to aging and cancer. *Nature Genetics*, *44*(6), 651–658. https://doi.org/10.1038/ng.2270

Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., … Farh, K. K.-H. (2019). Predicting Splicing from Primary Sequence with Deep Learning. *Cell*, *176*(3), 535-548.e24. https://doi.org/10.1016/J.CELL.2018.12.015

Jaiswal, S., Fontanillas, P., Flannick, J., Manning, A., Grauman, P. V, Mar, B. G., … Ebert, B. L. (2014). Age-related clonal hematopoiesis associated with adverse outcomes. *The New England Journal of Medicine*, *371*(26), 2488–2498. https://doi.org/10.1056/NEJMoa1408617

Jaiswal, S., Natarajan, P., Silver, A. J., Gibson, C. J., Bick, A. G., Shvartz, E., … Ebert, B. L. (2017). Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular Disease. *New England Journal of Medicine*, *377*(2), 111–121. https://doi.org/10.1056/nejmoa1701719

Jamuar, S. S., Lam, A.-T. N., Kircher, M., D'Gama, A. M., Wang, J., Barry, B. J., … Walsh, C. A. (2014). Somatic Mutations in Cerebral Cortical Malformations. *New England Journal of Medicine*, *371*(8), 733–743. https://doi.org/10.1056/NEJMoa1314432

Jenkins, K. J., Correa, A., Feinstein, J. A., Botto, L., Britt, A. E., Daniels, S. R., … American Heart Association Council on Cardiovascular Disease in the Young. (2007). Noninherited Risk Factors and Congenital Cardiovascular Defects: Current Knowledge. *Circulation*, *115*(23), 2995–3014. https://doi.org/10.1161/CIRCULATIONAHA.106.183216

Jin, S. C., Homsy, J., Zaidi, S., Lu, Q., Morton, S., DePalma, S. R., … Brueckner, M. (2017). Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. *Nature Genetics*, *49*(11), 1593–1601. https://doi.org/10.1038/ng.3970

Kaltman, J. R., Di, H., Tian, Z., & Rychik, J. (2005). Impact of congenital heart disease on cerebrovascular blood flow dynamics in the fetus. *Ultrasound in Obstetrics and Gynecology*, *25*(1), 32–36. https://doi.org/10.1002/uog.1785

KANTER, R. J., & GARSON, A. (1997). Atrial Arrhythmias During Chronic Follow-Up of Surgery for Complex Congenital Heart Disease. *Pacing and Clinical Electrophysiology*, *20*(2), 502–511. https://doi.org/10.1111/j.1540-8159.1997.tb06207.x

Kennedy, M. P., Omran, H., Leigh, M. W., Dell, S., Morgan, L., Molina, P. L., … Knowles, M. R. (2007). Congenital Heart Disease and Other Heterotaxic Defects in a Large Cohort of Patients With Primary Ciliary Dyskinesia. *Circulation*, *115*(22), 2814–2821. https://doi.org/10.1161/CIRCULATIONAHA.106.649038

Khairy, P., Dore, A., Talajic, M., Dubuc, M., Poirier, N., Roy, D., & Mercier, L.-A. (2006). Arrhythmias in adult congenital heart disease. *Expert Review of Cardiovascular Therapy*, *4*(1), 83–95. https://doi.org/10.1586/14779072.4.1.83

Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., … Kathiresan, S. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, *50*(9), 1219–1224. https://doi.org/10.1038/s41588-018-0183-z

Kim, D. S., Kim, J. H., Burt, A. A., Crosslin, D. R., Burnham, N., Kim, C. E., … Jarvik, G. P. (2016). Burden of potentially pathologic copy number variants is higher in children with isolated congenital heart disease and significantly impairs covariate-adjusted transplant-free survival. *The Journal of Thoracic and Cardiovascular Surgery*, *151*(4), 1147-1151.e4. https://doi.org/10.1016/j.jtcvs.2015.09.136

King, D. A., Jones, W. D., Crow, Y. J., Dominiczak, A. F., Foster, N. A., Gaunt, T. R., … Hurles, M. E. (2015). Mosaic structural variation in children with developmental disorders. *Human Molecular Genetics*, *24*(10), 2733–2745. https://doi.org/10.1093/hmg/ddv033

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., … Wilson, R. K. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, *22*(3), 568–576. https://doi.org/10.1101/gr.129684.111

König, A., Happle, R., Bornholdt, D., Engel, H., & Grzeschik, K. H. (2000). Mutations in the NSDHL gene, encoding a 3beta-hydroxysteroid dehydrogenase, cause CHILD syndrome. *American Journal of Medical Genetics*, *90*(4), 339–346. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10710235

Krupp, D. R., Barnard, R. A., Duffourd, Y., Evans, S. A., Mulqueen, R. M., Bernier, R., … O'Roak, B. J. (2017). Exonic Mosaic Mutations Contribute Risk for Autism Spectrum Disorder. *The American Journal of Human Genetics*, *101*(3), 369–390. https://doi.org/10.1016/j.ajhg.2017.07.016

Kurahashi, H., Akagi, K., Inazawa, J., Ohta, T., Niikawa, N., Kayatani, F., … Nishisho, I. (1995). Isolation and characterization of a novel gene deleted in DiGeorge syndrome. *Human Molecular Genetics*, *4*(4), 541–549. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7633402

Kurek, K. C., Luks, V. L., Ayturk, U. M., Alomari, A. I., Fishman, S. J., Spencer, S. A., … Warman, M. L. (2012). Somatic Mosaic Activating Mutations in PIK3CA Cause CLOVES Syndrome. *The American Journal of Human Genetics*, *90*(6), 1108–1115. https://doi.org/10.1016/j.ajhg.2012.05.006

Lai, Z., Markovets, A., Ahdesmaki, M., Chapman, B., Hofmann, O., McEwen, R., … Dry, J. R. (2016). VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Research*, *44*(11), e108–e108. https://doi.org/10.1093/nar/gkw227

Larson, D. E., Harris, C. C., Chen, K., Koboldt, D. C., Abbott, T. E., Dooling, D. J., … Ding, L. (2012). SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics*, *28*(3), 311–317. https://doi.org/10.1093/bioinformatics/btr665

Lauriat, T. L., Shiue, L., Haroutunian, V., Verbitsky, M., Ares, M., Ospina, L., & McInnes, L. A. (2008). Developmental expression profile ofquaking, a candidate gene for schizophrenia, and its target genes in human prefrontal cortex and hippocampus shows regional specificity. *Journal of Neuroscience Research*, *86*(4), 785–796. https://doi.org/10.1002/jnr.21534

Laurie, C. C., Laurie, C. A., Rice, K., Doheny, K. F., Zelnick, L. R., McHugh, C. P., … Weir, B. S. (2012). Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nature Genetics*, *44*(6), 642–650. https://doi.org/10.1038/ng.2271

Lee, J. H., Huynh, M., Silhavy, J. L., Kim, S., Dixon-Salazar, T., Heiberg, A., … Gleeson, J. G. (2012). De novo somatic mutations in components of the PI3K-AKT3-mTOR pathway cause hemimegalencephaly. *Nature Genetics*, *44*(8), 941–945. https://doi.org/10.1038/ng.2329

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., … Consortium, E. A. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), 285–291. https://doi.org/10.1038/nature19057

Lerman, B. B., Dong, B., Stein, K. M., Markowitz, S. M., Linden, J., & Catanzaro, D. F. (1998). Right ventricular outflow tract tachycardia due to a somatic cell mutation in G protein subunitalphai2. *Journal of Clinical Investigation*, *101*(12), 2862–2868. https://doi.org/10.1172/JCI1582

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics (Oxford, England)*, *27*(21), 2987–2993. https://doi.org/10.1093/bioinformatics/btr509

Li, Y., Klena, N. T., Gabriel, G. C., Liu, X., Kim, A. J., Lemke, K., … Lo, C. W. (2015). Global genetic analysis in mice unveils central role for cilia in congenital heart disease. *Nature*, *521*(7553), 520–524. https://doi.org/10.1038/nature14269

Lim, E. T., Uddin, M., De Rubeis, S., Chan, Y., Kamumbu, A. S., Zhang, X., … Walsh, C. A. (2017). Rates, distribution and implications of postzygotic mosaic mutations in autism spectrum disorder. *Nature Neuroscience*, *20*(9), 1217–1224. https://doi.org/10.1038/nn.4598

Lindhurst, M. J., Parker, V. E. R., Payne, F., Sapp, J. C., Rudge, S., Harris, J., … Semple, R. K. (2012). Mosaic overgrowth with fibroadipose hyperplasia is caused by somatic activating mutations in PIK3CA. *Nature Genetics*, *44*(8), 928–933. https://doi.org/10.1038/ng.2332

Lonigro, R. J., Grasso, C. S., Robinson, D. R., Jing, X., Wu, Y.-M., Cao, X., … Chinnaiyan, A. M. (2011). Detection of Somatic Copy Number Alterations in Cancer Using Targeted Exome Capture Sequencing. *Neoplasia*, *13*(11), 1019-IN21. https://doi.org/10.1593/NEO.111252

Maertens, O., De Schepper, S., Vandesompele, J., Brems, H., Heyns, I., Janssens, S., … Messiaen, L. (2007). Molecular Dissection of Isolated Disease Features in Mosaic Neurofibromatosis Type 1. *The American Journal of Human Genetics*, *81*(2), 243–251. https://doi.org/10.1086/519562

Mahle, W. T., Tavani, F., Zimmerman, R. A., Nicolson, S. C., Galli, K. K., Gaynor, J. W., … Kurth, C. D. (2002). An MRI study of neurological injury before and after congenital heart surgery. *Circulation*, *106*(12 Suppl 1), I109-14. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12354718

Manheimer, K. B., Richter, F., Edelmann, L. J., D'Souza, S. L., Shi, L., Shen, Y., … Gelb, B. D. (2018). Robust identification of mosaic variants in congenital heart disease. *Human Genetics*, *137*(2), 183–193. https://doi.org/10.1007/s00439-018-1871-6

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., … Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*(7265), 747–753. https://doi.org/10.1038/nature08494

Marelli, A., Miller, S. P., Marino, B. S., Jefferson, A. L., & Newburger, J. W. (2016). Brain in Congenital Heart Disease Across the Lifespan. *Circulation*, *133*(20), 1951–1962. https://doi.org/10.1161/CIRCULATIONAHA.115.019881

Marino, B. S., Lipkin, P. H., Newburger, J. W., Peacock, G., Gerdes, M., Gaynor, J. W., … American Heart Association Congenital Heart Defects Committee, Council on Cardiovascular Disease in the Young, Council on Cardiovascular Nursing, and Stroke Council. (2012). Neurodevelopmental Outcomes in Children With Congenital Heart Disease: Evaluation and Management. *Circulation*, *126*(9), 1143–1172. https://doi.org/10.1161/CIR.0b013e318265ee8a

McDonald, J., Wooderchak-Donahue, W. L., Henderson, K., Paul, E., Morris, A., & Bayrak-Toydemir, P. (2018). Tissue-specific mosaicism in hereditary hemorrhagic telangiectasia: Implications for genetic testing in families. *American Journal of Medical Genetics, Part A*, *176*(7), 1618–1621. https://doi.org/10.1002/ajmg.a.38695

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Messiaen, L., Vogt, J., Bengesser, K., Fu, C., Mikhail, F., Serra, E., … Kehrer-Sawatzki, H. (2011). Mosaic type-1 NF1 microdeletions as a cause of both generalized and segmental neurofibromatosis type-1 (NF1). *Human Mutation*, *32*(2), 213–219. https://doi.org/10.1002/humu.21418

Miller, S. P., McQuillen, P. S., Hamrick, S., Xu, D., Glidden, D. V., Charlton, N., … Vigneron, D. B. (2007). Abnormal Brain Development in Newborns with Congenital Heart Disease. *New England Journal of Medicine*, *357*(19), 1928–1938. https://doi.org/10.1056/NEJMoa067393

Mone, S. M., Gillman, M. W., Miller, T. L., Herman, E. H., & Lipshultz, S. E. (2004). Effects of environmental exposures on the cardiovascular system: prenatal period through adolescence. *Pediatrics*, *113*(4 Suppl), 1058–1069. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15060200

Moorman, A. (2003). Development of the Heart: (1) Formation of the Cardiac Chambers and Arterial Trunks. *Heart*, *89*(7), 806–814. https://doi.org/10.1136/heart.89.7.806

Nawa, M., & Matsuoka, M. (2013). KCTD20, a relative of BTBD10, is a positive regulator of Akt. *BMC Biochemistry*, *14*(1), 27. https://doi.org/10.1186/1471-2091-14-27

Noveroske, J. K., Lai, L., Gaussin, V., Northrop, J. L., Nakamura, H., Hirschi, K. K., & Justice, M. J. (2002). Quaking is essential for blood vessel development. *Genesis (New York, N.Y. : 2000)*, *32*(3), 218–230. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11892011

Øyen, N., Poulsen, G., Boyd, H. A., Wohlfahrt, J., Jensen, P. K. A., & Melbye, M. (2009). National time trends in congenital heart defects, Denmark, 1977-2005. *American Heart Journal*, *157*(3), 467-473.e1. https://doi.org/10.1016/j.ahj.2008.10.017

Poduri, A., Evrony, G. D., Cai, X., Elhosary, P. C., Beroukhim, R., Lehtinen, M. K., … Walsh, C. A. (2012). Somatic Activation of AKT3 Causes Hemispheric Developmental Brain Malformations. *Neuron*, *74*(1), 41–48. https://doi.org/10.1016/j.neuron.2012.03.010

Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., … DePristo, M. A. (2018). A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology*, *36*(10), 983–987. https://doi.org/10.1038/nbt.4235

Prabhu, S., Jenny, B., James, H., & Provenzano, S. (2015). Mosaic 22q11.2 Deletion and Tetralogy of Fallot With Absent Pulmonary Valve. *World Journal for Pediatric and Congenital Heart Surgery*, *6*(2), 342–345. https://doi.org/10.1177/2150135114561686

Preuss, C., Capredon, M., Wünnemann, F., Chetaille, P., Prince, A., Godard, B., … Andelfinger, G. (2016). Family Based Whole Exome Sequencing Reveals the Multifaceted Role of Notch Signaling in Congenital Heart Disease. *PLOS Genetics*, *12*(10), e1006335. https://doi.org/10.1371/journal.pgen.1006335

Priest, J. R., Gawad, C., Kahlig, K. M., Yu, J. K., O'Hara, T., Boyle, P. M., … Ashley, E. A. (2016). Early somatic mosaicism is a rare cause of long-QT syndrome. *Proceedings of the National Academy of Sciences*, *113*(41), 11555–11560. https://doi.org/10.1073/pnas.1607187113

Purcell, S. M., Moran, J. L., Fromer, M., Ruderfer, D., Solovieff, N., Roussos, P., … Sklar, P. (2014). A polygenic burden of rare disruptive mutations in schizophrenia. *Nature*, *506*(7487), 185–190. https://doi.org/10.1038/nature12975

Raffel, L. J., Mohandas, T., Rimoin, D. L., Opitz, J. M., & Reynolds, J. F. (1986). Chromosomal mosaicism in the Killian/Teschler-Nicola syndrome. *American Journal of Medical Genetics*, *24*(4), 607–611. https://doi.org/10.1002/ajmg.1320240404

Ramsdell, A. F. (2005). Left–right asymmetry and congenital cardiac defects: Getting to the heart of the matter in vertebrate left–right axis determination. *Developmental Biology*, *288*(1), 1–20. https://doi.org/10.1016/J.YDBIO.2005.07.038

Ramu, A., Noordam, M. J., Schwartz, R. S., Wuster, A., Hurles, M. E., Cartwright, R. A., & Conrad, D. F. (2013). DeNovoGear: de novo indel and point mutation discovery and phasing. *Nature Methods*, *10*(10), 985–987. https://doi.org/10.1038/nmeth.2611

Reich, D. E., & Lander, E. S. (2001). On the allelic spectrum of human disease. *Trends in Genetics : TIG*, *17*(9), 502–510. https://doi.org/10.1016/s0168-9525(01)02410-6

Reller, M. D., Strickland, M. J., Riehle-Colarusso, T., Mahle, W. T., & Correa, A. (2008). Prevalence of Congenital Heart Defects in Metropolitan Atlanta, 1998-2005. *The Journal of Pediatrics*, *153*(6), 807–813. https://doi.org/10.1016/j.jpeds.2008.05.059

Ren, K., Yuan, J., Yang, M., Gao, X., Ding, X., Zhou, J., … Zhang, J. (2014). KCTD10 Is Involved in the Cardiovascular System and Notch Signaling during Early Embryonic Development. *PLoS ONE*, *9*(11), e112275. https://doi.org/10.1371/journal.pone.0112275

Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S. R. F., Wilkie, A. O. M., … Lunter, G. (2014). Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature Genetics*, *46*(8), 912–918. https://doi.org/10.1038/ng.3036

Ripatti, S., Tikkanen, E., Orho-Melander, M., Havulinna, A. S., Silander, K., Sharma, A., … Kathiresan, S. (2010). A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses. *The Lancet*, *376*(9750), 1393–1400. https://doi.org/10.1016/S0140-6736(10)61267-6

Rivière, J.-B., Mirzaa, G. M., O'Roak, B. J., Beddaoui, M., Alcantara, D., Conway, R. L., … Dobyns, W. B. (2012). De novo germline and postzygotic mutations in AKT3, PIK3R2 and PIK3CA cause a spectrum of related megalencephaly syndromes. *Nature Genetics*, *44*(8), 934–940. https://doi.org/10.1038/ng.2331

Roth, A., Ding, J., Morin, R., Crisan, A., Ha, G., Giuliany, R., … Shah, S. P. (2012). JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. *Bioinformatics*, *28*(7), 907–913. https://doi.org/10.1093/bioinformatics/bts053

Ruggieri, M., & Huson, S. M. (2001). The clinical and diagnostic implications of mosaicism in the neurofibromatoses. *Neurology*, *56*(11), 1433–1443. https://doi.org/10.1212/WNL.56.11.1433

Sala Frigerio, C., Lau, P., Troakes, C., Deramecourt, V., Gele, P., Van Loo, P., … De Strooper, B. (2015). On the identification of low allele frequency mosaic mutations in the brains of Alzheimer's disease patients. *Alzheimer's & Dementia*, *11*(11), 1265–1276. https://doi.org/10.1016/j.jalz.2015.02.007

Sallman, D. A., Komrokji, R., Vaupel, C., Cluzeau, T., Geyer, S. M., McGraw, K. L., … Padron, E. (2016). Impact of TP53 mutation variant allele frequency on phenotype and outcomes in myelodysplastic syndromes. *Leukemia*, *30*(3), 666–673. https://doi.org/10.1038/leu.2015.304

Sampson, J., Jacobs, K., Yeager, M., Chanock, S., & Chatterjee, N. (2011). Efficient study design for next generation sequencing. *Genetic Epidemiology*, *35*(4), n/a-n/a. https://doi.org/10.1002/gepi.20575

Satterstrom, F. K., Kosmicki, J. A., Wang, J., Breen, M. S., Rubeis, S. De, An, J.-Y., … Buxbaum, J. D. (2019). Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *BioRxiv*, 484113. https://doi.org/10.1101/484113

Saunders, C. T., Wong, W. S. W., Swamy, S., Becq, J., Murray, L. J., & Cheetham, R. K. (2012). Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics*, *28*(14), 1811–1817. https://doi.org/10.1093/bioinformatics/bts271

Schott, J., Benson, D. W., Basson, C. T., Pease, W., Silberbach, G. M., Moak, J. P., … Seidman, J. G. (1998). Congenital Heart Disease Caused by Mutations in the Transcription Factor NKX2-5. *Science*, *281*(5373), 108–111. https://doi.org/10.1126/science.281.5373.108

Shirley, M. D., Tang, H., Gallione, C. J., Baugher, J. D., Frelin, L. P., Cohen, B., … Pevsner, J. (2013). Sturge–Weber Syndrome and Port-Wine Stains Caused by Somatic Mutation in *GNAQ*. *New England Journal of Medicine*, *368*(21), 1971–1979. https://doi.org/10.1056/NEJMoa1213507

Sifrim, A., Hitz, M.-P., Wilsdon, A., Breckpot, J., Turki, S. H. Al, Thienpont, B., … Hurles, M. E. (2016). Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing. *Nature Genetics*, *48*(9), 1060–1065. https://doi.org/10.1038/ng.3627

Slough, J., Cooney, L., & Brueckner, M. (2008). Monocilia in the embryonic mouse heart suggest a direct role for cilia in cardiac morphogenesis. *Developmental Dynamics*, *237*(9), 2304–2314. https://doi.org/10.1002/dvdy.21669

Smith, C. L., Blake, J. A., Kadin, J. A., Richardson, J. E., Bult, C. J., & Mouse Genome Database Group. (2018). Mouse Genome Database (MGD)-2018: knowledgebase for the laboratory mouse. *Nucleic Acids Research*, *46*(D1), D836–D842. https://doi.org/10.1093/nar/gkx1006

Smith, K. S., Yadav, V. K., Pei, S., Pollyea, D. A., Jordan, C. T., & De, S. (2016). SomVarIUS: somatic variant identification from unpaired tissue samples. *Bioinformatics*, *32*(6), 808–813. https://doi.org/10.1093/bioinformatics/btv685

Soubrier, F., Chung, W. K., Machado, R., Grünig, E., Aldred, M., Geraci, M., … Humbert, M. (2013). Genetics and Genomics of Pulmonary Arterial Hypertension. *Journal of the American College of Cardiology*, *62*(25), D13–D21. https://doi.org/10.1016/J.JACC.2013.10.035

Stevens, K. N., Hakonarson, H., Kim, C. E., Doevendans, P. A., Koeleman, B. P. C., Mital, S., … Gruber, P. J. (2010). Common Variation in ISL1 Confers Genetic Susceptibility for Human Congenital Heart Disease. *PLoS ONE*, *5*(5), e10855. https://doi.org/10.1371/journal.pone.0010855

Stosser, M. B., Lindy, A. S., Butler, E., Retterer, K., Piccirillo-Stosser, C. M., Richard, G., & McKnight, D. A. (2018). High frequency of mosaic pathogenic variants in genes causing epilepsy-related neurodevelopmental disorders. *Genetics in Medicine*, *20*(4), 403–410. https://doi.org/10.1038/gim.2017.114

Sun, J. X., He, Y., Sanford, E., Montesion, M., Frampton, G. M., Vignot, S., … Yelensky, R. (2018). A computational approach to distinguish somatic vs. germline origin of genomic alterations from deep sequencing of cancer specimens without a matched normal. *PLOS Computational Biology*, *14*(2), e1005965. https://doi.org/10.1371/journal.pcbi.1005965

Symoens, S., Steyaert, W., Demuynck, L., De Paepe, A., Diderich, K. E. M., Malfait, F., & Coucke, P. J. (2017). Tissue-specific mosaicism for a lethal osteogenesis imperfecta

COL1A1 mutation causes mild OI/EDS overlap syndrome. *American Journal of Medical Genetics, Part A*, *173*(4), 1047–1050. https://doi.org/10.1002/ajmg.a.38135

Tanaka, N., Izawa, K., Saito, M. K., Sakuma, M., Oshima, K., Ohara, O., … Heike, T. (2011). High incidence of NLRP3 somatic mosaicism in patients with chronic infantile neurologic, cutaneous, articular syndrome: Results of an international multicenter collaborative study. *Arthritis & Rheumatism*, *63*(11), 3625–3632. https://doi.org/10.1002/art.30512

Tong, X., Zu, Y., Li, Z., Li, W., Ying, L., Yang, J., … Zhang, B. (2014). Kctd10 regulates heart morphogenesis by repressing the transcriptional activity of Tbx5a in zebrafish. *Nature Communications*, *5*, 1–10. https://doi.org/10.1038/ncomms4153

van der Linde, D., Konings, E. E. M., Slager, M. A., Witsenburg, M., Helbing, W. A., Takkenberg, J. J. M., & Roos-Hesselink, J. W. (2011). Birth Prevalence of Congenital Heart Disease Worldwide. *Journal of the American College of Cardiology*, *58*(21), 2241–2247. https://doi.org/10.1016/j.jacc.2011.08.025

Wallis, G. A., Starman, B. J., Zinn, A. B., & Byers, P. H. (1990). Variable expression of osteogenesis imperfecta in a nuclear family is explained by somatic mosaicism for a lethal point mutation in the alpha 1(I) gene (COL1A1) of type I collagen in a parent. *American Journal of Human Genetics*, *46*(6), 1034–1040. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2339700

Weinstein, M. M., Kang, T., Lachman, R. S., Bamshad, M., Nickerson, D. A., Krakow, D., & Cohn, D. H. (2016). Somatic mosaicism for a lethal *TRPV4* mutation results in non-lethal metatropic dysplasia. *American Journal of Medical Genetics Part A*, *170*(12), 3298–3302. https://doi.org/10.1002/ajmg.a.37942

Weismann, C. G., Hager, A., Kaemmerer, H., Maslen, C. L., Morris, C. D., Schranz, D., … Gelb, B. D. (2005). PTPN11 mutations play a minor role in isolated congenital heart disease. *American Journal of Medical Genetics Part A*, *136A*(2), 146–151. https://doi.org/10.1002/ajmg.a.30789

Wiedemann, H.-R., Burgio, G. R., Aldenhoff, P., Kunze, J., Kaufmann, H. J., & Schirg, E. (1983). The proteus syndrome. *European Journal of Pediatrics*, *140*(1), 5–12. https://doi.org/10.1007/BF00661895

Wilm, A., Aw, P. P. K., Bertrand, D., Yeo, G. H. T., Ong, S. H., Wong, C. H., … Nagarajan, N. (2012). LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Research*, *40*(22), 11189–11201. https://doi.org/10.1093/nar/gks918

Xie, M., Lu, C., Wang, J., McLellan, M. D., Johnson, K. J., Wendl, M. C., … Ding, L. (2014). Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nature Medicine*, *20*(12), 1472–1478. https://doi.org/10.1038/nm.3733

Xu, C. (2018). A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Computational and Structural Biotechnology Journal, 16*, 15–24. https://doi.org/10.1016/j.csbj.2018.01.003

Yamamoto, G. L., Aguena, M., Gos, M., Hung, C., Pilch, J., Fahiminiya, S., … Bertola, D. R. (2015). Rare variants in SOS2 and LZTR1 are associated with Noonan syndrome. *Journal of Medical Genetics, 52*(6), 413–421. https://doi.org/10.1136/jmedgenet-2015-103018

Yang, Q., Chen, H., Correa, A., Devine, O., Mathews, T. J., & Honein, M. A. (2006). Racial differences in infant mortality attributable to birth defects in the United States, 1989–2002. *Birth Defects Research Part A: Clinical and Molecular Teratology, 76*(10), 706–713. https://doi.org/10.1002/bdra.20308

Zaidi, S., & Brueckner, M. (2017). Genetics and Genomics of Congenital Heart Disease. *Circulation Research*, *120*(6), 923–940. https://doi.org/10.1161/CIRCRESAHA.116.309140

Zaidi, S., Choi, M., Wakimoto, H., Ma, L., Jiang, J., Overton, J. D., … Lifton, R. P. (2013). De novo mutations in histone-modifying genes in congenital heart disease. *Nature, 498*(7453), 220–223. https://doi.org/10.1038/nature12141

Appendix

For Tables S1 to S12, please refer to the Supplementary Material in our bioRxiv preprint (https://www.biorxiv.org/content/10.1101/733105v1)

For all code used in this dissertation, please refer to our GitHub repositories: (https://github.com/ShenLab/mosaicism)