Computational approaches identify novel risk loci and interactions in heart defects

by
Maureen Pittman

DISSERTATION
Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY

in

Biological and Medical Informatics

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:
Katherine Pollard
9B0DDB91029D4F1...

Katherine Pollard
Chair

DocuSigned by:
Deepak Srivastava
[illegible]4DE...

Deepak Srivastava

DocuSigned by:
Arthur Jeremy Willsey
89C6079C878F433...

Arthur Jeremy Willsey

Committee Members

Dedicated to the village that raised me: Pittmans, Searles, Marleys, Tighes, and all.

ACKNOWLEDGEMENTS

I am indebted to all the many people who made this graduate work possible, and everyone who made life delightful while I did it.

First I thank my family, without whose support I would not have been able to pursue (much less achieve) a PhD. I’m grateful to my dad Faran for teaching me skepticism, candor, and curiosity about all the world’s critters. My mom Marguerite taught me how to read, and more importantly how to love doing it. By watching her example I learned the most important virtues that pushed me through twenty-two years of schooling: discipline, sacrifice, and love. And of course my siblings Helen, Thomas, and Rachel played a huge role in raising me into the person I am today. Among countless other things, I thank them for modeling success, blazing a path through an otherwise wild childhood and letting me tag along to read the big kid books. Pittmaniacs, I wouldn’t have made it this far without your example, support, and understanding.

Other family members in need of shouting out are ELOPSOcons Chris, Annemarieke, and Brady for the fun, the support, and the D&D. Thanks especially to Chris, who with Helen has been feeding me since middle school. I am grateful also to Hannah for inspiring me with her creativity, from age five to age thirty, and for sharing comics, jobs, classes, book recs, outfits, a room for four years and an apartment for two more. To all the extended relatives who helped raise me, especially my mom’s sisters: your labor and your love are endlessly appreciated.

There have been too many scientific mentors throughout my pre-graduate years to thank them all, but specific mention must go to Dr. Carolyn Fisher, whose high school chemistry class at Brookland-Cayce was the first class that inspired me to consider a career in science. I owe many thanks as well to Richard Kamens and Shabbir Gheewala, who introduced me to the process of scientific research and discovery. Terry Furey’s classes at UNC Chapel Hill sparked my interest in genetics and coding, further cultivated by Holly Mortensen, who advised my first independent research project and first-author manuscript submission. I am grateful to them all (and many more) for shaping the foundation of my scientific understanding.

In my time at UCSF and Gladstone, I have been repeatedly blown away by the scientific excellence of my peers and instructors, who nevertheless are extremely friendly and down-to-earth people. I thank my senior students in the Biological and Medical Informatics program who welcomed me and brought me up to speed where necessary, especially Tamas Nagy and Jason Town; and to my peers who helped with problem sets, discussed the latest biomedical breakthroughs, and tempted me away from lab to try new restaurants in the city. Thanks especially to the faculty who remained invested in student success and wellbeing throughout my time in the program, especially Tony Capra and Ryan Hernandez.

I was especially fortunate to be affiliated with Gladstone Institutes, where I had access to the time, attention, and feedback of titans across multiple fields of disease research. I am grateful for the guidance of Benoit Bruneau, whose expertise and insight into the epigenetics of heart development were invaluable to the completion of this work. I must also acknowledge my talented co-authors, especially Barbara Gonzalez Teran and Kihyun Lee, who as postdoctoral scholars in the Srivastava lab taught me so much about biomedical research. Thanks to both of them for showing me the art and craftsmanship of a well-designed experiment, and the ins and outs of leading and publishing a large-scale independent research project. I am a better and more versatile scientist thanks to their mentorship.

I was lucky to have a committee that consistently advocated for my needs and my development as an independent scientist. Dr. Deepak Srivastava joined my committee and remains involved in my research, despite running his large and active research lab (and also running Gladstone Institutes as president). The resources and feedback he provided throughout my graduate degree shaped this project every step of the way. I am so grateful that Deepak has continued to invest time and energy into my career, as he helps usher the final manuscript from my graduate work into existence. I also thank committee member Dr. Jeremy Willsey for his remarkable mentorship, engagement, and enthusiasm. I left every one of our meetings bursting with energy and excitement about my project, no matter how tired I'd felt at the start of it.

More than anyone else, the people who defined my experience in grad school were the members of the Pollard Lab. Thanks to all the alumni who established the spirit of the lab as I arrived to it, especially Patrick, Geoff, Pawel, Katheleen, and Jason; and current research scientists and postdocs Sean, Chunyu, Byron, Shuzhen, Zhirui, and Veronika. Special thanks go to Dr. Abigail Lind, my UNC bud, neighbor and carpooler, and role model for young womanhood in science. I'm so grateful for the graduate students who joined the lab with me and shortly after: I will miss having Annamarie, Calla, Laura, and Miriam to ask for advice and stress over grant applications and presentations together. Thanks also to the upcoming students Amanda, Katie G, and Shiron, who are doing amazing work and inspire me to greater science and mentorship. After five-and-a-half years of Pollard lab journal clubs, lab meetings, and interactions, I am a better scientist, communicator, and friend. Thank you all.

Of course the Pollard group can only attract such great people because Katie Pollard herself sets the example for inclusivity and curiosity. Beyond the way-cool science that first attracted me to her lab, Katie is beloved by all of her trainees for her award-winning (for good reason) mentorship. As my PI, Katie has been steady as a rock, calm and empathetic, and has always put the needs and success of her trainees first. I faced plenty of inner struggles during my PhD: insecurities about my data, my work, my capabilities, my future. Katie always greeted my concerns with patience and understanding. Despite being director of GIDB, mother of a vibrant family, and wearing plenty of other hats I probably don't know about, she has always made a point to be present and responsive to any questions and needs. I am so grateful to her for making this work possible, and for defining my proficiencies as a scientist. I can't thank you enough, Katie. I am extremely proud to have graduated from the Pollard lab.

Finally, endless thanks and love to Federico, who kept me cheerful, sane, and under a roof for all five-plus years in San Francisco. This thesis would not have been written without his vital services as weighted blanket, pep-talker, and rubber duck.

CONTRIBUTIONS

The content of this dissertation is adapted from manuscripts both published and under review, as well as unpublished work. The analyses and conclusions included derive exclusively from my graduate work except where otherwise specified at the start of each chapter. For additional context, see the following works:

Gonzalez-Teran, Barbara, **Maureen Pittman**, Franco Felix, Reuben Thomas, Desmond Richmond-Buccola, Ruth Hüttenhain, Krishna Choudhary, et al. 2022. “Transcription Factor Protein Interactomes Reveal Genetic Determinants in Heart Disease.” *Cell* 185 (5): 794–814.e30.

Pittman, Maureen*, Kihyun Lee*, Deepak Srivastava, and Katherine S. Pollard. 2022. “An Oligogenic Inheritance Test Detects Risk Genes and Their Interactions in Congenital Heart Defects and Developmental Comorbidities.” *bioRxiv*. https://doi.org/10.1101/2022.04.08.487704.

Mengyao Yu, Andrew Harper, Matthew Aguirre, **Maureen Pittman**, Catherine Tcheandjieu, A. Dulguun, C.Grace, A. Goel, M. Farrall, K. Xiao, J. Engreitz, K.S. Pollard, H. Watkins, and J. Priest. “Genetic Determinants of Interventricular Septal Anatomy and the Risk of Ventricular Septal Defects and Hypertrophic Cardiomyopathy.” medRxiv. https://doi.org/10.1101/2021.04.19.21255650.

*indicates co-first authorship.

Computational approaches identify novel risk loci and interactions in heart defects

Maureen Pittman

ABSTRACT

Congenital heart defects (CHD) occur in nearly one percent of live births each year and are the leading cause of defect-associated infant mortality. In spite of the growing size of disease cohorts, the molecular underpinnings of most cases remain unexplained. Given its high recurrence rate in families, we expect much of this contribution to be found within patient genomes, but extensive genetic heterogeneity limits our ability to statistically confirm risk loci. Previously-validated causal mutations occur in a wide range of genes that encode for proteins in signaling and migration, chromatin remodelers that induce lineage specification, and transcription factors regulating the expression of these genes. In order to identify cryptic risk loci, my thesis has focused on creating novel computational approaches to overcome statistical challenges and broaden our understanding of the mechanisms that can lead to CHD. By integrating protein-protein interaction networks of cardiac transcription factors with whole exome sequencing, I showed that interactors are enriched for rare and de novo mutations in CHD patients. I developed a variant prioritization scheme for de novo variants, which identified a GLYR1 mutation that destabilizes its interaction with cardiac transcription factor GATA4. I describe GCOD, a novel algorithm that uses probabilistic modeling to identify sets of genes predicted to interact in the etiology of CHD, including a novel genetic interaction between GATA6 and POR. Finally, in addition to coding mutations, I aimed to assess whether disruption to chromatin organization contributes to disease by characterizing three CHD patient variants that I predicted would alter the regulatory landscape of heart-relevant genes. My work has increased our repertoire of known and suspected disease loci in CHD and related developmental co-morbidities, and provided evidence of oligogenic combinations and disrupted genome folding as a mechanism in CHD.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AP-MS - affinity purification and mass spectrometry

ASD - autism spectrum disorder

CADD - combined annotation-dependent depletion

CDS - coding sequence

CHD - congenital heart defect

CTCF - CCCTC-binding factor

cTF - cardiac transcription factor

CM - cardiomyocytes

CPs - cardiac progenitors

DD - developmental delay

DNV - *de novo* variant

FDR - false discovery rate

GATA - GATA binding proteins, named for their sequence motif

GT-PPI - GATA4-TBX5 combined interactome

HHE - high heart expression

IP - immunoprecipitation

KO - knock-out

LoF - loss of function

MAF - minor allele frequency

OR - odds ratio

PCGC - Pediatric Cardiac Genomics Consortium

pLI - probability of loss-of-function intolerance

SNV - single nucleotide variant

SV - structural variant

TAD - topologically-associating domain

TF - transcription factor

tpm - transcripts per million

VPS - variant prioritization score

WT - wild type

CHAPTER 1: INTRODUCTION

As DNA sequencing technology becomes more accessible and affordable, and the number of sequenced genomes grows, so too has our power to draw conclusions about the relationship between genetic variation and phenotypic outcome. One of the clearest and most urgent applications therein is to advance our understanding of the molecular underpinnings of disease, empowering medical researchers to develop therapeutics and offer personalized treatment informed by patient genetic variation. This thesis encompasses my research towards this goal, with a focus on the role of DNA folding and gene-gene interactions in the context of congenital heart defects. The following introductory chapter provides a basic summary of the relevant background concepts.

1.1 The role of chromatin conformation and transcription factors in gene expression

In the human body, each cell contains a copy of the 3.2 billion base pairs of DNA that make up our individual genome. While often represented as a linear string of nucleotides (written as the letters A, C, T, and G), in reality our DNA is wrapped around proteins and other molecules into complex three-dimensional shapes, in a structure collectively called chromatin. Chromatin's ability to be transiently organized into higher-order shapes is key to our multicellular nature - after all, if all cells have the same DNA instruction manual, how can we account for morphological differences across cell types, or our body's ability to respond to changes in its environment? Chromatin organization allows for the differential expression of genes in different spatio-temporal contexts, partly based on which DNA is accessible to be transcribed and ultimately translated into functional proteins.

The three-dimensional organization of the genome is therefore critical for transcriptional control, and this is accomplished through mechanisms at multiple scales. At the nuclear level, chromosomes localize according to gene density, such that gene-dense chromosomes occupy a central nuclear region, whereas gene-poor chromosomes localize to the nuclear periphery

(Cremer and Cremer 2001). At the chromosome level, active/accessible regions cluster together in physical space separately from the inactive/inaccessible regions where transcription is not occurring (Lieberman-Aiden et al. 2009). In a given cellular context, DNA regions co-associate in either the A (active) or B (inactive) compartment, in a pattern that alternates along the linear genome (Imakaev et al. 2012; Smith et al. 2016). DNA scaffolding proteins called histones are marked with different chemical modifications depending on compartment and transcriptional activity, and these active or repressive histone marks represent an important avenue of gene regulation (Allfrey, Faulkner, and Mirsky 1964; Bannister and Kouzarides 2011); that is, modifications to histone acetylation and methylation induced by changes in cellular context can convert a gene from being repressed to actively transcribed and vice versa (Kimura 2013). To broadly summarize, actively transcribed genes tend to cluster together in three-dimensional space within the nucleus, marked by chemical modifications that expose gene promoters to transcriptional machinery.

Another key mechanism of gene regulatory control is transcription factor (TF) binding. TFs are proteins that initiate, increase, or otherwise regulate the transcription of associated genes, usually by binding to the gene promoter and/or associated enhancer sequences (Banerji, Olson, and Schaffner 1983; Palstra and Grosveld 2012). TFs typically function in *trans,* that is, they are expressed and translated in order to localize to their target genes elsewhere in the genome (Reuveni et al. 2018). Enhancers can regulate multiple genes, can be selectively accessible depending on histone marks and other factors regarding cellular context, and are distal from the gene(s) they regulate (Moorthy et al. 2017; Calo and Wysocka 2013; Palstra and Grosveld 2012; Levine 2010; Riethoven 2010). It is thought that enhancers function partially through physical proximity (Bulger and Groudine 2011; Schoenfelder et al. 2015), and so they typically occur with their associated gene(s) in a topologically-associating domain (TAD), or a region characterized by higher DNA interaction frequency within the domain relative to loci outside of that domain (Nora et al. 2012; Dixon et al. 2012). TAD boundaries also function as insulators between genes and enhancer sequences with incompatible regulatory profiles (Furlong and Levine 2018).

Thus, enhancers extend regulatory control in a complex interplay between multiple factors in *cis* and *trans*. In *cis,* enhancers regulate gene expression via physical contact with the promoter and the transcription factors that bind both sites. Enhancers are limited in the contacts they can make by local (*cis*) DNA accessibility and 3D organization, for example to promoters within the same TAD. In *trans*, enhancer-promoter contact is itself regulated by the expression of transcription factors elsewhere in the genome, often in combinatorial fashion, leading to intricate regulatory networks of DNA accessibility, TF expression, and downstream transcription effects therein.

1.2 Transcriptional programs control cell and tissue fate during human development

During embryonic development, a single fertilized cell must replicate and differentiate into the set of complex tissues and organs that make up a human body. This is predictably an extremely complex process, requiring spatio-temporal control of cellular migration, internal and external cellular signaling, apoptosis and proliferation, and chromatin remodeling (Scarpa and Mayor 2016; Cooper 2000; Meier, Finch, and Evan 2000; Ho and Crabtree 2010; Gilbert 2000; Casamassimi and Ciccodicola 2019). Section 1.1 summarizes how trans-acting transcription factors (TFs) work together with cis-acting enhancers, insulators, and histone modification to accomplish regulatory control. These principles guide embryogenesis and cell fate specification.

From chromatin accessibility to gene expression to morphology, differentiated cells display phenotypic differences reflective of their eventual role in the body (Marstrand and Storey 2014; Mohammed et al. 2017). Cell identity is regulated in large part by TF expression, as demonstrated in systems where over-expression of specific TFs induces trans-differentiation across distinct cell types (Takahashi and Yamanaka 2006; Takahashi et al. 2007; Takeuchi and Bruneau 2009). One example that illustrates the interplay between chromatin accessibility, TF expression, and cell identity is the BAF chromatin remodeling complex (also called SWI/SNF), which is crucial in the regulation of early embryonic development (Wang et al. 1996; Euskirchen, Auerbach, and Snyder 2012; Heyao Zhang et al. 2021). The BAF complex repositions DNA-wrapped histones to render genomic regions accessible or inaccessible to transcription (Kwon et al. 1994; Wang et al. 1996; He et al. 2020), a process which has been especially studied in multipotency and proliferation during embryonic development (Lazar et al. 2020; van der Vaart et al. 2020; Laubscher et al. 2021). An essential aspect of BAF activity is its capability to bind various combinations of transcription factors, allowing for targeted genome localization dependent on TF availability (Toto, Puri, and Albini 2016; Tseng, Cabot, and Cabot 2017; Sun et al. 2018; Barisic et al. 2019).

In summary, the integrity of chromatin conformation and transcription factor (TF) regulatory networks are crucial to a zygote's successful development into a viable human embryo. During differentiation and specification, chromatin remodelers achieve context-specific action and genome localization by particular combinations of TF binding (Buchler, Gerland, and Hwa 2003; Kato et al. 2004; Vandel et al. 2019; Charest et al. 2020). The concept of combinatorial regulatory logic will be particularly relevant in the third chapter of this thesis, in which I describe oligogenic variant combinations that lead to disease.

1.3 Key pathways and genes involved in heart morphogenesis and congenital defects

In the developing embryo, cells that eventually give rise to the heart begin as mesodermal germ cells that express the transcription factor MESP1, which is required for migration of heart precursor cells and cardiac specification (Kitajima et al. 2000; Chan et al. 2013; Ajima et al. 2021). The greater part of these cardiac progenitor cells migrate to form the cardiac crescent, where they begin to express core cardiac transcription factors like GATA4 (Rossi et al. 2001; Oka et al. 2006) and TBX5 (Chapman et al. 1996; Bruneau et al. 1999). These TFs are highly expressed in cardiomyocytes, and moreover are capable of inducing cardiomyocyte identity in cardiac fibroblasts along with MEF2 (Ieda et al. 2010; Inagawa et al. 2012; Qian et al. 2012).

Unsurprisingly, damaging mutations in these and other cardiac TF genes cause congenital malformations in the heart (Werner et al. 2016; Tomita-Mitchell et al. 2007; Mori and Bruneau 2004; Qiao et al. 2017; McElhinney et al. 2003; Pierpont et al. 2018), presumably due to the disruption of transcriptional pathways regulated by those genes. Similarly, pathologies of chromatin remodeling and histone modification have been implicated in heart defects, like the BAF complex described above, the p300/CREB-binding protein histone acetyltransferase (H. M. Chan and La Thangue 2001; Ghosh 2020), and others reviewed in Lim, Foo, and Chen 2021. Clearly gene regulatory networks are necessary for cardiogenesis, and identifying the particular cellular processes and components that defect-associated TFs regulate will provide valuable insight into mechanisms of heart development and disease.

Previously-identified targets include several cilia genes (Li et al. 2015; Klena, Gibbs, and Lo 2017), which are additionally implicated in renal and brain congenital phenotypes (Marley and von Zastrow 2012; Guo et al. 2015; Gabriel, Pazour, and Lo 2018). Proper cilia formation is necessary for cellular motility and directional migration, as are TF-regulated cell signaling pathways like WNT and NOTCH (Foulquier et al. 2018; Bray 2006; Ji et al. 2020). Cardiac TFs also regulate cell-cell adhesion genes (Soini et al. 2018) and structural proteins like actin filaments in the sarcomere (Potthoff et al. 2007). Finally, mitochondrial activity and oxidative metabolism genes are necessary to heart formation (Hom et al. 2011; Cheong et al. 2020), and are partially regulated by chromatin remodeler SRCAP in the developing heart (Xu et al. 2021). In summary, cardiac TFs regulate a broad range of molecular processes that are necessary for the migration and adhesion of cells into tissues that will express the proteins necessary for heart function.

1.4 The contribution of genetic variation to congenital heart defects (CHD)

Given the breadth of fine-tuned regulatory and molecular processes that are essential in cardiogenesis, the opportunities to get it wrong are varied and numerous. This is reflected in the extreme genetic heterogeneity of defects associated with heart development (Zaidi and Brueckner 2017).

Congenital heart defects (CHD) are the most common birth defect, affecting 1% of live births every year (Marelli et al. 2007). Heritability estimates ranging between 70-90% (Cripe et al. 2004; McBride et al. 2005; Hinton et al. 2007; Pierpont et al. 2018) and a high family recurrence rate (Gill et al. 2003; Øyen et al. 2009) suggest that most defects can be explained in large part by patient genetics. However, despite increasingly large cohorts with exome sequencing like the Pediatric Cardiac Genomics Consortium (PCGC), the complex genetic architecture of CHD limits our ability to identify causal variants and genes (Homsy et al. 2015; Jin et al. 2017). It is estimated that cohorts of approximately 10,000 parent-proband trios would be needed for whole-exome sequencing to detect 80% of genes contributing to haplo-insufficient syndromic CHD alone (Sifrim et al. 2016), highlighting the need for new strategies to identify potentially causative genomic loci.

From the perspective of patients and family, fewer than 30% of CHD cases have a known underlying cause (Pierpont et al. 2018), and so most families are denied the opportunities that genetic explanation can afford with respect to symptoms management, prognosis, early intervention for later-onset comorbidities, and family planning. Some unexplained cases presumably involve damaging variants in cryptic risk genes, thus far undiscovered due to too few observations of patient mutations in each gene (given the plethora of possible disease loci). In the second chapter of this thesis, I will describe an approach that uses cell-type specific protein-protein interaction data to identify novel genes participating in heart development. I will also describe the computational framework I developed to prioritize potentially causal variants in those genes, leading to the discovery of a variant in the novel CHD gene, GLYR1 (Gonzalez-Teran et al. 2022).

Even in cases with a suspected monogenic explanation, some of those “causal” CHD variants were inherited from parents without a heart defect, suggesting they are insufficient on their own to cause disease (Pierpont et al. 2018; Pediatric Cardiac Genomics Consortium et al. 2013). Such variants are said to display incomplete penetrance, i.e. only some carriers have the phenotype; or variable expressivity, i.e. the same mutation causes different symptoms or severity in different patients (Coll et al. 2017; Kingdom and Wright 2022). In addition to environmental factors, a potential explanation for these phenomena is oligogenic inheritance, or a type of inheritance in which a few variants are required in combination to cause a particular phenotype (Kousi and Katsanis 2015).

Previous work has validated an instance of oligogenic inheritance in CHD (Gifford et al. 2019), as well as speculated a role for oligogenic inheritance in other developmental disorders like autism spectrum disorder (ASD) (Schaaf et al. 2011; Wenger et al. 2016). While these studies yielded mechanistic insights into specific gene and variant combinations, they relied on known risk genes and existing functional information to propose testable hypotheses. Alternatively, one could enumerate and prioritize all damaging variant combinations in an automated and statistically rigorous manner, but this strategy is infeasible due to combinatorial explosion of the number of tests (Edwards and Glass 2000). Recent advances in the field include the Digenic method, which reduces the combinatorial search space by aggregating rare variants at the gene level (Kerner et al. 2020), as does RareComb (Pounraja and Girirajan 2022), an algorithm that tests for greater frequency of gene combinations in cases compared to controls. However, neither of these computational methods incorporate parental sequencing data, which are especially useful in reducing false positives for simplex families in which an affected proband is born to unaffected parents (Ewens 1999). To address this unexplored approach I developed GCOD, a trio-based probabilistic model of variant transmission that identifies groups of genes in which rare variants are transmitted together more often than expected by chance (Pittman et al. 2022), described in the third chapter of this thesis.

The mechanisms and methods investigated in chapters two and three involve coding variants, i.e. variants inside genes that affect the amino acid sequence of translated proteins. Another important consideration is that some genetic risk is carried in the non-coding genome, which whole-exome sequencing cannot successfully capture. For example, regulatory enhancers are often found in intronic and intergenic regions, and given the developing heart's sensitivity to timing and gene dosage (Zaidi and Brueckner 2017; Kasah, Oddy, and Basson 2018; Hui Zhang, Liu, and Tian 2019), gene regulatory disruption seems a plausible mechanism. In fact, none of the ~20 identified GWAS SNPs occur in exonic coding regions (Klemm et al. 2013; Lahm et al. 2021). Instead they are likely to alter the activity of cis-regulatory elements like enhancers and insulators. Large structural variants (SVs) and copy are promising candidates for studying cis-regulatory elements, as well as their relationship to the 3D structure of the genome, because they have the potential to drastically rearrange local chromatin structure (Spielmann, Lupiáñez, and Mundlos 2018; Shanta et al. 2020).

Our group previously found that while unaffected controls show a clear depletion of CNVs at insulator regions across the genome, individuals diagnosed with developmental delay (DD) showed no bias in the genomic location of deletions (Fudenberg and Pollard 2019). Based on these findings and high rates of co-morbidity between DD and CHD (Rollins and Newburger 2014), we hypothesized that the disruption of enhancer-promoter contacts as a result of chromatin rearrangement can contribute to the etiology of CHD. However, non-coding variants yield less straightforward interpretation than their exonic counterparts, requiring new methods to prioritize and interpret variant effects. Our group previously developed the deep neural net model Akita, which was trained to predict chromatin contact frequency along a 1Mb region in five cell types by minimizing the loss on experimental Hi-C data (Fudenberg, Kelley, and Pollard 2020). In chapter four of this thesis, I use Akita to predict the effects of CHD patient SVs on local chromatin contact frequency, discovering three deletions that are predicted to alter regulatory interactions near heart-related genes. I include as-yet unpublished data in a cell line engineered with a CHD patient deletion, demonstrating the utility of predictive models in non-coding variant interpretation and suggesting a role for 3D chromatin rearrangement in CHD.

1.5 Summary and scope of this thesis

To summarize, the etiology of congenital heart defects is known to be primarily genetic; however, many risk genes remain cryptic, the specific underpinnings of most cases are unknown, and potential disease mechanisms like oligogenic interaction and high-effect non-coding variation remain underexplored. Novel datasets, like high-resolution micro-C chromatin contact data and affinity-purification of endogenous tagging of TFs, have enabled computational researchers to develop new models and statistical approaches to light up the dark corners of this disease.

My doctoral dissertation comprises the following chapters: 1) this introduction, to summarize the context of the work; 2) the integration of de novo exonic variants with transcription factor protein interactomes to discover novel risk genes and variants in CHD; 3) GCOD, the statistical model and software that predicts gene interactions from trio data; 4) the use of chromatin interaction frequency predictions to prioritize structural variants in CHD; and 5) a summary and discussion of the major findings and limitations of the work described here.

REFERENCES

Ajima, Rieko, Yuko Sakakibara, Noriko Sakurai-Yamatani, Masafumi Muraoka, and Yumiko Saga. 2021. “Formal Proof of the Requirement of MESP1 and MESP2 in Mesoderm Specification and Their Transcriptional Control via Specific Enhancers in Mice.” *Development* 148 (20). https://doi.org/10.1242/dev.194613.

Allfrey, V. G., R. Faulkner, and A. E. Mirsky. 1964. “ACETYLATION AND METHYLATION OF HISTONES AND THEIR POSSIBLE ROLE IN THE REGULATION OF RNA SYNTHESIS*.” *Proceedings of the National Academy of Sciences* 51 (5): 786–94.

Banerji, J., L. Olson, and W. Schaffner. 1983. “A Lymphocyte-Specific Cellular Enhancer Is Located Downstream of the Joining Region in Immunoglobulin Heavy Chain Genes.” *Cell* 33 (3): 729–40.

Bannister, Andrew J., and Tony Kouzarides. 2011. “Regulation of Chromatin by Histone Modifications.” *Cell Research* 21 (3): 381–95.

Barisic, Darko, Michael B. Stadler, Mario Iurlaro, and Dirk Schübeler. 2019. “Mammalian ISWI and SWI/SNF Selectively Mediate Binding of Distinct Transcription Factors.” *Nature* 569 (7754): 136–40.

Bray, Sarah J. 2006. “Notch Signalling: A Simple Pathway Becomes Complex.” *Nature Reviews. Molecular Cell Biology* 7 (9): 678–89.

Bruneau, B. G., M. Logan, N. Davis, T. Levi, C. J. Tabin, J. G. Seidman, and C. E. Seidman. 1999. “Chamber-Specific Cardiac Expression of Tbx5 and Heart Defects in Holt-Oram Syndrome.” *Developmental Biology* 211 (1): 100–108.

Buchler, Nicolas E., Ulrich Gerland, and Terence Hwa. 2003. “On Schemes of Combinatorial Transcription Logic.” *Proceedings of the National Academy of Sciences of the United States of America* 100 (9): 5136–41.

Bulger, Michael, and Mark Groudine. 2011. “Functional and Mechanistic Diversity of Distal

Transcription Enhancers." *Cell* 144 (3): 327–39.

Calo, Eliezer, and Joanna Wysocka. 2013. "Modification of Enhancer Chromatin: What, How, and Why?" *Molecular Cell* 49 (5): 825–37.

Casamassimi, Amelia, and Alfredo Ciccodicola. 2019. "Transcriptional Regulation: Molecules, Involved Mechanisms, and Misregulation." *International Journal of Molecular Sciences* 20 (6). https://doi.org/10.3390/ijms20061281.

Chan, H. M., and N. B. La Thangue. 2001. "p300/CBP Proteins: HATs for Transcriptional Bridges and Scaffolds." *Journal of Cell Science* 114 (Pt 13): 2363–73.

Chan, Sunny Sun-Kin, Xiaozhong Shi, Akira Toyama, Robert W. Arpke, Abhijit Dandapat, Michelina Iacovino, Jinjoo Kang, et al. 2013. "Mesp1 Patterns Mesoderm into Cardiac, Hematopoietic, or Skeletal Myogenic Progenitors in a Context-Dependent Manner." *Cell Stem Cell* 12 (5): 587–601.

Chapman, D. L., N. Garvey, S. Hancock, M. Alexiou, S. I. Agulnik, J. J. Gibson-Brown, J. Cebra-Thomas, R. J. Bollag, L. M. Silver, and V. E. Papaioannou. 1996. "Expression of the T-Box Family Genes, Tbx1-Tbx5, during Early Mouse Development." *Developmental Dynamics: An Official Publication of the American Association of Anatomists* 206 (4): 379–90.

Charest, Julien, Thomas Daniele, Jingkui Wang, Aleksandr Bykov, Ariane Mandlbauer, Mila Asparuhova, Josef Röhsner, Paula Gutiérrez-Pérez, and Luisa Cochella. 2020. "Combinatorial Action of Temporally Segregated Transcription Factors." *Developmental Cell* 55 (4): 483–99.e7.

Cheong, Agnes, Danielle Archambault, Rinat Degani, Elizabeth Iverson, Kimberly D. Tremblay, and Jesse Mager. 2020. "Nuclear-Encoded Mitochondrial Ribosomal Proteins Are Required to Initiate Gastrulation." *Development* 147 (10). https://doi.org/10.1242/dev.188714.

Coll, Monica, Alexandra Pérez-Serra, Jesus Mates, Bernat Del Olmo, Marta Puigmulé, Anna

Fernandez-Falgueras, Anna Iglesias, et al. 2017. “Incomplete Penetrance and Variable Expressivity: Hallmarks in Channelopathies Associated with Sudden Cardiac Death.” *Biology* 7 (1). https://doi.org/10.3390/biology7010003.

Cooper, Geoffrey M. 2000. *Signaling in Development and Differentiation*. Sinauer Associates.

Cremer, T., and C. Cremer. 2001. “Chromosome Territories, Nuclear Architecture and Gene Regulation in Mammalian Cells.” *Nature Reviews. Genetics* 2 (4): 292–301.

Cripe, Linda, Gregor Andelfinger, Lisa J. Martin, Kerry Shooner, and D. Woodrow Benson. 2004. “Bicuspid Aortic Valve Is Heritable.” *Journal of the American College of Cardiology* 44 (1): 138–43.

Dixon, Jesse R., Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S. Liu, and Bing Ren. 2012. “Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions.” *Nature* 485 (7398): 376–80.

Edwards, R., and L. Glass. 2000. “Combinatorial Explosion in Model Gene Networks.” *Chaos* 10 (3): 691–704.

Euskirchen, Ghia, Raymond K. Auerbach, and Michael Snyder. 2012. “SWI/SNF Chromatin-Remodeling Factors: Multiscale Analyses and Diverse Functions.” *The Journal of Biological Chemistry* 287 (37): 30897–905.

Ewens, Warren J. 1999. “Statistical Aspects of the Transmission/Disequilibrium Test (TDT).” *Lecture Notes-Monograph Series / Institute of Mathematical Statistics* 33: 77–94.

Foulquier, Sébastien, Evangelos P. Daskalopoulos, Gentian Lluri, Kevin C. M. Hermans, Arjun Deb, and W. Matthijs Blankesteijn. 2018. “WNT Signaling in Cardiac and Vascular Disease.” *Pharmacological Reviews* 70 (1): 68–141.

Fudenberg, Geoff, David R. Kelley, and Katherine S. Pollard. 2020. “Predicting 3D Genome Folding from DNA Sequence with Akita.” *Nature Methods* 17 (11): 1111–17.

Fudenberg, Geoff, and Katherine S. Pollard. 2019. “Chromatin Features Constrain Structural Variation across Evolutionary Timescales.” *Proceedings of the National Academy of*

Sciences of the United States of America 116 (6): 2175–80.

Furlong, Eileen E. M., and Michael Levine. 2018. “Developmental Enhancers and Chromosome Topology.” *Science* 361 (6409): 1341–45.

Gabriel, George C., Gregory J. Pazour, and Cecilia W. Lo. 2018. “Congenital Heart Defects and Ciliopathies Associated With Renal Phenotypes.” *Frontiers in Pediatrics* 6 (June): 175.

Ghosh, Asish K. 2020. “p300 in Cardiac Development and Accelerated Cardiac Aging.” *Aging and Disease* 11 (4): 916–26.

Gifford, Casey A., Sanjeev S. Ranade, Ryan Samarakoon, Hazel T. Salunga, T. Yvanka de Soysa, Yu Huang, Ping Zhou, et al. 2019. “Oligogenic Inheritance of a Human Heart Disease Involving a Genetic Modifier.” *Science* 364 (6443): 865–70.

Gilbert, Scott F. 2000. *The Developmental Mechanics of Cell Specification*. Sinauer Associates.

Gill, Harinder K., Miranda Splitt, Gurleen K. Sharland, and John M. Simpson. 2003. “Patterns of Recurrence of Congenital Heart Disease: An Analysis of 6,640 Consecutive Pregnancies Evaluated by Detailed Fetal Echocardiography.” *Journal of the American College of Cardiology* 42 (5): 923–29.

Gonzalez-Teran, Barbara, Maureen Pittman, Franco Felix, Reuben Thomas, Desmond Richmond-Buccola, Ruth Hüttenhain, Krishna Choudhary, et al. 2022. “Transcription Factor Protein Interactomes Reveal Genetic Determinants in Heart Disease.” *Cell* 185 (5): 794–814.e30.

Guo, Jiami, Holden Higginbotham, Jingjun Li, Jackie Nichols, Josua Hirt, Vladimir Ghukasyan, and E. S. Anton. 2015. “Developmental Disruptions Underlying Brain Abnormalities in Ciliopathies.” *Nature Communications* 6 (July): 7857.

He, Shuang, Zihan Wu, Yuan Tian, Zishuo Yu, Jiali Yu, Xinxin Wang, Jie Li, Bijun Liu, and Yanhui Xu. 2020. “Structure of Nucleosome-Bound Human BAF Complex.” *Science* 367 (6480): 875–81.

Hinton, Robert B., Jr, Lisa J. Martin, Meredith E. Tabangin, Mjaye L. Mazwi, Linda H. Cripe, and

D. Woodrow Benson. 2007. “Hypoplastic Left Heart Syndrome Is Heritable.” *Journal of the American College of Cardiology* 50 (16): 1590–95.

Ho, Lena, and Gerald R. Crabtree. 2010. “Chromatin Remodelling during Development.” *Nature* 463 (7280): 474–84.

Hom, Jennifer R., Rodrigo A. Quintanilla, David L. Hoffman, Karen L. de Mesy Bentley, Jeffery D. Molkentin, Shey-Shing Sheu, and George A. Porter Jr. 2011. “The Permeability Transition Pore Controls Cardiac Mitochondrial Maturation and Myocyte Differentiation.” *Developmental Cell*.

Homsy, Jason, Samir Zaidi, Yufeng Shen, James S. Ware, Kaitlin E. Samocha, Konrad J. Karczewski, Steven R. DePalma, et al. 2015. “De Novo Mutations in Congenital Heart Disease with Neurodevelopmental and Other Congenital Anomalies.” *Science* 350 (6265): 1262–66.

Ieda, Masaki, Ji-Dong Fu, Paul Delgado-Olguin, Vasanth Vedantham, Yohei Hayashi, Benoit G. Bruneau, and Deepak Srivastava. 2010. “Direct Reprogramming of Fibroblasts into Functional Cardiomyocytes by Defined Factors.” *Cell* 142 (3): 375–86.

Imakaev, Maxim, Geoffrey Fudenberg, Rachel Patton McCord, Natalia Naumova, Anton Goloborodko, Bryan R. Lajoie, Job Dekker, and Leonid A. Mirny. 2012. “Iterative Correction of Hi-C Data Reveals Hallmarks of Chromosome Organization.” *Nature Methods* 9 (10): 999–1003.

Inagawa, Kohei, Kazutaka Miyamoto, Hiroyuki Yamakawa, Naoto Muraoka, Taketaro Sadahiro, Tomohiko Umei, Rie Wada, et al. 2012. “Induction of Cardiomyocyte-like Cells in Infarct Hearts by Gene Transfer of Gata4, Mef2c, and Tbx5.” *Circulation Research* 111 (9): 1147–56.

Jin, Sheng Chih, Jason Homsy, Samir Zaidi, Qiongshi Lu, Sarah Morton, Steven R. DePalma, Xue Zeng, et al. 2017. “Contribution of Rare Inherited and de Novo Variants in 2,871 Congenital Heart Disease Probands.” *Nature Genetics* 49 (11): 1593–1601.

Ji, Weizhen, Dina Ferdman, Joshua Copel, Dustin Scheinost, Veronika Shabanova, Martina Brueckner, Mustafa K. Khokha, and Laura R. Ment. 2020. “De Novo Damaging Variants Associated with Congenital Heart Diseases Contribute to the Connectome.” *Scientific Reports* 10 (1): 7046.

Kasah, Sahrunizam, Christopher Oddy, and M. Albert Basson. 2018. “Autism-Linked CHD Gene Expression Patterns during Development Predict Multi-Organ Disease Phenotypes.” *Journal of Anatomy* 233 (6): 755–69.

Kato, Mamoru, Naoya Hata, Nilanjana Banerjee, Bruce Futcher, and Michael Q. Zhang. 2004. “Identifying Combinatorial Regulation of Transcription Factors and Binding Motifs.” *Genome Biology* 5 (8): R56.

Kerner, Gaspard, Matthieu Bouaziz, Aurélie Cobat, Benedetta Bigio, Andrew T. Timberlake, Jacinta Bustamante, Richard P. Lifton, Jean-Laurent Casanova, and Laurent Abel. 2020. “A Genome-Wide Case-Only Test for the Detection of Digenic Inheritance in Human Exomes.” *Proceedings of the National Academy of Sciences of the United States of America* 117 (32): 19367–75.

Kimura, Hiroshi. 2013. “Histone Modifications for Human Epigenome Analysis.” *Journal of Human Genetics* 58 (7): 439–45.

Kingdom, Rebecca, and Caroline F. Wright. 2022. “Incomplete Penetrance and Variable Expressivity: From Clinical Studies to Population Cohorts.” *Frontiers in Genetics* 13 (July): 920390.

Kitajima, S., A. Takagi, T. Inoue, and Y. Saga. 2000. “MesP1 and MesP2 Are Essential for the Development of Cardiac Mesoderm.” *Development* 127 (15): 3215–26.

Klemm, A., P. Flicek, T. Manolio, and L. Hindorff. 2013. “The NHGRI GWAS Catalog, a Curated Resource of SNP-Trait Associations.” *Nucleic Acids*. https://academic.oup.com/nar/article-abstract/42/D1/D1001/1062755.

Klena, Nikolai T., Brian C. Gibbs, and Cecilia W. Lo. 2017. “Cilia and Ciliopathies in Congenital

Heart Disease." *Cold Spring Harbor Perspectives in Biology* 9 (8). https://doi.org/10.1101/cshperspect.a028266.

Kousi, Maria, and Nicholas Katsanis. 2015. "Genetic Modifiers and Oligogenic Inheritance." *Cold Spring Harbor Perspectives in Medicine* 5 (6). https://doi.org/10.1101/cshperspect.a017145.

Kwon, H., A. N. Imbalzano, P. A. Khavari, R. E. Kingston, and M. R. Green. 1994. "Nucleosome Disruption and Enhancement of Activator Binding by a Human SW1/SNF Complex." *Nature* 370 (6489): 477–81.

Lahm, Harald, Meiwen Jia, Martina Dreßen, Felix Wirth, Nazan Puluca, Ralf Gilsbach, Bernard D. Keavney, et al. 2021. "Congenital Heart Disease Risk Loci Identified by Genome-Wide Association Study in European Patients." *The Journal of Clinical Investigation* 131 (2). https://doi.org/10.1172/JCI141837.

Laubscher, Dominik, Berkley E. Gryder, Benjamin D. Sunkel, Thorkell Andresson, Marco Wachtel, Sudipto Das, Bernd Roschitzki, et al. 2021. "BAF Complexes Drive Proliferation and Block Myogenic Differentiation in Fusion-Positive Rhabdomyosarcoma." *Nature Communications* 12 (1): 6924.

Lazar, John E., Sandra Stehling-Sun, Vivek Nandakumar, Hao Wang, Daniel R. Chee, Nicholas P. Howard, Reyes Acosta, et al. 2020. "Global Regulatory DNA Potentiation by SMARCA4 Propagates to Selective Gene Expression Programs via Domain-Level Remodeling." *Cell Reports* 31 (8): 107676.

Levine, Mike. 2010. "Transcriptional Enhancers in Animal Development and Evolution." *Current Biology: CB* 20 (17): R754–63.

Lieberman-Aiden, Erez, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, et al. 2009. "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome." *Science* 326 (5950): 289–93.

Lim, Tingsen Benson, Sik Yin Roger Foo, and Ching Kit Chen. 2021. “The Role of Epigenetics in Congenital Heart Disease.” *Genes* 12 (3). https://doi.org/10.3390/genes12030390.

Li, You, Nikolai T. Klena, George C. Gabriel, Xiaoqin Liu, Andrew J. Kim, Kristi Lemke, Yu Chen, et al. 2015. “Global Genetic Analysis in Mice Unveils Central Role for Cilia in Congenital Heart Disease.” *Nature* 521 (7553): 520–24.

Marelli, Ariane J., Andrew S. Mackie, Raluca Ionescu-Ittu, Elham Rahme, and Louise Pilote. 2007. “Congenital Heart Disease in the General Population: Changing Prevalence and Age Distribution.” *Circulation* 115 (2): 163–72.

Marley, Aaron, and Mark von Zastrow. 2012. “A Simple Cell-Based Assay Reveals That Diverse Neuropsychiatric Risk Genes Converge on Primary Cilia.” *PloS One* 7 (10): e46647.

Marstrand, Troels T., and John D. Storey. 2014. “Identifying and Mapping Cell-Type-Specific Chromatin Programming of Gene Expression.” *Proceedings of the National Academy of Sciences of the United States of America* 111 (6): E645–54.

McBride, Kim L., Ricardo Pignatelli, Mark Lewin, Trang Ho, Susan Fernbach, Andres Menesses, Wilbur Lam, et al. 2005. “Inheritance Analysis of Congenital Left Ventricular Outflow Tract Obstruction Malformations: Segregation, Multiplex Relative Risk, and Heritability.” *American Journal of Medical Genetics. Part A* 134A (2): 180–86.

McElhinney, Doff B., Elizabeth Geiger, Joshua Blinder, D. Woodrow Benson, and Elizabeth Goldmuntz. 2003. “NKX2.5 Mutations in Patients with Congenital Heart Disease.” *Journal of the American College of Cardiology* 42 (9): 1650–55.

Meier, P., A. Finch, and G. Evan. 2000. “Apoptosis in Development.” *Nature* 407 (6805).

Mohammed, Hisham, Irene Hernando-Herraez, Aurora Savino, Antonio Scialdone, Iain Macaulay, Carla Mulas, Tamir Chandra, et al. 2017. “Single-Cell Landscape of Transcriptional Heterogeneity and Cell Fate Decisions during Mouse Early Gastrulation.” *Cell Reports* 20 (5): 1215–28.

Moorthy, Sakthi D., Scott Davidson, Virlana M. Shchuka, Gurdeep Singh, Nakisa Malek-Gilani, Lida Langroudi, Alexandre Martchenko, Vincent So, Neil N. Macpherson, and Jennifer A. Mitchell. 2017. “Enhancers and Super-Enhancers Have an Equivalent Regulatory Role in Embryonic Stem Cells through Regulation of Single or Multiple Genes.” *Genome Research* 27 (2): 246–58.

Mori, Alessandro D., and Benoit G. Bruneau. 2004. “TBX5 Mutations and Congenital Heart Disease: Holt-Oram Syndrome Revealed.” *Current Opinion in Cardiology* 19 (3): 211–15.

Nora, Elphège P., Bryan R. Lajoie, Edda G. Schulz, Luca Giorgetti, Ikuhiro Okamoto, Nicolas Servant, Tristan Piolot, et al. 2012. “Spatial Partitioning of the Regulatory Landscape of the X-Inactivation Centre.” *Nature* 485 (7398): 381–85.

Oka, Toru, Marjorie Maillet, Alistair J. Watt, Robert J. Schwartz, Bruce J. Aronow, Stephen A. Duncan, and Jeffery D. Molkentin. 2006. “Cardiac-Specific Deletion of Gata4 Reveals Its Requirement for Hypertrophy, Compensation, and Myocyte Viability.” *Circulation Research* 98 (6): 837–45.

Øyen, Nina, Gry Poulsen, Heather A. Boyd, Jan Wohlfahrt, Peter K. A. Jensen, and Mads Melbye. 2009. “Recurrence of Congenital Heart Defects in Families.” *Circulation* 120 (4): 295–301.

Palstra, Robert-Jan, and Frank Grosveld. 2012. “Transcription Factor Binding at Enhancers: Shaping a Genomic Regulatory Landscape in Flux.” *Frontiers in Genetics* 3 (September): 195.

Pediatric Cardiac Genomics Consortium, Bruce Gelb, Martina Brueckner, Wendy Chung, Elizabeth Goldmuntz, Jonathan Kaltman, Juan Pablo Kaski, et al. 2013. “The Congenital Heart Disease Genetic Network Study: Rationale, Design, and Early Results.” *Circulation Research* 112 (4): 698–706.

Pierpont, Mary Ella, Martina Brueckner, Wendy K. Chung, Vidu Garg, Ronald V. Lacro, Amy L. McGuire, Seema Mital, et al. 2018. “Genetic Basis for Congenital Heart Disease: Revisited: A Scientific Statement From the American Heart Association.” *Circulation* 138

(21): e653–711.

Pittman, Maureen, Kihyun Lee, Deepak Srivastava, and Katherine S. Pollard. 2022. “An Oligogenic Inheritance Test Detects Risk Genes and Their Interactions in Congenital Heart Defects and Developmental Comorbidities.” *bioRxiv*. https://doi.org/10.1101/2022.04.08.487704.

Potthoff, Matthew J., Michael A. Arnold, John McAnally, James A. Richardson, Rhonda Bassel-Duby, and Eric N. Olson. 2007. “Regulation of Skeletal Muscle Sarcomere Integrity and Postnatal Muscle Function by Mef2c.” *Molecular and Cellular Biology* 27 (23): 8143–51.

Pounraja, Vijay Kumar, and Santhosh Girirajan. 2022. “A General Framework for Identifying Oligogenic Combinations of Rare Variants in Complex Disorders.” *Genome Research*, March. https://doi.org/10.1101/gr.276348.121.

Qian, Li, Yu Huang, C. Ian Spencer, Amy Foley, Vasanth Vedantham, Lei Liu, Simon J. Conway, Ji-Dong Fu, and Deepak Srivastava. 2012. “In Vivo Reprogramming of Murine Cardiac Fibroblasts into Induced Cardiomyocytes.” *Nature* 485 (7400): 593–98.

Qiao, Xiao-Hui, Fei Wang, Xian-Ling Zhang, Ri-Tai Huang, Song Xue, Juan Wang, Xing-Biao Qiu, Xing-Yuan Liu, and Yi-Qing Yang. 2017. “MEF2C Loss-of-Function Mutation Contributes to Congenital Heart Defects.” *International Journal of Medical Sciences* 14 (11): 1143–53.

Reuveni, Eli, Dmitry Getselter, Oded Oron, and Evan Elliott. 2018. “Differential Contribution of Cis and Trans Gene Transcription Regulatory Mechanisms in Amygdala and Prefrontal Cortex and Modulation by Social Stress.” *Scientific Reports* 8 (1): 6339.

Riethoven, Jean-Jack M. 2010. “Regulatory Regions in DNA: Promoters, Enhancers, Silencers, and Insulators.” *Methods in Molecular Biology* 674: 33–42.

Rollins, Caitlin K., and Jane W. Newburger. 2014. “Neurodevelopmental Outcomes in Congenital Heart Disease.” *Circulation* 130 (14): e124–26.

Rossi, J. M., N. R. Dunn, B. L. Hogan, and K. S. Zaret. 2001. “Distinct Mesodermal Signals, Including BMPs from the Septum Transversum Mesenchyme, Are Required in Combination for Hepatogenesis from the Endoderm.” *Genes & Development* 15 (15): 1998–2009.

Scarpa, Elena, and Roberto Mayor. 2016. “Collective Cell Migration in Development.” *The Journal of Cell Biology* 212 (2): 143–55.

Schaaf, Christian P., Aniko Sabo, Yasunari Sakai, Jacy Crosby, Donna Muzny, Alicia Hawes, Lora Lewis, et al. 2011. “Oligogenic Heterozygosity in Individuals with High-Functioning Autism Spectrum Disorders.” *Human Molecular Genetics* 20 (17): 3366–75.

Schoenfelder, Stefan, Mayra Furlan-Magaril, Borbala Mifsud, Filipe Tavares-Cadete, Robert Sugar, Biola-Maria Javierre, Takashi Nagano, et al. 2015. “The Pluripotent Regulatory Circuitry Connecting Promoters to Their Long-Range Interacting Elements.” *Genome Research* 25 (4): 582–97.

Shanta, Omar, Amina Noor, Human Genome Structural Variation Consortium (HGSVC), and Jonathan Sebat. 2020. “The Effects of Common Structural Variants on 3D Chromatin Structure.” *BMC Genomics* 21 (1): 95.

Sifrim, Alejandro, Marc-Phillip Hitz, Anna Wilsdon, Jeroen Breckpot, Saeed H. Al Turki, Bernard Thienpont, Jeremy McRae, et al. 2016. “Distinct Genetic Architectures for Syndromic and Nonsyndromic Congenital Heart Defects Identified by Exome Sequencing.” *Nature Genetics* 48 (9): 1060–65.

Smith, Emily M., Bryan R. Lajoie, Gaurav Jain, and Job Dekker. 2016. “Invariant TAD Boundaries Constrain Cell-Type-Specific Looping Interactions between Promoters and Distal Elements around the CFTR Locus.” *American Journal of Human Genetics* 98 (1): 185–201.

Soini, Tea, Katja Eloranta, Marjut Pihlajoki, Antti Kyrönlahti, Oyediran Akinrinade, Noora Andersson, Jouko Lohi, Mikko P. Pakarinen, David B. Wilson, and Markku Heikinheimo.

2018. “Transcription Factor GATA4 Associates with Mesenchymal-like Gene Expression in Human Hepatoblastoma Cells.” *Tumour Biology: The Journal of the International Society for Oncodevelopmental Biology and Medicine* 40 (7): 1010428318785498.

Spielmann, Malte, Darío G. Lupiáñez, and Stefan Mundlos. 2018. “Structural Variation in the 3D Genome.” *Nature Reviews. Genetics* 19 (7): 453–67.

Sun, Xin, Swetansu K. Hota, Yu-Qing Zhou, Stefanie Novak, Dario Miguel-Perez, Danos Christodoulou, Christine E. Seidman, et al. 2018. “Cardiac-Enriched BAF Chromatin-Remodeling Complex Subunit Baf60c Regulates Gene Expression Programs Essential for Heart Development and Function.” *Biology Open* 7 (1). https://doi.org/10.1242/bio.029512.

Takahashi, Kazutoshi, Koji Tanabe, Mari Ohnuki, Megumi Narita, Tomoko Ichisaka, Kiichiro Tomoda, and Shinya Yamanaka. 2007. “Induction of Pluripotent Stem Cells from Adult Human Fibroblasts by Defined Factors.” *Cell* 131 (5): 861–72.

Takahashi, Kazutoshi, and Shinya Yamanaka. 2006. “Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors.” *Cell* 126 (4): 663–76.

Takeuchi, Jun K., and Benoit G. Bruneau. 2009. “Directed Transdifferentiation of Mouse Mesoderm to Heart Tissue by Defined Factors.” *Nature* 459 (7247): 708–11.

Tomita-Mitchell, A., C. L. Maslen, C. D. Morris, V. Garg, and E. Goldmuntz. 2007. “GATA4 Sequence Variants in Patients with Congenital Heart Disease.” *Journal of Medical Genetics* 44 (12): 779–83.

Toto, Paula Coutinho, Pier Lorenzo Puri, and Sonia Albini. 2016. “SWI/SNF-Directed Stem Cell Lineage Specification: Dynamic Composition Regulates Specific Stages of Skeletal Myogenesis.” *Cellular and Molecular Life Sciences: CMLS* 73 (20): 3887–96.

Tseng, Yu-Chun, Birgit Cabot, and Ryan A. Cabot. 2017. “ARID1A, a Component of SWI/SNF Chromatin Remodeling Complexes, Is Required for Porcine Embryo Development.”

Molecular Reproduction and Development 84 (12): 1250–56.

Vaart, Aniek van der, Molly Godfrey, Vincent Portegijs, and Sander van den Heuvel. 2020. “Dose-Dependent Functions of SWI/SNF BAF in Permitting and Inhibiting Cell Proliferation in Vivo.” *Science Advances* 6 (21): eaay3823.

Vandel, Jimmy, Océane Cassan, Sophie Lèbre, Charles-Henri Lecellier, and Laurent Bréhélin. 2019. “Probing Transcription Factor Combinatorics in Different Promoter Classes and in Enhancers.” *BMC Genomics* 20 (1): 103.

Wang, W., J. Côté, Y. Xue, S. Zhou, P. A. Khavari, S. R. Biggar, C. Muchardt, et al. 1996. “Purification and Biochemical Heterogeneity of the Mammalian SWI-SNF Complex.” *The EMBO Journal* 15 (19): 5370–82.

Wenger, Tara L., Charlly Kao, Donna M. McDonald-McGinn, Elaine H. Zackai, Alice Bailey, Robert T. Schultz, Bernice E. Morrow, Beverly S. Emanuel, and Hakon Hakonarson. 2016. “The Role of mGluR Copy Number Variation in Genetic and Environmental Forms of Syndromic Autism Spectrum Disorder.” *Scientific Reports* 6 (January): 19372.

Werner, Petra, Brande Latney, Matthew A. Deardorff, and Elizabeth Goldmuntz. 2016. “MESP1 Mutations in Patients with Congenital Heart Defects.” *Human Mutation* 37 (3): 308–14.

Xu, Mingjie, Jie Yao, Yingchao Shi, Huijuan Yi, Wukui Zhao, Xinhua Lin, and Zhongzhou Yang. 2021. “The SRCAP Chromatin Remodeling Complex Promotes Oxidative Metabolism during Prenatal Heart Development.” *Development* 148 (8).

Zaidi, Samir, and Martina Brueckner. 2017. “Genetics and Genomics of Congenital Heart Disease.” *Circulation Research* 120 (6): 923–40.

Zhang, Heyao, Xuepeng Wang, Jingsheng Li, Ronghua Shi, and Ying Ye. 2021. “BAF Complex in Embryonic Stem Cells and Early Embryonic Development.” *Stem Cells International* 2021 (January): 6668866.

Zhang, Hui, Lingjuan Liu, and Jie Tian. 2019. “Molecular Mechanisms of Congenital Heart Disease in down Syndrome.” *Genes & Diseases* 6 (4): 372–77.

CHAPTER 2: Transcription factor protein interactomes reveal genetic determinants in heart disease

This chapter is adapted from my contributions to Gonzalez-Teran et al., 2022*. Other contributions to the work enclosed here are below:

- D.R.-B, B.E.J.M., and B.G.-T. generated the knockout hiPSC lines. D.R.-B. and B.G.-T. performed WT, GATA4, and TBX5 knockout CP and CM differentiations.
- B.R.C., R.H., and B.G.-T. defined the appropriate affinity purification strategy for cTFs. M.M. and B.G.-T. performed GATA4 and TBX5 affinity purification and sample preparation for mass spectrometry and GT-PPI classification.
- F.F. and B.G.-T. performed GATA4 and GLYR1 silencing in CPs and luciferase reporter assays. A.P. designed and generated the Glyr1 P495L mouse line. C.Y.L., Y.H., M.W.C. and B.G.-T performed Glyr1 P495L mouse studies.
- Thanks to B.G.-T., D.S. and K.S.P. who drafted text featured in this chapter.

*Gonzalez-Teran, Barbara, Maureen Pittman, Franco Felix, Reuben Thomas, Desmond Richmond-Buccola, Ruth Hüttenhain, Krishna Choudhary, Elisabetta Moroni, Mauro W Costa, Yu Huang, Arun Padmanabhan, Michael Alexanian, Clara Youngna Lee, Bonnie E J Maven, Kaitlen Samse-Knapp, Sarah U Morton, Michael McGregor, Casey A Gifford, J G Seidman, Christine E Seidman, Bruce D Gelb, Giorgio Colombo, Bruce R Conklin, Brian L Black, Benoit G Bruneau , Nevan J Krogan, Katherine S Pollard, Deepak Srivastava. 2022. “Transcription Factor Protein Interactomes Reveal Genetic Determinants in Heart Disease.” Cell 185 (5): 794–814.e30.

2.1 ABSTRACT

Congenital heart disease (CHD) is present in 1% of live births, yet identification of causal mutations remains challenging. We hypothesized that genetic determinants for CHDs may lie in the protein interactomes of transcription factors whose mutations cause CHDs. Defining the interactomes of two transcription factors haploinsufficient in CHD, *GATA4* and *TBX5*, within human cardiac progenitors, and integrating the results with nearly 9,000 exomes from proband-parent trios revealed an enrichment of de novo missense variants associated with CHD within the interactomes. Scoring variants of interactome members based on residue, gene, and proband features identified likely CHD-causing genes, including the epigenetic reader *GLYR1*. GLYR1 and GATA4 widely co-occupied and co-activated cardiac developmental genes, and the identified *GLYR1* missense variant disrupted interaction with GATA4, impairing in vitro and in vivo function in mice. This integrative proteomic and genetic approach provides a framework for prioritizing and interrogating genetic variants in heart disease.

2.2 BACKGROUND

Birth defects are complex developmental phenotypes affecting 6% of births worldwide, yet their genetic roots are multifactorial and difficult to ascertain (Christianson, Howson, and Modell 2005; Deciphering Developmental Disorders Study 2015). Particularly challenging are rare disorders and more common complex defects with high allelic and locus heterogeneity. In recent years, whole-exome sequencing has accelerated our understanding of such disorders, including the most common birth defect, congenital heart disease (CHD) (Zaidi et al. 2013; Homsy et al. 2015; Jin et al. 2017; Richter et al. 2020). De novo monogenic aberrations were found to collectively contribute to 10% of CHD cases, whereas rare inherited and copy number variants have been identified in 1% and 25% of cases, respectively (Zaidi and Brueckner 2017). Additionally, polygenic and oligogenic inheritance models, where multiple genetic variants with epistatic relationships are implicated, have been proposed as mechanistic explanations for certain complex phenotypes. A recent study highlighted the involvement of genetic modifiers in human cardiac disease (Gifford et al. 2019), but the net contribution of oligogenic inheritance remains to be determined. Despite the growing catalog of human genome variants, the cause of over 50% of CHD cases remains unknown (Zaidi and Brueckner 2017).

A barrier to a complete understanding of CHD's etiology is its immense genetic heterogeneity. Estimates based on de novo mutations alone indicate that more than 390 genes may contribute to CHD pathogenesis (Homsy et al. 2015). This heterogeneity reduces the statistical power of CHD risk gene analysis with the cohorts currently available. It is estimated that cohorts of approximately 10,000 parent-proband trios would be needed for whole-exome sequencing to detect 80% of genes contributing to haplo-insufficient syndromic CHD (Sifrim et al. 2016), highlighting the need for alternative strategies to identify CHD risk genes and to prioritize for potentially causative variants.

Many diseases demonstrate tissue-restricted phenotypes but are rarely explained by mutations in genes with tissue-specific expression (Hekselman and Yeger-Lotem 2020). For example, cardiac malformations have been linked to variants in tissue-enriched cardiac transcription factors (cTFs) that are expressed more widely. Such cTFs typically form complexes with other tissue-enriched and ubiquitous proteins to orchestrate specific developmental gene programs (Lambert et al. 2018). cTF missense variants may disrupt specific interactions with other proteins, affecting their transcriptional cooperativity and causing disease (Ang et al. 2016; Moskowitz et al. 2011; Waldron et al. 2016). This observation suggests a functional relevance for cTF interactors in genetic disorders, including CHD. In agreement, Barshir et al. (Barshir et al. 2014) observed that disease causal genes are often widely expressed across tissues but with a tendency to exhibit more tissue-specific protein-protein interactions in diseased versus unaffected tissues. In CHD specifically, an excess of protein-altering de novo variants from the Pediatric Cardiac Genomics Consortium's (PCGC) cohort were found in ubiquitously expressed chromatin regulators that partner with cTFs to regulate the expression of key developmental genes (Zaidi et al. 2013). This led us to hypothesize that protein-protein interactors of cTFs associated with CHD may be enriched in disease-associated proteins, even if these proteins are not tissue specific.

GATA4 and TBX5 are two essential cTFs (Kuo et al. 1997; Bruneau et al. 1999, 2001; Oka et al. 2006) and among the first identified monogenic etiologies of familial CHD. Heterozygous pathogenic variations in TBX5 are a cause of septation defects and other forms of CHD in the setting of Holt-Oram syndrome (Basson et al. 1997; Yi Li et al. 1997). Heterozygous variations in GATA4 also cause atrial and ventricular septal defects, as well as pulmonary stenosis and outflow tract abnormalities (Garg et al. 2003; Rajagopal et al. 2007; Tomita-Mitchell et al. 2007). Subsequent studies have demonstrated that TBX5 and GATA4 cooperatively interact on DNA throughout the genome to regulate heart development (Ang et al. 2016; Luna-Zurita et al. 2016). Disruption of the physical interaction between these cTFs or with other specific co-factors by missense variants can impair transcriptional cooperativity and lineage specification, and ultimately cause cardiac malformations (Ang et al. 2016; Garg et al. 2003; Maitra et al. 2010; Waldron et al. 2016). Therefore, the identification of human GATA4 and TBX5 (GT) protein interactors during cardiogenesis could highlight disease mechanisms and aid in predicting the impact of protein-coding variants in CHD.

Here, I and co-authors designed an integrated proteomics and human genetics approach that dissects the protein-protein interactors of endogenous GATA4 and TBX5 in human cardiac progenitor cells, in order to identify and prioritize potential disease genes harboring CHD-associated variants. We used this approach to reveal aspects of cardiac gene regulation, which can be extended to the genetic underpinnings of many human diseases.

2.3 RESULTS

2.3.1 Identification of the GATA4 and TBX5 protein interactomes in cardiac progenitors

To identify the GATA4 and TBX5 protein interactome (GT-PPI) in human induced pluripotent stem-cell-derived cardiac progenitors (CPs), antibodies against each endogenous cTF were used for affinity purification and mass spectrometry (AP-MS) (**Figure 2.1**). Using CRISPR-Cas9-gRNA ribonucleoproteins, co-authors generated clonal TBX5 or GATA4 homozygous knockout (KO) hiPSC lines as negative controls. These control lines were differentiated to CP and cardiomyocyte (CM) stages, and the absence of the respective cTF expression was confirmed. Consistent with previous reports (Kathiriya et al. 2021; Luna-Zurita et al. 2016; Narita, Bielinska, and Wilson 1997), GATA4 and TBX5 KO cells were able to differentiate into CMs, albeit with delayed beating and reduced differentiation efficiency.

GATA4 or TBX5 mass spectrometry data were generated by co-authors from three replicates of nuclei-enriched day 6 hiPSC-derived CPs from wild-type (WT) or KO samples treated with RNase and DNase (**Figure 2.1**).

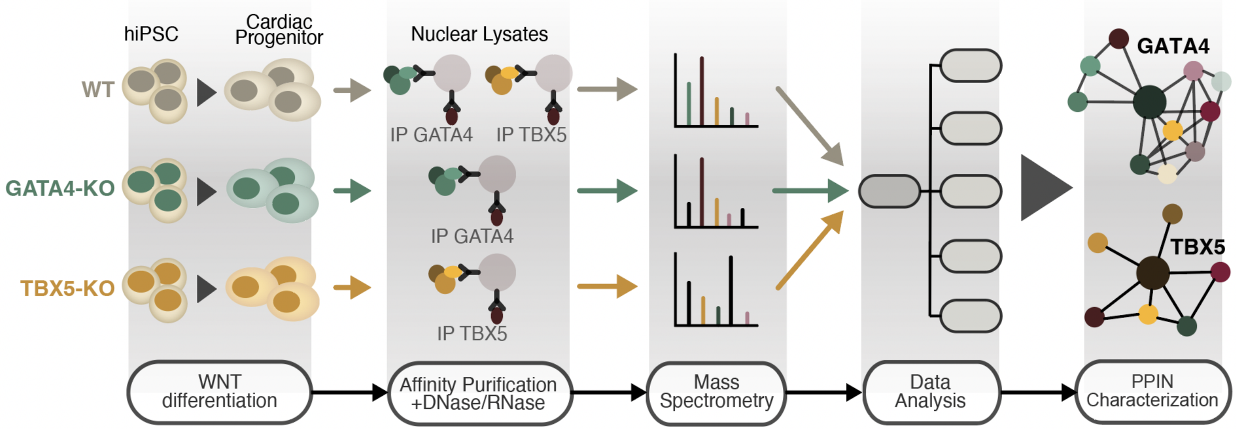


Figure 2.1: GATA4 and TBX5 AP-MS strategy from hiPSC-derived cardiac progenitors with gene knockout lines as negative controls.

Using the SAINTq software and algorithm (Teo et al. 2016), I obtained an initial list of GT interactors in WT CPs by scoring the proteins identified in WT AP-MS experiments to their corresponding KO control line. For further stringency, I additionally filtered based on nuclear localization and co-expression in the same cell types as the bait protein. Proteins whose mRNA was downregulated in the KO cells compared to WT were excluded (**see Methods 2.5.1: Selection of interactome proteins**). This approach yielded 272 proteins in total, which comprised several of the previously reported GATA4 and TBX5 interactors as well as novel interactors (Waldron et al. 2016; Enane et al. 2017; Padmanabhan et al. 2020). Mutations in several of these interactors have been previously associated with cardiac malformations, highlighting the potential of our approach for disease-gene discovery (**Figure 2.2, A-B**).

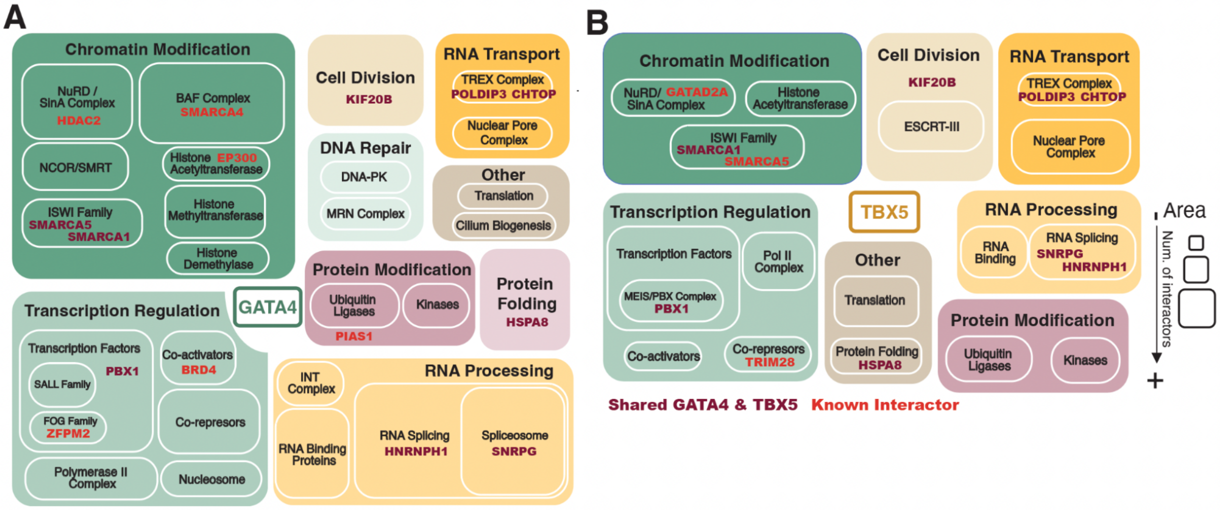


Figure 2.2: GATA4 **(A)** and TBX5 **(B)** interacting protein categories with boxed areas proportional to the number of interactors in each. Proteins interacting with both GATA4 and TBX5 (blue) or previously reported interactors (red) are highlighted.

Consistent with the interdependence of GATA4 and TBX5 during cardiac development, their networks showed some overlap, but the bulk of the detected interactors were unique to each cTF. Both networks were enriched in proteins involved in similar biological processes (**Figure 2.2, A-B**). The top two most represented processes were transcription regulation and chromatin modification (**Figure 2.3, A**), as expected from the cTFs' well-established functions in gene regulation. Both known and previously unreported low-abundance TFs were found to interact with GATA4 and/or TBX5 (e.g., ZFPM1, ZNF787, SALL3, ZNF219, and MAB21L2) demonstrating the sensitivity of the AP-MS approach. Chromatin modifiers (~25% or 15% of

GATA4 or TBX5 interactors, respectively) predominantly belonged to ATP-dependent complexes, and I found several histone-modifying enzymes in the GATA4-PPI (**Figure 2.2, A-B**) (Enane et al. 2017). A number of RNA processing and splicing proteins, as well as members of the nuclear pore complex, were also identified (**Figures 2.2, A-B**). The GT-PPIs mostly included proteins expressed ubiquitously, with a small number of tissue-enriched and cell-type enriched interactors (**Figure 2.3, B**).

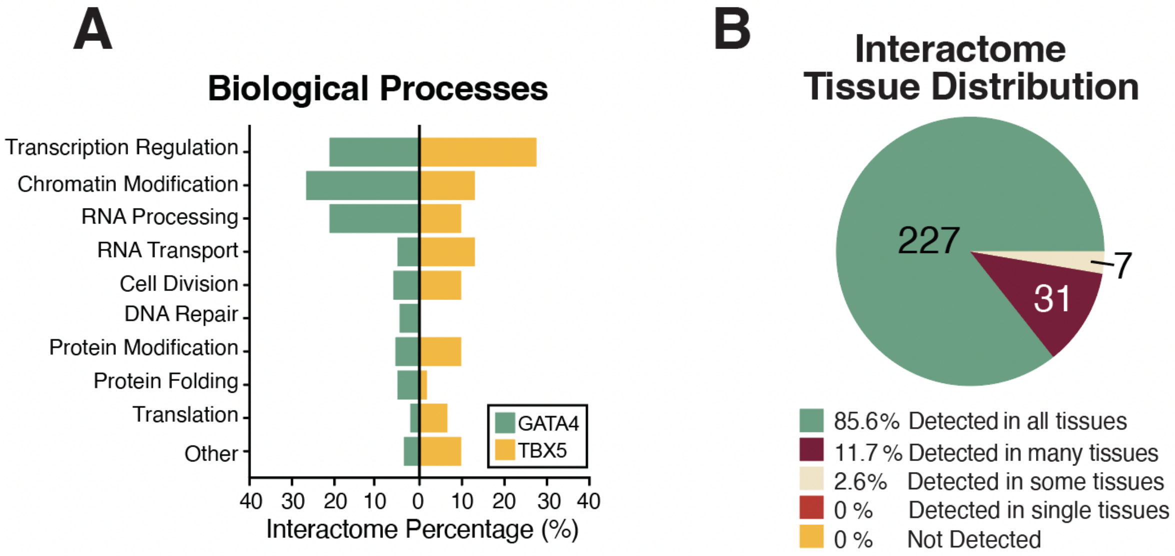


Figure 2.3: (A) Distribution of GATA4 and TBX5 PPIs in biological processes, as annotated in Figure 2.2. **(B)** Tissue expression distribution of GATA4 and TBX5 interactors across the six Human Protein Atlas categories based on transcript detection (tpm ≥ 1) in all 37 analyzed tissues.

2.3.2 GATA4:TBX5 interactome is enriched in proteins harboring de novo variants in CHD

To determine whether the GT interactors identified in human CPs might help predict genetic risk factors for CHD, I assessed their intersection with de novo variants (DNVs) and very rare (minor allele frequency < 10^{-5}) inherited loss-of-function (LoF) variants found in CHD probands from the PCGC. In addition to a previously published cohort of parent-offspring CHD trios and control trios (Jin et al. 2017), I processed and included variant data from an additional 419 CHD probands and their parents for a total of over 3,000 trios. A permutation-based statistical test was used to analyze the frequency of variants in GT-interacting proteins among the CHD probands compared to the control group (see **Methods 2.5.4: Permutation-based tests**). Briefly, the observed odds ratio (OR) of finding a DNV in an interactome gene was adjusted by a factor correcting for synonymous mutation frequency (adjusted OR), then compared to a distribution of odds ratios in which the case/control status of the dataset was permuted (permuted ORs) (**Figure 2.4, A**). The analysis indicated that protein-altering DNVs were significantly more likely to be found within GT interactors in the CHD cohort relative to the control cohort (adjusted OR GATA4-PPI: 5.59 and Bonferroni-adjusted p value 0.001; adjusted OR TBX5-PPI: 4.34 and Bonferroni-adjusted p value 0.0096). By contrast, very rare inherited LoF variants occurred in GT-PPI proteins with the same frequency in control and CHD groups (**Figure 2.4, B**).

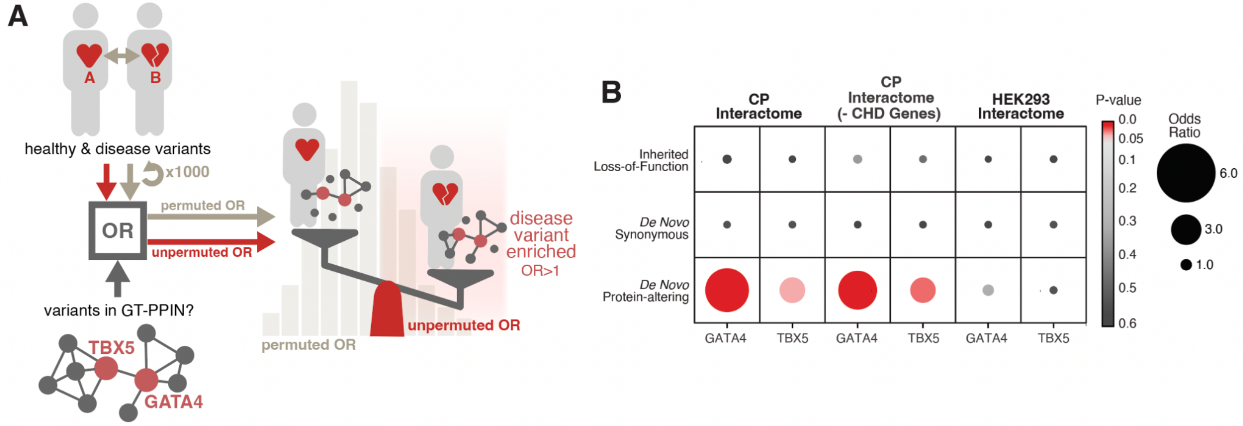


Figure 2.4: (A) Permutation-based statistical test design to analyze enrichment in genetic variants from a CHD cohort relative to a control cohort in GATA4 or TBX5 PPIs (odds ratio, OR). **(B)** Results of permutation-based test in (A) for genomic variation indicated from PCGC CHD and control cohorts within the GATA4 or TBX5 inter- actomes in cardiac progenitors (CP interactome), or after removing proteins involved in human or mouse cardiac malformations (CP interactome heart dev. unknown). The same analysis is shown for HEK293s (HEK293 interactome).

To determine whether the enrichment was predominantly driven by genes previously known to be involved in cardiac development, I removed a published curated list of genes involved in human or mouse cardiac malformations from the dataset (Jin et al. 2017) and repeated the permutation-based analysis (**Table 2.1**). Still an enrichment was found in proteins harboring protein-altering DNVs from CHD probands in both GATA4 and TBX5 interactomes (**Figure 2.4, B**). Similar trends were observed holding out a smaller list of 144 published human CHD genes (Izarzugaza et al. 2020).

Table 2.1: Odds ratios (OR) and p-values for permutation-based tests of protein-protein interactome (PPI) genes, restricted by novel to CHD (n-CHD) and Heart Development (n-HD).

TF interactome	Known Gene Status	Case PPI variant count	Case non-PPI variant count	Control PPI variant count	Control non-PPI variant count	Syn-adjusted OR	P-value (Bonferroni-corrected)
GATA4	n-HD	36	2341	15	1302	2.95	0.0024
GATA4	n-CHD	40	2362	14	1317	3.16	0.0008
TBX5	n-HD	11	2366	3	1314	2.60	0.592
TBX5	n-CHD	11	2391	3	1328	2.63	0.216

Although our AP-MS analysis was conducted in human CP cells for endogenous TBX5 and GATA4, most PPIs have been identified in less biologically relevant cells and upon overexpression. To assess the importance of biological context, co-authors generated GT-PPIs in kidney cells (HEK293) overexpressing human GATA4 or TBX5, again filtering based on nuclear localization, and subjected them to the same permutation analysis with the CHD and control cohorts. There was no significant enrichment in proteins harboring CHD-associated protein-altering DNVs for HEK cell interactomes (**Figure 2.4, B**). The GT-PPI overlap between cell types was small, with only 20 GATA4 and 13 TBX5-interactors shared, highlighting the importance of endogenous tissue-specific protein-protein interactions in elucidating the genetic underpinnings of diseases.

In a complementary analysis to test whether genes in the GT-PPI were enriched for protein-altering DNVs in CHD probands, I permuted the list of interactors and tallied the number of variants found in each gene set. This allowed us to compare the null distribution of the number of variants found in otherwise-comparable non-GT-PPI genes to what was observed in interactome genes. For each gene in the GT-PPI, I selected other genes that had comparable de novo mutability scores (Samocha et al. 2014), then further narrowed the list of matches based on similarity of expression levels in cardiac progenitor cells (de Soysa et al. 2019). The observed number of protein-altering DNVs was significantly higher in GT-PPI genes compared to permuted selections of non-interactome genes with similar mutability and expression (Bonferroni-adjusted p = 0.009).

Having demonstrated that the GT-PPI was enriched in protein-altering variants found in CHD patients, we aimed to assess the likelihood that the GT-PPI variants contribute to disease. Using combined annotation-dependent depletion (CADD) scores, I found that GT-PPI protein-altering variants found in the CHD cases were more likely predicted to be deleterious than the rest of protein-altering DNVs in CHD cases outside the GT interactome (**Figure 2.5**).

2.3.3 GATA4:TBX5 interactors with protein-altering DNVs unveil CHD candidate genes with characteristic features of disease genes

I next determined whether the candidate CHD genes identified in the GT-PPI exhibited features that could increase their likelihood of causing disease compared to the remaining non-interactome genes mutated in CHD probands. Extreme intolerance to LoF variation and haploinsufficiency are common features of genes associated with developmental disorders (Fuller et al. 2019). Remarkably, most candidate CHD genes in the PPI were extremely intolerant to LoF variation (probability of being intolerant to LoF [pLI] > 0.9) and exhibited significantly higher pLI and haplo-insufficiency scores than genes outside the interactome with protein-altering DNVs (**Figure 2.6, A**). Another feature of disease genes is an increased tendency for their products to interact with one another when their mutations result in similar phenotypes (Goh et al. 2007). Based on iRefIndex database information (Razick, Magklaras, and Donaldson 2008), I found that the proteins encoded by our candidate genes had a higher connectivity degree with other proteins found to be mutated in the CHD cohort, as well as with a curated list of proteins involved in mouse/human cardiac malformations (Jin et al. 2017) than proteins outside the interactome with protein-altering DNVs (**Figure 2.6, B-C**).

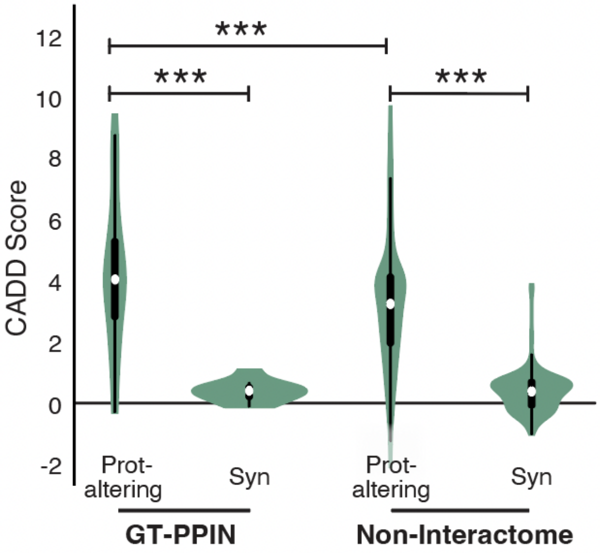


Figure 2.5: Violin plot for the Combined Annotation-Dependent Depletion (CADD) scores of protein-altering or synonymous (Syn) variants found in the CHD cohort affecting proteins within the GT-PPI or proteins outside the interactome. White dot = median; black lines = interquartile range (thick) or 1.53 the inter- quartile range (thin). Two-sided Mann-Whitney-Wilcoxon test with Bonferroni correction for p values; ***p value < 0.001.

GT interactors with protein-altering proband DNVs exhibited higher expression in the developing heart than genes with protein-altering DNVs outside the GT-PPIs (**Figure 2.6, D**), but they generally displayed a broad expression pattern across most cell types (**Figure 2.7, A**) and largely involved proteins relevant to chromatin biology (**Figure 2.7, B**). Other biological processes with unexplored roles in CHD were affected, such as RNA splicing and protein folding (**Figure 2.7, B**).

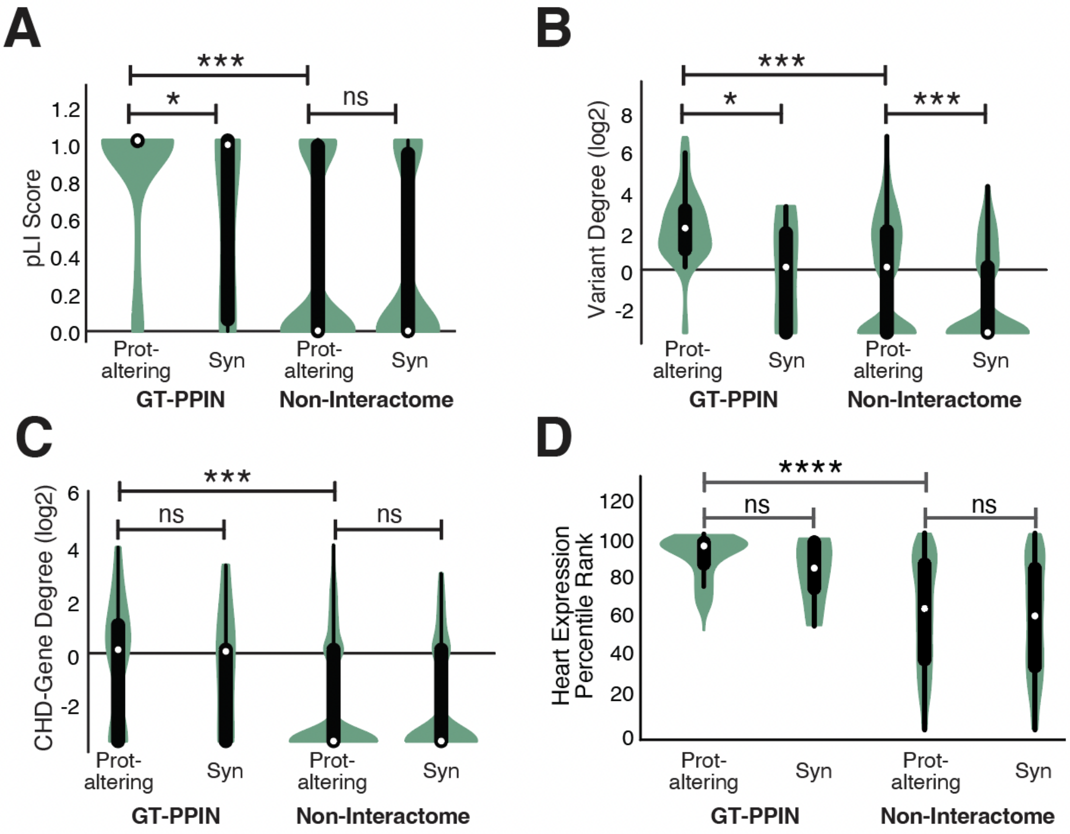


Figure 2.6: De novo variants in GATA4 and TBX5 interactomes exhibit features typical of disease genes. **(A–D)** Violin plots for the distribution of **(A)** intolerance to LoF (pLI Score), **(B)** degree of connectivity with all protein-altering DNVs found in the CHD cohort, **(C)** degree of connectivity with proteins encoded by genes involved in mouse/human cardiac malformations, and **(D)** expression percentile rank in the developing heart (E14.5) for genes harboring synonymous (Syn) or protein-altering DNVs found in the CHD cohort and affecting proteins inside the GT interactome (GT-PPI) or outside the interactome (non-interactome). White dot = median; black lines = interquartile range (thick) and 1.53 the interquartile range (thin). Two-sided Mann-Whitney-Wilcoxon test with Bonferroni correction for p values; ***p value < 0.001, **p value < 0.01, *p value < 0.05, and ns: non-significant.

We next investigated the specific types of protein-altering de novo CHD variants corresponding to proteins in the GT-PPIs. Among the 272 proteins in the GT-PPI, I identified 20 LoF DNVs and 53 missense DNVs present in CHD cases. The odds of a DNV occurring in a GT-PPI gene was substantially greater in CHD probands compared to controls for both LoF (adj. OR 4.96) and missense DNVs (adj. OR: 3.76). LoF DNVs preferentially affected genes involved in human and mouse cardiac malformations, whereas the bulk of GT-PPI genes with

CHD-missense DNVs had not previously been linked to cardiac development or CHD. The contribution of de novo splice variants could not be determined due to their low counts in interactome genes from cases and controls.

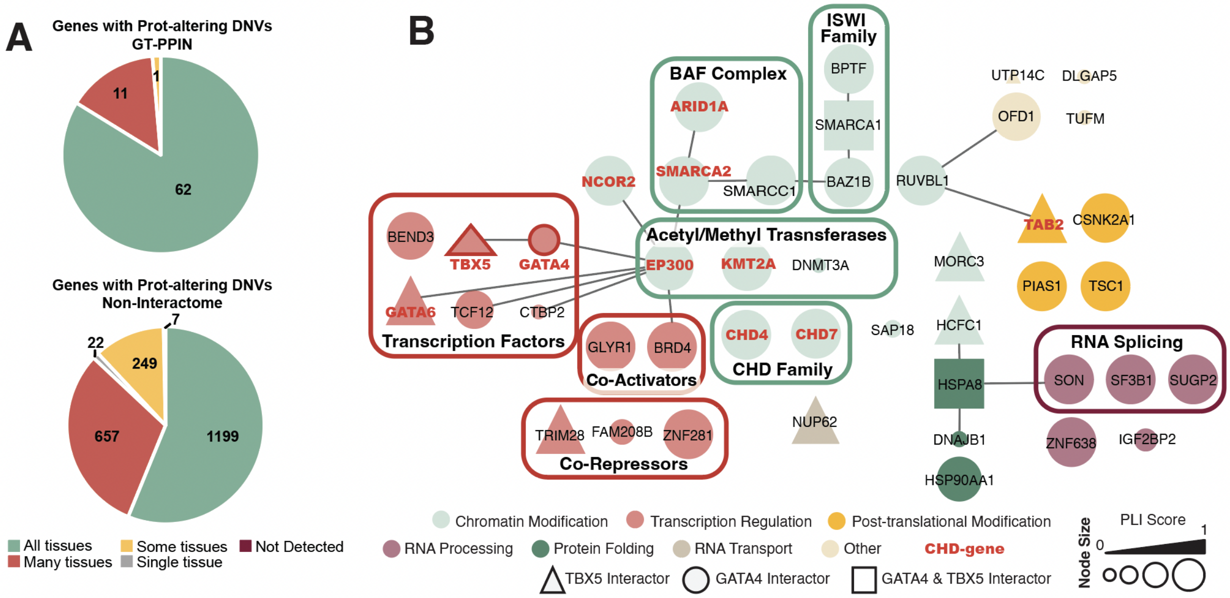


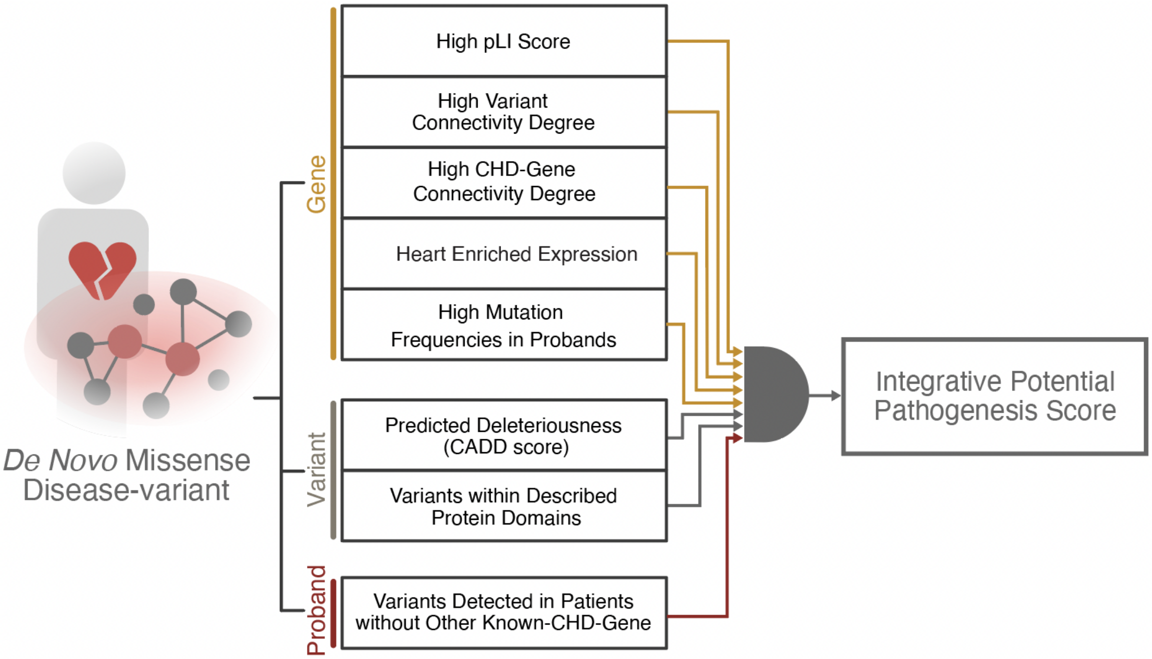
Figure 2.7: (A) Pie chart of tissue expression distribution of GT-PPI or non-interactome genes harboring protein-altering DNVs across the six Human Protein Atlas categories. **(B)** Interactome CHD candidate genes represented as a network after integration with PPI information from iRefIndex database. Nodes colored based on manually annotated biological processes and protein families/complexes grouped in boxed areas. Node size reflects probability of loss-of-function intolerance (pLI) scores. Node shape reflects belonging to TBX5 (triangle), GATA4 (circle), or GATA4&TBX5 (square) networks. Red highlights proteins encoded by genes involved in human CHD. Edges represent protein-protein interactions from iRefIndex database (Razick et al., 2008).

2.3.4 An integrative method for scoring variants identifies specific GT interactors as candidate genes for CHD

The GT-PPI framework combined with trio sequencing allowed me to significantly reduce the number of candidate variants in individual genomes to 20 LoF and 53 missense DNVs in genes encoding protein partners of cTFs that may contribute to CHD. However, even after this significant filtering step, the interpretation of missense variants remains a challenge and requires methods to prioritize those that could substantially impact human phenotypes.

Many variant prioritization methods have been described to date, and most integrate widely accepted variant and gene features to rank potential candidate variants based on the combined evidence of the variant's predicted deleterious effect on protein function, the harboring gene's accumulated mutational damage, and its biological relatedness to known CHD-causing genes (Eilbeck, Quinlan, and Yandell 2017; Köhler et al. 2008; Rentzsch et al. 2019; Sevim Bayrak et al. 2020). However, the development of a score that would work universally is theoretically difficult, and a common finding of many genetic studies is that gene-set specific rules for pathogenicity are required for proper evaluation (Eilbeck, Quinlan, and Yandell 2017). In addition, most of these methods were designed for singleton sequencing studies and fail to incorporate proband pedigree information that can aid prioritizing variants with potential greater effect within an individual (Farwell et al. 2015). Thus, we developed an integrative pipeline customized for the CHD trio whole-exome-sequencing dataset to calculate a variant prioritization score for the 53 missense DNVs mapped to our GT-PPI. This scoring method has two steps: (1) variant prioritization based on the consolidation of annotations from a combination of widely used gene and variant metrics to assess variant deleteriousness, together with the gene's frequency of mutation within our dataset, and (2) re-weighting based on occurrence at a known functional residue/domain and on the presence of other potentially causal variants in the same proband (**Figure 2.8**).

Specifically, at the gene level, a higher score indicates (1) a gene's low tolerance to LoF variation, (2) connectivity to a high number of proteins involved in cardiac malformations (Jin et al., 2017) and to PCGC CHD proband variant-harboring proteins based on publicly available PPI information, (3) high cardiac expression compared to other tissues, and (4) high number of PCGC variants within the gene relative to coding sequence (CDS) length. At the residue level, a higher score indicates (5) a variant's increased likelihood of being deleterious (CADD score) and (6) occurrence at a functional residue or protein domain. At the proband level, a higher score indicates that (7) the background genetic variation of this individual does not include DNVs or rare inherited LoF variants in genes known to be involved in cardiac malformations and includes none or fewer variants in other GT-PPI genes. The individual features are combined by rank sum and weighted where applicable (see **Methods 2.5.6: Variant scoring**) (**Figure 2.8**). The resulting score is represented with respect to the gene's percentile of expression in the developing heart (E14.5), a feature previously shown to be effective for variant filtering in CHD by the PCGC (Zaidi et al. 2013; Homsy et al. 2015; Jin et al. 2017; Sevim Bayrak et al. 2020).



De Novo Missense Disease-variant
Gene
High pLI Score
High Variant Connectivity Degree
High CHD-Gene Connectivity Degree
Heart Enriched Expression
High Mutation Frequencies in Probands
Variant
Predicted Deleteriousness (CADD score)
Variants within Described Protein Domains
Proband
Variants Detected in Patients without Other Known-CHD-Gene
Integrative Potential Pathogenesis Score

Figure 2.8: Variant prioritization score strategy (see Methods: 2.5.6 Variant Scoring).

I applied this scoring method to previously identified variants implicated in CHD (Basson et al. 1997; Garg et al. 2003; Furtado et al. 2017) and found that the method ranked these reference monogenic variants more highly than the few mutations demonstrated to partially contribute to CHD and cause oligogenic disease (**Figure 2.9**) (Gifford et al. 2019), even when the mutations affected the same gene. Furthermore, among the top-scoring interactome variants, there were several in proteins known to cause cardiac malformations, consistent with the relevance of this score for identifying gene variants with potential for contributing to disease (**Figure 2.9**).

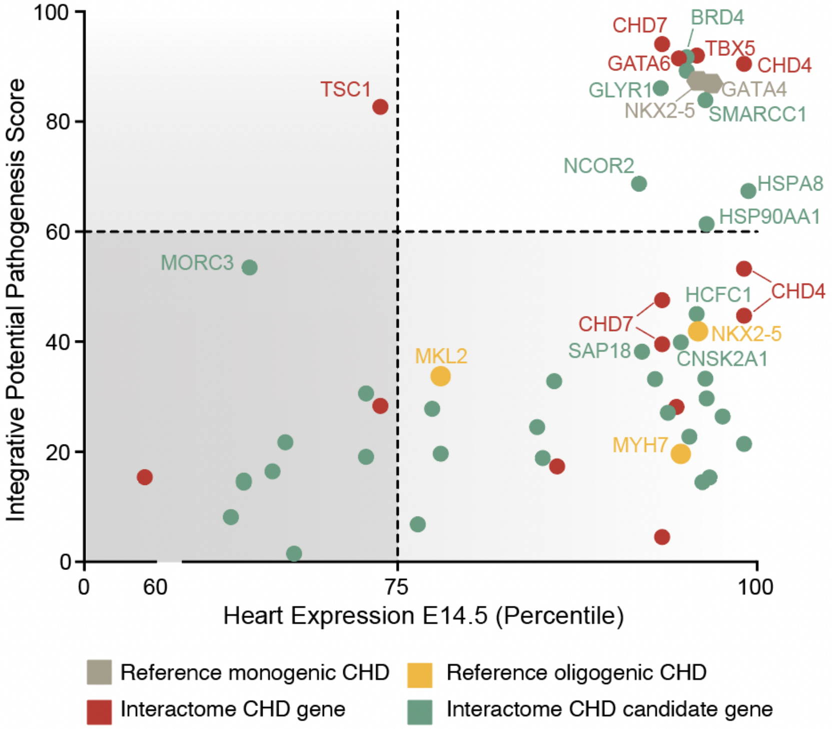


Figure 2.9: Variant prioritization scores for interactome missense DNVs in described CHD genes (red) or in CHD candidate genes (green) plotted against the corresponding genes' expression percentile rank in the developing heart (E14.5). Published mutations with strong contribution (gray) or partial contribution (yellow) to CHD are included as references.

In order to test whether higher variant prioritization scores indeed translated to greater functional impact of variants, we evaluated the effect of multiple variants on cofactor activity in a luciferase reporter assay using a luciferase reporter containing the PPARGC1a promoter, which is strongly activated by GATA4 (Padmanabhan et al. 2020). We selected NKX2-5, a reference gene with one high and one low scored variant; CHD7, a CHD gene encoding a GATA4 interactor with four identified missense DNVs; and BRD4 and SMARCC1, CHD candidate genes and GATA4 interactors, each with two identified missense DNVs in CHD patients. For the GATA4 interactors—CHD7, SMARCC1, and BRD4—each variant's impact on transcriptional activity was tested in the presence of GATA4 with a luciferase assay. We found that variants with a higher prioritization score exerted a greater effect on the encoded protein's transcriptional activity (**Figure 2.10**).

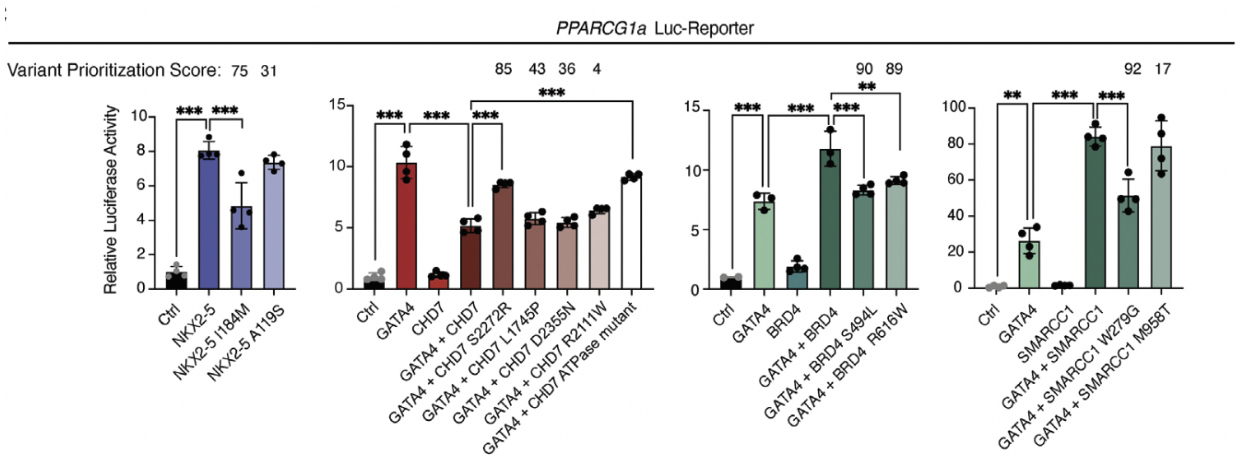


Figure 2.10: Biochemical evaluation by luciferase assays of the functional impact for variant alleles with different prioritization scores within NKX2-5, CHD7, BRD4, or SMARCC1. The CHD7 ATPase mutant is used as positive control for CHD7 loss of function (Liu et al. 2014). One-way ANOVA coupled with Tukey post hoc test: ***p value < 0.001; **p value < 0.01.

Next, I evaluated the benefit of the GATA4 and TBX5 PPI incorporation as a filtering strategy for the identification of CHD candidate genes by applying the variant prioritization method to all de novo missense variants from probands found in both interactome and non-interactome genes. The variant prioritization score's 75th percentile was 23 or 22 points higher (score range 0–99) for GT-PPI missense DNVs than for variants in genes outside the GT-PPI network or all unfiltered missense DNVs, respectively (**Figure 2.11, A**). Moreover, 41.5% of interactome missense DNVs ranked within the top quartile of all DNV prioritization scores, and within the top quartile of Developing Heart Expression percentile (Zaidi et al. 2013), compared to just 12.4% of unfiltered missense DNVs (**Figure 2.11, B)**.

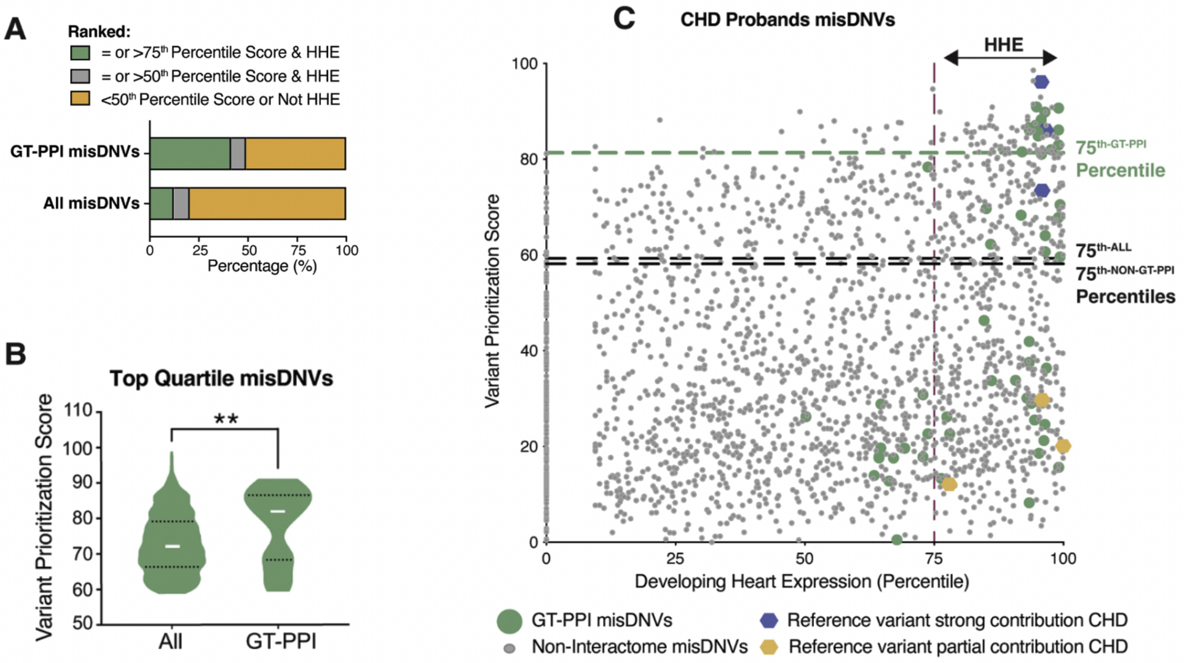


Figure 2.11: (A) Percentage of (All) versus interactome (GT-PPI) missense DNVs (misDNVs) in genes within the top quartile of Developing Heart Expression (High Heart Expressed genes, HHE) and the top quartile of Variant Prioritization Score (VPS) (green), the top quartile of Developing Heart Expression and the top half of VPS (gray), or below the 75th percentile of Developing Heart Expression or in the bottom half of VPS (orange). **(B)** Average VPS for all misDNVs and GT-PPI misDNVs within the top quartile of Developing Heart Expression and Variant Prioritization Score. The white line represents the median, the black lines the interquartile range. Unpaired Student's t test: ∗∗p value < 0.01. **(C)** Variant prioritization scores for all de novo missense variants from probands found in both interactome (green) and non-interactome (gray) genes plotted against the corresponding genes' expression percentile rank in the developing heart (E14.5), (Zaidi et al., 2013). Published mutations with monogenic contribution (blue) or partial contribution (orange) to CHD are included as references. Variant prioritization score's 75th percentile is higher for GT-PPI missense DNVs than for non-interactome variants (NON-GT-PPI) and all unfiltered missense DNVs. Genes within the top quartile of expression in the developing heart are indicated as High Heart expressed (HHE)**.**

Among the CHD candidate missense DNVs, the majority affected interactome proteins highly expressed in the developing heart, with only 25% occurring in GT interactors outside the top quartile of expression (**Figure 2.11, C**). The genes with lower heart expression generally also exhibited low variant prioritization scores, except for the tuberous sclerosis gene, TSC1, which is associated with cardiac rhabdomyomas (Hinton et al. 2014). On the other hand, missense DNVs in GT interactors highly expressed in the developing heart exhibited a broad range of prioritization scores, with a potentially highly pathogenic cluster of variants ranking close to the published reference variants with known strong contribution to CHD, and a more dispersed group of variants scoring similarly to the few reference variants with known partial contribution to CHD. Among the missense DNVs with the highest scores, which we hypothesized to be more significant contributors, there were four variants in GT interactors with previously described monogenic contribution to human cardiac defects (TBX5, GATA6, CHD4, and CHD7), and six variants within proteins with yet undescribed functions in human congenital heart malformations (BRD4 x2, SMARCC1, GLYR1, CSNK2A1, and SAP18) (**Figure 2.9**).

BRD4, GLYR1, and SMARCC1 are chromatin modifiers, in concordance with the observed enrichment of CHD-associated DNVs in genes involved in this process (Zaidi et al. 2013), and were detected as GATA4 interactors, which was validated by co-immunoprecipitation (Gonzalez-Teran et al. 2022). While GLYR1 and SMARCC1 were previously unknown to interact with GATA4, the Srivastava group has reported a role for a BRD4-GATA4 protein module in the regulation of cardiac mitochondrial homeostasis and showed that deletion of BRD4 during embryonic development resulted in embryonic lethality with signs of cardiac dysfunction (Padmanabhan et al. 2020). Although the specific contribution of *SMARCC1* to CHD is yet uncertain, its encoded protein, BAF155, is a component of the BAF complex, which orchestrates many aspects of heart development (Hota and Bruneau 2016). The GLYR1 DNV occurred in a patient with atrioventricular septal defects, left ventricle outflow tract obstruction, and pulmonary stenosis, a spectrum of cardiac malformations observed in humans with GATA4 mutations. However, the role of GLYR1 in most tissues (including the heart) remains unexplored.

I therefore investigated the genetic landscape of the GLYR1 variant carrier and identified three rare LoF and 62 rare missense variants inherited from their asymptomatic parents, while no other DNVs were found in this proband. Interestingly, one of these inherited missense variants occurred in GATA6, encoding a GATA factor that genetically interacts and is partially redundant, with GATA4 in cardiac development (Xin et al. 2006). Although these inherited LoF and missense variants are present in the asymptomatic parents and are therefore unlikely to be sufficient to cause cardiac malformations, future studies may assess if any could contribute together with GLYR1 to the cardiac malformations observed in this patient.

2.3.5. The CHD variant in GLYR1 destabilizes its physical interaction with GATA4

GLYR1, also known as NDF, NPAC, or NP60, is a chromatin reader involved in chromatin modification and regulation of gene expression through nucleosome demethylation (Fang et al. 2013; Fei et al. 2018; Marabelli et al. 2019). The GLYR1 missense CHD DNV I identified leads to the substitution of a highly conserved proline with a leucine at amino acid (aa) 496 within the β-hydroxyacid dehydrogenase domain, described to mediate the interaction between GLYR1 monomers (Marabelli et al. 2019; Montefiori et al. 2019). Co-immunoprecipitation assays demonstrated that the GLYR1 P496L DNV destabilized its physical interaction with GATA4 (**Figure 2.12, A**). Since previous studies indicated a role for GLYR1 in transcriptional regulation, we probed whether GLYR1 co-regulates gene expression together with GATA4 and found that co-transfection of GATA4 and GLYR1 increased Nppa-luciferase reporter activity by approximately 15-fold, compared with an 8-fold activation with GATA4 alone. Synergistic transactivation of the Ccnd2-luciferase reporter by GLYR1 and GATA4 was similarly observed and in both cases was attenuated by the GLYR1 P496L mutation (**Figures 2.12, B**).

Additional analyses performed by co-authors found that GATA4 and GLYR1 co-bind a defined set of heart development genes and co-regulate their expression (Gonzalez-Teran et al. 2022). In order to further assess the biological importance of the GLYR1 P496L variant in vivo, they generated a mouse line harboring a P495L single nucleotide variant in GLYR1 (Glyr1P495L/+), homologous to human P496L, using CRISPR-Cas9 mediated genome editing. Although all genotypes were born at the expected mendelian ratios, 54% of the

Glyr1P495L/P495L and 15.5% of the heterozygous (Glyr1P495L/+) mice displayed postnatal lethality between days 0 and 1, compared to only 4.4% of the WT littermates (p = 0.02 at P1). Thus, our findings provide evidence for the biological importance of GLYR1 in cardiac development and demonstrate a deleterious effect of the P495L variant in vivo.

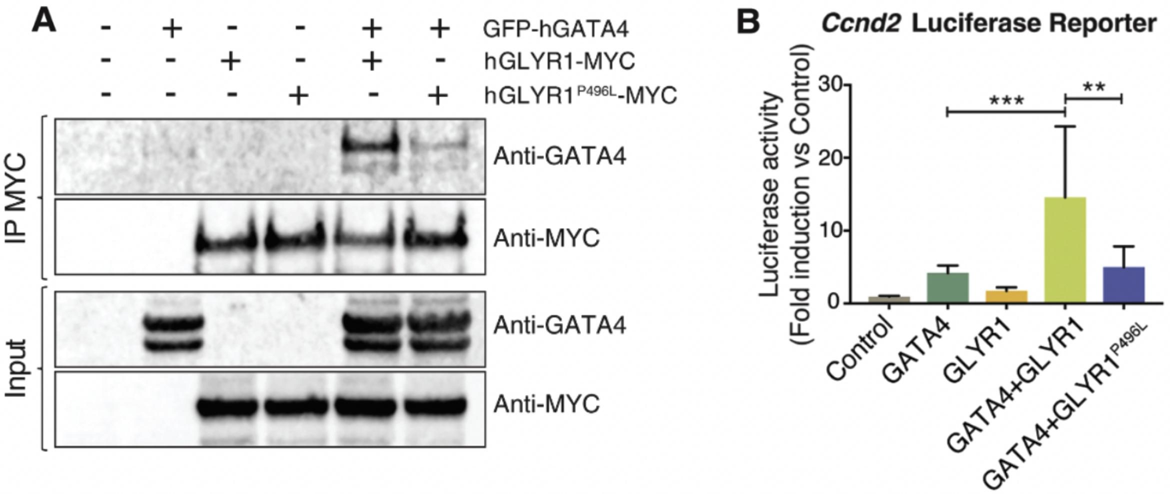


Figure 2.12: (A) The ability of GLYR1 WT or P496L mutant to interact with GATA4 by immunoprecipitation (IP) of GLYR1-MYC and immunoblotting with indicated antibodies. **(B)** Luciferase reporter assay in HeLa cells showing activation of the luciferase reporter upon addition of plasmids encoding indicated proteins. Equal amount of total transfected DNA per condition was adjusted with empty vector. (n = 3 independent experiments). One-way ANOVA coupled with Tukey post hoc test: $**$p value < 0.01, $***$p value < 0.001.

2.4 DISCUSSION

I integrated an analysis of the protein-protein interaction network of CHD-associated TFs with human whole-exome-sequencing data to inform the genetic underpinnings of CHD. A hypothesis-free PPI reconstruction for two essential cTFs, GATA4 and TBX5, identified known and previously unreported functional relationships. DNVs in GT-PPIs occurred with significantly greater frequency in CHD patients than healthy controls. Additionally, a consolidative computational framework devised to prioritize variants in GT-interacting proteins identified numerous candidate disease genes, including GLYR1, a ubiquitously expressed epigenetic reader. My co-authors demonstrated that the GLYR1 CHD proband variant P496L disrupted the interaction with GATA4 and co-activation of cardiac developmental genes. The importance of the GLYR1 variant and the GATA4-GLRY1 interaction in cardiac development was further confirmed in a mouse model. These findings indicate that the use of tissue- and disease-specific PPIs may partially overcome the genetic heterogeneity of CHDs and help prioritize the potential impact of de novo missense variants present in disease.

The integration of PPI information from publicly available databases with human genetic data has been previously used to prioritize disease candidate variants based on network topological measures (Köhler et al. 2008; Greene et al. 2015; Priest et al. 2016; Bryois et al. 2020; Izarzugaza et al. 2020). Since most of the available PPI datasets have been reconstructed in non-physiological settings and cell types not relevant to the disease of interest, some of these methods incorporated RNA expression information to generate predicted "tissue-specific" networks to reduce the number of candidate variants (Magger et al. 2012; Barshir et al. 2014). However, whether a protein-protein interaction indeed occurs in the tissue depends on additional factors, and co-expression of both partners is only a necessary initial requirement but not a guarantee for the interaction to occur. Even after the application of these prioritization strategies, the large number of highly ranked candidate variants makes it challenging to identify likely contributing mutations in the absence of additional biologically meaningful information. In contrast, the approach described here allowed us to capture ubiquitously expressed CHD candidate genes that might have tissue-specific effects due to their interaction with tissue-enriched factors. This is of importance, as the majority of known disease genes are broadly expressed across multiple human tissues.

In contrast to single-gene enrichment approaches, the network-enrichment analysis allows the detection of rare CHD candidate genes, but it does so without resolving the relative contributions of specific variants. Hence, downstream prioritization of candidate disease variants is needed to rank the likelihood that specific variants contribute to CHD. For this purpose, the integrative scoring method I developed combines commonly used disease-variant prioritization metrics, including diverse and complementary biological information at the gene and variant levels, together with proband pedigree information. The integration of proband genomic information regarding the co-occurrence of variants in known CHD genes with metrics that predict variant deleteriousness and gene-level parameters allowed prioritization of variants with potentially higher contribution to CHD. Functional investigation will be needed to test whether the identified CHD candidate genes are essential in heart development and to determine the causal nature of the associated variants, as we have done here with *GLYR1*. In the future, high-throughput screening methods similar to our integrative PPI-genetic variant scoring pipeline will aid in assessing the vast genomic variation catalog provided by the increasing number of large-scale sequencing studies.

Our integrative proteomics and human genetics approach revealed GLYR1 as a GATA4 interactor in CPs that constitutes a strong candidate gene for CHD. During CM differentiation, physical interaction between GATA4 and GLYR1 may be one of the mechanisms explaining how GLYR1 can bind a specific subset of heart development genes. Disruption of this co-regulation in the context of the P496L variant has detrimental effects in CM differentiation that may contribute to cardiac malformations. Indeed, the genetic interaction observed in mice compound heterozygous for GATA4 and GLYR1 P496L, with a high incidence of atrioventricular septal defects, is in agreement with the GLYR1 variant playing a role in CHD.

Overall, this work has identified interactors of TFs essential for cardiac development, provided a ranked list of candidate disease variants potentially contributing to CHD, and revealed biology of gene regulation related to cardiac disease. Notably, this tissue- and disease-specific TF network-based approach could be applied with slight modifications of the variant prioritization scoring to other genetic disorders for which large-scale sequencing data are available to highlight disease mechanisms and provide a powerful filter for interrogating the genetic basis of disease.

The variant prioritization scoring developed in this study is limited in that it has been customized specifically for the CHD variant dataset from trio whole-exome sequencing and designed as a complementary method to our interactome filtering approach. Although the principles could be widely applicable to other genetic diseases, disease- and dataset-specific modifications would be necessary for its application in different disease contexts. Further, this study focuses on very rare variants, which are normally depleted from the population, as is the case for the P496L variant in GLYR1. Indeed, as a *de novo* variant, the likelihood of observing P496L in another sequenced individual is small. Future studies will be needed to determine whether additional variants in GLYR1 contribute to other cases of CHD or to other diseases in humans.

2.5 MATERIALS AND METHODS

2.5.1: Selection of interactome proteins

Cell culture and mass spectrometry analysis were performed by [contributions] as described in publication (Gonzalez-Teran et al. 2022), yielding peptide, protein, and site-level spectral counts. I analyzed count data using the artMS package (Jimenez-Morales et al. 2020) in R followed by protein-protein interaction scoring by the SAINTq software (Teo et al. 2016) to identify significantly-interacting proteins for GATA4 and TBX5 baits. Default parameters for both softwares were used except where indicated here: To create the GATA4 interactome, I analyzed at the protein level and select proteins that interact at a BFDR cutoff of ≤0.001; to create the TBX5 interactome, I analyzed at the peptide level and selected those that interact at a BFDR cutoff of ≤0.05. Intensity data from the control (knockout) cell lines was normalized per SAINTq configuration options such that the average total intensity in each bait purification was equal to the average total intensity across the control experiments.

To focus on transcriptionally-relevant interactions, I additionally filtered proteins by those that appear in the nuclear compartment, those that are expressed at detectable levels in at least one of the same cell types as the bait, and proteins whose gene expression was significantly lower in the control line but did not have a greater than 0.5 log-fold change drop in intensity.

Nuclear compartment: nuclear compartment genes were identified using the Cytoscape package BiNGO (Shannon et al. 2003; Maere, Heymans, and Kuiper 2005) with additional manual curation from literature. Cell type co-expression: single-cell RNA-seq data from deSoysa et al. (2019) was used to determine if an interaction was likely to occur, given co-expression in the same cell type. Briefly, seven cell type populations were selected from mesoderm and neural crest cells in the developing heart (multipotent Isl1+ progenitors, endothelial or endocardial cells, epicardium, myocardium, neural crest-derived mesenchyme, paraxial mesoderm and lateral plate mesoderm) (de Soysa et al. 2019). A bait protein was considered to be expressed in one of these cell types if the transcripts per million (tpm) for the bait gene were greater than 0.05 tpm. Prey proteins were considered to be potentially physiologically relevant interactors if they were detected at any level in one of the same cell types as the bait.

Controlling for differential gene expression: protein hits that were considered likely false positives based on lower expression in the control cell lines, without concomitant reduction in protein intensity, were removed from the interactome list. This is intended to control for genes that are expressed less in the controls due to bait knockout, but whose APMS protein intensities do not change (suggesting the protein pulled down was background rather than an interactor). Significant differential gene expression was determined in R using the edgeR package (Robinson et al., 2010); normalized protein intensities were averaged in all control experiments and bait experiments. Proteins with significantly reduced expression in control (FDR ≤ 0.05) with less than a 0.5 log-fold change drop in intensity were not considered to be interactors.

2.5.2: Gene expression tissue distribution and specificity

I used the categories for gene expression tissue distribution and tissue specificity defined by the Tissue Atlas within the Human Protein Atlas to classify interactome gene groups (https://www.proteinatlas.org/humanproteome/tissue/tissue+specific). These classifications are based on transcriptomics analysis across all major organs and tissue types in the human body, where all putative 19670 protein coding genes have been classified with regard to abundance and distribution of transcribed mRNA molecules (Uhlén et al. 2015).

Specificity illustrates the number of genes with elevated or non-elevated expression. Elevated expression includes three sub-category types:

- Tissue enriched: At least four-fold higher mRNA level in a particular tissue compared to any other tissues.
- Group enriched: At least four-fold higher average mRNA level in a group of 2-5 tissues compared to any other tissue.
- Tissue enhanced: At least four-fold higher mRNA level in a particular tissue compared to the average level in all other tissues.

Distribution, on the other hand, visualizes how many genes that do or do not have detectable levels (tpm ≥ 1) of transcribed mRNA molecules. Elevated genes are categorized as:

- Detected in single: Detected in a single tissue
- Detected in some: Detected in more than one but less than one third of tissues
- Detected in many: Detected in at least a third but not all tissues
- Detected in all: Detected in all tissues

2.5.3: Variant calling

Whole Exome Sequencing data from 2645 CHD trios and 1789 control trios was processed as described and published in (Jin et al. 2017). We include Whole Exome Sequencing data from 419 additional CHD trios recruited to the Pediatric Cardiac Genomics Consortium (PCGC), processed by the HMS pipeline as described in Jin et al., 2017. I filtered protein-coding mutations based on a Mapping Quality score > 59 and Genotype Quality > 90, then annotated qualifying variants using ANNOVAR. De novo variants were called using the

TrioDeNovo program (Wei et al. 2015), and accepted if the minor allele frequency (MAF) and read-depth criteria described in Homsy et al., 2015 are met. Namely, the in-cohort MAF of the variant must be below 4x10−4, with a minimum of 5 alternative reads and 10 total reads in the proband, and a minimum of 10 reference reads in the parents (with a maximum alternate allele ratio of 3.5%).

2.5.4: Permutation-based tests

Case-control permutation: I tested the adjusted odds ratio of observing a *de novo* mutation in an interactome gene in CHD probands relative to controls. I ran 10,000 permutations in which case/control status was randomly shuffled to generate a null distribution of permuted odds ratios (ORs). This was performed for protein-altering (non-synonymous) *de novo* mutations, synonymous *de novo* mutations, and rare inherited loss-of-function mutations (at minor allele frequency 10^{-5}) (Jin et al., 2017) on the GATA4 and TBX5 interactomes generated from both cardiac progenitor and HEK293 APMS experiments. The raw p value for each test is equal to the proportion of random shuffles with a permuted OR greater than or equal to the observed OR. Raw p-values were adjusted for multiple testing using the Bonferroni correction.

We observed that some genes appeared to have been more deeply sequenced in control individuals, while other genes showed the opposite trend. This is not unexpected, as control individuals in the Jin et al. dataset were sequenced for a different study and at different institutions from PCGC individuals. Therefore, to control for regional biases in sequencing between the case and control studies, I created an adjusted odds ratio metric that multiplies synonymous and protein-altering variant ORs by a factor restricting the synonymous odds ratio to 1 (null expectation). This correction was performed for the observed odds ratio and the odds ratios calculated in each permutation of *de novo* variants. To determine whether this signal was driven by already-identified CHD risk genes, I repeated the analysis after removing *de novo* variants occurring in known known Human/Mouse CHD genes (sourced from Jin et al., 2017, Supplemental Dataset 2: 253 Curated Genes) (Jin et al., 2017) as well as after removing a curated list of 144 human CHD-genes (Izarzugaza et al., 2020).

Gene set permutation: For each gene in the GT-PPI, I identified non-interactome genes that are expressed at similar levels in WT CP cells, and have comparable mutability scores as calculated by (Samocha et al. 2014). A gene is considered a match if its mutability score (expected number of *de novo* mutations in this gene per chromosome per generation) is equivalent when rounded to the order of one hundred-thousandth. I further filtered the list of matches based on similarity of expression levels in WTC-11 cardiac progenitor cells, such that the measured transcripts per million (tpm) is equivalent to the order of one one-hundredth (de Soysa et al. 2019). For genes with fewer than 100 possible matches, we relax these requirements by ± 0.5×10^N (where N yields the relevant order of magnitude), and remove any genes from the analysis in the case of < 10 matches. For 1000 permutations, we permute each interactome gene into one from its list of comparable non-interactome genes to compare the total count of variants found in CHD cases from the GT-PPI interactome versus those across all permuted gene-sets.

2.5.5: Characterization of CHD candidate variants and genes

All *de novo* variants and harboring genes observed in CHD probands and matched controls were assessed for the following properties: CADD score, pLI score, variant degree, CHD-gene degree, heart expression percentile rank, haploinsufficiency, and number of mutations per kilobase. The residue-level CADD score (Rentzsch et al. 2019) estimates the likely deleteriousness of a variant based on conservation data. pLI score indicates the predicted loss-of-function intolerance of the gene, scaled between 0 and 1, and was sourced from gnomAD version 2.1.1 (Karczewski et al. 2019). Similarly, haploinsufficiency predicts the deleteriousness of having only a single functional copy of a gene. I used the predicted haploinsufficiency values from Huang 2010 (Huang et al. 2010). The CHD-gene degree counts the number of protein-protein interactions that the gene shares with previously-identified CHD risk genes, while the variant degree counts the number of protein-protein interactions shared with other genes that had *de novo* variants (DNVs) in a CHD proband. These node degree counts were normalized by the total number of connections observed in the gene, and are based on known mammalian protein-protein interactions in iRefIndex version 15.0 (Razick, Magklaras, and Donaldson 2008). Finally, the number of mutations per kilobase measures the number of times a *de novo* or rare loss of function variant was observed in a CHD proband, normalized by the coding length of that gene. I used a Mann-Whitney U test with Bonferroni correction to assess whether protein-altering DNVs in interactome genes differ significantly from those in non-interactome genes with respect to these properties, as well as whether protein-altering DNVs in cases differ from those found in controls.

2.5.6: Variant Scoring

All protein-altering de novo missense variants occurring in GT-interacting genes and observed in CHD probands were ranked based on a series of gene-level, residue-level, and patient-level properties. A mutations per kilobase value was determined for each gene, based on the number of protein-altering de novo and rare loss-of-function mutations found in CHD probands in the PCGC, normalized by the CDS length of that gene in the gnomAD database (Karczewski et al. 2020). pLI score indicates the predicted loss-of-function intolerance of the gene, scaled between 0 and 1 where 1 is more intolerant. pLI data was sourced from gnomAD version 2.1.1 (Karczewski et al. 2020). CHD-gene degree, variant degree, and mutations per kilobase values were calculated as described above based on known mammalian protein-protein interactions in iRefIndex version 15.0 (Razick, Magklaras, and Donaldson 2008) (see **2.5.5: Characterization of CHD candidate variants and genes**). Expression specificity was calculated using data from median transcripts-per-million (tpm) as published in GTEx version 8.1.1.9 (GTEx Consortium et al. 2017). Average median tpm was calculated for heart tissues (adult atrium, adult left ventricle) and all other available tissues with the exception of testis. The specificity score is then defined as the average tpm in heart tissues normalized by average tpm across all tissues.

For each of these properties, the variants were ranked based on their relative scores. Ties were resolved by taking the average value of the would-be ranks. Missing data was imputed to the median value of the given property. Gene-level rankings (mutations per kilobase, pLI score, CHD-gene degree, variant degree, and expression specificity) and residue-level rankings (CADD score) were separately averaged and then summed. This average rank sum was then additionally weighted by two factors to capture aspects of their proband-level and protein contexts.

First, if the proband had additional mutations in other interactome genes or other previously-identified CHD genes, we reduced the variant's weight. Specifically, we multiply the rank-sum score by the lowest-applicable factor if they meet any of the conditions in **Table 2.2.**

Table 2.2: Factors for weighting variants based on proband background genetics.

Factor	Conditions
0.75	Proband has another rare (MAF 10−5) inherited loss-of-function OR missense de novo variant in an interactome gene OR proband has an inherited missense damaging variant in a known CHD gene.
0.50	Proband has a predicted-damaging de novo mutation in an interactome gene or rare inherited loss-of-function mutation in a previously-identified CHD gene
0.25	Proband has a de novo missense mutation in a previously-identified CHD gene
0.10	Proband had a de novo missense mutation in a previously-identified CHD gene, and that variant was predicted-damaging or led to protein loss-of-function.

To summarize, the variant is down-weighted in cases where it is likely that another mutation in the proband is causing or contributing to the CHD phenotype.

Second, if the de novo variant leads to protein loss-of-function, or if it occurred in a known protein domain (and therefore is suspected to interfere with protein activity), the variant rank-sum was transmitted as-is. Otherwise, the variant's rank-sum was multiplied by 0.5.

2.6 DATA AND SOFTWARE ACCESS

PCGC variants are available under dbGaP Study Accession: phs000571.v6.p2 for qualifying researchers. Code is available at https://github.com/mepittman/ctf-apms.

REFERENCES

Ang, Yen-Sin, Renee N. Rivas, Alexandre J. S. Ribeiro, Rohith Srivas, Janell Rivera, Nicole R. Stone, Karishma Pratt, et al. 2016. “Disease Model of GATA4 Mutation Reveals Transcription Factor Cooperativity in Human Cardiogenesis.” *Cell* 167 (7): 1734–49.e22.

Barshir, Ruth, Omer Shwartz, Ilan Y. Smoly, and Esti Yeger-Lotem. 2014. “Comparative Analysis of Human Tissue Interactomes Reveals Factors Leading to Tissue-Specific Manifestation of Hereditary Diseases.” *PLoS Computational Biology* 10 (6): e1003632.

Basson, Craig T., David R. Bachinsky, Robert C. Lin, Tatjana Levi, Jacob A. Elkins, Johann Soults, David Grayzel, et al. 1997. “Mutations in Human Cause Limb and Cardiac Malformation in Holt-Oram Syndrome.” *Nature Genetics* 15 (1): 30–35.

Bruneau, B. G., M. Logan, N. Davis, T. Levi, C. J. Tabin, J. G. Seidman, and C. E. Seidman. 1999. “Chamber-Specific Cardiac Expression of Tbx5 and Heart Defects in Holt-Oram Syndrome.” *Developmental Biology* 211 (1): 100–108.

Bruneau, B. G., G. Nemer, J. P. Schmitt, F. Charron, L. Robitaille, S. Caron, D. A. Conner, et al. 2001. “A Murine Model of Holt-Oram Syndrome Defines Roles of the T-Box Transcription Factor Tbx5 in Cardiogenesis and Disease.” *Cell* 106 (6): 709–21.

Bryois, Julien, Nathan G. Skene, Thomas Folkmann Hansen, Lisette J. A. Kogelman, Hunna J. Watson, Zijing Liu, Eating Disorders Working Group of the Psychiatric Genomics Consortium, et al. 2020. “Genetic Identification of Cell Types Underlying Brain Complex Traits Yields Insights into the Etiology of Parkinson’s Disease.” *Nature Genetics* 52 (5): 482–93.

Christianson, A., C. Howson, and B. Modell. 2005. “March of Dimes: Global Report on Birth Defects, the Hidden Toll of Dying and Disabled Children.” https://www.semanticscholar.org/paper/35d6f76146f47b7a79ecbe1d2e6098689db81eb2 .

Deciphering Developmental Disorders Study. 2015. “Large-Scale Discovery of Novel Genetic Causes of Developmental Disorders.” *Nature* 519 (7542): 223–28.

Eilbeck, Karen, Aaron Quinlan, and Mark Yandell. 2017. “Settling the Score: Variant Prioritization and Mendelian Disease.” *Nature Reviews. Genetics* 18 (10): 599–612.

Enane, Francis O., Wai Ho Shuen, Xiaorong Gu, Ebrahem Quteba, Bartlomiej Przychodzen, Hideki Makishima, Juraj Bodo, et al. 2017. “GATA4 Loss of Function in Liver Cancer Impedes Precursor to Hepatocyte Transition.” *The Journal of Clinical Investigation* 127 (9): 3527–42.

Fang, Rui, Fei Chen, Zhenghong Dong, Di Hu, Andrew J. Barbera, Erin A. Clark, Jian Fang, et al. 2013. “LSD2/KDM1B and Its Cofactor NPAC/GLYR1 Endow a Structural and Molecular Model for Regulation of H3K4 Demethylation.” *Molecular Cell* 49 (3): 558–70.

Farwell, Kelly D., Layla Shahmirzadi, Dima El-Khechen, Zöe Powis, Elizabeth C. Chao, Brigette Tippin Davis, Ruth M. Baxter, et al. 2015. “Enhanced Utility of Family-Centered Diagnostic Exome Sequencing with Inheritance Model–based Analysis: Results from 500 Unselected Families with Undiagnosed Genetic Conditions.” *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 17 (7): 578–86.

Fei, Jia, Haruhiko Ishii, Marten A. Hoeksema, Franz Meitinger, George A. Kassavetis, Christopher K. Glass, Bing Ren, and James T. Kadonaga. 2018. “NDF, a Nucleosome-Destabilizing Factor That Facilitates Transcription through Nucleosomes.” *Genes & Development* 32 (9-10): 682–94.

Fuller, Zachary L., Jeremy J. Berg, Hakhamanesh Mostafavi, Guy Sella, and Molly Przeworski. 2019. “Measuring Intolerance to Mutation in Human Genetics.” *Nature Genetics* 51 (5): 772–76.

Furtado, Milena B., D. Jo Merriner, Silke Berger, Danielle Rhodes, Duangporn Jamsai, and Moira K. O’Bryan. 2017. “Mutations in the Katnb1 Gene Cause Left-Right Asymmetry and Heart Defects.” *Developmental Dynamics: An Official Publication of the American*

Association of Anatomists 246 (12): 1027–35.

Garg, Vidu, Irfan S. Kathiriya, Robert Barnes, Marie K. Schluterman, Isabelle N. King, Cheryl A. Butler, Caryn R. Rothrock, et al. 2003. “GATA4 Mutations Cause Human Congenital Heart Defects and Reveal an Interaction with TBX5.” *Nature* 424 (6947): 443–47.

Gifford, Casey A., Sanjeev S. Ranade, Ryan Samarakoon, Hazel T. Salunga, T. Yvanka de Soysa, Yu Huang, Ping Zhou, et al. 2019. “Oligogenic Inheritance of a Human Heart Disease Involving a Genetic Modifier.” *Science* 364 (6443): 865–70.

Goh, Kwang-Il, Michael E. Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási. 2007. “The Human Disease Network.” *Proceedings of the National Academy of Sciences of the United States of America* 104 (21): 8685–90.

Gonzalez-Teran, Barbara, Maureen Pittman, Franco Felix, Reuben Thomas, Desmond Richmond-Buccola, Ruth Hüttenhain, Krishna Choudhary, et al. 2022. “Transcription Factor Protein Interactomes Reveal Genetic Determinants in Heart Disease.” *Cell* 185 (5): 794–814.e30.

Greene, Casey S., Arjun Krishnan, Aaron K. Wong, Emanuela Ricciotti, Rene A. Zelaya, Daniel S. Himmelstein, Ran Zhang, et al. 2015. “Understanding Multicellular Function and Disease with Human Tissue-Specific Networks.” *Nature Genetics* 47 (6): 569–76.

GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, et al. 2017. “Genetic Effects on Gene Expression across Human Tissues.” *Nature* 550 (7675): 204–13.

Hekselman, Idan, and Esti Yeger-Lotem. 2020. “Mechanisms of Tissue and Cell-Type Specificity in Heritable Traits and Diseases.” *Nature Reviews. Genetics* 21 (3): 137–50.

Hinton, Robert B., Ashwin Prakash, Robb L. Romp, Darcy A. Krueger, Timothy K. Knilans, and International Tuberous Sclerosis Consensus Group. 2014. “Cardiovascular Manifestations of Tuberous Sclerosis Complex and Summary of the Revised Diagnostic

Criteria and Surveillance and Management Recommendations from the International Tuberous Sclerosis Consensus Group." *Journal of the American Heart Association* 3 (6): e001493.

Homsy, Jason, Samir Zaidi, Yufeng Shen, James S. Ware, Kaitlin E. Samocha, Konrad J. Karczewski, Steven R. DePalma, et al. 2015. "De Novo Mutations in Congenital Heart Disease with Neurodevelopmental and Other Congenital Anomalies." *Science* 350 (6265): 1262–66.

Hota, Swetansu K., and Benoit G. Bruneau. 2016. "ATP-Dependent Chromatin Remodeling during Mammalian Development." *Development* 143 (16): 2882–97.

Huang, Ni, Insuk Lee, Edward M. Marcotte, and Matthew E. Hurles. 2010. "Characterising and Predicting Haploinsufficiency in the Human Genome." *PLoS Genetics* 6 (10): e1001154.

Izarzugaza, Jose M. G., Sabrina G. Ellesøe, Canan Doganli, Natasja Spring Ehlers, Marlene D. Dalgaard, Enrique Audain, Gregor Dombrowsky, et al. 2020. "Systems Genetics Analysis Identifies Calcium-Signaling Defects as Novel Cause of Congenital Heart Disease." *Genome Medicine* 12 (1): 76.

Jimenez-Morales, D., Rosa Campos A, J. Von Dollen, and D. Swaney. 2020. *artMS: Analytical R Tools for Mass Spectrometry* (version R package version 1.6.5). http://artms.org.

Jin, Sheng Chih, Jason Homsy, Samir Zaidi, Qiongshi Lu, Sarah Morton, Steven R. DePalma, Xue Zeng, et al. 2017. "Contribution of Rare Inherited and de Novo Variants in 2,871 Congenital Heart Disease Probands." *Nature Genetics* 49 (11): 1593–1601.

Karczewski, Konrad J., Laurent C. Francioli, Grace Tiao, Beryl B. Cummings, Jessica Alföldi, Qingbo Wang, Ryan L. Collins, et al. 2020. "The Mutational Constraint Spectrum Quantified from Variation in 141,456 Humans." *Nature* 581 (7809): 434–43.

Kathiriya, Irfan S., Kavitha S. Rao, Giovanni Iacono, W. Patrick Devine, Andrew P. Blair, Swetansu K. Hota, Michael H. Lai, et al. 2021. "Modeling Human TBX5 Haploinsufficiency Predicts Regulatory Networks for Congenital Heart Disease."

Developmental Cell 56 (3): 292–309.e9.

Köhler, Sebastian, Sebastian Bauer, Denise Horn, and Peter N. Robinson. 2008. “Walking the Interactome for Prioritization of Candidate Disease Genes.” *American Journal of Human Genetics* 82 (4): 949–58.

Kuo, C. T., E. E. Morrisey, R. Anandappa, K. Sigrist, M. M. Lu, M. S. Parmacek, C. Soudais, and J. M. Leiden. 1997. “GATA4 Transcription Factor Is Required for Ventral Morphogenesis and Heart Tube Formation.” *Genes & Development* 11 (8): 1048–60.

Lambert, Samuel A., Arttu Jolma, Laura F. Campitelli, Pratyush K. Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, Jussi Taipale, Timothy R. Hughes, and Matthew T. Weirauch. 2018. “The Human Transcription Factors.” *Cell* 172 (4): 650–65.

Liu, Yuelong, Cristina Harmelink, Yin Peng, Yunjia Chen, Qin Wang, and Kai Jiao. 2014. “CHD7 Interacts with BMP R-SMADs to Epigenetically Regulate Cardiogenesis in Mice.” *Human Molecular Genetics* 23 (8): 2145–56.

Luna-Zurita, Luis, Christian U. Stirnimann, Sebastian Glatt, Bogac L. Kaynak, Sean Thomas, Florence Baudin, Md Abul Hassan Samee, et al. 2016. “Complex Interdependence Regulates Heterotypic Transcription Factor Distribution and Coordinates Cardiogenesis.” *Cell* 164 (5): 999–1014.

Maere, Steven, Karel Heymans, and Martin Kuiper. 2005. “BiNGO: A Cytoscape Plugin to Assess Overrepresentation of Gene Ontology Categories in Biological Networks.” *Bioinformatics* 21 (16): 3448–49.

Magger, Oded, Yedael Y. Waldman, Eytan Ruppin, and Roded Sharan. 2012. “Enhancing the Prioritization of Disease-Causing Genes through Tissue Specific Protein Interaction Networks.” *PLoS Computational Biology* 8 (9): e1002690.

Maitra, Meenakshi, Sara N. Koenig, Deepak Srivastava, and Vidu Garg. 2010. “Identification of GATA6 Sequence Variants in Patients with Congenital Heart Defects.” *Pediatric Research* 68 (4): 281–85.

Marabelli, Chiara, Biagina Marrocco, Simona Pilotto, Sagar Chittori, Sarah Picaud, Sara Marchese, Giuseppe Ciossani, et al. 2019. “A Tail-Based Mechanism Drives Nucleosome Demethylation by the LSD2/NPAC Multimeric Complex.” *Cell Reports* 27 (2): 387–99.e7.

Montefiori, Marco, Simona Pilotto, Chiara Marabelli, Elisabetta Moroni, Mariarosaria Ferraro, Stefano A. Serapian, Andrea Mattevi, and Giorgio Colombo. 2019. “Impact of Mutations on NPAC Structural Dynamics: Mechanistic Insights from MD Simulations.” *Journal of Chemical Information and Modeling* 59 (9): 3927–37.

Moskowitz, Ivan P., Jun Wang, Michael A. Peterson, William T. Pu, Alexander C. Mackinnon, Leif Oxburgh, Gerald C. Chu, et al. 2011. “Transcription Factor Genes Smad4 and Gata4 Cooperatively Regulate Cardiac Valve Development. [corrected].” *Proceedings of the National Academy of Sciences of the United States of America* 108 (10): 4006–11.

Narita, N., M. Bielinska, and D. B. Wilson. 1997. “Wild-Type Endoderm Abrogates the Ventral Developmental Defects Associated with GATA-4 Deficiency in the Mouse.” *Developmental Biology* 189 (2): 270–74.

Oka, Toru, Marjorie Maillet, Alistair J. Watt, Robert J. Schwartz, Bruce J. Aronow, Stephen A. Duncan, and Jeffery D. Molkentin. 2006. “Cardiac-Specific Deletion of Gata4 Reveals Its Requirement for Hypertrophy, Compensation, and Myocyte Viability.” *Circulation Research* 98 (6): 837–45.

Padmanabhan, Arun, Michael Alexanian, Ricardo Linares-Saldana, Bárbara González-Terán, Gaia Andreoletti, Yu Huang, Andrew J. Connolly, et al. 2020. “BRD4 (Bromodomain-Containing Protein 4) Interacts with GATA4 (GATA Binding Protein 4) to Govern Mitochondrial Homeostasis in Adult Cardiomyocytes.” *Circulation* 142 (24): 2338–55.

Priest, James R., Kazutoyo Osoegawa, Nebil Mohammed, Vivek Nanda, Ramendra Kundu, Kathleen Schultz, Edward J. Lammer, et al. 2016. “De Novo and Rare Variants at

Multiple Loci Support the Oligogenic Origins of Atrioventricular Septal Heart Defects." *PLoS Genetics* 12 (4): e1005963.

Rajagopal, Satish K., Qing Ma, Dita Obler, Jie Shen, Ani Manichaikul, Aoy Tomita-Mitchell, Kari Boardman, et al. 2007. "Spectrum of Heart Disease Associated with Murine and Human GATA4 Mutation." *Journal of Molecular and Cellular Cardiology* 43 (6): 677–85.

Razick, Sabry, George Magklaras, and Ian M. Donaldson. 2008. "iRefIndex: A Consolidated Protein Interaction Database with Provenance." *BMC Bioinformatics* 9 (September): 405.

Rentzsch, Philipp, Daniela Witten, Gregory M. Cooper, Jay Shendure, and Martin Kircher. 2019. "CADD: Predicting the Deleteriousness of Variants throughout the Human Genome." *Nucleic Acids Research* 47 (D1): D886–94.

Richter, Felix, Sarah U. Morton, Seong Won Kim, Alexander Kitaygorodsky, Lauren K. Wasson, Kathleen M. Chen, Jian Zhou, et al. 2020. "Genomic Analyses Implicate Noncoding de Novo Variants in Congenital Heart Disease." *Nature Genetics* 52 (8): 769–77.

Samocha, Kaitlin E., Elise B. Robinson, Stephan J. Sanders, Christine Stevens, Aniko Sabo, Lauren M. McGrath, Jack A. Kosmicki, et al. 2014. "A Framework for the Interpretation of de Novo Mutation in Human Disease." *Nature Genetics* 46 (9): 944–50.

Sevim Bayrak, Cigdem, Peng Zhang, Martin Tristani-Firouzi, Bruce D. Gelb, and Yuval Itan. 2020. "De Novo Variants in Exomes of Congenital Heart Disease Patients Identify Risk Genes and Pathways." *Genome Medicine* 12 (1): 9.

Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. 2003. "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks." *Genome Research* 13 (11): 2498–2504.

Sifrim, Alejandro, Marc-Phillip Hitz, Anna Wilsdon, Jeroen Breckpot, Saeed H. Al Turki, Bernard Thienpont, Jeremy McRae, et al. 2016. "Distinct Genetic Architectures for Syndromic and Nonsyndromic Congenital Heart Defects Identified by Exome Sequencing." *Nature*

Genetics 48 (9): 1060–65.

Soysa, T. Yvanka de, Sanjeev S. Ranade, Satoshi Okawa, Srikanth Ravichandran, Yu Huang, Hazel T. Salunga, Amelia Schricker, Antonio Del Sol, Casey A. Gifford, and Deepak Srivastava. 2019. “Single-Cell Analysis of Cardiogenesis Reveals Basis for Organ-Level Developmental Defects.” *Nature* 572 (7767): 120–24.

Teo, Guoci, Hiromi Koh, Damian Fermin, Jean-Philippe Lambert, James D. R. Knight, Anne-Claude Gingras, and Hyungwon Choi. 2016. “SAINTq: Scoring Protein-Protein Interactions in Affinity Purification - Mass Spectrometry Experiments with Fragment or Peptide Intensity Data.” *Proteomics* 16 (15-16): 2238–45.

Tomita-Mitchell, A., C. L. Maslen, C. D. Morris, V. Garg, and E. Goldmuntz. 2007. “GATA4 Sequence Variants in Patients with Congenital Heart Disease.” *Journal of Medical Genetics* 44 (12): 779–83.

Uhlén, Mathias, Linn Fagerberg, Björn M. Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, et al. 2015. “Proteomics. Tissue-Based Map of the Human Proteome.” *Science* 347 (6220): 1260419.

Waldron, Lauren, Jeffrey D. Steimle, Todd M. Greco, Nicholas C. Gomez, Kerry M. Dorr, Junghun Kweon, Brenda Temple, et al. 2016. “The Cardiac TBX5 Interactome Reveals a Chromatin Remodeling Network Essential for Cardiac Septation.” *Developmental Cell* 36 (3): 262–75.

Wei, Qiang, Xiaowei Zhan, Xue Zhong, Yongzhuang Liu, Yujun Han, Wei Chen, and Bingshan Li. 2015. “A Bayesian Framework for de Novo Mutation Calling in Parents-Offspring Trios.” *Bioinformatics* 31 (9): 1375–81.

Xin, Mei, Christopher A. Davis, Jeffery D. Molkentin, Ching-Ling Lien, Stephen A. Duncan, James A. Richardson, and Eric N. Olson. 2006. “A Threshold of *GATA4* and *GATA6* Expression Is Required for Cardiovascular Development.” *Proceedings of the National Academy of Sciences of the United States of America* 103 (30): 11189–94.

Yi Li, Quan, Ruth A. Newbury-Ecob, Jonathan A. Terrett, David I. Wilson, Andrew R. J. Curtis, Cheong Ho Yi, Tom Gebuhr, et al. 1997. “Holt-Oram Syndrome Is Caused by Mutations in TBX5, a Member of the Brachyury (T) Gene Family.” *Nature Genetics* 15 (1): 21–29.

Zaidi, Samir, and Martina Brueckner. 2017. “Genetics and Genomics of Congenital Heart Disease.” *Circulation Research* 120 (6): 923–40.

Zaidi, Samir, Murim Choi, Hiroko Wakimoto, Lijiang Ma, Jianming Jiang, John D. Overton, Angela Romano-Adesman, et al. 2013. “De Novo Mutations in Histone-Modifying Genes in Congenital Heart Disease.” *Nature* 498 (7453): 220–23.

CHAPTER 3: A trio-based probabilistic model of variant transmission finds risk genes and their interactions in developmental disease

Adapted from Pittman & Lee et al. 2022*, a preprint posted to biorxiv and currently under revision. I carried out all work in this chapter save for the following contributions:

- K.L., F.F., and A.L. generated the GATA6 and POR knock-down mouse lines and crosses.
- M.C. performed histology analysis.

3.1 ABSTRACT

Exome sequencing of thousands of families has revealed many risk genes for congenital heart defects (CHD), yet most cases cannot be explained by a single causal mutation. Even within the same family, individuals carrying a particular mutation in a known risk gene often demonstrate variable phenotypes, indicating the presence of genetic modifiers. To explore oligogenic causes of CHD without assessing billions of variant combinations, we developed an efficient, simulation-based method to detect gene sets that carry co-occurring damaging variants in probands at a higher rate than expected given parental genotypes. We implemented this approach in software called Gene Combinations in Oligogenic Disease (GCOD) and applied it to a cohort of 3377 CHD trios with exome sequencing. This analysis detected 353 high-confidence risk genes in 202 pairs that have co-occurring variants in multiple probands but rarely or never appear in combination in their unaffected parents. Stratifying analyses by specific heart phenotypes and considering gene combinations of higher orders yielded an additional 2613 potentially-contributing genes. Genes found in oligogenic sets cluster in pathways related to heart development and suggest new molecular disease mechanisms, such as *de novo* nucleotide biosynthesis. Mouse models of the newly-identified digenic pair *GATA6-POR* confirm that these genes interact to regulate heart development, and provide support for the hypothesis that oligogenic combinations drive disease in some cases of CHD. As genome sequencing is applied to more families and other disorders, GCOD will enable detection of increasingly large, novel gene combinations, shedding light on combinatorial causes of genetic diseases.

3.2 BACKGROUND

Many diseases are genetically heterogeneous such that damaging variants at different genetic loci can lead to a common phenotype (Veenstra-Vanderweele, Christian, and Cook 2004; Akhirome et al. 2017). Complex genetic architecture limits our ability to identify causal variants and genes, despite increasingly large cohorts with exome or whole genome sequencing. For instance, congenital heart defects (CHD) have heritability estimates of 70-90% (Cripe et al. 2004; McBride et al. 2005; Hinton et al. 2007; Pierpont et al. 2018), but just 8% of probands have a damaging variant in a known CHD gene that is not shared by an unaffected family member (Jin et al. 2017). This indicates that although most defects can be explained in large part by patient genetics, our current knowledge is insufficient to pinpoint the full genetic origin and mechanism of most individual cases. Some unexplained cases presumably involve damaging variants in non-coding regions, where consequences are less straightforward than amino acid substitution in a protein (Ellingford et al. 2022). Others are likely caused by damaging variants in cryptic risk genes. Some of these genes remain undiscovered due to the incomplete penetrance of damaging variants found within them, which reduces our statistical power to detect them; at the same time, previously-identified risk genes and variants may themselves exhibit variable expressivity such that carriers differ in the presence and severity of disease symptoms (Parker and Landstrom 2021; Kingdom and Wright 2022).

A potential explanation for these phenomena is oligogenic inheritance, or a type of inheritance in which a few variants are required in combination to cause a particular phenotype (Kousi and Katsanis 2015). For example, an unaffected father may carry a variant that is tolerated in most genetic backgrounds, and is therefore unlikely to be considered a candidate for causing disease; however, if that variant is transmitted to a child along with a *de novo* or maternally-derived damaging allele in the same or a related gene, this could destabilize key pathways and cause disease where the single variant would not.

Previous work has validated an instance of oligogenic inheritance in CHD (Priest et al. 2016; Gifford et al. 2019), as well as speculated a role for oligogenic inheritance in other developmental disorders like autism spectrum disorder (ASD) (Schaaf et al. 2011; Wenger et al. 2016). While these studies yielded mechanistic insights into specific gene and variant combinations, they relied on known risk genes and existing functional information to propose testable hypotheses. Alternatively, one could enumerate and prioritize all damaging variant combinations in an automated and statistically rigorous manner. But this strategy is complicated by combinatorial explosion: given that each human genome typically contains 40,000 to 200,000 variants at <0.5% minor allele frequency (MAF) (1000 Genomes Project Consortium et al. 2015), an analysis of all possible rare variant pairs yields between 8 billion to nearly 20 trillion combinations per individual.

Nevertheless, there are existing methods that scan the genome for possible oligogenic disease combinations. ORVAL accepts a list of variants as input, and uses gene annotations to assess their potential for oligogenic combinations (Renaux et al. 2019). However, this tool only offers qualitative hypotheses for a single individual, which limits applicability across probands. In contrast, the Digenic method reduces the combinatorial search space by aggregating rare variants at the gene level and assesses the statistical significance of each gene pair based on its prevalence in a disease cohort (Kerner et al. 2020), but this method is limited to oligogenic combinations of only two genes ("digenic"). Another recent advance in the field is RareComb (Pounraja and Girirajan 2022), an algorithm that tests for greater frequency of gene combinations in cases compared to controls. However, neither of these computational methods incorporate parental sequencing data, which are especially useful in reducing false positives for simplex families in which an affected proband is born to unaffected parents (Ewens 1999). It logically follows that the highest-effect causal variants are likely transmitted separately from each parent or include at least one *de novo* mutation; thus, incorporating parental transmission data is potentially a powerful tool to reduce false positives, increase confidence in the phenotypic relevance of variants, and correct for population structure.

To address this gap, we developed the tool Gene Combinations in Oligogenic Disease (GCOD), an algorithmic framework that uses simulation experiments to identify disease gene sets that carry rare damaging variants significantly more often than expected given parental genotypes. GCOD automatically assesses every digenic pair observed in multiple probands, as well as sets of three, four, or more genes where applied. Filtering based on information from unaffected parents limits the number of gene sets tested, which increases statistical power and accelerates computation. Applying GCOD to exome sequencing from families with CHD, we identified potential novel risk genes, variant combinations, and pathways in disease. These results expand candidate causal mechanisms in developmental phenotypes and provide evidence that oligogenic combinations of high-effect variants are potentially causal in CHD.

3.3 RESULTS

3.3.1 Computational identification of pathogenic gene sets

We developed an algorithm to predict pathogenic gene sets using exome or whole genome sequencing data and implemented it in software we called GCOD. The method involves four steps (**Figure 3.1**): select variants of interest based on constraint or other criteria, enumerate sets of genes with co-occurring variants, count their observed frequency across probands, and evaluate the statistical significance of these counts given the parental genotypes. An optional fifth step repeats the process for healthy siblings or pseudo-siblings (Yu and Deng 2011) as a control. GCOD can be viewed as an extension of the Transmission Disequilibrium Test (TDT) in which the unit being tested is an observed co-occurrence of variants in a group of genes, rather than observed transmissions at a single locus (Spielman, McGinnis, and Ewens 1993) or within a single gene (He et al. 2017). GCOD also differs from the TDT in its use of a nonparametric simulation to assess significance rather than a chi-squared statistic, allowing us to detect rare gene combinations with very few expected observations and without making assumptions about the underlying data. Details and options for each step of GCOD follow.

First, users provide a list of damaging variants meeting some criteria of interest from each individual in the dataset (**Figure 3.1, A**). GCOD then enumerates all combinations of genes that contain co-occurring damaging variants in each proband (**Figure 3.1, B**), retaining only those combinations observed in multiple families (referred to as “candidate sets”). By

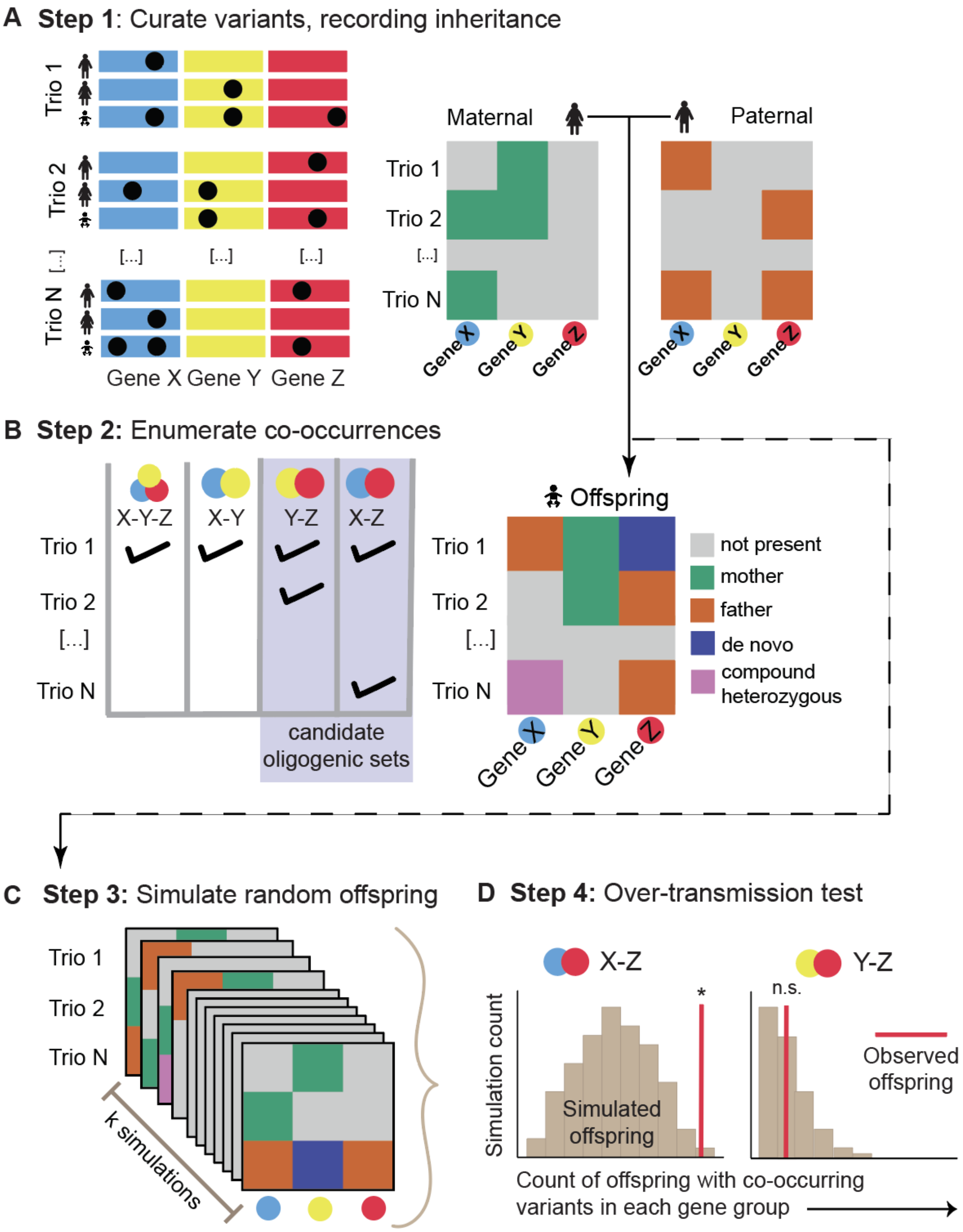


Figure 3.1: The GCOD approach to identify sets of genes that interact in oligogenic disease. (A) Users submit a list of variants of interest, or define criteria for some level of “rare” or “predicted damaging.” These variants are summarized in two binary gene-by-trio matrices that denote whether at least one damaging allele was present in each of the maternal and paternal genomes, as well as an offspring matrix (**B, right**) that encodes damaging variant

presence and inheritance. The variant(s) in each gene can be only maternally-derived, only paternally-derived, compound heterozygous, or de novo. GCOD enumerates the candidate sets of genes for which variants occur in multiple probands (**B, left**), by default only including combinations not transmitted from a single parent (i.e. the provenance of the variant combination in the offspring includes either paternal+maternal, de novo+[*], or compound heterozygous+[*]). **(C)** A user-defined number of simulations are performed for each pair of parents; each variant has a 50% chance of being passed to the simulated offspring, and each gene has a pre-defined per-chromosome probability of de novo mutation (Samocha et al. 2014). This yields a distribution of simulated carriers for each candidate set. **(D)** For a given candidate set, the simulated distribution is compared to the observed number of oligogenic transmissions in probands to compute a p-value.

changing the definition of damaging, different scenarios can be explored, from small gene sets with very rare co-occurring variants of high impact to larger gene sets that include small-effect, possibly-damaging variants that could modify the effects of other primary disease-causing variant(s) or otherwise contribute to a polygenic signal.

We designed GCOD to automatically consider three tiers of variant severity (**Table 3.1**), defined by residue-level characteristics such as minor allele frequency (MAF) and CADD score (Rentzsch et al. 2019), gene-level cutoffs for gnomAD observed-expected z-score (Karczewski et al. 2020), as well as gene shet score (Cassa et al. 2017) and minimum expression level in relevant cell types (J. Cao et al. 2020). The variant severity tiers are nested such that more severe variants are included in each of the decreasingly strict tiers.

Table 3.1: Thresholds for the default three tiers of variant severity.

	Strict	Base-damaging	Base
MAF upper bound	0.01	0.05	0.05
Variant types included	-de novo -loss-of-function -missense damaging	-de novo -loss-of-function -missense damaging	-de novo -loss-of-function -missense damaging -missense
Constraint (variants meet at least one condition)	CADD Phred > 15 or S-het > 0.4 or Mis-z or lof-z score > 2.5	CADD Phred > 10 or S-het > 0.25 or Mis-z or lof-z score > 1.5	CADD Phred > 10 or S-het > 0.25 or Mis-z or lof-z score > 1.5
Gene expression	TPM > 0.5	TPM > 0.5	NA

By default, a candidate gene set is only considered in a particular proband if the qualifying variants were derived from at least two sources (that is, “oligogenic transmission” where a healthy parent did not also carry that same combination of variants), though users can remove this requirement. The count of observed oligogenic transmissions comprise the test statistic, which is compared against the null distribution of counts over thousands of simulations for an empirically-derived p-value.

To reduce the number of tested hypotheses and the multiple testing correction burden where possible, we implemented an algorithm that begins with digenic pairs and merges these up to the “highest-order” set, defined as the largest unique candidate set observed in two or more probands. Lower-order combinations are added to the analysis if an additional proband harbors this candidate set and not of any larger sets. For example, two probands carrying damaging variants in seven genes share 21 unique gene pairs, 35 unique gene trios and quartets, 21 unique gene pentads, and so on (**Figure 3.2, Methods: Highest-order gene set enumeration**). If no other probands carry co-occurring variants in the set, GCOD tests only the seven-gene set and none of the 119 lower-order combinations. If a third proband were to carry one of the gene pairs but none of the higher-order candidate sets, the pair (with a count of three) is tested along with the seven-gene set (with a count of two). This algorithm produces a co-occurrence count for each highest-order candidate set, recursively checking for additional transmissions at each level descending to pairs.

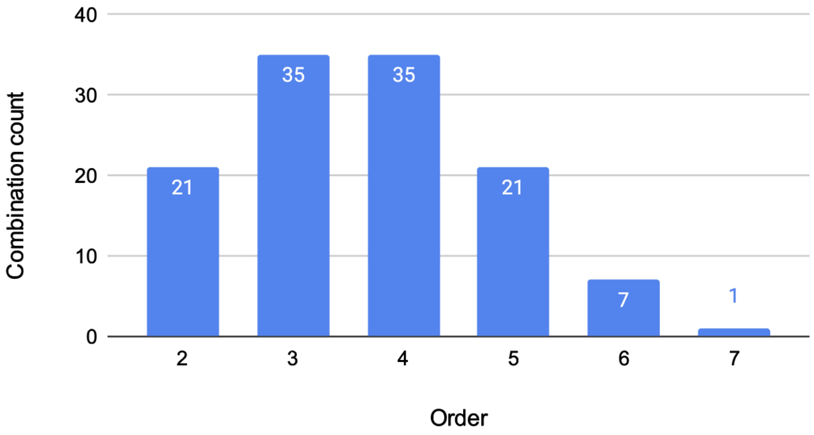


Figure 3.2: Unique combinations of a seven-gene group. If a proband pair shares seven genes with candidate variant hits, there are 21 unique gene pairs, 35 unique gene trios, and so on. The highest-order groups in this hypothetical scenario only include the one seven-gene group to test. Additional lower-order combinations are only considered if a third proband also harbors candidate variants within those genes.

In the third step, GCOD performs simulations based on parental genotypes to assess the statistical significance of the candidate set co-occurrence counts (**Figure 3.1, C**). The goal of the simulations is to estimate the distribution of oligogenic co-occurrences expected by chance assuming no transmission disequilibrium between the gene group and disease.

In each iteration, one hypothetical offspring is simulated for each parental pair as follows. Damaging parental variants have a 50% chance of being passed to the simulated offspring. We treat variants independently and do not model linkage disequilibrium because GCOD records the presence/absence of a damaging variant in a gene (rather than the number of variants per gene), and given that linkage disequilibrium decays by 70 kilobases in most regions of the human genome, genes are very rarely in the same LD block (Bosch et al. 2009). Additionally, variants must be derived from at least two sources such that a candidate set is unlikely to be entirely inherited as a single LD block (i.e., a variant from the other parent or mutated de novo must have been observed). After simulating inherited variants, GCOD simulates de novo variants using previous estimates of per-gene de novo mutation probabilities (Samocha et al. 2014). Here we use the term “oligogenic transmission” to describe an event in which a variant combination occurs in the offspring that was not seen in either parent, i.e. comprises at least one variant from each parent and/or a de novo mutation.

For each simulated cohort, we tally the number of times a given gene set was transmitted with oligogenic inheritance across simulated offspring. Repeating over many iterations generates a null distribution of oligogenic transmissions against which we compare the observed number in actual probands in the fourth step. We calculate the probability of no over-transmission for each candidate gene set as the proportion of iterations with a simulated count equal to or greater than the number of probands (**Figure 3.1, D**). GCOD returns the full list of candidate sets along with the p-values associated with their transmission to probands (unadjusted, Benjamini-Hochberg corrected, and Bonferroni corrected).

In an optional fifth step, GCOD repeats this entire procedure for sibling controls. If healthy siblings have been sequenced, these may be used. Otherwise, GCOD generates a genotype for a pseudo-sibling of each proband, a hypothetical sibling who has inherited the parental alleles that were not transmitted to the affected individual (Yu and Deng 2011) (**Figure 3.3**). Siblings, unlike pseudo-siblings, carry some variants inherited by the proband. They also can be assessed for the phenotype, whereas pseudo-siblings cannot. Siblings without a diagnosis and pseudo-siblings are typically presumed to be unaffected, though this is not always true; siblings must be deeply phenotyped past the typical age of onset to rule out disease, and pseudo-siblings have unknown phenotypes. Hence, it can be helpful to use both when sibling data is available and to interpret results with this caveat in mind. Collectively, the sibling controls serve several purposes. Significant candidate sets from probands can be filtered to exclude any that also appear in (pseudo-)siblings. A comparison of the number of significant sets for probands versus (pseudo-)siblings can also be used to assess overall evidence that a phenotype is likely oligogenic, where probands would be expected to carry more over-transmitted gene sets than (pseudo-)siblings. Finally, the types of genes in significant sets can be compared between probands and (pseudo-)siblings, for example, using gene set enrichment, pathway analysis, or literature review. These analyses help to refine and prioritize significant sets from step three. We refer to the resulting statistically significant gene sets as “oligogenic sets”.

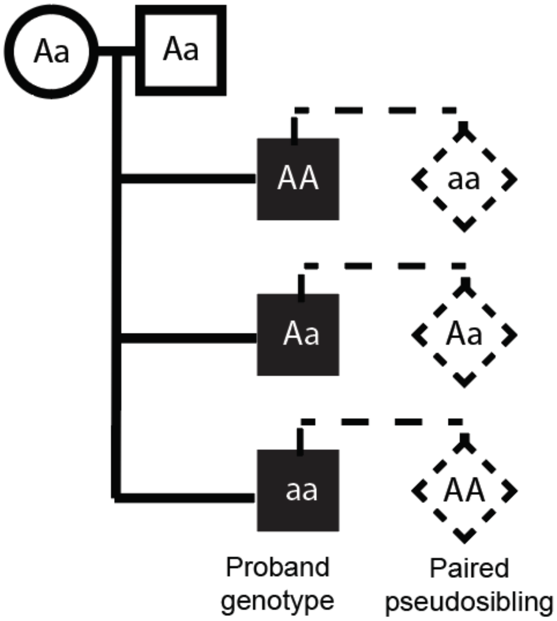


Figure 3.3: Pseudo-sibling creation. For each input variant locus, the pseudo-sibling inherits the alleles that were not transmitted to the observed proband.

3.3.2 CHD probands carry more oligogenic gene sets than familial controls

To discover oligogenic disease genes for CHD, we applied GCOD to whole exome sequencing data from 3377 trios in the Pediatric Cardiac Genomics Consortium (PCGC) (Pediatric Cardiac Genomics Consortium et al. 2013; Hoang et al. 2018; Morton et al. 2021). In each family, the PCGC generated exome variants for two asymptomatic parents and an affected child. We used parental data to compute pseudo-sibling genotypes for each trio (Yu and Deng 2011) (**Figure 3.3**). We ran GCOD at three tiers of variant severity (**Table 3.1**): the Strict tier which includes only very rare mutations with a high probability of impairing protein function, plus two more lenient tiers of variant severity (Base-damaging and Base). Only oligogenic transmissions of a variant set were considered. To reduce false positives caused by a single driver gene with spurious co-transmitted partners, we additionally filtered oligogenic sets to those with a logistic regression interaction coefficient greater than 1.0 (**Methods: Logistic regression**). In short, this step aims to restrict our analysis to sets in which a higher proportion of individuals with damaging mutations in the full oligogenic set have CHD compared to individuals with damaging mutations in each gene separately.

The GCOD approach revealed 202 oligogenic gene pairs comprising 353 CHD risk genes at the Strict and Base-damaging tiers. At each variant tier, we identified more oligogenic gene pairs in CHD probands compared to their pseudo-siblings (**Figure 3.4, A**). These differences were all statistically significant (Binomial p ≤ 1.4x10-23). At the Strict tier, for example, we found 179 gene pairs significantly over-transmitted to probands, compared to just 38 gene pairs over-transmitted to their pseudo-siblings (Bonferroni-adjusted $p < 0.05$) (**Table 3.2**). A similar nearly 5-fold enrichment was seen at the Base-damaging tier. Oligogenic pairs are present in a substantial number of probands, with 9.2% of PCGC probands carrying at least one pair even at the Strict tier. Together these results provide new evidence that oligogenic inheritance plays a role in the etiology of some CHD cases.

As expected, the number of oligogenic pairs increased with more lenient definitions of damaging variants for both probands and pseudo-siblings (**Table 3.2**), with the Base tier yielding 22515 oligogenic pairs in probands (versus 2783 in pseudo-siblings). This many significant gene pairs means that 96.6% of probands carry at least one oligogenic pair at the Base tier. These gene pairs likely comprise a more diffuse polygenic risk compared to the 179 pairs at the Strict tier, which may play a more central role in the disease of the 312 probands who carry them. We therefore focus the subsequent analyses on the oligogenic sets from the Strict and Base-damaging tiers, which are expected to have larger pathogenic effects per occurrence.

We next computed the highest-order gene sets shared among PCGC probands using Strict and Base-damaging variants (**Methods: Highest-order gene sets**). Highest-order sets

Table 3.2: GCOD counts of significantly over-transmitted oligogenic pairs and genes, categorized by membership to the list of curated human/mouse CHD genes in Jin et al. 2017.

Tier	**Disease Status**	**N Oligogenic Pairs**	**N unknown genes in Oligogenic Pairs (N Total Genes)**	**N (% of cohort) that carry at least one pair**	**Number of Single Genes**	**Number of CHD known Single Genes**
Strict	Proband	179	295 (317)	312 (9.2%)	133	8
	Pseudo-sibling	38	70 (70)	79 (2.3%)	0	0
Base-damaging	Proband	189	312 (334)	328 (9.7%)	136	12
	Pseudo-sibling	38	70 (71)	76 (2.2%)	0	0
Base	Proband	22515	7627 (7782)	3279 (96.9%)	363	10
	Pseudo-sibling	1893	2498 (2541)	1693 (50.1%)	51	0

ranged from groups of two genes to groups of twelve genes for both probands and pseudo-siblings. As with gene pairs, we find a significantly greater count of over-transmitted sets in probands compared to pseudo-siblings (both binomial $p < 1 \times 10^{-150}$, **Figure 3.4, B**). The highest-order oligogenic sets at the Base-damaging and Strict level yield an additional 1662 potential contributing risk genes.

We next investigated whether the oligogenic pairs identified in PCGC probands include any CHD genes that were not found in smaller datasets and single-gene tests. First, we annotated which oligogenic sets contained one or more of the 253 curated human or mouse CHD genes reported by Jin et al. 2017. While 18% of Base-damaging and Strict proband oligogenic pairs contain at least one of these CHD genes, the vast majority of risk genes captured by GCOD were not previously known (312 of 353 genes, **Table 3.2**).

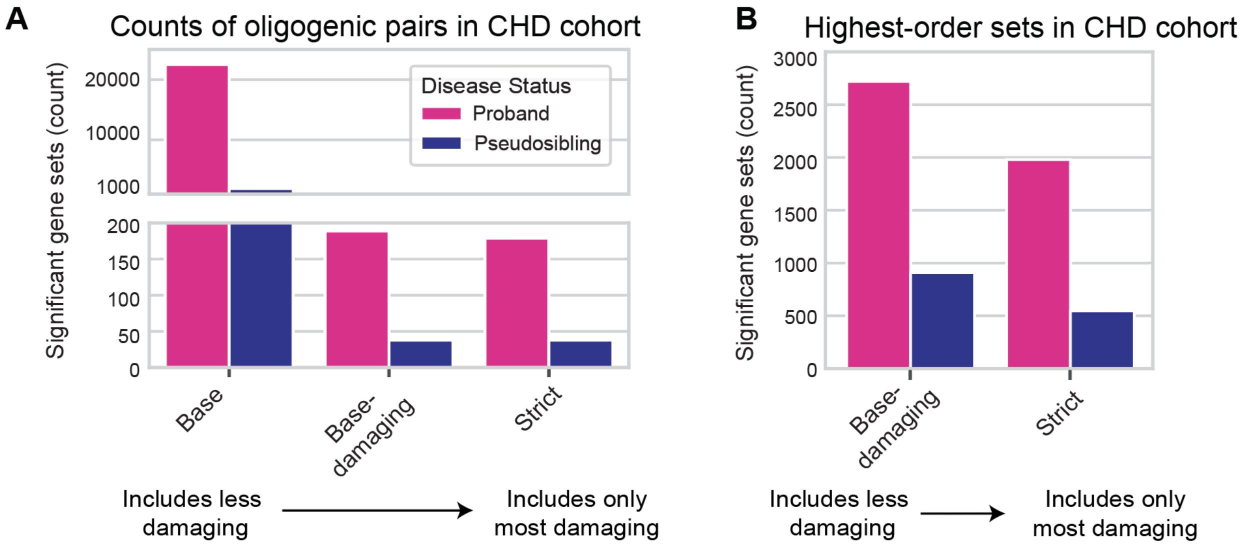


Figure 3.4: Counts of significant oligogenic groups among CHD probands (pink) and pseudo-siblings (blue) at each tier of variant severity. **(A)** Counts of digenic pairs. **(B)** Counts of highest-order gene sets. Note that computing highest-order interactions at the Base tier was infeasible.

Next, we directly assessed the additional genes discovered by testing pairs versus single genes within the GCOD simulation framework (**Methods: Single-gene transmission simulation test**). This analysis is distinct from using known CHD genes, because prior studies used a variety of statistical tests, cohorts, and sample sizes. GCOD detects about a third as many genes in single-gene mode compared to testing oligogenic pairs at the Strict and Base-damaging tiers, and this difference is even more pronounced at the Base tier. In all tiers, fewer known CHD genes are detected in single-gene mode compared to when oligogenic sets are run. These findings underscore the additional signal GCOD is able to capture by considering oligogenic transmissions.

3.3.3 Genes in proband oligogenic sets are found together in canonical cardiac gene sets

We have so far shown that proband cohorts have a greater number of oligogenic gene sets, and that these sets are enriched for known CHD risk genes. To test whether the CHD oligogenic pairs identified by GCOD comprise phenotype-relevant genes, we performed a Gene

Ontology (GO) analysis. Strict tier oligogenic genes were enriched for many GO terms related to heart development, such as “atrioventricular canal development” and “cardiac atrium morphogenesis” (**Figure 3.5**). In contrast, no GO terms were enriched in pseudo-sibling genes at the Strict or Base-damaging tiers. We therefore conclude that genes comprising proband oligogenic pairs are relevant to CHD, and genes over-transmitted to pseudo-siblings are largely false positives unrelated to any particular phenotype, confirming their utility as controls.

The aforementioned analysis combines oligogenic sets to find common functions among all putative risk genes. We further hypothesized that genes within the same disease-associated oligogenic set would share functional annotations and also co-occur in pathways and protein complexes. To test this, we calculated the odds of an oligogenic combination being found in a proband and co-occurring in a Molecular Signatures Database (MSigDB) ontology gene set (Liberzon et al. 2015), and found a significant association for the combined GO collections (odds ratio = 1.6, Fisher exact p = 2.2e-33) and for the Human Phenotype Ontology (Köhler et al. 2021) collections related to the heart (odds ratio = 3.8, Fisher exact p = 3.5e-12). A similar enrichment was observed for the PCNet composite database of gene-gene interaction networks (S.-Y. Cao et al. 2021) (odds ratio = 1.8, Fisher exact p = 4.8e-05). These results indicate that the genes comprising CHD oligogenic pairs in the PCGC cohort are functionally associated, perhaps indicating that the interaction of their gene products is necessary for shared functions in the context of disease pathogenesis.

In summary, we have enumerated gene sets for which multiple CHD probands inherited an oligogenic combination of variants, and tested whether each gene set was transmitted together more often than expected. We showed that the genes in the resulting oligogenic sets are enriched for heart development functions and tend to occur together in canonical gene networks. By leveraging rare variation from separate sources (two parents or inherited plus de novo), GCOD captures additional known and putative disease genes that other rare variant aggregation methods do not.

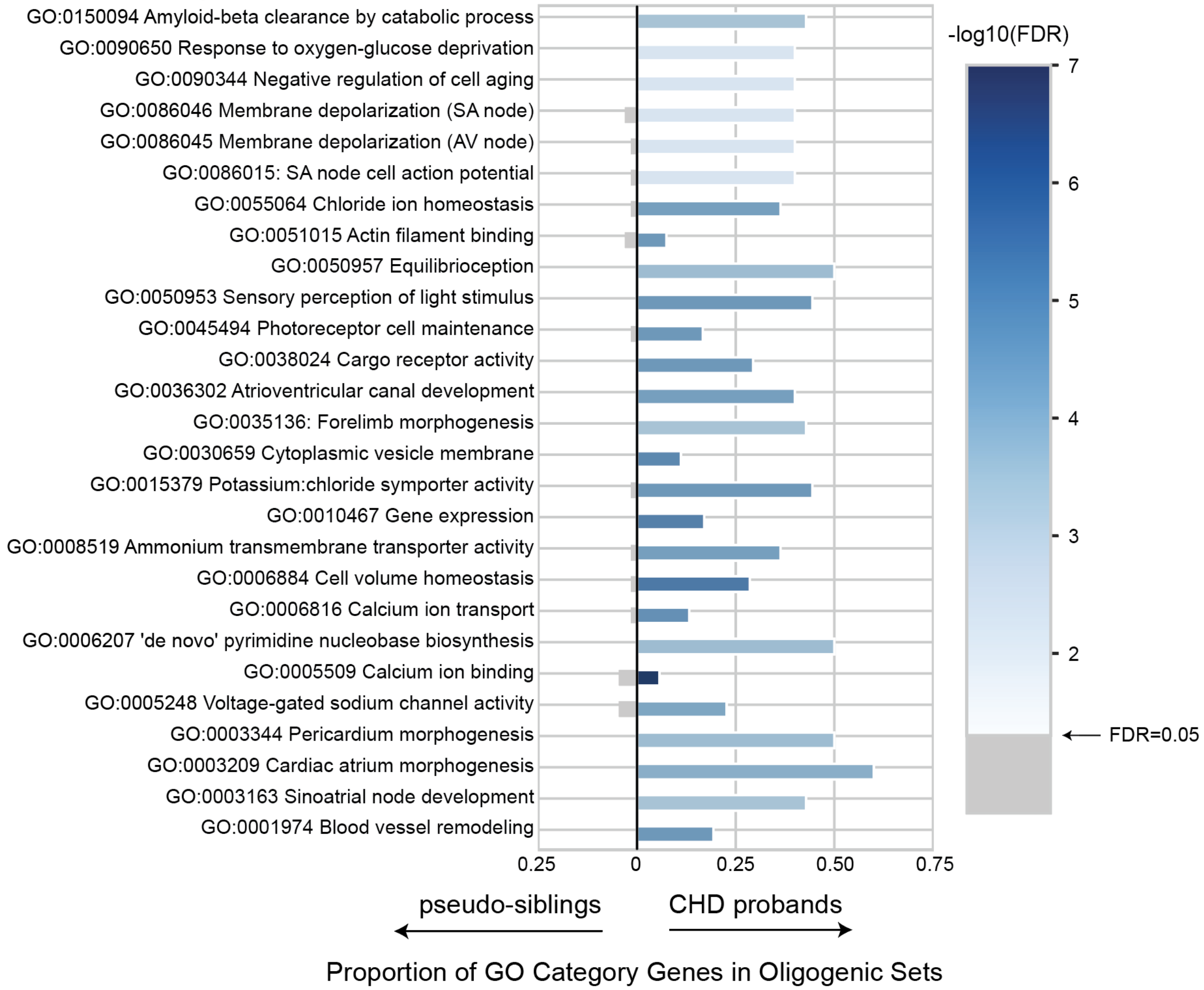


Figure 3.5: Top 15 GO terms with highest fold enrichment and 15 GO terms with lowest FDR for genes in Strict oligogenic pairs (total 27 terms: “Potassium:chloride symporter activity”, “Sensory perception of light stimulus”, and “Atrioventricular canal development” are among the top categories for both fold enrichment and FDR significance.)

3.3.4 Novel risk genes and interactions in CHD

While up to this point we have analyzed oligogenic transmission frequency across the entirety of the exomes of the 3382 trios in the PCGC, these probands have been diagnosed with myriad distinct diagnoses. To explore the possibility that similar phenotypes are caused by more similar genetic etiology, we selected five diagnoses with relatively consistent phenotypic criteria (**Table 3.3**): Ebstein’s anomaly, pulmonary atresia with intact ventricular septum (PA-IVS), truncus arteriosus, tetralogy of Fallot, and hypoplastic left heart syndrome (HLHS). HLHS with mitral atresia and aortic atresia (HLHS-MA-AA) or mitral stenosis and aortic atresia (HLHS-MS-AA) were also analyzed separately. We hypothesized that these smaller analyses might have improved power to detect oligogenic sets that are specific to one or a few diagnoses but do not replicate in the entire cohort.

Table 3.3. Description of CHD sub-diagnoses and sample counts of patients analyzed.

CHD diagnosis	Proband Count (989)	Description
Hypoplastic Left Heart Syndrome (HLHS) - all subtypes	427	The left ventricle is underdeveloped and too small, often with mitral and aortic valve atresia (absence of opening) or stenosis (narrowing).
HLHS - mitral atresia and aortic atresia (HLHS-MA-AA)	104	A severe form of HLHS accompanied by atresia of both mitral and aortic valves.
HLHS - mitral stenosis and aortic atresia (HLHS-MS-AA)	41	A rare version of HLHS characterized by atresia of the aortic valve, while the mitral valve opens but is narrowed.
Truncus Arteriosus (TA)	240	The aorta and pulmonary artery fail to separate completely during development and the corresponding conotruncal ventricular septum does not form (ventricular septal defect).
Tetralogy of Fallot	202	Tetralogy of Fallot comprises four co-occurring defects: ventricular septal defect, pulmonary valve stenosis, an aorta displaced over the ventricular septum, and increased muscle mass (hypertrophy) of the right ventricle.
Pulmonary Atresia, intact ventricular septum (PA-IVS)	66	Atresia of the pulmonary valve, with an intact ventricular septum (no ventricular septal defect).
Ebstein’s anomaly	54	A displacement of the tricuspid valve into the right ventricle that often leads to valve regurgitation and right ventricular dysfunction.

Counts of oligogenic sets including digenic pairs and highest-order sets in phenotypically similar sub-groups corroborated our finding from the full PCGC cohort that probands harbor significantly more oligogenic sets compared to pseudo-siblings. Many of these sets were found in the broader PCGC as a whole, but we find 160, 181, and 7960 additional diagnosis-specific oligogenic sets at the Strict, Base-damaging, and Base tiers respectively. Many of these were found in the largest diagnostic categories, like HLHS (all subtypes) and Tetralogy of Fallot. However, even at the Strict level we detect the oligogenic pair CAPN9-MGA, a rare combination which was transmitted to two of just 66 individuals with PA-IVS. MGA's role as a transcription factor in cell differentiation, and CAPN9's known association with Noonan syndrome

We cross-referenced canonical protein complexes sourced from the CORUM database (Giurgiu et al. 2019) with our oligogenic sets from all of PCGC and from separate diagnoses, finding 34 proband oligogenic sets in which two or more genes physically interact. For example, the products of the CREBBP and EP300 genes participate in the p300-CBP-p270-SWI/SNF complex, which regulates histone acetyltransferase activity during development (Chan and La Thangue 2001). Variants in these genes were transmitted oligogenically to two individuals in combination with the COL6A3 gene (COL6A3-CREBBP-EP300, 2 probands, p = 0.016). These three genes were also over-transmitted independently with other genes, which led to the discovery of a network of interconnected oligogenic sets that collectively harbor multiple hits to three complexes: p300-CBP-p270-SWI/SNF, AML1-HIPK2-p300, and the pRb2/p130-multimolecular complex (**Figure 3.6, A**).

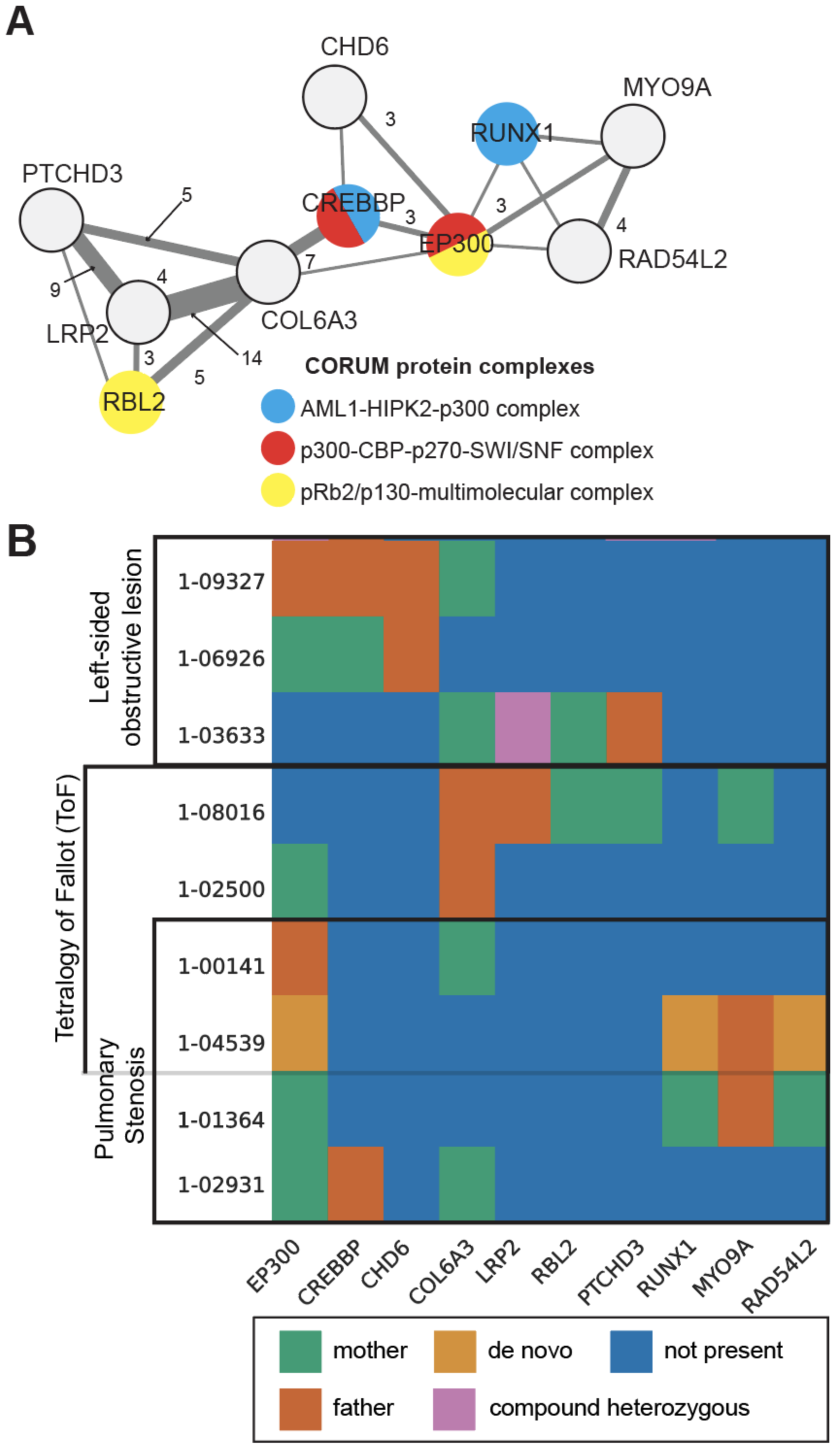


Figure 3.6: Interconnected oligogenic network of protein complex genes. **(A)** Network visualization of genes in oligogenic sets containing one or more of the genes coding for members of AML1-HIPK2-p300 complex, p300-CBP-p270-SWI/SNF complex, or pRb2/p130-multimolecular complex. Node colors represent functional annotations and edge width indicates the number of probands with co-occurring mutations in each gene pair. Note that the edges can include observations that were inherited from a single parent (non-oligogenic transmission) if the gene pair is part of a higher-order oligogenic set. Non-annotated edges

represent a co-occurrence count of two probands. **(B)** Variant inheritance matrix for genes in (A). Columns indicate genes, and rows indicate probands that harbor damaging mutations in each gene. Cell color represents a proband's inheritance of variant(s) in the specified gene column: blue if the proband does not carry a damaging mutation in that gene; green, orange, and pink for mother, father, and both parents respectively; and gold for de novo mutations. Multiple probands sharing a diagnostic category are indicated by black boxes.

EP300 co-occurs with RUNX1 in the AML1-HIPK2-p300 complex, and although the digenic pair is observed in one unaffected parent, the highest-order oligogenic set including MYO9A from an alternate source is only observed in CHD probands, both diagnosed with pulmonary stenosis (**Figure 3.6, B;** EP300-MYO9A-RAD54L2-RUNX1, 2 probands, $p<0.0001$). RUNX1 phosphorylates EP300 and CREBBP to activate acetyltransferase activity (Aikawa et al. 2006). The co-occurrence of multiple combinations of transcriptional regulators and chromatin modifiers suggests that they interact in gene regulatory networks that are necessary for normal heart development. Genes not annotated with these functions, such as COL6A3 and MYO9A, could be target genes or have upstream or downstream regulatory functions. For example, the fibrillar collagen gene COL6A3 has a role in TGFβ signaling as previously characterized in cancers (Huang et al. 2018; Dankel et al. 2020). Given the involvement of TGFβ signaling in epithelial–mesenchymal transformations in cardiovascular development (Azhar et al. 2003) and its coactivation by CBP/p300 (Janknecht, Wells, and Hunter 1998), co-occurring mutations in COL6A3, CREBBP, and EP300 are likely pathogenic due to perturbation of this pathway.

Notably, twelve probands carried predicted-damaging variants in both the MYO18B gene and the SACS gene, which encode for the myosin-18B and sacsin proteins respectively (**Figure 3.7, A**). Ten of these observations were oligogenic transmissions in which the variants did not co-occur in an unaffected parent. The other two probands inherited MYO18B-SACS from their respective mothers, but additionally inherited mutations in KCNH6 from the fathers (**Figure 3.7, B**). We conclude that there is an underappreciated role for the SACS gene in heart development, likely related to the organization and activity of the cytoskeleton. This is consistent with previous research indicating that sacsin regulates filament assembly, though this was determined in neurofilaments (Gentil et al. 2019). The SACS protein also acts as a chaperone in the binding of LRP1B (Marschang et al. 2004).

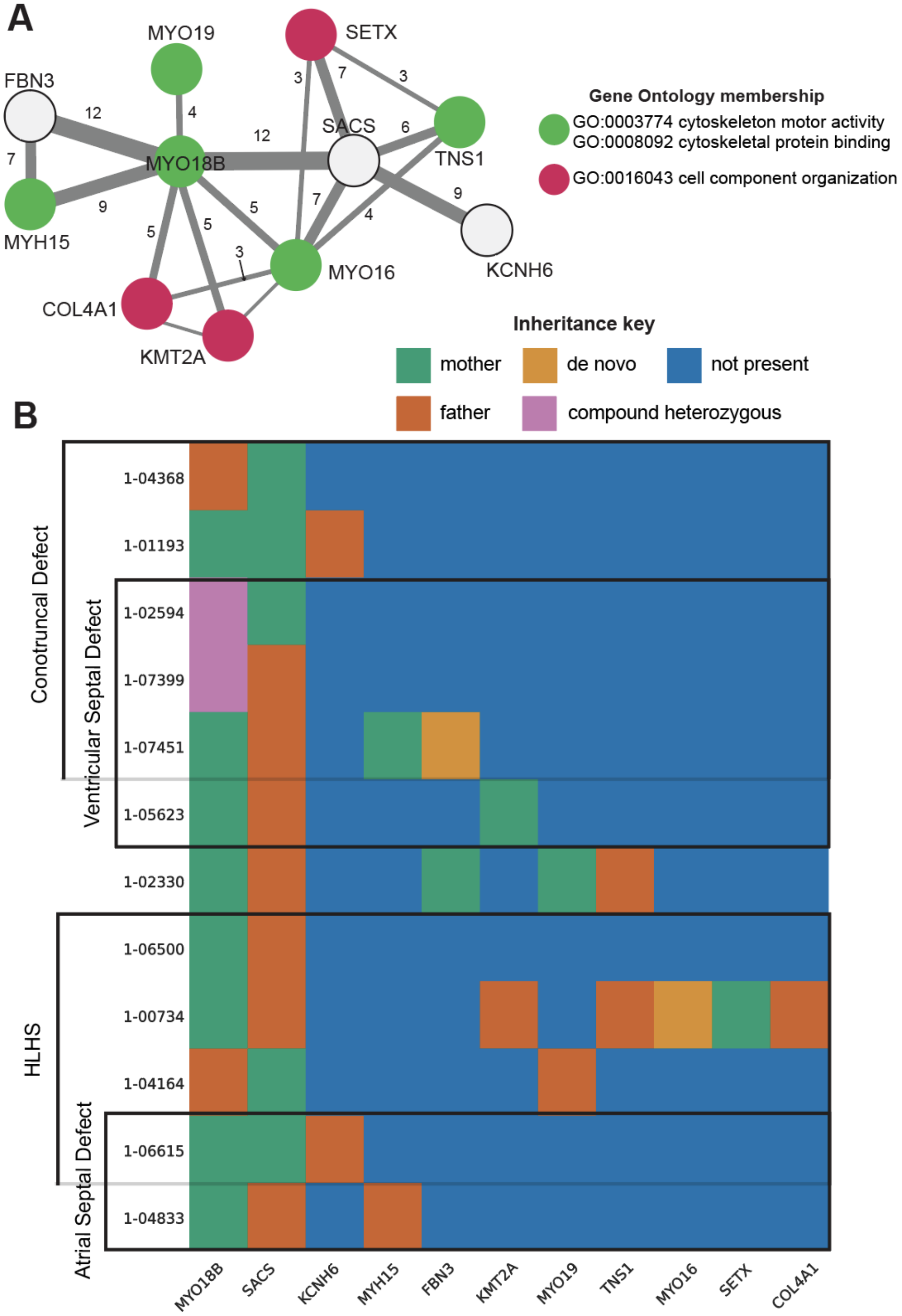


Figure 3.7: Interconnected oligogenic sets centered on the MYO18B-SACS digenic pair. **(A)** Network visualization of interconnected oligogenic sets, centered on the MYO18B-SACS digenic pair. Genes annotated with ontology terms related to cytoskeleton activity and cellular component organization are indicated in green and pink respectively. **(B)** Variant inheritance matrix for genes in (A). Columns indicate genes, and rows indicate probands that harbor

damaging mutations in each gene. Cell color represents a proband's inheritance of variant(s) in the specified gene column: blue if the proband does not carry a damaging mutation in that gene; green, orange, and pink for mother, father, and both parents respectively; and gold for de novo mutations. Multiple probands sharing a diagnostic category are indicated by black boxes.

LRP1B was found to be significantly over-transmitted with SACS at the Base tier of variant severity (4 probands, $p < 0.0001$). Furthermore, oligogenic transmission of LRP1B variants with both MYH6 and SRCAP variants at the Strict tier was observed in three families. SRCAP is the Snf2-related CREBBP activator protein, and is known to be involved in histone modification and oxidative metabolism during prenatal heart development (Xu et al. 2021), and MYH6 is a major component in the thick filaments of the sarcomere (Mahdavi, Periasamy, and Nadal-Ginard 1982; Warkman et al. 2012). Oligogenic transmission of Base-damaging/Strict variants in the MYH6-SRCAP pair itself occurred in nine CHD probands ($p < 0.0001$). Other potential genetic modifiers of this interaction include the oxidase DUOX1 (transmitted with MYH6-SRCAP in 4 of 9 probands) and the membrane-repair gene DYSF (2 of 9 probands). This result suggests a genetic interaction of cell membrane proteins like LRP1B and its chaperone SACS with genes involved in cell motility and oxidative metabolism.

3.3.5 Compound heterozygous knockdown of GATA6 and POR increase incidence of cardiac defects and early death in mice

A digenic pair of particular interest was that of *GATA6* and *POR*, which was improbably found in three probands (**Figure 3.8, A**) all with truncus arteriosus (TA), a defect of the outflow tract in which the pulmonary artery and the aorta fail to separate (Hutson and Kirby 2007; Keyte and Hutson 2012). TA pathology is consistent with malfunction of normal neural crest cell migration during formation of the outflow tract (Neeb et al. 2013). The transcription factor *GATA6* has been shown to play a role in regulating the formation of the cardiac outflow tract (Koutsourakis et al. 1999); specifically, the conditional deletion of *GATA6* in neural crest-derived smooth muscle recapitulates several heart malformations, including persistent truncus arteriosus and double-outlet right ventricle (Lepore et al. 2006). While a few human polymorphisms associated with heart defects have been found in the *GATA6* gene (Maitra et al. 2010), we found instances of predicted-damaging *GATA6* mutations in unaffected parents in the PCGC dataset, and such mutations are not rare among public datasets (Karczewski et al. 2020). Therefore a single mutated copy of *GATA6* is not a fully penetrant cause of CHD.

Its digenic counterpart, P450 oxidoreductase (POR), transports electrons from NADPH to the cytochrome P450 enzymes. *POR* null mice die during fetal development (Shen, O'Leary, and Kasper 2002; Otto et al. 2003), and severe *POR* mutations A287P (in European ancestry) and R457H (in Japanese ancestry) cause the Antley-Bixler skeletal malformation syndrome (ABS) (Miller et al. 2011). Heart defects occur in 21% of ABS cases (Oh et al. 2017), suggesting an as-yet-unrecognized role for *POR* in heart development. Given its role in retinoic acid metabolism (Zlotnik et al. 2022), and the importance of RA signaling in outflow tract and neural crest formation (Li, Pashmforoush, and Sucov 2010; Keyte and Hutson 2012), we hypothesize that *POR*'s interaction with *GATA6* is necessary for normal heart development, and the compound damaging mutations found in these patients' genomes caused or contributed to their TA symptoms.

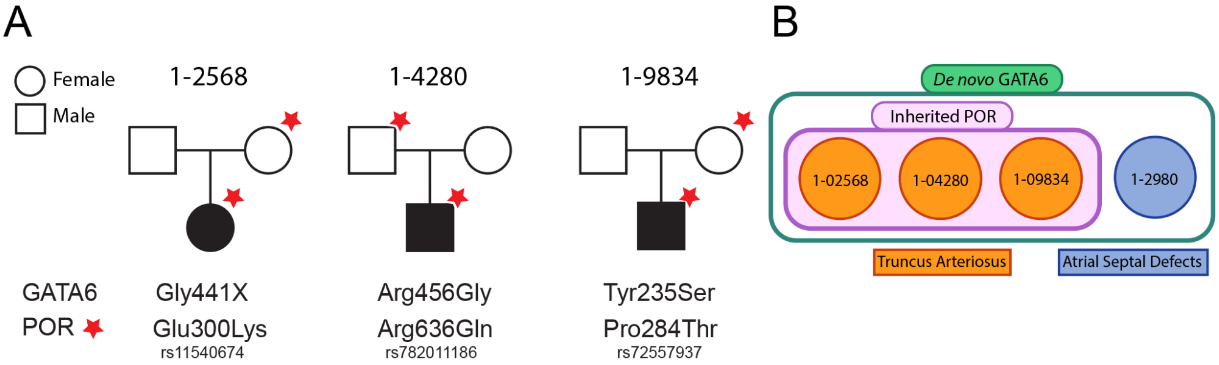


Figure 3.8: GATA6 de novo variants in CHD patients. **(A):** Trio pedigrees for TA probands with a GATA6 de novo mutation. All three TA patients also carry a predicted-damaging inherited mutation in POR. **(B)** Four CHD patients carry a de novo GATA6 mutation (green box), but the more severe TA phenotype is limited to individuals also carrying a POR mutation (orange).

Interestingly, another proband in the PCGC carries a GATA6 *de novo* mutation without a damaging POR variant, but this individual presented with atrial septal defects instead of the more severe TA phenotype (**Figure 3.8, B**). We therefore investigated whether compound heterozygosity of *GATA6* and *POR* knockdown leads to heart defects more often or more severely than just *GATA6* or just *POR* knockdown, indicating a genetic interaction between the two genes during heart development. To evaluate the *in vivo* functional consequence of the individual and combined heterozygotic loss of these genes, we created transgenic heterozygous knockdown mouse lines (C57BL/6J background), then crossed them to examine the effects of GATA6 and POR knockdown individually and in combination.

Gata6-/+ Por-/+ compound heterozygous mice were born below Mendelian ratios (5 observed, 8 expected) though this is not a statistically significant finding. Notably, compound knockdown mice experienced atrial- and ventricular-septal defects at a rate of approximately 50% (n = 17, **Figure 3.9, Table 3.4**), compared with the WT, single-hit Gata6-/+, and single-hit Por-/+ mice that experienced no clear defects in histological analyses (n = 6, 11, and 9, respectively).

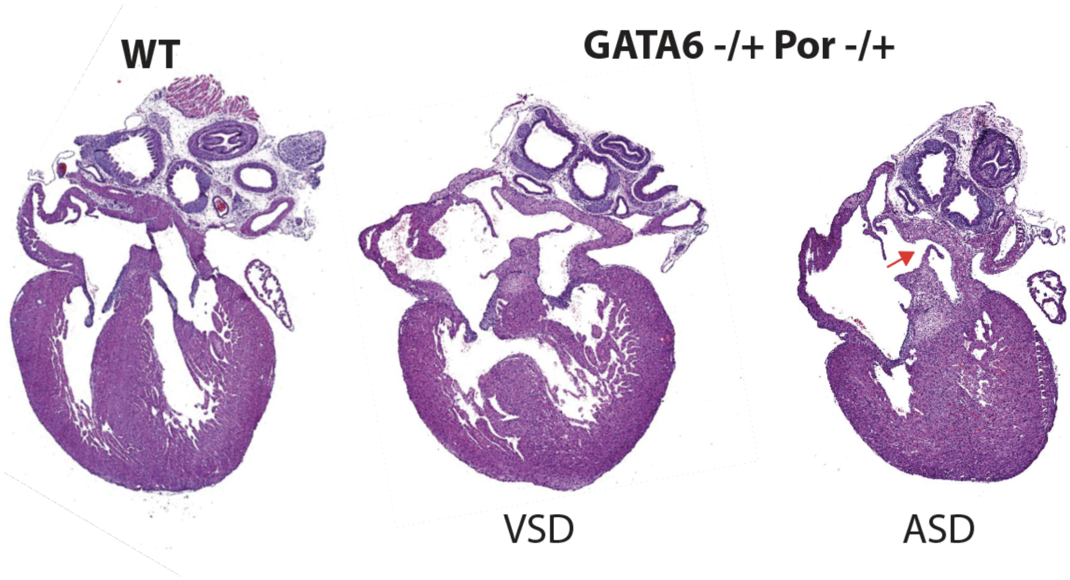
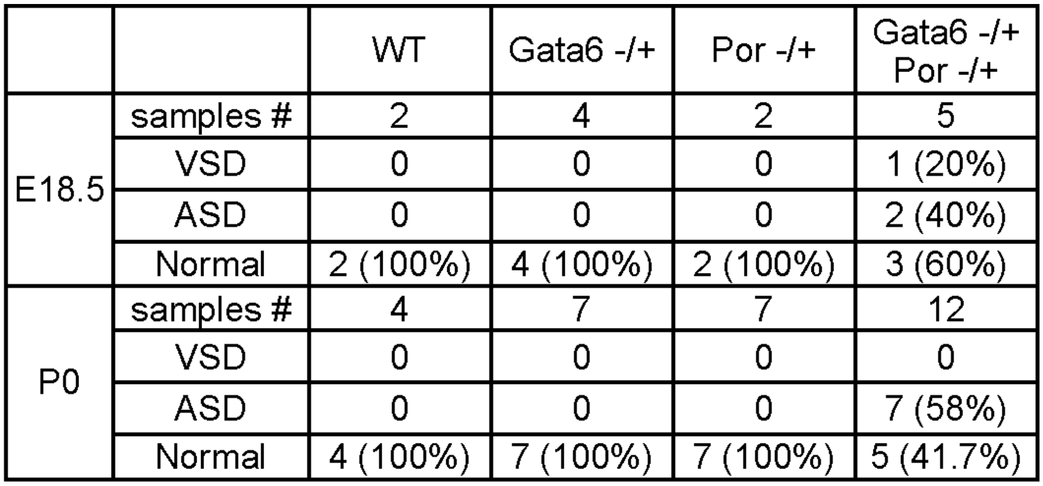


Figure 3.9: Representative histology samples from Gata6-/+ Por-/+ compound heterozygous mice showing VSD or ASD.

Table 3.4: Incidence of atrial-septal defects (ASD) and ventricular-septal defects (VSD) in GATA6 and POR knockdown mice.



		WT	Gata6 -/+	Por -/+	Gata6 -/+ Por -/+
E18.5	samples #	2	4	2	5
	VSD	0	0	0	1 (20%)
	ASD	0	0	0	2 (40%)
	Normal	2 (100%)	4 (100%)	2 (100%)	3 (60%)
P0	samples #	4	7	7	12
	VSD	0	0	0	0
	ASD	0	0	0	7 (58%)
	Normal	4 (100%)	7 (100%)	7 (100%)	5 (41.7%)

3.4 DISCUSSION

Developmental phenotypes affecting the heart and brain have profound impacts on the lives of patients and their families, and understanding the underlying cause in each case can reveal additional disease risks to monitor as the patient ages to adulthood. Evidence of oligogenic transmission of developmental disease has increased in recent years (Priest et al. 2016; Alsemari et al. 2018; Gifford et al. 2019; Mkaouar et al. 2021), necessitating a fast and accurate method to determine gene combinations of interest from patient data. We provide a computational framework called GCOD to test for the significance of oligogenic gene set transmission and report 202, 1023, and 922 significantly over-transmitted oligogenic pairs in CHD, ASD, and kidney disease respectively. These findings collectively contribute 2966 novel potential risk genes and modifiers from 7889 unique predicted-damaging higher-order sets across all three datasets. This software is unique compared to other recent developments in the field of oligogenic combination detection (Kerner et al. 2020; Pounraja and Girirajan 2022) because it uses family genotypes to limit statistical inference to gene sets with oligogenic transmission, thereby reducing computation and increasing our power to detect true phenotype associations.

Given that ten percent of CHD probands carry predicted-damaging variants in an oligogenic set, our method potentially explains the genetic underpinnings of a substantial minority of cases with previously cryptic causal variants (**Table 3.2**). Experimental validation is required to determine how many of these candidate variant sets are indeed sufficient to cause a heart defect, or whether some level of additional background genetic risk is necessary. This is especially key as 97% of probands also carry combinations of less damaging variants that collectively confer some risk of CHD. Integrating a polygenic risk score approach (Wendt et al. 2020; Isgut et al. 2021) could be used to prioritize oligogenic sets that lead to disease even in otherwise relatively low-risk genomes.

We have established the concept of the “highest-order” gene set, which allows users to examine gene groups of three, four, and larger sets. This increases power by minimizing the number of individual tests. Such an analysis could be computationally prohibitive, but the GCOD algorithm efficiently leverages family sequencing data to filter out combinations seen in healthy parents. Parental sequencing also mitigates the effects of population stratification, as gene set significance is conditional only on variant transmission from parents to offspring, and does not rely on relative frequencies in a population of potentially different ancestry.

Finally, our mouse model incorporating mouse crosses of *Gata6* and *Por* compound heterozygous knockdown shows that GCOD can identify and prioritize disease-relevant combinations from trio data. We include the caveat that *GATA6* alone might lead to disease in some cases, given the incidence of ASD in a patient not carrying a *POR* mutation. Furthermore, our mouse model did not show the expected penetrant Truncus Arteriosus phenotype, but instead sporadic incidence of VSD and ASD. More molecular assays and experiments are needed to understand the context of this genetic interaction, and its relationship to heart development and TA.

Our method identified several protein complexes and functions associated with heart development, and further expanded these canonical sets to include potentially-interacting genes and risk modifiers. For example, oligogenic combinations including SACS hint at key protein interactions between the SACS filamental organizing protein with various myosin and membrane-related gene products. Exploring the function and interactions of such gene products could further elucidate cytoskeletal activity in the developing heart. Furthermore, genes that appear in CHD oligogenic sets are enriched for appropriate functions like atrioventricular canal development, as well as more novel associations like de novo pyrimidine synthesis and perception of light stimulus. As was recently explored by (Morton et al. 2021), individuals born with heart defects are at risk for additional complications later in life, including cancers. We similarly hypothesize a role for known bladder cancer gene COL6A3 in CHD (Huang et al. 2018). We recover several retinitis pigmentosa risk genes, suggesting future studies to determine if the eye health of CHD patients should be monitored as they age to adulthood.

Our simulation method to assess significance of oligogenic sets relies on having sequencing data from enough families to observe multiple occurrences of rare gene combinations. However, we expect that clinicians working with one or just a few family pedigrees could make use of the first two steps of GCOD alone. Variant and gene combinations segregating with the disease but not observed in healthy family members can be enumerated to develop hypotheses for disease etiology where no clear Mendelian inheritance pattern exists.

GCOD as implemented here is not an exhaustive search for oligogenic sets in these families, as it does not take into account common variants (which are expected to confer less risk, but undoubtedly play a role in some cases). It is also important to note that kinship controls, while offering advantages in a conservative analysis, are likely not entirely unaffected by the risk variants they carry and may exhibit subclinical phenotypes. This is an especially important caveat in the SFARI autism cohort, as sets were filtered to those never observed in an unaffected sibling; however, the protein products of sibling gene sets might indeed interact and collaboratively cause disease contingent on other factors like sex or differences in environmental exposure between siblings.

Since GCOD is limited to gene combinations seen in a minimum of two probands, we recommend that users hoping to characterize individual cases of potentially oligogenic pathogenesis use a complementary tool such as ORVAL (Renaux et al. 2019) to identify other contributing risk genes and variants in the particular case(s) of interest. Finally, GCOD is more sensitive to de novo variation due to the fact that an inherited event has a 50% probability of occurring, while most de novo events are exceedingly rare; results are therefore biased towards discovering oligogenic sets where a non-synonymous de novo variant was observed.

Overall, we have created a framework to identify combinations of genes transmitted in CHD. We have provided researchers with these lists of prioritized, high-confidence gene and variant combinations for high-throughput screens or more precise mechanistic experiments. GCOD is freely available for applications to other diseases and species, in hopes of moving the field toward a more complete understanding of gene interactions during development, and how genetic variants combine in disease. This work is published with open access to biorxiv (Pittman et al. 2022) and is currently under revision.

3.5 METHODS

3.5.1: Variant calling and filtering

Whole Exome Sequencing data from the 3382 CHD trios was processed as described in Jin et al. 2017 and Morton et al. 2021. Briefly, protein-coding mutations were filtered based on a Mapping Quality score > 59 and Genotype Quality > 90. De novo variants were called using the TrioDeNovo program (Wei et al. 2015) and accepted if the minor allele frequency (MAF) and read-depth criteria described in Homsy et al. 2015 are met. Namely, the in-cohort MAF of the variant must be below 4x10-4, with a minimum of 5 alternative reads and 10 total reads in the proband, and a minimum of 10 reference reads in the parents (with a maximum alternate allele ratio of 3.5%). Variants in the SFARI autism cohort were called as described in Krumm et al. 2015. In addition to the criteria above, we filtered this set for inherited variants with a Genotype

Quality Mean > 85 and plausible inheritance pattern (e.g. if a parent is 1/1 and a child is 0/0 without a de novo variant at this locus, the variant is disqualified). De novo variants for SFARI probands and unaffected siblings were provided by Iossifov et al. 2014. All variants were called and are reported in GRCh37 coordinates.

For both datasets, we annotated variants using ANNOVAR version date 2021-06-08 (Wang, Li, and Hakonarson 2010). We restricted our analysis to variants annotated as <5% Minor Allele Frequency (MAF) in either the gnomAD (Karczewski et al. 2020) or ExAC non-psychiatric (Lek et al. 2016) databases, and below an in-cohort MAF of 10%. GCOD users can change these cutoffs where desired. Variants predicted to be damaging by at least one of MetaSVM, FatHMM, and SIFT models were considered in the Strict and Base-damaging tiers of variant severity (Liu et al. 2016; Rogers et al. 2017). Filtering criteria of variant severity is specified in Supplemental Table 1. We used gnomAD observed/expected (Karczewski et al. 2020), Shet (Zhu, Zhang, and Sha 2018), CADD scores (Rentzsch et al. 2019), and a minimum expression of tpm > 0.5 in any cardiac or brain cell type (for CHD and ASD genes, respectively) included in the DESCARTES developmental gene expression database (J. Cao et al. 2020). Cell types considered are listed below.

Cardiac cell types: 'Heart-Cardiomyocytes', 'Heart-CLC_IL5RA positive cells', 'Heart-ELF3_AGBL2 positive cells', 'Heart-Endocardial cells', 'Heart-Epicardial fat cells', 'Heart-Erythroblasts', 'Heart-Lymphatic endothelial cells', 'Heart-Lymphoid cells', 'Heart-Megakaryocytes', 'Heart-Myeloid cells', 'Heart-SATB2_LRRC7 positive cells', 'Heart-Schwann cells', 'Heart-Smooth muscle cells', 'Heart-Stromal cells', 'Heart-Vascular endothelial cells', 'Heart-Visceral neurons.'

Brain cell types: 'Cerebellum-Astrocytes', 'Cerebellum-Granule neurons', 'Cerebellum-Inhibitory interneurons', 'Cerebellum-Microglia', 'Cerebellum-Oligodendrocytes', 'Cerebellum-Purkinje neurons', 'Cerebellum-SLC24A4_PEX5L positive cells', 'Cerebellum-Unipolar brush cells', 'Cerebellum-Vascular endothelial cells', 'Cerebrum-Astrocytes', 'Cerebrum-Excitatory neurons', 'Cerebrum-Inhibitory neurons', 'Cerebrum-Limbic system neurons', 'Cerebrum-Megakaryocytes', 'Cerebrum-Microglia', 'Cerebrum-Oligodendrocytes', 'Cerebrum-SKOR2_NPSR1 positive cells', 'Cerebrum-Vascular endothelial cells.'

3.5.2: Pseudo-sibling genotype generation

Pseudo-sibling genotypes were generated from the two parental alleles not transmitted to the proband as described in Yu and Deng 2011. In the rare case in which a proband de novo mutation occurred at a locus where a parent carried a qualifying variant, the pseudo-sibling inherited the parental rare allele if the proband did not. De novo variants in pseudo-siblings were randomly assigned to genes based on previously-derived protein-coding non-synonymous (‘prot’) mutability (Samocha et al. 2014). Since all such variants are automatically included at the Strict variant tier, GCOD does not predict specific amino acid substitutions to further qualify the de novo variant’s CADD, shet, or predicted-damaging status.

3.5.3: Enumeration of candidate digenic pairs

In order to efficiently enumerate gene pairs in which multiple probands harbor co-occurring mutations meeting user criteria, we create three gene-by-family matrices where a cell value indicates the presence of such a variant in the mother, father, and offspring. This information is stored in compressed sparse row format. The values in maternal and paternal matrices indicate the number of alleles each parent carries (later used to calculate the probability of transmitting at least one of these variants across k simulations), while the values in the proband matrix M store information about the inheritance of variants in a given gene. We use the following key:

1 = proband harbors at least one variant in this gene; all variants were transmitted from the mother.

2 = proband harbors at least one variant in this gene; all variants were transmitted from the father.

3 = proband harbors compound heterozygous variants in this gene, i.e. each parent transmitted at least one variant in this gene.

4 = proband harbors at least one de novo variant in this gene.

Information about variant provenance is used during group enumeration and offspring simulations to determine whether a given combination comprises an oligogenically-inherited variant set: that is, the set of offspring variants in this gene combination are not also collectively carried by one of the unaffected parents. Explicitly, for a gene-pair GeneA-GeneB and its occurrence in Trio1, one of the provenance values at MTrio1, GeneA or MTrio1, GeneB must be >2, or else MTrio1, GeneA != MTrio1, GeneB, for the observation to be counted as an oligogenic transmission.

Gene columns in M are dropped if fewer than two probands (or siblings) carried a mutation at or above the relevant variant criteria. For remaining genes, all possible combinations of digenic pairs are enumerated and checked against the matrix for multiple co-occurrences. If the provenance matrix M indicates an oligogenic transmission of a given set of variants in a digenic pair, and this is true of at least two probands (or siblings), the gene pair is considered a "candidate set."

3.5.4: Highest-order gene set enumeration

We defined the concept of a "highest-order" gene set, which comprises the largest unique candidate set observed in two or more probands (Supplemental Figure 1). To efficiently enumerate these sets, we begin with the previously-calculated digenic pairs and the list of offspring harboring them. A highest-order candidate set necessarily contains at least one qualifying digenic pair; therefore we reduce the computational search space by only examining pairs of offspring already known to carry a common digenic candidate. For every pair of probands sharing a digenic candidate, all additional damaging variant-harboring genes in common between the two probands comprise a highest-order set. In other words, the gene pair is expanded by adding all other shared genes, for every pairwise combination of probands that share the digenic pair. We refer to these groups as the “maximum common gene-sets” among pairs of probands.

The final step is to check whether any lower-order sets for a given pair of probands is found in one or more additional probands. For example, Proband1 and Proband2 may have a highest-order set of GeneA-GeneB-GeneC-GeneD, while Proband1 and Proband3 have a highest-order set of GeneA-GeneB-GeneD-GeneE. The initial algorithmic pass only records the following from pairwise proband comparisons:

{ [(Proband1, Proband2) : (GeneA-GeneB-GeneC-GeneD) ,
(Proband1, Proband3) : (GeneA-GeneB-GeneD-GeneE) ,
(Proband2, Proband3) : (GeneA-GeneB-GeneD-GeneX-GeneY)] }.

However, the set GeneA-GeneB-GeneD could have been transmitted oligogenically to all three probands, comprising a unique candidate set with its own transmission count = 3. GCOD therefore sorts the maximum pairwise common sets by the number of participating genes, beginning with the largest gene-set and sequentially checking against smaller gene-sets common between other proband pairs. If the union of genes is greater than 1, and if the variants comprising that smaller set were transmitted oligogenically, then the subset becomes an additional highest-order candidate entry. Thus, since the A-B-D combination is a subset of A-B-C-D, A-B-D-E, and A-B-D-X-Y in the example above, we append GeneA-GeneB-GeneD as a candidate oligogenic set in the highest-order analysis (as long as the transmission of those variants meet criteria for oligogenic transmission in all probands).

3.5.5: Logistic regression filtering

Before simulation analysis, candidate sets are filtered based on their interaction coefficient in a logistic regression model. Disease status is predicted by n+1 variables, where n is the number of genes in an oligogenic candidate set. For all parents and probands (and sequenced siblings where applicable), the presence/absence of a qualifying variant in each individual gene, as well as whether all genes in the combination harbor a variant in that individual, are denoted by 0/1. We use the binomial glm() function in R (version 3.6.1,(R Core Team 2020)) with 50 maximum iterations to test whether the coefficient of the gene interaction is greater than 1. This indicates that the combination itself is broadly associated with disease in this dataset, and it removes spurious combinations that are driven by a single gene within the set.

3.5.6: Oligogenic transmission simulation test

For each family, parental genotypes at loci participating in candidate sets are aggregated and used to simulate 10,000 random offspring. One of each allele from father and mother is randomly selected to comprise the simulated offspring genotype at each locus, and additional de novo mutations are simulated based on previously-derived mutability constants(Samocha et al. 2014). A gene-by-family matrix and parental provenance matrices (as described above in “Digenic pairs enumeration”) is created for each of the 10,000 simulated cohorts. For each oligogenic candidate set, we index the constituent genes and record which simulated offspring carry that combination, contingent on oligogenic transmission. For each candidate set, this generates a distribution of co-occurrence counts based on parental genotypes under the null hypothesis of no disease association (i.e., random transmission). The raw p-value is determined by the proportion of iterations with a simulated count as large or larger than the proband test statistic. GCOD returns the test statistic (number of probands observed with the oligogenic combination) and raw, Benjamini-Hochberg FDR-corrected, and Bonferroni-corrected p-values for each candidate set. In this study, we use a significance threshold of Bonferroni-corrected p-value less than 0.05.

3.5.7: Single-gene transmission simulation test

From the 10,000 simulated gene-by-family matrices described in the “Oligogenic transmission simulation test” above, we additionally create marginal null distributions of the number of probands harboring variants in each gene and use these to compute p-values for individual genes. Significant p-values indicate the binary presence of any qualifying variant in a gene is significantly higher in probands than randomly simulated offspring. Note that this is distinct from a traditional Transmission Disequilibrium Test in that the count of binary presence/absence across probands is the test statistic here, instead of all variant transmissions within the given gene.

3.5.8: Gene ontology analysis

We performed Gene Ontology (GO) analyses using the GOATOOLS package in Python3 (version 1.1.6,(Klopfenstein et al. 2018)), using the GOEnrichmentStudyNS() function with default settings (alpha = 0.05, correction method = fdr_bh for Benjamini-Hochberg FDR-adjusted p-values). When reporting top enriched terms, we focused on GO categories with more than five gene members because small categories are prone to false positives.

3.5.9: Oligogenic set network discovery and depiction

We selected specific oligogenic sets in CHD probands based on criteria detailed below. To visualize gene network nodes represent genes and are colored according to functional annotations. Edges indicate that genes carry damaging variants in the same proband, with edge width representing the number of probands with co-occurring mutations in that gene pair. Note that the edge counts can include probands in which the gene pair was inherited from one parent (non-oligogenic transmission), under the condition that a higher-order oligogenic set includes that transmission (i.e., at least one variant in another gene in the higher-order set was inherited from the other parent or mutated *de novo*). Edges are not drawn between nodes unless the two genes appear together in at least one significant oligogenic set. Oligogenic sets are reported in the accompanying text, as well as Supplemental Table 2 for the full cohort and Supplemental Table 4 for groups specific to a particular diagnosis. For brevity, only select significant sets of interest are included in the provenance matrices in these supplemental tables.

For Figure 3B, we discovered two oligogenic sets in which genes are known to physically interact in a canonical protein complex. We selected all other significant oligogenic sets containing at least one gene in these complexes, and visualized them using genes for nodes and co-occurrences as edges. For Figure 3D, we sought an oligogenic set with several counts of oligogenic transmissions but rarely any variant combinations seen in unaffected parents, discovering the MYO18B-SACS combination transmitted oligogenically 10 out of 12 times. We additionally incorporated genes appearing in significant oligogenic sets in any of these 12 probands. Figure 4A shows all significant sets containing ARSG in the CHD and ASD cohorts (6 pairs), and Figure 4B shows all significant sets containing HERC1 in the ASD cohort (3 pairs).

3.6 DATA AND SOFTWARE ACCESS

PCGC variants are available under dbGaP Study Accession: phs000571.v6.p2 for qualifying researchers. GCOD software, including Jupyter notebooks with detailed analyses and use cases, is provided at https://github.com/mepittman/gcod.

REFERENCES

1000 Genomes Project Consortium, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Erik P. Garrison, Hyun Min Kang, Jan O. Korbel, et al. 2015. “A Global Reference for Human Genetic Variation.” *Nature* 526 (7571): 68–74.

Akhirome, Ehiole, Nephi A. Walton, Julie M. Nogee, and Patrick Y. Jay. 2017. “The Complex Genetic Basis of Congenital Heart Defects.” *Circulation Journal: Official Journal of the Japanese Circulation Society* 81 (5): 629–34.

Alsemari, Abdulaziz, Mohanned Alsuhaibani, Rawabi Alhathlool, and Bayan Mamdouh Ali. 2018. “Potential Oligogenic Disease of Mental Retardation, Short Stature, Spastic Paraparesis, and Osteopetrosis.” *The Application of Clinical Genetics* 11 (November): 129–34.

Azhar, Mohamad, Jo El J. Schultz, Ingrid Grupp, Gerald W. Dorn 2nd, Pierre Meneton, Daniel G. M. Molin, Adriana C. Gittenberger-de Groot, and Thomas Doetschman. 2003. “Transforming Growth Factor Beta in Cardiovascular Development and Function.” *Cytokine & Growth Factor Reviews* 14 (5): 391–407.

Bosch, Elena, Hafid Laayouni, Carlos Morcillo-Suarez, Ferran Casals, Andrés Moreno-Estrada, Anna Ferrer-Admetlla, Michelle Gardner, et al. 2009. “Decay of Linkage Disequilibrium within Genes across HGDP-CEPH Human Samples: Most Population Isolates Do Not Show Increased LD.” *BMC Genomics* 10 (July): 338.

Cao, Junyue, Diana R. O’Day, Hannah A. Pliner, Paul D. Kingsley, Mei Deng, Riza M. Daza, Michael A. Zager, et al. 2020. “A Human Cell Atlas of Fetal Gene Expression.” *Science* 370 (6518). https://doi.org/10.1126/science.aba7721.

Cao, Si-Yuan, Hui-Liang Shen, Lun Luo, Shu-Jie Chen, and Chunguang Li. 2021. “PCNet: A Structure Similarity Enhancement Method for Multispectral and Multimodal Image Registration.” *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/2106.05124.

Cassa, Christopher A., Donate Weghorn, Daniel J. Balick, Daniel M. Jordan, David Nusinow, Kaitlin E. Samocha, Anne O'Donnell-Luria, et al. 2017. "Estimating the Selective Effects of Heterozygous Protein-Truncating Variants from Human Exome Data." *Nature Genetics* 49 (5): 806–10.

Chan, H. M., and N. B. La Thangue. 2001. "p300/CBP Proteins: HATs for Transcriptional Bridges and Scaffolds." *Journal of Cell Science* 114 (Pt 13): 2363–73.

Cripe, Linda, Gregor Andelfinger, Lisa J. Martin, Kerry Shooner, and D. Woodrow Benson. 2004. "Bicuspid Aortic Valve Is Heritable." *Journal of the American College of Cardiology* 44 (1): 138–43.

Dankel, Simon N., Elise Grytten, Jan-Inge Bjune, Hans Jørgen Nielsen, Arne Dietrich, Matthias Blüher, Jørn V. Sagen, and Gunnar Mellgren. 2020. "COL6A3 Expression in Adipose Tissue Cells Is Associated with Levels of the Homeobox Transcription Factor PRRX1." *Scientific Reports* 10 (1): 20164.

Ellingford, Jamie M., Joo Wook Ahn, Richard D. Bagnall, Diana Baralle, Stephanie Barton, Chris Campbell, Kate Downes, et al. 2022. "Recommendations for Clinical Interpretation of Variants Found in Non-Coding Regions of the Genome." *Genome Medicine* 14 (1): 73.

Ewens, Warren J. 1999. "Statistical Aspects of the Transmission/Disequilibrium Test (TDT)." *Lecture Notes-Monograph Series / Institute of Mathematical Statistics* 33: 77–94.

Gentil, Benoit J., Gia-Thanh Lai, Marie Menade, Roxanne Larivière, Sandra Minotti, Kalle Gehring, J-Paul Chapple, Bernard Brais, and Heather D. Durham. 2019. "Sacsin, Mutated in the Ataxia ARSACS, Regulates Intermediate Filament Assembly and Dynamics." *FASEB Journal: Official Publication of the Federation of American Societies for Experimental Biology* 33 (2): 2982–94.

Gifford, Casey A., Sanjeev S. Ranade, Ryan Samarakoon, Hazel T. Salunga, T. Yvanka de Soysa, Yu Huang, Ping Zhou, et al. 2019. "Oligogenic Inheritance of a Human Heart Disease Involving a Genetic Modifier." *Science* 364 (6443): 865–70.

Giurgiu, Madalina, Julian Reinhard, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Gisela Fobo, Goar Frishman, Corinna Montrone, and Andreas Ruepp. 2019. “CORUM: The Comprehensive Resource of Mammalian Protein Complexes-2019.” *Nucleic Acids Research* 47 (D1): D559–63.

He, Zongxiao, Di Zhang, Alan E. Renton, Biao Li, Linhai Zhao, Gao T. Wang, Alison M. Goate, Richard Mayeux, and Suzanne M. Leal. 2017. “The Rare-Variant Generalized Disequilibrium Test for Association Analysis of Nuclear and Extended Pedigrees with Application to Alzheimer Disease WGS Data.” *American Journal of Human Genetics* 100 (2): 193–204.

Hinton, Robert B., Jr, Lisa J. Martin, Meredith E. Tabangin, Mjaye L. Mazwi, Linda H. Cripe, and D. Woodrow Benson. 2007. “Hypoplastic Left Heart Syndrome Is Heritable.” *Journal of the American College of Cardiology* 50 (16): 1590–95.

Hoang, Thanh T., Elizabeth Goldmuntz, Amy E. Roberts, Wendy K. Chung, Jennie K. Kline, John E. Deanfield, Alessandro Giardini, et al. 2018. “The Congenital Heart Disease Genetic Network Study: Cohort Description.” *PloS One* 13 (1): e0191319.

Huang, Yan, Gang Li, Kai Wang, Zhongyi Mu, Qingpeng Xie, Hongchen Qu, Hang Lv, and Bin Hu. 2018. “Collagen Type VI Alpha 3 Chain Promotes Epithelial-Mesenchymal Transition in Bladder Cancer Cells via Transforming Growth Factor β (TGF-β)/Smad Pathway.” *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research* 24 (August): 5346–54.

Hutson, Mary R., and Margaret L. Kirby. 2007. “Model Systems for the Study of Heart Development and Disease. Cardiac Neural Crest and Conotruncal Malformations.” *Seminars in Cell & Developmental Biology* 18 (1): 101–10.

Isgut, Monica, Jimeng Sun, Arshed A. Quyyumi, and Greg Gibson. 2021. “Highly Elevated Polygenic Risk Scores Are Better Predictors of Myocardial Infarction Risk Early in Life than Later.” *Genome Medicine* 13 (1): 13.

Janknecht, R., N. J. Wells, and T. Hunter. 1998. “TGF-Beta-Stimulated Cooperation of Smad Proteins with the Coactivators CBP/p300.” *Genes & Development* 12 (14): 2114–19.

Jin, Sheng Chih, Jason Homsy, Samir Zaidi, Qiongshi Lu, Sarah Morton, Steven R. DePalma, Xue Zeng, et al. 2017. “Contribution of Rare Inherited and de Novo Variants in 2,871 Congenital Heart Disease Probands.” *Nature Genetics* 49 (11): 1593–1601.

Karczewski, Konrad J., Laurent C. Francioli, Grace Tiao, Beryl B. Cummings, Jessica Alföldi, Qingbo Wang, Ryan L. Collins, et al. 2020. “The Mutational Constraint Spectrum Quantified from Variation in 141,456 Humans.” *Nature* 581 (7809): 434–43.

Kerner, Gaspard, Matthieu Bouaziz, Aurélie Cobat, Benedetta Bigio, Andrew T. Timberlake, Jacinta Bustamante, Richard P. Lifton, Jean-Laurent Casanova, and Laurent Abel. 2020. “A Genome-Wide Case-Only Test for the Detection of Digenic Inheritance in Human Exomes.” *Proceedings of the National Academy of Sciences of the United States of America* 117 (32): 19367–75.

Keyte, Anna, and Mary Redmond Hutson. 2012. “The Neural Crest in Cardiac Congenital Anomalies.” *Differentiation; Research in Biological Diversity* 84 (1): 25–40.

Kingdom, Rebecca, and Caroline F. Wright. 2022. “Incomplete Penetrance and Variable Expressivity: From Clinical Studies to Population Cohorts.” *Frontiers in Genetics* 13 (July): 920390.

Köhler, Sebastian, Michael Gargano, Nicolas Matentzoglu, Leigh C. Carmody, David Lewis-Smith, Nicole A. Vasilevsky, Daniel Danis, et al. 2021. “The Human Phenotype Ontology in 2021.” *Nucleic Acids Research* 49 (D1): D1207–17.

Kousi, Maria, and Nicholas Katsanis. 2015. “Genetic Modifiers and Oligogenic Inheritance.” *Cold Spring Harbor Perspectives in Medicine* 5 (6). https://doi.org/10.1101/cshperspect.a017145.

Koutsourakis, M., A. Langeveld, R. Patient, R. Beddington, and F. Grosveld. 1999. “The Transcription Factor GATA6 Is Essential for Early Extraembryonic Development.”

Development 126 (4): 723–32.

Lepore, John J., Patricia A. Mericko, Lan Cheng, Min Min Lu, Edward E. Morrisey, and Michael S. Parmacek. 2006. “GATA-6 Regulates Semaphorin 3C and Is Required in Cardiac Neural Crest for Cardiovascular Morphogenesis.” *The Journal of Clinical Investigation* 116 (4): 929–39.

Liberzon, Arthur, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. 2015. “The Molecular Signatures Database (MSigDB) Hallmark Gene Set Collection.” *Cell Systems* 1 (6): 417–25.

Li, Peng, Mohammad Pashmforoush, and Henry M. Sucov. 2010. “Retinoic Acid Regulates Differentiation of the Secondary Heart Field and TGFbeta-Mediated Outflow Tract Septation.” *Developmental Cell* 18 (3): 480–85.

Mahdavi, V., M. Periasamy, and B. Nadal-Ginard. 1982. “Molecular Characterization of Two Myosin Heavy Chain Genes Expressed in the Adult Heart.” *Nature* 297 (5868): 659–64.

Maitra, Meenakshi, Sara N. Koenig, Deepak Srivastava, and Vidu Garg. 2010. “Identification of GATA6 Sequence Variants in Patients with Congenital Heart Defects.” *Pediatric Research* 68 (4): 281–85.

Marschang, Peter, Jochen Brich, Edwin J. Weeber, J. David Sweatt, John M. Shelton, James A. Richardson, Robert E. Hammer, and Joachim Herz. 2004. “Normal Development and Fertility of Knockout Mice Lacking the Tumor Suppressor Gene LRP1b Suggest Functional Compensation by LRP1.” *Molecular and Cellular Biology* 24 (9): 3782–93.

McBride, Kim L., Ricardo Pignatelli, Mark Lewin, Trang Ho, Susan Fernbach, Andres Menesses, Wilbur Lam, et al. 2005. “Inheritance Analysis of Congenital Left Ventricular Outflow Tract Obstruction Malformations: Segregation, Multiplex Relative Risk, and Heritability.” *American Journal of Medical Genetics. Part A* 134A (2): 180–86.

Miller, Walter L., Vishal Agrawal, Duanpen Sandee, Meng Kian Tee, Ningwu Huang, Ji Ha Choi, Kari Morrissey, and Kathleen M. Giacomini. 2011. “Consequences of POR Mutations and

Polymorphisms." *Molecular and Cellular Endocrinology* 336 (1-2): 174–79.

Mkaouar, Rahma, Lamia Cherif Ben Abdallah, Chokri Naouali, Saida Lahbib, Zinet Turki, Sahar Elouej, Yosra Bouyacoub, et al. 2021. "Oligogenic Inheritance Underlying Incomplete Penetrance of PROKR2 Mutations in Hypogonadotropic Hypogonadism." *Frontiers in Genetics* 12 (September): 665174.

Morton, Sarah U., Akiko Shimamura, Peter E. Newburger, Alexander R. Opotowsky, Daniel Quiat, Alexandre C. Pereira, Sheng Chih Jin, et al. 2021. "Association of Damaging Variants in Genes With Increased Cancer Risk Among Patients With Congenital Heart Disease." *JAMA Cardiology* 6 (4): 457–62.

Neeb, Zachary, Jacquelyn D. Lajiness, Esther Bolanis, and Simon J. Conway. 2013. "Cardiac Outflow Tract Anomalies." *Wiley Interdisciplinary Reviews. Developmental Biology* 2 (4): 499–530.

Oh, Jongwon, Ju Sun Song, Jong Eun Park, Shin Yi Jang, Chang Seok Ki, and Duk Kyung Kim. 2017. "A Case of Antley-Bixler Syndrome With a Novel Likely Pathogenic Variant (c.529G>C) in the POR Gene." *Annals of Laboratory Medicine* 37 (6): 559–62.

Otto, Diana M. E., Colin J. Henderson, Dianne Carrie, Megan Davey, Thomas E. Gundersen, Rune Blomhoff, Ralf H. Adams, Cheryll Tickle, and C. Roland Wolf. 2003. "Identification of Novel Roles of the Cytochrome p450 System in Early Embryogenesis: Effects on Vasculogenesis and Retinoic Acid Homeostasis." *Molecular and Cellular Biology* 23 (17): 6103–16.

Parker, Lauren E., and Andrew P. Landstrom. 2021. "Genetic Etiology of Left-Sided Obstructive Heart Lesions: A Story in Development." *Journal of the American Heart Association* 10 (2): e019006.

Pediatric Cardiac Genomics Consortium, Bruce Gelb, Martina Brueckner, Wendy Chung, Elizabeth Goldmuntz, Jonathan Kaltman, Juan Pablo Kaski, et al. 2013. "The Congenital Heart Disease Genetic Network Study: Rationale, Design, and Early Results." *Circulation*

Research 112 (4): 698–706.

Pierpont, Mary Ella, Martina Brueckner, Wendy K. Chung, Vidu Garg, Ronald V. Lacro, Amy L. McGuire, Seema Mital, et al. 2018. “Genetic Basis for Congenital Heart Disease: Revisited: A Scientific Statement From the American Heart Association.” *Circulation* 138 (21): e653–711.

Pounraja, Vijay Kumar, and Santhosh Girirajan. 2022. “A General Framework for Identifying Oligogenic Combinations of Rare Variants in Complex Disorders.” *Genome Research*, March. https://doi.org/10.1101/gr.276348.121.

Priest, James R., Kazutoyo Osoegawa, Nebil Mohammed, Vivek Nanda, Ramendra Kundu, Kathleen Schultz, Edward J. Lammer, et al. 2016. “De Novo and Rare Variants at Multiple Loci Support the Oligogenic Origins of Atrioventricular Septal Heart Defects.” *PLoS Genetics* 12 (4): e1005963.

Renaux, Alexandre, Sofia Papadimitriou, Nassim Versbraegen, Charlotte Nachtegael, Simon Boutry, Ann Nowé, Guillaume Smits, and Tom Lenaerts. 2019. “ORVAL: A Novel Platform for the Prediction and Exploration of Disease-Causing Oligogenic Variant Combinations.” *Nucleic Acids Research* 47 (W1): W93–98.

Rentzsch, Philipp, Daniela Witten, Gregory M. Cooper, Jay Shendure, and Martin Kircher. 2019. “CADD: Predicting the Deleteriousness of Variants throughout the Human Genome.” *Nucleic Acids Research* 47 (D1): D886–94.

Samocha, Kaitlin E., Elise B. Robinson, Stephan J. Sanders, Christine Stevens, Aniko Sabo, Lauren M. McGrath, Jack A. Kosmicki, et al. 2014. “A Framework for the Interpretation of de Novo Mutation in Human Disease.” *Nature Genetics* 46 (9): 944–50.

Schaaf, Christian P., Aniko Sabo, Yasunari Sakai, Jacy Crosby, Donna Muzny, Alicia Hawes, Lora Lewis, et al. 2011. “Oligogenic Heterozygosity in Individuals with High-Functioning Autism Spectrum Disorders.” *Human Molecular Genetics* 20 (17): 3366–75.

Shen, Anna L., Kathleen A. O’Leary, and Charles B. Kasper. 2002. “Association of Multiple

Developmental Defects and Embryonic Lethality with Loss of Microsomal NADPH-Cytochrome P450 Oxidoreductase.” *The Journal of Biological Chemistry* 277 (8): 6536–41.

Spielman, R. S., R. E. McGinnis, and W. J. Ewens. 1993. “Transmission Test for Linkage Disequilibrium: The Insulin Gene Region and Insulin-Dependent Diabetes Mellitus (IDDM).” *American Journal of Human Genetics* 52 (3): 506–16.

Veenstra-Vanderweele, Jeremy, Susan L. Christian, and Edwin H. Cook Jr. 2004. “Autism as a Paradigmatic Complex Genetic Disorder.” *Annual Review of Genomics and Human Genetics* 5: 379–405.

Warkman, Andrew S., Samantha A. Whitman, Melanie K. Miller, Robert J. Garriock, Catherine M. Schwach, Carol C. Gregorio, and Paul A. Krieg. 2012. “Developmental Expression and Cardiac Transcriptional Regulation of Myh7b, a Third Myosin Heavy Chain in the Vertebrate Heart.” *Cytoskeleton* 69 (5): 324–35.

Wendt, Frank R., Carolina Muniz Carvalho, Gita A. Pathak, Joel Gelernter, and Renato Polimanti. 2020. “Polygenic Risk for Autism Spectrum Disorder Associates with Anger Recognition in a Neurodevelopment-Focused Phenome-Wide Scan of Unaffected Youths from a Population-Based Cohort.” *PLoS Genetics* 16 (9): e1009036.

Wenger, Tara L., Charlly Kao, Donna M. McDonald-McGinn, Elaine H. Zackai, Alice Bailey, Robert T. Schultz, Bernice E. Morrow, Beverly S. Emanuel, and Hakon Hakonarson. 2016. “The Role of mGluR Copy Number Variation in Genetic and Environmental Forms of Syndromic Autism Spectrum Disorder.” *Scientific Reports* 6 (January): 19372.

Xu, Mingjie, Jie Yao, Yingchao Shi, Huijuan Yi, Wukui Zhao, Xinhua Lin, and Zhongzhou Yang. 2021. “The SRCAP Chromatin Remodeling Complex Promotes Oxidative Metabolism during Prenatal Heart Development.” *Development* 148 (8). https://doi.org/10.1242/dev.199026.

Yu, Zhaoxia, and Li Deng. 2011. “Pseudosibship Methods in the Case-Parents Design.”

Statistics in Medicine 30 (27): 3236–51.

Zlotnik, Dor, Tatiana Rabinski, Aviv Halfon, Shira Anzi, Inbar Plaschkes, Hadar Benyamini, Yuval Nevo, et al. 2022. “P450 Oxidoreductase Regulates Barrier Maturation by Mediating Retinoic Acid Metabolism in a Model of the Human BBB.” *Stem Cell Reports* 17 (9): 2050–63.

CHAPTER 4: Computational prediction of chromatin interaction frequency helps prioritize structural variants

This chapter is adapted from my grant proposal entitled “High-throughput computational modeling to assess the role of 3D genome folding in human congenital anomalies,” which was awarded the Ruth L. Kirschstein National Research Service Award #F31HL156439. Section 4.3.4 is adapted from my work included in a manuscript* in press at *Circulation: Genomic and Precision Medicine,* in addition to the preprint cited here. Contributions to work included in this chapter not performed by me are specified below.

Section 4.3.3

- Dr. Jianhua Chu at RUCDR infinity-biologix performed CRISPR cell editing.
- Dr. Jinshing Wu (Bruneau group) performed cell culture and differentiation of cells.

Section 4.3.4*

- J.P. and M.A. performed GWAS, SNP mapping, and association analyses.
- J.P. contributed to the interpretation and annotation of Figure 4.8, and wrote text from which I adapted section 4.3.4.

*Mengyao Yu, Andrew Harper, Matthew Aguirre, **Maureen Pittman**, Catherine Tcheandjieu, A. Dulguun, C.Grace, A. Goel, M. Farrall, K. Xiao, J. Engreitz, K.S. Pollard, H. Watkins, and J. Priest. “Genetic Determinants of Interventricular Septal Anatomy and the Risk of Ventricular Septal Defects and Hypertrophic Cardiomyopathy.” medRxiv. https://doi.org/10.1101/2021.04.19.21255650.

4.1 ABSTRACT

The majority of genetic studies in CHD have focused on variation within the protein-coding exome; however, most GWAS disease-risk loci fall in non-coding regions, and it is presumed that some of these represent important regulators of gene expression such as cis-acting enhancers and insulators. I hypothesized that genetic variants that alter 3D genome folding contribute to the etiology of CHD by disrupting the contacts of key cis-regulatory elements in development. I found that structural variation at TAD boundaries and other non-coding regulatory regions is found more commonly in CHD patients, and interestingly their parents as well, relative to controls. Case variants are further distinguished from controls' by the level of contact disruption specifically near gene loci and in transcriptionally-active regions of relevant cell types. An annotation-agnostic deep learning approach developed in our group, called Akita (Fudenberg, Kelley, and Pollard 2020), predicted chromatin contact changes as a result of genetic variants in CHD patients, which were recapitulated in iPSC-derived cardiac progenitor cell types. I showed that a common structural variant associated with decreased interventricular septum size strongly alters 3D contacts near the *KANSL1* gene. Collectively, these findings elucidate 3D genome organization as a previously underappreciated source of regulatory disruption in CHD, provides interpretation for variants associated with disease, and validates the use of Akita for high-throughput prediction of chromatin contact frequency.

4.2 BACKGROUND

The three-dimensional organization of the genome is critical for transcriptional control through mechanisms like cis-acting enhancers and insulators (Riethoven 2010). In the context of human heart and brain development, the disruption of finely-tuned transcriptional networks is likely to have acute consequences (Won et al. 2016; Carullo and Day 2019; Richter et al. 2020; Yuan, Scott, and Wilson 2021). That intuition has led to studies like the one discussed in Chapter 2 of this thesis, in which transcription factors and their interaction networks were investigated for relevance to congenital heart defects (CHD) (Gonzalez-Teran et al. 2022). This and the majority of other studies in CHD have focused on rare variation within the protein-coding exome (Homsy et al. 2015; Sifrim et al. 2016; Jin et al. 2017). However, most common disease-risk loci are mapped to non-coding regions: roughly half of statistically associated SNPs are non-coding intronic, with an additional one third found in non-coding intergenic regions (Klemm et al. 2013; F. Zhang and Lupski 2015). Given that the basis of fewer than 40% of CHD cases is known, but largely expected to reside within patient genomes (Akhirome et al. 2017), non-coding variation represents a promising avenue to further elucidate the molecular underpinnings of CHD.

While non-coding phenotype-associated single-nucleotide variants (SNVs) are thought to modulate the binding efficiency of transcription factors at their DNA sequence motifs (Albert and Kruglyak 2015; Degtyareva, Antontseva, and Merkulova 2021), structural variants (SVs) like duplications, deletions, inversions, and translocations have the potential to cause larger structural rearrangements that affect local chromatin conformation near non-coding cis-regulatory regions (Feuk, Carson, and Scherer 2006; Shanta et al. 2020). Individuals with CHD have a higher frequency of rare and *de novo* structural deletions and duplications relative to unaffected controls, suggesting that these SVs could play a role in heart defect phenotypes (Glessner et al. 2014). A potentially relevant chromatin structure is the Topologically Associating Domain (TAD), which refers to a level of chromatin organization characterized by higher contact frequency within the domain relative to loci outside of that domain (Nora et al. 2012; Dixon et al. 2012). Previous studies have revealed that merged TADs are often drivers of pathogenic gene expression in cancer and some mammalian limb malformations (Lupiáñez et al. 2015; Gong et al. 2018; Krefting, Andrade-Navarro, and Ibn-Salem 2018; Kragesteen et al. 2018; Fudenberg and Pollard 2019; Claringbould and Zaugg 2021), but it remains unanswered whether such disruptions lead to CHD.

In support of this hypothesis, Fudenberg and Pollard found that while unaffected controls show a clear depletion of SVs at TAD boundary regions across the genome, individuals diagnosed with developmental delay (DD) and autism showed no bias in the genomic location of SVs (Fudenberg and Pollard 2019). SVs at TAD boundaries leading to loss of necessary cis-regulatory contact, or gain of ectopic enhancer-promoter contact, could therefore plausibly underlie cases of these developmental disorders. Interestingly, CHD is comorbid with DD at a rate of up to fifty percent in severe CHD (Marino et al. 2012; Rollins and Newburger 2014). Furthermore, some developmental phenotypes characterized by heart defects are caused by mutations in proteins required for structural integrity of TAD boundaries, like NIPBL and HDAC in Cornelia de Lange syndrome (Piché et al. 2019). I therefore hypothesized that ectopic and disrupted cis-regulatory contacts by structural variation at TAD boundaries can contribute to the genetic basis of developmental disorders, including CHD.

Such mechanisms have been difficult to predict and observe directly, but with the recent influx of high-resolution chromatin contact data and computational tools for its prediction, it may now be possible to characterize the extent to which this occurs in human disease and uncover specific examples. One such tool is the model Akita, which was trained by a deep neural network to predict Hi-C chromatin contact maps from one megabase (Mb) of DNA sequence (Fudenberg, Kelley, and Pollard 2020). Comparisons of Akita's predicted maps have highlighted 3D genome folding changes implicated in cell biology and eukaryotic evolution (Kaaij et al. 2019; Fudenberg, Kelley, and Pollard 2020; McArthur and Capra 2021; McArthur et al. 2022; Gunsalus, Keiser, and Pollard 2022).

In this chapter, I will describe my work investigating CHD patient SVs for signals of chromatin rearrangement, as well as the application of Akita to prioritize variants at scale and interpret their potential molecular effects. This work informed the creation of a cell line in iPSC-derived mesoderm to examine the consequences of a deletion at the locus of cardiac genes *MESP1* and *MESP2*. I also applied Akita to a common genomic inversion associated with the diameter of the interventricular septum, providing molecular hypotheses for disease association via misregulation of *KANSL1*. As researchers generate ever more high-throughput genomic data, both linear and three-dimensional, there is a unique opportunity to use these and related methods to derive novel insights into the etiology of CHD, as well as extend it to other diseases and biological questions.

4.3 RESULTS

4.3.1 Key insulators are enriched for structural variants in CHD patients

First I sought to determine whether the structural variants of CHD patients are more likely to occur at TAD boundaries compared to controls, as shown in DD and autism (Fudenberg and Pollard 2019). To demonstrate this, I obtained SV calls from the Pediatric Cardiac Genomics Consortium (PCGC), consisting of array hybridization data from 2645 CHD proband-parent trios (Hoang et al. 2018). DD structural variant calls were sourced from a meta-analysis of 29,085 affected probands and 19,584 unaffected controls (Coe et al. 2014), **Figure 4.1, A**). Individuals with known chromosomal syndromes and causal protein-coding mutations were removed to maximize novel discovery. I used a previously described null model that assumes that structural variants are not preferentially abundant or depleted at any particular loci (Fudenberg and Pollard 2019). The actual observed number of altered base pairs in that region is then compared to the null expectation in both cases and controls.

While control deletions were significantly depleted at the transcription start sites (TSSs) of highly-expressed genes and the CTCF clusters that often demarcate TAD boundaries (Fudenberg and Pollard 2019), **Figure 4.1, A**), I found that *parents* of CHD probands show less depletion of deleted base pairs in these regions compared to controls (**Figure 4.1, B)**, while CHD probands themselves show a surprising enrichment for deleted basepairs in more highly-expressed TSSs and strongly-bound CTCF sites (**Figure 4.1, C**). In fact this pattern resembles that of deletions found in cancer genomes (Fudenberg and Pollard 2019; Forbes et al. 2010)), which are enriched for structural rearrangement at TAD boundaries. These preliminary results suggest that compromising the integrity of 3D genome folding at the TAD level is associated with CHD as well, and that some of this risk is inherited.

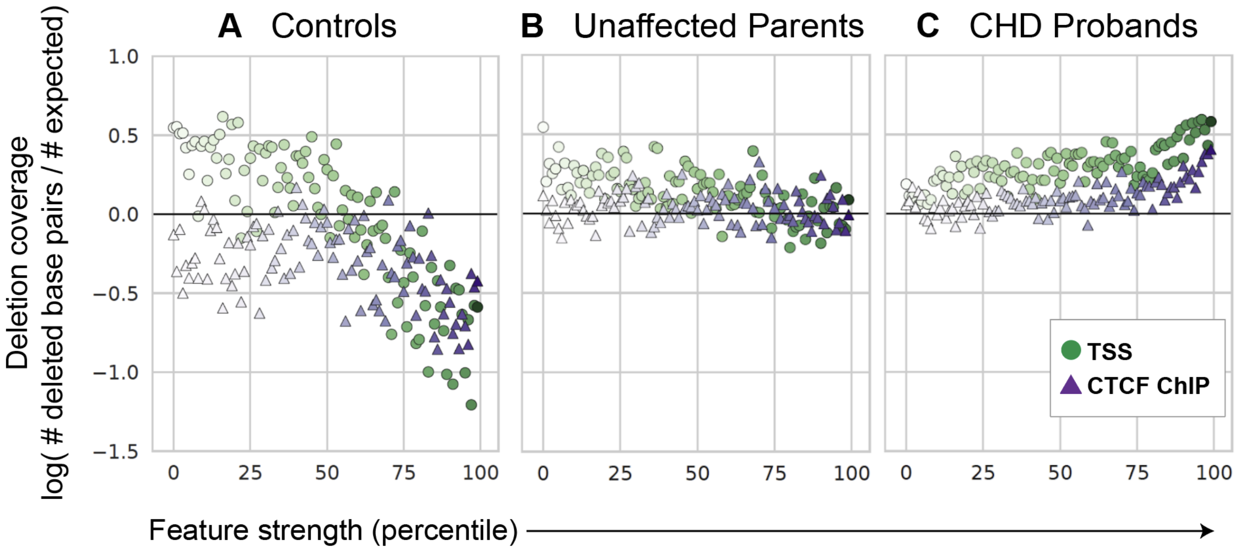


Figure 4.1: Coverage of deleted base pairs in **(A)** controls from (Coe et al. 2014), **(B)** CHD parents, and **(C)** CHD patients at key regulatory regions. Green circles represent transcription start sites (TSSs) of genes at each percentile of expression across tissues in GTEx (GTEx Consortium et al. 2017), arranged along the x-axis and also shaded according to expression strength. Purple triangles represent CTCF peaks along the genome, also stratified and shaded according to strength: in this case ChIP-peak binding strength from ENCODE (ENCODE Project Consortium 2012). Deletion coverage was calculated as previously described (Fudenberg and Pollard 2019), expressed as the log of the ratio between the number of base pairs deleted in those regions and what was expected by chance. See **Methods: Coverage calculation** for additional details.

4.3.2 Locus-informed SV prioritization predicts three deletions disrupting cis-regulatory elements near heart-relevant genes

Prior work found that altering TAD boundaries disrupts DNA folding organization (Lupiáñez et al. 2015), but it is unclear which base pairs and combinations of CTCF binding clusters are key to, and which SVs are subject to, this phenomenon. Additionally, some important features of genome folding during development might not rely on CTCF-mediated activity, like certain repetitive and transposable elements (Y. Zhang et al. 2019; Gunsalus, Keiser, and Pollard 2022). To further characterize the nature of chromatin contact disruption in human variation and disease, I asked directly whether SVs in the genomes of CHD patients cause different levels of contact disruption compared to those found in healthy genomes. While it would be infeasible to experimentally determine the chromatin contact changes of each SV, which would have required 30,000 transgenic or genetically-edited lines and at least as many Hi-C sequencing assays, recent advancements in DNA sequence models allowed us to predict these consequences *in silico*. I used the model Akita (Fudenberg, Kelley, and Pollard 2020), trained using a convolutional neural network on high-resolution chromatin contact data, to predict the Hi-C maps from one million base pairs (1 Mb) of DNA sequence centered around each SV. I predicted maps for both the human reference genome (hg38 assembly) and for that sequence altered *in silico* by the SV, for both private structural variants from 19,584 controls in (Coe et al. 2014) and *de novo* SVs found in 1571 CHD patients (unpublished array data).

In order to quantify the extent of chromatin contact disruption per SV, I used a modified mean squared error (MSE) metric. Disruption between case SVs and control SVs was marginally different (not pictured), so I further weighted this score based on published information about local gene expression and chromatin marks. MSE at each genomic bin (one bin = 2024 base pairs) was multiplied by a factor indicating gene expression in heart tissue (GTEx Consortium et al. 2017) and chromatin activity state in the human fetal heart as predicted by chromHMM (Ernst and Kellis 2017) (See **Methods 4.5.3: Locus-aware prioritization of SVs**). These metrics are referred to as expression-aware disruption and activity-aware disruption, respectively.

Applying these metrics to CHD and control deletions, I found that CHD *de novo* SVs tend to cause more disruption of contacts near genes highly expressed in heart tissue, as well as active chromatin regions (**Figure 4.2**). There is more zero-inflation of control SV disruption scores compared to case SVs, suggesting that a greater proportion of structural variation avoids large-scale reorganization near heart-expressed genes in healthy genomes (**Figure 4.2, A**).

However, it is important to note that many control SVs are themselves quite disruptive even in heart-relevant loci, reaching similar levels to SVs found in cases. Context-specific analysis will be required to determine if any given disruptive mutation is actually disease-causing.

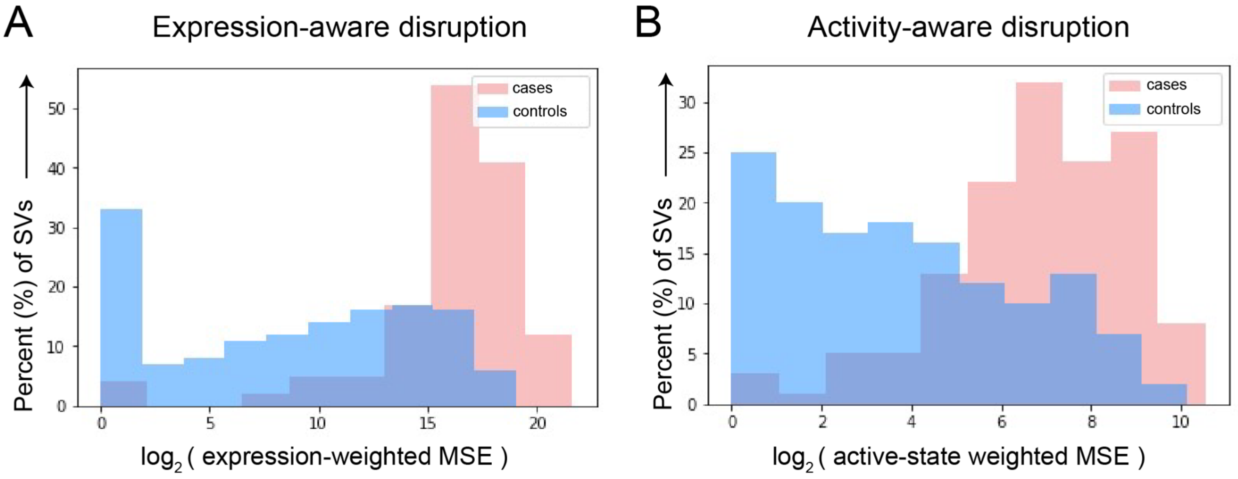


Figure 4.2: (A) Expression- and **(B)** activity-aware disruption scores of case and control SVs. The x-axis indicates the $\log_2$ of the scored difference between reference and SV sequence, weighted by local features. The proportion of SVs with that log-score are indicated by bars, blue for control and pink for cases.

Having observed that case SVs are predicted to cause more disruption, and especially in regions of active expression and active chromatin markers in heart tissue, I thus examined the maps of variants with the highest expression- and activity-aware scores. Three of these SVs occurred near a gene with a known relevant role in cardiac development: an 84 kb deletion that merges two domains in a highly active region encompassing *COX20* and *HNRNPU* (hg38 chr1:244785827-244869791, **Figure 4.3, A**); a small 68-base pair deletion near *STRA6* that merges two small TADs (hg38 chr15:89737238-89795480, **Figure 4.3, B**); and finally, a 58kb deletion that partially deletes mesoderm genes *MESP1/2* and removes a boundary between CHD gene *KIF7* and a neighboring domain (hg38 chr15:73973054- 73973122, **Figure 4.4**).

COX20 plays a role in the assembly of cytochrome C oxidase, and *HNRNPU* is an RNA processing gene that has been associated with developmental brain disorders (Balasubramanian 2022). Neither gene has known function in the heart, so it is unclear what consequences the heterozygous deletion of their locus would have in CHD, but notably a homozygous deletion CRISPR-edited cell line was not viable (Jianhua Chu, unpublished data). The deletion leads to the merging of two TADs that are strongly insulated in wild type (**Figure 4.3, A**), with ectopic contact between ciliary gene *EFCAB2* and the domain containing *DESI2* and various RNA pseudo-genes (not pictured). Further experimental validation and interpretation is required to understand what role the deleted genes and ectopic *EFCAB2* contacts could play in CHD. This is true as well of the ectopic contacts predicted downstream of the *STRA6* locus as a result of a small non-coding deletion (**Figure 4.3, B**).

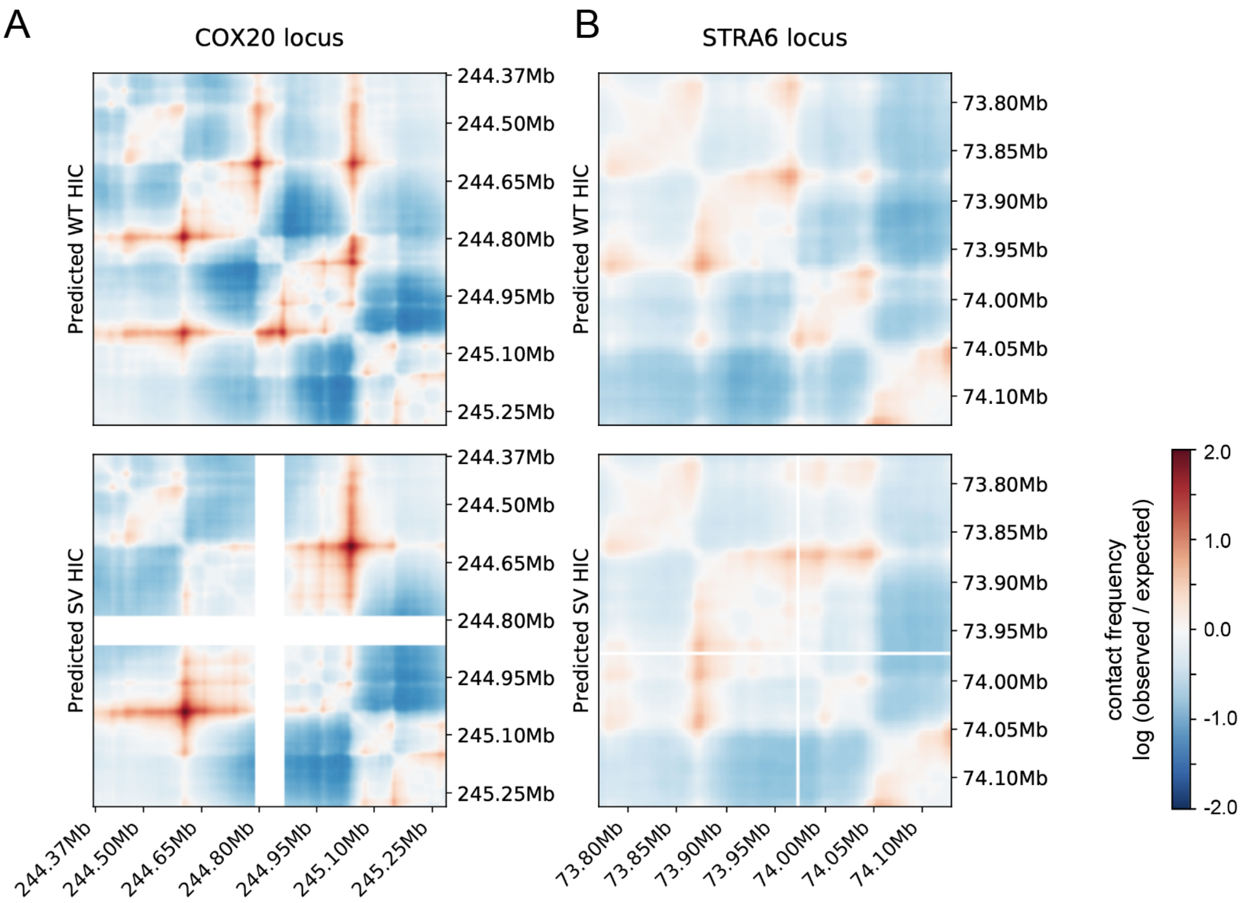


Figure 4.3: (A) Predicted Hi-C maps for reference sequence (upper panel) and patient deletion sequence (lower panel) at the *COX20* locus. Colors represent predicted contact frequency, normalized to the log of hypothetical observed contact over expected, with warm colors indicating higher-than-expected contact based on linear genomic distance and cool colors indicating lower-than-expected. **(B)** Predicted Hi-C maps for reference sequence (upper panel) and sequence with a small 68 base pair deletion near the *STRA6* locus.

In the *MESP* deletion region, Akita predicts a strong ectopic contact between the CHD risk gene *KIF7* and a large region that includes vesicle transport gene *AP3S2* and many ENCODE candidate cis-regulatory elements (Sheffield et al. 2013).

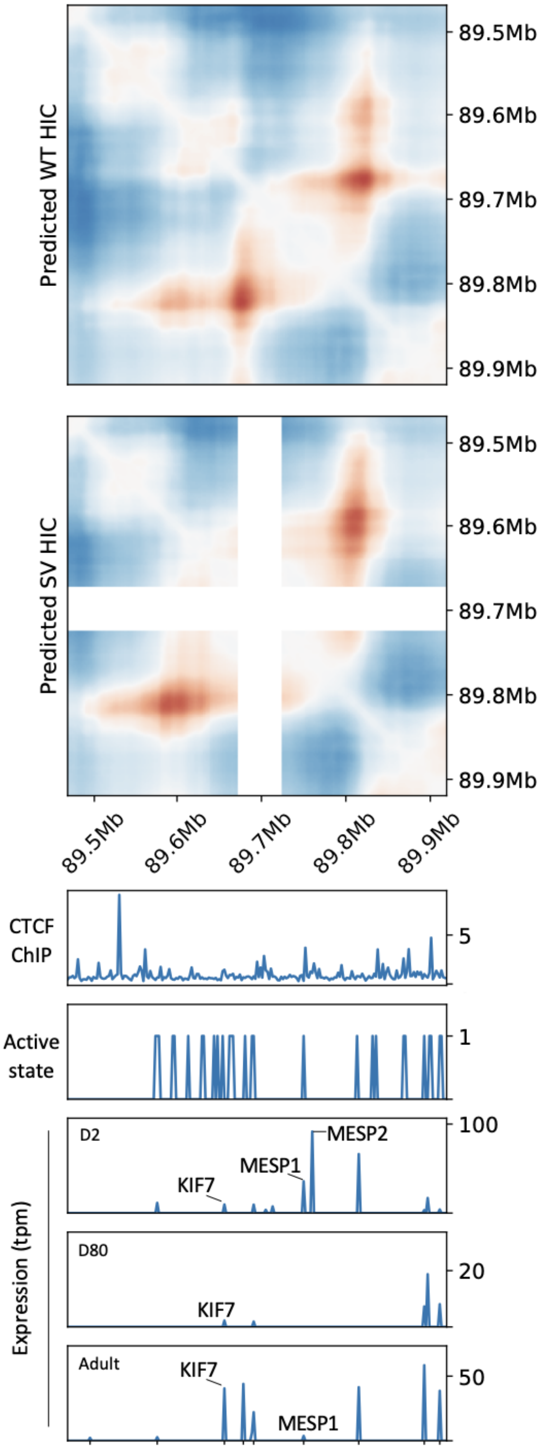


Figure 4.4: 58kb deletion encompassing the MESP locus. (A) Akita prediction of reference sequence (upper panel) and deletion sequence (lower panel). Colors represent predicted contact frequency, normalized to the log of hypothetical observed contact over expected, with warm colors indicating higher-than-expected contact frequency based on linear genomic distance and cool colors indicating lower-than-expected. Colormap legend from Figure 4.3 also applies here. **(B)** CTCF binding strength in this region, sourced from Roadmap Epigenomics ChIP-seq assays in fetal heart tissue (Roadmap Epigenomics Consortium et al. 2015). **(C)** Active state prediction from ChromHMM model from Roadmap fetal heart tissue data. **(D)** RNA-seq expression measured in transcripts per million (tpm), at day 2 (D2, upper panel) and day 80 (D80, middle panel) of cardiomyocyte differentiation (Y. Zhang et al. 2019) and in adult heart tissue (GTEx Consortium et al. 2017, lower panel).

4.3.3 Human iPSCs with patient deletions recapitulate model predictions

As a proof of concept to demonstrate the accuracy of these deep-learning predictions on genomic profiles, we sought to experimentally reproduce the Akita-predicted effect of variants on 3D genome folding. We selected the variant at the *MESP* locus for initial engineering and sequencing (**Figure 4.4**). Based on its crucial activity during day 2 (D2) of hESC differentiation into mesoderm and eventually cardiomyocytes (**Figure 4.4, D**), we sequenced Hi-C contact frequency of cells at this stage of cardiac differentiation.

To determine if the Akita-predicted changes to genome folding are experimentally reproducible, and also produce further evidence pertaining to the pathological relevance of this variant, cells from a hESC WTC-11 cell line were engineered to carry the patient deletion using CRISPR/Cas9, then differentiated as described previously into cardiac mesoderm (Kattman et al. 2011). We performed Capture-C sequencing on these cells to measure chromatin contact frequency around this locus. Capture-C results recapitulated Akita's predictions for both the reference and alternate allele (**Figure 4.5, A-B**). Notably, Akita's predictions in this region were experimentally reproducible despite not being formed in the hESC-derived cardiac mesoderm experimental cell line, consistent with some previous observations of TAD stability across cell types (Dixon et al. 2012; Rao et al. 2014; Dixon et al. 2015). Together, these data support the claim that Akita can accurately model the effects of genetic variants, and furthermore suggests that the variant might cause dysregulated expression of the known CHD gene *KIF7* via increased chromatin contact with any of various ENCODE regulatory elements. In addition to the loss of cardiac specification genes *MESP1* and *MESP2* caused by the deletion, the altered expression of *KIF7* could be a possible pathological mechanism by which this variant is causative in CHD.

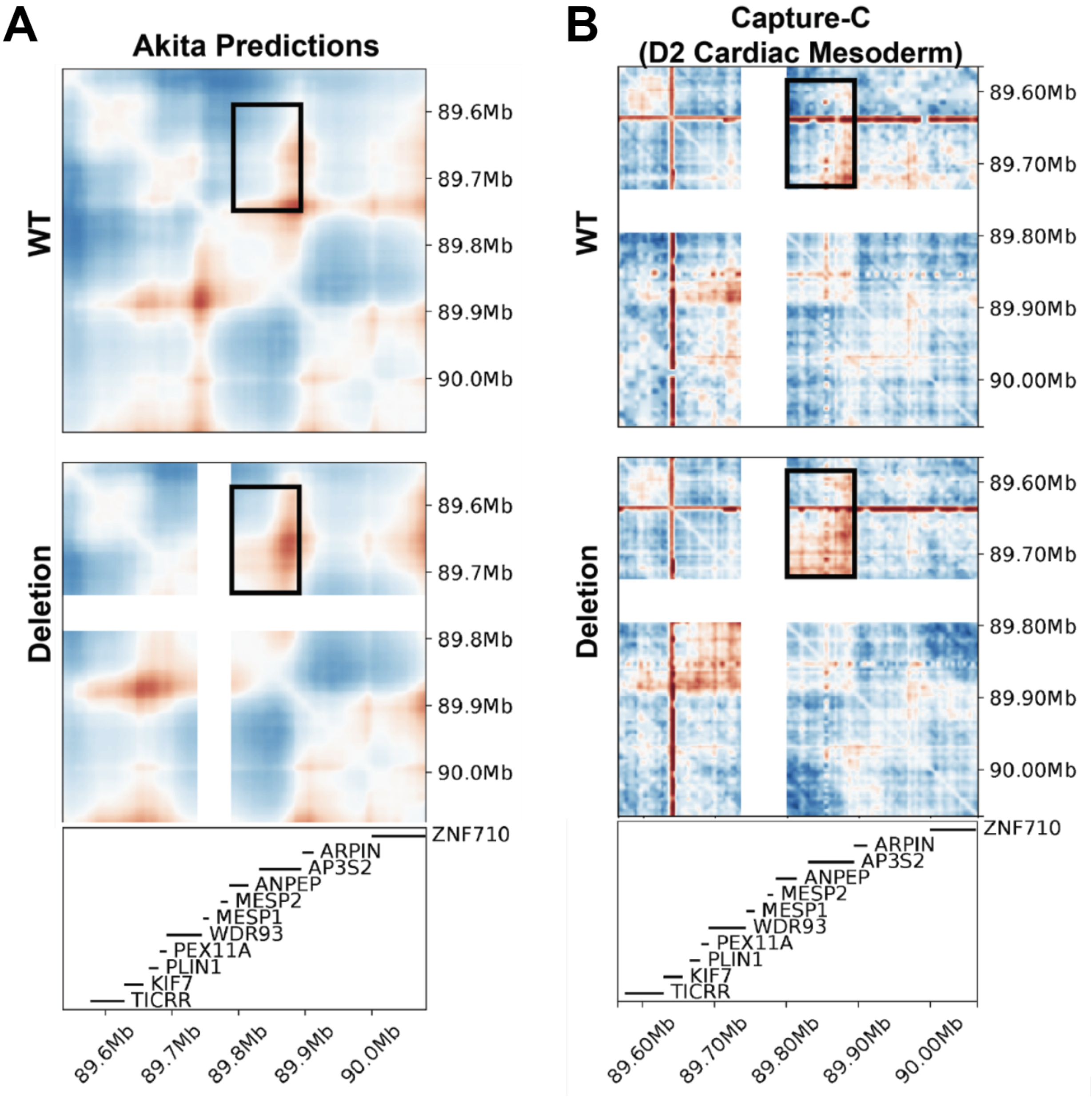


Figure 4.5: Deep learning genomic profile predictions are experimentally reproducible. Colors represent predicted contact frequency, normalized to the log of hypothetical observed contact over expected, with warm colors indicating higher-than-expected contact frequency based on linear genomic distance and cool colors indicating lower-than-expected. Colormap legend from Figure 4.3 also applies here. **(A)** Akita predictions for the reference genome (top) and with a 58kb deletion (bottom) in chromosome 15 near the MESP locus. **(B)** Capture-C results from isogenic hESC-derived cardiac mesoderm cells (2 days after differentiation) without deletion (top) and in a cell line with engineered patient deletion (bottom). White bars indicate the deleted region, also shown in WT for reference. Red bars represent masked repetitive elements. Black boxes highlight a region containing KIF7, AP3S2, and many ENCODE cCREs (not shown) within which there are gained interactions (A) predicted by Akita and (B) observed experimentally.

4.3.4 A common SV increases risk for CHD through disruption of genome folding

In mammalian hearts, the interventricular septum (IVS) separates deoxygenated pulmonary blood flow from oxygenated systemic blood flow within the cardiac ventricles. When the process of IVS formation is disrupted during embryonic development, infants are born with a ventricular septal defect (VSD), the most common heart defect found in children (Dakkak and Oliver 2022). A single monogenic explanation is rarely found in cases of VSD, meaning that the majority of these affected individuals do not have a hypothesized genetic cause in spite of the disease's considerable heritability (Jin et al. 2017).

In order to study how genetic variation might lead to phenotypic variability without relying solely on case/control cohorts of a rare disease, co-authors developed a simple standardized and automated measure of the IVS cross-sectional area derived from MRI imaging of the heart (Yu et al. 2021). Applied to 31,587 individuals with sequencing and MRI imaging data in the UK Biobank (Shah et al. 2020), Mendelian randomization analysis found that having a smaller IVS cross-section increases risk for VSD by approximately a factor of 2 per standard deviation (SD = 166 mm^2, Yu et al. 2021). Among other genomic insights, authors discovered that a common SV incorporating *KANSL1* is strongly associated with decreased IVS cross-sectional area at diastole, implying increased risk of VSD.

The structural variant, sometimes denoted as CPX_17_4670 (Collins et al. 2020) and located at chr17:45571611-46261810 (GRCh38.p13 assembly), is a duplication-flanked inversion across three genes, with a minor allele frequency (MAF) of 18% in European populations (4% in African populations, <0.1% in East Asian populations, Collins et al. 2020). I applied the Akita model to the 1 Mb region centered on this variant, predicting first the chromatin organization at the reference sequence (**Figure 4.6, A**). Akita predicts a strong interaction between the *KANSL1* promoter and regions including the *MAPT* gene body and *KANSL1-AS1* in the reference genotype (**Figure 4.6, A,** black arrows). In contrast, DNA sequence containing the common inversion is not predicted to form these interactions (**Figure 4.6, B,** white arrows).

Several candidate regulatory enhancers are predicted by the GeneHancer database in this region (Fishilevich et al. 2017). Furthermore, a stripe along the TSS of the long intergenic non-coding RNA *LINC02210*, representing high contact frequency between it and the locus containing the *KANSL1* TSS, *ARL17B, LRRC37A, KANSL1-AS1,* as well as the locus containing *MAPT, MAPT-AS1*, and *CRHR1*, is lost as a result of this SV (**Figure 4.6**, boxes). Given that enhancers and non-coding RNAs are known to play an important role in regulation of expression during differentiation, likely by their effects on 3D organization and often in *cis* *(Khosraviani, Ostrowski, and Mekhail 2019; Tachiwana, Yamamoto, and Saitoh 2020)*, disrupted contacts comprise a plausible mechanism for altered expression of these genes, possibly in a tissue-specific manner.

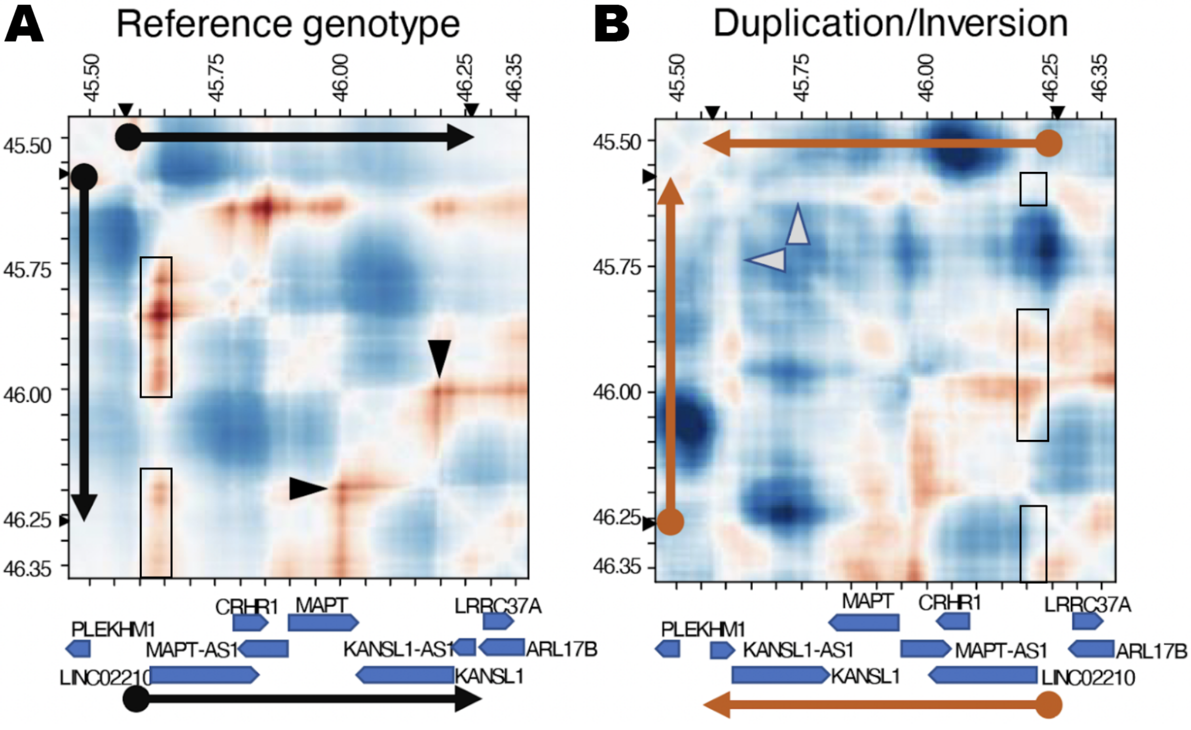


Figure 4.6: Two-dimensional plots of the Akita model of genome folding with GRCh38 coordinates on the X and Y axes, with relevant genes and strand orientation indicated below by blue boxes. Colors represent predicted contact frequency, normalized to the log of hypothetical observed contact over expected, with warm colors indicating higher-than-expected contact frequency based on linear genomic distance and cool colors indicating lower-than-expected. Colormap legend from Figure 4.3 also applies here. The reference genotype predicts a strong

3-dimensional chromatin interaction (black triangle) between the KANSL1 promoter and regions including the MAPT gene body and KANSL1-AS1. The common duplication/inversion 17:45571611-46261810 (orientation indicated by the orange arrow) is predicted to completely disrupt the interaction (corresponding locations within the inversion indicated by gray triangles).

A gene-based burden test for rare SVs described previously (Aguirre, Rivas, and Priest 2019) found that rare *KANSL1* duplications are strongly associated with decreased IVS size (Priest et al. 2016). *KANSL1* is a component of the WDR5-MLL1 histone modifying complex (Dias et al. 2014), so its association with developmental phenotypes is unsurprising. Haploinsufficiency (i.e. loss-of-function of just one copy) of *KANSL1* is causal for Koolen-de Vries syndrome, which includes a VSD phenotype among others (Koolen et al. 2016). We therefore speculate that improper expression of *KANSL1* in pathogenic spatio-temporal contexts leads to a smaller IVS and increased risk for VSD. Experimental characterization of these variants and their effects on chromatin organization and molecular phenotypes are needed to further elucidate the specifics of *KANSL1* regulation and its relationship to 3D genome organization during cardiogenesis.

4.4 DISCUSSION

Non-coding variation, particularly variants that alter chromatin organization near genes with complex regulatory logic, have the potential to be relevant in the context of congenital phenotypes and cancers (Moore 2009; Fernández-Medarde and Santos 2011; Lee et al. 2015; Bernhart et al. 2016). Here, I have expanded upon previous work in the lab establishing that SVs near gene regulatory features are associated with developmental phenotypes (Fudenberg and Pollard 2019), and extending that observation from developmental delay and autism to congenital heart defects. I further demonstrated use of the trained network model Akita (Fudenberg, Kelley, and Pollard 2020) to prioritize patient variants for experimental exploration. I found that Akita correctly predicts the consequences of large boundary-disrupting variants, validating its use for hypothesis generation and experimental design, as well as its application to the interpretation of a common human polymorphism.

Other contemporary studies dovetail with our findings: in a recent preprint, 5 of 8 experimentally-tested TAD boundary deletions caused increased embryonic lethality or other developmental phenotypes in mice. In particular, they found these phenotypes near known cardiac genes like Smad3/Smad6 and Tbx5 (Rajderkar et al. 2021). Combined with my findings on TAD boundary deletions in human SVs from CHD patients, we conclude that the 3D organization of the genome is a key aspect of the regulation of human heart formation. Ascertainment of variants predicted to cause its disruption during clinical screening could shed light on as-yet unexplained CHD cases. Tools like Akita will be key to finding such instances, and to help learn and interpret the cis-regulatory grammar underlying these phenomena.

There are caveats associated with using computational models to predict effects of SVs and SNVs. In particular, Akita was trained using Hi-C and micro-C data from five cell types, but was largely limited in its ability to predict cell type-specific differences (Fudenberg, Kelley, and Pollard 2020). Though most TAD boundaries are thought to be invariant across cell types, smaller-scale interactions within them can be cell type specific in a manner that is important to disease etiology, like enhancer-promoter loops (Smith et al. 2016; McArthur and Capra 2021). Future work in this space will incorporate cell type information like chromatin marks, accessibility, and RNA expression for the interpretation of disease variants.

While it was previously shown that controls are depleted for base pair deletion at highly-expressed genes and CTCF binding clusters, and DD recapitulates the null model with no depletion, Figures 4.1 and 4.2 suggest that the SVs in CHD patients are *enriched* for deleted base pairs in these regions, resembling the signature of SVs in cancer genomes. This was in surprising contrast to DD, which appeared to follow the null model rather than showing enrichment over it (Fudenberg and Pollard 2019). One obvious caveat is the possibility of processing differences across the datasets used. Additionally, some CHD deletion calls from the array data used here failed to replicate in whole exome sequencing data (2 failed of 7 examined). With these caveats stated, my observations are consistent with the 1.6 to 2-fold increased cancer risk observed in survivors of CHD at all ages (Lee et al. 2015; Gurvitz et al. 2016; Mandalenakis et al. 2019). Some specific risk genes have been identified already (Morton et al. 2021), but further investigation is required to understand the full suite of mechanisms underlying this relationship.

4.5 METHODS

4.5.1 Structural variant calls

Deletions from healthy controls, patients with developmental delay (DD), and tumors were processed as described in (Fudenberg and Pollard 2019). Briefly, variants from 11,256 controls and 29,083 DD patients were sourced from (Coe et al. 2014). Deletion calls from 14,908 tumors were obtained from COSMIC (Forbes et al. 2010, release v84), and liftOver (Hinrichs et al. 2006) was used to convert coordinates from hg38 to hg19.

Deletions from CHD patients and their parents were compiled from multiple sequencing methods and sources. SVs analyzed from CHD patients and parents in Figures 4.1 and 4.2 were sourced from 1536 individuals (512 trios) in the PCGC using array hybridization on one of the Illumina Omni-1M or the Illumina Omni-2.5M. Collaborators called CNVs for each subject using the hidden Markov model algorithm PennCNV (Wang et al. 2007) using custom parameters for population frequency of B-allele (PFB) and GC model. To maximize disease CNV discovery, CNVs with a minor allele frequency > 1% were removed. A more stringently-filtered subset of this data was published in (Glessner et al. 2014). CHD patient SVs were sourced from collaborators using various SV calling methods (unpublished data).

4.5.2 Coverage by feature strength estimates

This method was originally developed by (Fudenberg and Pollard 2019), updated with the latest versions of input data by me, with thanks to Geoff Fudenberg for files and code. TSS analyses were performed using GTEx v7 release (GTEx Consortium et al. 2017), where the strength of a TSS in GTEx was quantified as the sum of the gene’s expression across all tissues except testes. CTCF feature strength is defined as its aggregate binding across samples, where broad- and narrowPeak CTCF ChIP-seq files (generated across multiple cell types and processed as described previously in (Kellis et al. 2014)) were downloaded from ENCODE (ENCODE Project Consortium 2012) and aggregated.

Observed deletion coverage over expected deletion coverage is calculated as:

$$(\sum_{i \in k} N_i) \div (N_{total} \sum_{i \in k} \frac{S_i}{S_{total}}),$$

where i indexes genomic regions within a particular feature class k (quantile of TSS expression or CTCF binding strength), S_i is the size of region i, N_i is the number of base pairs deleted in region i, S_{total} is genome size, and N_{total} is the number of deleted base pairs genome-wide. The UCSC hg19 gap file was used to exclude regions that are more prone to variant artifacts (Hinrichs et al. 2006; Coe et al. 2014). Bootstrap estimates for deletion coverage as a function of feature strength were generated by sampling 1000 times from the full list of observed coverage-percentile strength pairs with replacement and computing averages in sliding windows of ±5 percentiles. The mean value over 1000 bootstraps is reported.

4.5.3 Prediction and scoring of candidate SVs

Akita (Fudenberg, Kelley, and Pollard 2020) was used to predict Hi-C chromatin contact maps, using the 9-14 model parameters (*Get_model.sh at Master · Calico/basenji* n.d.). For each SV, a 1Mb window of the genome was selected such that the midpoint of the SV is located at the central genomic bin of the prediction. The hg19 reference sequence was used as input into Akita to predict wild type chromatin contact frequency in the region, which was repeated for the same sequence with SV base pairs deleted. Because Akita makes predictions with a fixed input size, I removed the DNA sequence that was deleted in the patient and symmetrically extended the start and end of the 1Mb region to maintain input size.

To assess local disruption of DNA contacts, I used the sum of mean-squared error (MSE) across the full region, given by:

$$\sum_{i,j=1}^{n} (W_{i,j} - M_{i,j})^2,$$

where *W* is the Hi-C map predicted from “wild type” reference sequence, *M* is the map predicted from “mutant” SV sequence, and (*i,j*) represent coordinates of the genomic bins from 1 to *n*. In this model, input sequence is 2^{20} base pairs that have been convolved into 512 total bins, with 2048 base pairs represented per bin. We do not include the marginal 64 bins in scoring, since the model cannot include sequence information upstream/downstream from this region, reducing the accuracy of folding predictions in those bins.

4.5.4 Locus-aware prioritization - expression data and chromatin state

For locus-aware prioritization, I intersected the MSE of predicted Hi-C contact frequency at each genomic bin with expression of TSSs and chromatin state prediction. MSE scores are multiplied by a factor indicating TSS/chromatin activity in that bin, and are then summed across the full 1 Mb map for an expression- or activity-aware score per SV:

$$\sum_{i=1}^{n} a_i * d_i,$$

where a_i is the expression or activity state value at bin *i*, and d_i is the disruption or MSE between the two matrices at genomic bin *i*. TSSs and active chromatin state loci were extended ± 2 Akita bins (4096 base pairs upstream and downstream) in hopes of encapsulating nearby regulatory elements.

To calculate expression-aware disruption, I used the sum of gene expression (transcripts per million, tpm) in GTEx heart tissues: 'Artery - Aorta', 'Artery - Coronary', 'Heart - Atrial Appendage', and 'Heart - Left Ventricle.' The MSE at each genomic bin was multiplied by that sum of expression for any TSSs located within the bin or two bins up/downstream.

To calculate activity-aware disruption, I used chromatin state predictions from the chromHMM core 15-state model (Ernst and Kellis 2017) on fetal heart tissue data, as processed and published by the Roadmap Epigenomics Consortium (E083 Fetal Heart PrimaryHMM, Roadmap Epigenomics Consortium et al. 2015). Of the 15 possible chromatin states, I considered the following states to be active: 1_TssA, 2_TssAFlnk, 3_TxFlnk, 4_Tx, 5_TxWk, 6_EnhG, 7_Enh. Chromatin state predictions were further simplified to "active" (=1) and "not active" (=0) as their weighting factor, such that only the MSE of genomic bins predicted to be active in fetal heart tissue is included in this weighted score.

4.5.5 Engineering and sequencing of cell lines

Cells from a hESC WTC-11 cell line were engineered to carry the patient deletion using CRISPR/Cas9 (infinity biologiX protocol). Controls from the same cell line were subjected to identical reagents without genome editing, to produce an isogenic pair of cell lines. Basic quality control was performed to confirm successful editing, regular karyotyping, and pluripotency. The edited lines (deletion and control) were then differentiated as described previously into cardiac mesoderm (Kattman et al. 2011). Capture-C (Arima Genomics) was performed using probes designed to capture contacts ~500kb upstream and downstream of the SV region.

I processed and analyzed experimental reads using the 4DN processing pipeline and quality control cutoffs (Reiff et al. 2022). Specifically, I mapped the Capture-C reads to the GRCh38 reference genome using bwa (version 0.7.17), then filtered, sorted, and merged reads using pairtools (version 0.2.2). Cooler (version 0.8.3) was used to create Hi-C matrices at a resolution of 2048 base pairs per bin, and cooltools for visualization (Open2C et al. 2022).

4.6 DATA AND SOFTWARE ACCESS

Representative code will be posted as a series of jupyter notebooks, available at https://github.com/mepittman/sv-prediction.

REFERENCES

Aguirre, Matthew, Manuel A. Rivas, and James Priest. 2019. “Phenome-Wide Burden of Copy-Number Variation in the UK Biobank.” *American Journal of Human Genetics* 105 (2): 373–83.

Akhirome, Ehiole, Nephi A. Walton, Julie M. Nogee, and Patrick Y. Jay. 2017. “The Complex Genetic Basis of Congenital Heart Defects.” *Circulation Journal: Official Journal of the Japanese Circulation Society* 81 (5): 629–34.

Albert, Frank W., and Leonid Kruglyak. 2015. “The Role of Regulatory Variation in Complex Traits and Disease.” *Nature Reviews. Genetics* 16 (4): 197–212.

Balasubramanian, Meena. 2022. *HNRNPU-Related Neurodevelopmental Disorder*. University of Washington, Seattle.

Bernhart, Stephan H., Helene Kretzmer, Lesca M. Holdt, Frank Jühling, Ole Ammerpohl, Anke K. Bergmann, Bernd H. Northoff, et al. 2016. “Changes of Bivalent Chromatin Coincide with Increased Expression of Developmental Genes in Cancer.” *Scientific Reports* 6 (November): 37393.

Carullo, Nancy V. N., and Jeremy J. Day. 2019. “Genomic Enhancers in Brain Health and Disease.” *Genes* 10 (1). https://doi.org/10.3390/genes10010043.

Claringbould, Annique, and Judith B. Zaugg. 2021. “Enhancers in Disease: Molecular Basis and Emerging Treatment Strategies.” *Trends in Molecular Medicine* 27 (11): 1060–73.

Coe, Bradley P., Kali Witherspoon, Jill A. Rosenfeld, Bregje W. M. van Bon, Anneke T. Vulto-van Silfhout, Paolo Bosco, Kathryn L. Friend, et al. 2014. “Refining Analyses of Copy Number Variation Identifies Specific Genes Associated with Developmental Delay.” *Nature Genetics* 46 (10): 1063–71.

Collins, Ryan L., Harrison Brand, Konrad J. Karczewski, Xuefang Zhao, Jessica Alföldi, Laurent C. Francioli, Amit V. Khera, et al. 2020. “A Structural Variation Reference for Medical and

Population Genetics.” *Nature* 581 (7809): 444–51.

Dakkak, Wael, and Tony I. Oliver. 2022. *Ventricular Septal Defect*. StatPearls Publishing.

Degtyareva, Arina O., Elena V. Antontseva, and Tatiana I. Merkulova. 2021. “Regulatory SNPs: Altered Transcription Factor Binding Sites Implicated in Complex Traits and Diseases.” *International Journal of Molecular Sciences* 22 (12): 6454.

Dias, Jorge, Nhuong Van Nguyen, Plamen Georgiev, Aline Gaub, Janine Brettschneider, Stephen Cusack, Jan Kadlec, and Asifa Akhtar. 2014. “Structural Analysis of the KANSL1/WDR5/KANSL2 Complex Reveals That WDR5 Is Required for Efficient Assembly and Chromatin Targeting of the NSL Complex.” *Genes & Development* 28 (9): 929–42.

Dixon, Jesse R., Inkyung Jung, Siddarth Selvaraj, Yin Shen, Jessica E. Antosiewicz-Bourget, Ah Young Lee, Zhen Ye, et al. 2015. “Chromatin Architecture Reorganization during Stem Cell Differentiation.” *Nature* 518 (7539): 331–36.

Dixon, Jesse R., Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S. Liu, and Bing Ren. 2012. “Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions.” *Nature* 485 (7398): 376–80.

ENCODE Project Consortium. 2012. “An Integrated Encyclopedia of DNA Elements in the Human Genome.” *Nature* 489 (7414): 57–74.

Ernst, Jason, and Manolis Kellis. 2017. “Chromatin-State Discovery and Genome Annotation with ChromHMM.” *Nature Protocols* 12 (12): 2478–92.

Fernández-Medarde, Alberto, and Eugenio Santos. 2011. “Ras in Cancer and Developmental Diseases.” *Genes & Cancer* 2 (3): 344–58.

Feuk, Lars, Andrew R. Carson, and Stephen W. Scherer. 2006. “Structural Variation in the Human Genome.” *Nature Reviews. Genetics* 7 (2): 85–97.

Fishilevich, Simon, Ron Nudel, Noa Rappaport, Rotem Hadar, Inbar Plaschkes, Tsippi Iny Stein, Naomi Rosen, et al. 2017. “GeneHancer: Genome-Wide Integration of Enhancers and Target Genes in GeneCards.” *Database: The Journal of Biological Databases and Curation*

2017 (January). https://doi.org/10.1093/database/bax028.

Forbes, Simon A., Gurpreet Tang, Nidhi Bindal, Sally Bamford, Elisabeth Dawson, Charlotte Cole, Chai Yin Kok, et al. 2010. “COSMIC (the Catalogue of Somatic Mutations in Cancer): A Resource to Investigate Acquired Mutations in Human Cancer.” *Nucleic Acids Research* 38 (Database issue): D652–57.

Fudenberg, Geoff, David R. Kelley, and Katherine S. Pollard. 2020. “Predicting 3D Genome Folding from DNA Sequence with Akita.” *Nature Methods* 17 (11): 1111–17.

Fudenberg, Geoff, and Katherine S. Pollard. 2019. “Chromatin Features Constrain Structural Variation across Evolutionary Timescales.” *Proceedings of the National Academy of Sciences of the United States of America* 116 (6): 2175–80.

Get_model.sh at Master · Calico/basenji. n.d. Github. Accessed March 16, 2023. https://github.com/calico/basenji.

Glessner, Joseph T., Alexander G. Bick, Kaoru Ito, Jason Homsy, Laura Rodriguez-Murillo, Menachem Fromer, Erica Mazaika, et al. 2014. “Increased Frequency of de Novo Copy Number Variants in Congenital Heart Disease by Integrative Analysis of Single Nucleotide Polymorphism Array and Exome Sequence Data.” *Circulation Research* 115 (10): 884–96.

Gong, Yixiao, Charalampos Lazaris, Theodore Sakellaropoulos, Aurelie Lozano, Prabhanjan Kambadur, Panagiotis Ntziachristos, Iannis Aifantis, and Aristotelis Tsirigos. 2018. “Stratification of TAD Boundaries Reveals Preferential Insulation of Super-Enhancers by Strong Boundaries.” *Nature Communications* 9 (1): 542.

Gonzalez-Teran, Barbara, Maureen Pittman, Franco Felix, Reuben Thomas, Desmond Richmond-Buccola, Ruth Hüttenhain, Krishna Choudhary, et al. 2022. “Transcription Factor Protein Interactomes Reveal Genetic Determinants in Heart Disease.” *Cell* 185 (5): 794–814.e30.

GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, et al. 2017. “Genetic Effects on Gene Expression across Human Tissues.” *Nature* 550 (7675): 204–13.

Gunsalus, Laura M., Michael J. Keiser, and Katherine S. Pollard. 2022. “In Silico Discovery of Repetitive Elements as Key Sequence Determinants of 3D Genome Folding.” *bioRxiv*. https://doi.org/10.1101/2022.08.11.503410.

Gurvitz, Michelle, Raluca Ionescu-Ittu, Liming Guo, Mark J. Eisenberg, Michal Abrahamowicz, Louise Pilote, and Ariane J. Marelli. 2016. “Prevalence of Cancer in Adults With Congenital Heart Disease Compared With the General Population.” *The American Journal of Cardiology* 118 (11): 1742–50.

Hinrichs, A. S., D. Karolchik, R. Baertsch, G. P. Barber, G. Bejerano, H. Clawson, M. Diekhans, et al. 2006. “The UCSC Genome Browser Database: Update 2006.” *Nucleic Acids Research* 34 (Database issue): D590–98.

Hoang, Thanh T., Elizabeth Goldmuntz, Amy E. Roberts, Wendy K. Chung, Jennie K. Kline, John E. Deanfield, Alessandro Giardini, et al. 2018. “The Congenital Heart Disease Genetic Network Study: Cohort Description.” *PloS One* 13 (1): e0191319.

Homsy, Jason, Samir Zaidi, Yufeng Shen, James S. Ware, Kaitlin E. Samocha, Konrad J. Karczewski, Steven R. DePalma, et al. 2015. “De Novo Mutations in Congenital Heart Disease with Neurodevelopmental and Other Congenital Anomalies.” *Science* 350 (6265): 1262–66.

Jin, Sheng Chih, Jason Homsy, Samir Zaidi, Qiongshi Lu, Sarah Morton, Steven R. DePalma, Xue Zeng, et al. 2017. “Contribution of Rare Inherited and de Novo Variants in 2,871 Congenital Heart Disease Probands.” *Nature Genetics* 49 (11): 1593–1601.

Kaaij, Lucas J. T., Fabio Mohn, Robin H. van der Weide, Elzo de Wit, and Marc Bühler. 2019. “The ChAHP Complex Counteracts Chromatin Looping at CTCF Sites That Emerged from SINE Expansions in Mouse.” *Cell* 178 (6): 1437–51.e14.

Kattman, Steven J., Alec D. Witty, Mark Gagliardi, Nicole C. Dubois, Maryam Niapour, Akitsu

Hotta, James Ellis, and Gordon Keller. 2011. “Stage-Specific Optimization of Activin/nodal and BMP Signaling Promotes Cardiac Differentiation of Mouse and Human Pluripotent Stem Cell Lines.” *Cell Stem Cell* 8 (2): 228–40.

Kellis, Manolis, Barbara Wold, Michael P. Snyder, Bradley E. Bernstein, Anshul Kundaje, Georgi K. Marinov, Lucas D. Ward, et al. 2014. “Defining Functional DNA Elements in the Human Genome.” *Proceedings of the National Academy of Sciences of the United States of America* 111 (17): 6131–38.

Khosraviani, Negin, Lauren A. Ostrowski, and Karim Mekhail. 2019. “Roles for Non-Coding RNAs in Spatial Genome Organization.” *Frontiers in Cell and Developmental Biology* 7 (December): 336.

Klemm, A., P. Flicek, T. Manolio, and L. Hindorff. 2013. “The NHGRI GWAS Catalog, a Curated Resource of SNP-Trait Associations.” *Nucleic Acids*. https://academic.oup.com/nar/article-abstract/42/D1/D1001/1062755.

Koolen, David A., Rolph Pfundt, Katrin Linda, Gea Beunders, Hermine E. Veenstra-Knol, Jessie H. Conta, Ana Maria Fortuna, et al. 2016. “The Koolen-de Vries Syndrome: A Phenotypic Comparison of Patients with a 17q21.31 Microdeletion versus a KANSL1 Sequence Variant.” *European Journal of Human Genetics: EJHG* 24 (5): 652–59.

Kragesteen, Bjørt K., Malte Spielmann, Christina Paliou, Verena Heinrich, Robert Schöpflin, Andrea Esposito, Carlo Annunziatella, et al. 2018. “Dynamic 3D Chromatin Architecture Contributes to Enhancer Specificity and Limb Morphogenesis.” *Nature Genetics* 50 (10): 1463–73.

Krefting, Jan, Miguel A. Andrade-Navarro, and Jonas Ibn-Salem. 2018. “Evolutionary Stability of Topologically Associating Domains Is Associated with Conserved Gene Regulation.” *BMC Biology* 16 (1): 87.

Lee, Yu-Sheng, Yung-Tai Chen, Mei-Jy Jeng, Pei-Chen Tsao, Hsiu-Ju Yen, Pi-Chang Lee, Szu-Yuan Li, et al. 2015. “The Risk of Cancer in Patients with Congenital Heart Disease: A

Nationwide Population-Based Cohort Study in Taiwan.” *PloS One* 10 (2): e0116844.

Lupiáñez, Darío G., Katerina Kraft, Verena Heinrich, Peter Krawitz, Francesco Brancati, Eva Klopocki, Denise Horn, et al. 2015. “Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions.” *Cell* 161 (5): 1012–25.

Mandalenakis, Zacharias, Christina Karazisi, Kristofer Skoglund, Annika Rosengren, Georgios Lappas, Peter Eriksson, and Mikael Dellborg. 2019. “Risk of Cancer Among Children and Young Adults With Congenital Heart Disease Compared With Healthy Controls.” *JAMA Network Open* 2 (7): e196762–e196762.

Marino, Bradley S., Paul H. Lipkin, Jane W. Newburger, Georgina Peacock, Marsha Gerdes, J. William Gaynor, Kathleen A. Mussatto, et al. 2012. “Neurodevelopmental Outcomes in Children with Congenital Heart Disease: Evaluation and Management: A Scientific Statement from the American Heart Association.” *Circulation* 126 (9): 1143–72.

McArthur, Evonne, and John A. Capra. 2021. “Topologically Associating Domain Boundaries That Are Stable across Diverse Cell Types Are Evolutionarily Constrained and Enriched for Heritability.” *American Journal of Human Genetics* 108 (2): 269–83.

McArthur, Evonne, David C. Rinker, Erin N. Gilbertson, Geoff Fudenberg, Maureen Pittman, Kathleen Keough, Katherine S. Pollard, and John A. Capra. 2022. “Reconstructing the 3D Genome Organization of Neanderthals Reveals That Chromatin Folding Shaped Phenotypic and Sequence Divergence.” *bioRxiv*. https://doi.org/10.1101/2022.02.07.479462.

Moore, Sam W. 2009. “Developmental Genes and Cancer in Children.” *Pediatric Blood & Cancer* 52 (7): 755–60.

Morton, Sarah U., Akiko Shimamura, Peter E. Newburger, Alexander R. Opotowsky, Daniel Quiat, Alexandre C. Pereira, Sheng Chih Jin, et al. 2021. “Association of Damaging Variants in Genes With Increased Cancer Risk Among Patients With Congenital Heart Disease.” *JAMA Cardiology* 6 (4): 457–62.

Nora, Elphège P., Bryan R. Lajoie, Edda G. Schulz, Luca Giorgetti, Ikuhiro Okamoto, Nicolas Servant, Tristan Piolot, et al. 2012. “Spatial Partitioning of the Regulatory Landscape of the X-Inactivation Centre.” *Nature* 485 (7398): 381–85.

Open2C, Nezar Abdennur, Sameer Abraham, Geoffrey Fudenberg, Ilya M. Flyamer, Aleksandra A. Galitsyna, Anton Goloborodko, Maxim Imakaev, Betul A. Oksuz, and Sergey V. Venev. 2022. “Cooltools: Enabling High-Resolution Hi-C Analysis in Python.” *bioRxiv*. https://doi.org/10.1101/2022.10.31.514564.

Piché, Jessica, Patrick Piet Van Vliet, Michel Pucéat, and Gregor Andelfinger. 2019. “The Expanding Phenotypes of Cohesinopathies: One Ring to Rule Them All!” *Cell Cycle* 18 (21): 2828–48.

Priest, James R., Kazutoyo Osoegawa, Nebil Mohammed, Vivek Nanda, Ramendra Kundu, Kathleen Schultz, Edward J. Lammer, et al. 2016. “De Novo and Rare Variants at Multiple Loci Support the Oligogenic Origins of Atrioventricular Septal Heart Defects.” *PLoS Genetics* 12 (4): e1005963.

Rajderkar, Sudha, Iros Barozzi, Yiwen Zhu, Rong Hu, Yanxiao Zhang, Bin Li, Yoko Fukuda-Yuzawa, et al. 2021. “Topologically Associating Domain Boundaries Are Commonly Required for Normal Genome Function.” *bioRxiv*. https://doi.org/10.1101/2021.05.06.443037.

Rao, Suhas S. P., Miriam H. Huntley, Neva C. Durand, Elena K. Stamenova, Ivan D. Bochkov, James T. Robinson, Adrian L. Sanborn, et al. 2014. “A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping.” *Cell* 159 (7): 1665–80.

Reiff, Sarah B., Andrew J. Schroeder, Koray Kırlı, Andrea Cosolo, Clara Bakker, Luisa Mercado, Soohyun Lee, et al. 2022. “The 4D Nucleome Data Portal as a Resource for Searching and Visualizing Curated Nucleomics Data.” *Nature Communications* 13 (1): 1–11.

Richter, Felix, Sarah U. Morton, Seong Won Kim, Alexander Kitaygorodsky, Lauren K. Wasson, Kathleen M. Chen, Jian Zhou, et al. 2020. “Genomic Analyses Implicate Noncoding de

Novo Variants in Congenital Heart Disease." *Nature Genetics* 52 (8): 769–77.

Riethoven, Jean-Jack M. 2010. "Regulatory Regions in DNA: Promoters, Enhancers, Silencers, and Insulators." *Methods in Molecular Biology* 674: 33–42.

Roadmap Epigenomics Consortium, Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, et al. 2015. "Integrative Analysis of 111 Reference Human Epigenomes." *Nature* 518 (7539): 317–30.

Rollins, Caitlin K., and Jane W. Newburger. 2014. "Neurodevelopmental Outcomes in Congenital Heart Disease." *Circulation* 130 (14): e124–26.

Shah, Sonia, Albert Henry, Carolina Roselli, Honghuang Lin, Garðar Sveinbjörnsson, Ghazaleh Fatemifar, Åsa K. Hedman, et al. 2020. "Genome-Wide Association and Mendelian Randomisation Analysis Provide Insights into the Pathogenesis of Heart Failure." *Nature Communications* 11 (1): 163.

Shanta, Omar, Amina Noor, Human Genome Structural Variation Consortium (HGSVC), and Jonathan Sebat. 2020. "The Effects of Common Structural Variants on 3D Chromatin Structure." *BMC Genomics* 21 (1): 95.

Sheffield, Nathan C., Robert E. Thurman, Lingyun Song, Alexias Safi, John A. Stamatoyannopoulos, Boris Lenhard, Gregory E. Crawford, and Terrence S. Furey. 2013. "Patterns of Regulatory Activity across Diverse Human Cell Types Predict Tissue Identity, Transcription Factor Binding, and Long-Range Interactions." *Genome Research* 23 (5): 777–88.

Sifrim, Alejandro, Marc-Phillip Hitz, Anna Wilsdon, Jeroen Breckpot, Saeed H. Al Turki, Bernard Thienpont, Jeremy McRae, et al. 2016. "Distinct Genetic Architectures for Syndromic and Nonsyndromic Congenital Heart Defects Identified by Exome Sequencing." *Nature Genetics* 48 (9): 1060–65.

Smith, Emily M., Bryan R. Lajoie, Gaurav Jain, and Job Dekker. 2016. "Invariant TAD Boundaries Constrain Cell-Type-Specific Looping Interactions between Promoters and

Distal Elements around the CFTR Locus." *American Journal of Human Genetics* 98 (1): 185–201.

Tachiwana, Hiroaki, Tatsuro Yamamoto, and Noriko Saitoh. 2020. "Gene Regulation by Non-Coding RNAs in the 3D G

enome Architecture." *Current Opinion in Genetics & Development* 61 (April): 69–74.

Wang, Kai, Mingyao Li, Dexter Hadley, Rui Liu, Joseph Glessner, Struan F. A. Grant, Hakon Hakonarson, and Maja Bucan. 2007. "PennCNV: An Integrated Hidden Markov Model Designed for High-Resolution Copy Number Variation Detection in Whole-Genome SNP Genotyping Data." *Genome Research* 17 (11): 1665–74.

Won, Hyejung, Luis de la Torre-Ubieta, Jason L. Stein, Neelroop N. Parikshak, Jerry Huang, Carli K. Opland, Michael J. Gandal, et al. 2016. "Chromosome Conformation Elucidates Regulatory Relationships in Developing Human Brain." *Nature* 538 (7626): 523–27.

Yuan, Xuefei, Ian C. Scott, and Michael D. Wilson. 2021. "Heart Enhancers: Development and Disease Control at a Distance." *Frontiers in Genetics* 12 (March): 642975.

Yu, Mengyao, Andrew R. Harper, Matthew Aguirre, Maureen Pittman, Catherine Tcheandjieu, Dulguun Amgalan, Christopher Grace, et al. 2021. "Genetic Determinants of Interventricular Septal Anatomy and the Risk of Ventricular Septal Defects and Hypertrophic Cardiomyopathy." *medRxiv*. https://doi.org/10.1101/2021.04.19.21255650.

Zhang, Feng, and James R. Lupski. 2015. "Non-Coding Genetic Variants in Human Disease." *Human Molecular Genetics* 24 (R1): R102–10.

Zhang, Yanxiao, Ting Li, Sebastian Preissl, Maria Luisa Amaral, Jonathan D. Grinstein, Elie N. Farah, Eugin Destici, et al. 2019. "Transcriptionally Active HERV-H Retrotransposons Demarcate Topologically Associating Domains in Human Pluripotent Stem Cells." *Nature Genetics* 51 (9): 1380–88.

CHAPTER 5: CONCLUDING REMARKS

The genome enforces spatio-temporal regulatory control of gene expression in a complex interplay between multiple interlocking factors. In *cis* this is accomplished via histone modification (Allfrey, Faulkner, and Mirsky 1964; Bannister and Kouzarides 2011; Kimura 2013), the binding of structural proteins like CTCF (Rao et al. 2014; Furlong and Levine 2018), and physical interaction between enhancers and promoters mediated by TF binding (Banerji, Olson, and Schaffner 1983; Palstra and Grosveld 2012). In *trans,* TF networks directly activate the expression of key gene programs by binding to their specific motif sequences at gene promoters or enhancers (ENCODE Project Consortium 2012). TFs also indirectly regulate genome conformation via the expression of histone modifying genes (Janknecht, Wells, and Hunter 1998; Chan and La Thangue 2001; Xu et al. 2021; Laubscher et al. 2021) and additional regulatory elements like lncRNAs (Tachiwana, Yamamoto, and Saitoh 2020). My doctoral dissertation has focused on the creation and application of new computational tools to ask questions about how this system falls apart in congenital disease, and in turn what those details can reveal about human development.

5.1 Summary of findings

In chapter two of this thesis, I integrated patient variant data with proteomics to yield new insights into the genetic underpinnings of CHD. I helped define the list of interacting proteins for two essential cTFs, GATA4 and TBX5, and by comparing their interactomes in cardiac progenitor cells with those derived from HEK-293 cells, further demonstrated that a substantial number of disease-relevant interactions can only be detected in cell type specific and endogenous contexts. By showing that the interactomes were enriched for CHD patient *de novo* mutations compared to controls, even when considering only genes that hadn’t previously been validated, I have verified the use of proteomics strategies to nominate and prioritize cryptic risk genes. My consolidative computational framework identified numerous candidate variants, including a CHD patient's *de novo* missense mutation in the novel risk gene GLYR1 that disrupts its interaction with GATA4 and thus downstream co-activation of cardiac developmental genes. Collectively, these findings indicate that the use of tissue- and disease-specific PPIs may partially overcome the genetic heterogeneity of CHDs and help prioritize the potential impact of variants in disease.

Having shown the utility of expanding known interactions of protein products for novel variant and risk gene discovery, I developed a similarly-motivated approach from a different angle: testing gene sets for statistical evidence of interaction in CHD genomes. In that vein I describe the algorithm GCOD in the third chapter of this dissertation. This trio-based probabilistic model identified previously-known CHD genes, recapitulated known protein-complex relationships while proposing additional cardiac-relevant complex interactors (e.g. CREBBP-EP300 with COL6A3), identified a novel gene interaction between GATA6 and POR, and prioritized 202 gene pairs likely to interact in CHD. My findings highlight the probable role of rare variant combinations in driving CHD, and establishes a new pipeline to maximize the utility of parental sequencing where available to discover these interactions.

Finally, I determined that SVs near gene regulatory features are associated with developmental phenotypes, including CHD. I further demonstrated use of the trained network model Akita (Fudenberg, Kelley, and Pollard 2020) to prioritize patient variants for experimental exploration, and validated its accuracy in large TAD boundary deletions. Applying Akita to a common structural variant associated with disease, I proposed probable and testable hypotheses regarding the regulatory consequences of this genetic inversion.

My work has demonstrated the effectiveness of using new statistical and computational techniques to interrogate the molecular causes of congenital heart defects, particularly in contexts that elucidate and leverage gene, protein, and DNA interactions.

5.2 Limitations

In each section of this dissertation, I explored a different suite of methods appropriate to different inheritance patterns and genomic locations: monogenic exome variants, oligogenic exome variants, and non-coding structural variants. Any one of these methods might provide part of the picture, but the most accurate interpretation of any individual CHD case will require an integrative approach to understand which factors drive and contribute to disease. Caveats specific to each of the approaches are below.

The prioritization scheme discussed in Chapter 2 to relatively rank missense CHD variants is specific to our dataset and diagnostic question, and furthermore was designed as a complementary method to the experimental discovery of endogenous TF interactomes. Although the principles could be widely applicable to other genetic diseases, context-specific modifications to datasets used and assumptions made would be necessary. Four of the variants we tested by affinity assay were ranked consistently with my prioritization score, which is encouraging but insufficient to conclude that the relative ranks are directly translatable to pathogenicity.

The trio-based method to discover gene interactions discussed in Chapter 3 assumes that parents are unaffected, and we also had to assume that the controls used here (pseudo-siblings) would be unaffected, despite having no phenotypic data for them. Since it is a probabilistic model, GCOD is necessarily more sensitive to observations that include DNVs, which are much less likely to occur compared to the 50% probability of inherited variant transmission. Furthermore, this method can be computationally intensive for large variant datasets and requires sequenced parent genomes, limiting its applicability to many datasets. A future iteration of GCOD will benefit from the inclusion of non-coding predictions and known protein interactions.

The two aforementioned studies focus on very rare variants, which are by definition depleted from the population. With collaborators, we were able to experimentally confirm their molecular or *in vivo* effects, but future experimental studies and/or extremely large cohorts will be needed to determine whether the remaining variants identified by these methods truly contribute to disease. Furthermore, these methods are limited to exonic regions; I strongly recommend integration with pipelines that account for the contribution of common and non-coding variants when forming hypotheses about the genetic cause of a case of CHD.

My work in non-coding regions discovered a potential role for the rearrangement of 3D genome organization in CHD, and further provided predictions and interpretations of the likely underlying regulatory cause and effect. However, the predictive model Akita is limited in its ability to predict cell type-specific differences and smaller-scale interactions like enhancer-promoter loops (Smith et al. 2016; McArthur and Capra 2021). Models incorporating higher-resolution sequencing and additional data types are needed (as they become available) to realize the full potential of this approach.

5.3 Implications and future work

Developmental phenotypes affecting the heart have profound impacts on the lives of patients and their families, and understanding individual genetic pathologies can suggest additional interventions and monitoring as patients age to adulthood. Using patient data to statistically identify contributing genetic and molecular factors will improve our ability to predict and treat disease. Overall, my work has identified interactors of TFs and other genes essential for cardiac development, as well as regulatory consequences by 3D chromatin folding, revealing biology of gene regulation related to cardiac disease.

Given our conclusion that the discovered interaction networks of endogenous proteins provide an abundance of disease-relevant information, the fact that most available protein interactomes were reconstructed in non-physiological settings and cell types (Köhler et al. 2008; Greene et al. 2015; Priest et al. 2016; Bryois et al. 2020; Izarzugaza et al. 2020) suggests great opportunity to apply our framework to other diseases to highlight disease mechanisms and provide a powerful filter for interrogating the genetic basis of disease. We also established GLYR1 as a GATA4 interactor in CPs, revealing their interaction as a candidate mechanism for GLYR1's localization to a specific subset of heart development genes during CM differentiation. Further investigation is required to determine whether this variant is sufficient to cause disease in humans, or if patient genetic background plays a larger role.

My prioritization of patient variants using GATA4 and TBX5 interactomes amounts to a list of genetic lesions predicted to impact heart development with high confidence. We enclosed that information with our original publication in the hopes that researchers with the resources to do so will experimentally interrogate their mechanism and effects *in vitro* and *in vivo*. In that vein, a collaborator at our institution is preparing a mouse model for the next-score highest variant, in the chromatin modifier *SMARCC1*. Similarly, we hope that the prioritized list of gene and variant interactions found by GCOD will help hone the hypotheses and experiments of researchers studying these genes and their mechanisms in disease, as well as inform clinical sequencing interpretation.

My work additionally implies that Akita is a useful tool to predict and prioritize structural variation altering 3D genome organization, which increasingly seems a plausible contributor in developmental phenotypes including CHD. For example, a recent preprint found that 5 of 8 experimentally-tested TAD boundary deletions caused increased embryonic lethality or other developmental phenotypes in mice (Rajderkar et al. 2021). Ascertainment of variants predicted to cause the disruption of genome folding during clinical screening could shed light on as-yet unexplained CHD cases. Tools like Akita will be key to finding such instances, and to help learn and interpret the cis-regulatory grammar underlying these phenomena. Future work will incorporate additional information to maximize cell-type specificity, as well as benchmark methods to prioritize types of contact disruption at scale.

Combining these findings and methods in the laboratory will continue to enrich our understanding of heart development and its molecular underpinnings. In the clinic, as sequencing becomes ever more accessible and informative, they will improve our ability to explain individual cases of disease and make recommendations about treatment.

REFERENCES

Allfrey, V. G., R. Faulkner, and A. E. Mirsky. 1964. “ACETYLATION AND METHYLATION OF HISTONES AND THEIR POSSIBLE ROLE IN THE REGULATION OF RNA SYNTHESIS*.” *Proceedings of the National Academy of Sciences* 51 (5): 786–94.

Banerji, J., L. Olson, and W. Schaffner. 1983. “A Lymphocyte-Specific Cellular Enhancer Is Located Downstream of the Joining Region in Immunoglobulin Heavy Chain Genes.” *Cell* 33 (3): 729–40.

Bannister, Andrew J., and Tony Kouzarides. 2011. “Regulation of Chromatin by Histone Modifications.” *Cell Research* 21 (3): 381–95.

Bryois, Julien, Nathan G. Skene, Thomas Folkmann Hansen, Lisette J. A. Kogelman, Hunna J. Watson, Zijing Liu, Eating Disorders Working Group of the Psychiatric Genomics Consortium, et al. 2020. “Genetic Identification of Cell Types Underlying Brain Complex Traits Yields Insights into the Etiology of Parkinson’s Disease.” *Nature Genetics* 52 (5): 482–93.

Chan, H. M., and N. B. La Thangue. 2001. “p300/CBP Proteins: HATs for Transcriptional Bridges and Scaffolds.” *Journal of Cell Science* 114 (Pt 13): 2363–73.

ENCODE Project Consortium. 2012. “An Integrated Encyclopedia of DNA Elements in the Human Genome.” *Nature* 489 (7414): 57–74.

Fudenberg, Geoff, David R. Kelley, and Katherine S. Pollard. 2020. “Predicting 3D Genome Folding from DNA Sequence with Akita.” *Nature Methods* 17 (11): 1111–17.

Furlong, Eileen E. M., and Michael Levine. 2018. “Developmental Enhancers and Chromosome Topology.” *Science* 361 (6409): 1341–45.

Greene, Casey S., Arjun Krishnan, Aaron K. Wong, Emanuela Ricciotti, Rene A. Zelaya, Daniel S. Himmelstein, Ran Zhang, et al. 2015. “Understanding Multicellular Function and Disease with Human Tissue-Specific Networks.” *Nature Genetics* 47 (6): 569–76.

Izarzugaza, Jose M. G., Sabrina G. Ellesøe, Canan Doganli, Natasja Spring Ehlers, Marlene D. Dalgaard, Enrique Audain, Gregor Dombrowsky, et al. 2020. “Systems Genetics Analysis Identifies Calcium-Signaling Defects as Novel Cause of Congenital Heart Disease.” *Genome Medicine* 12 (1): 76.

Janknecht, R., N. J. Wells, and T. Hunter. 1998. “TGF-Beta-Stimulated Cooperation of Smad Proteins with the Coactivators CBP/p300.” *Genes & Development* 12 (14): 2114–19.

Kimura, Hiroshi. 2013. “Histone Modifications for Human Epigenome Analysis.” *Journal of Human Genetics* 58 (7): 439–45.

Köhler, Sebastian, Sebastian Bauer, Denise Horn, and Peter N. Robinson. 2008. “Walking the Interactome for Prioritization of Candidate Disease Genes.” *American Journal of Human Genetics* 82 (4): 949–58.

Laubscher, Dominik, Berkley E. Gryder, Benjamin D. Sunkel, Thorkell Andresson, Marco Wachtel, Sudipto Das, Bernd Roschitzki, et al. 2021. “BAF Complexes Drive Proliferation and Block Myogenic Differentiation in Fusion-Positive Rhabdomyosarcoma.” *Nature Communications* 12 (1): 6924.

McArthur, Evonne, and John A. Capra. 2021. “Topologically Associating Domain Boundaries That Are Stable across Diverse Cell Types Are Evolutionarily Constrained and Enriched for Heritability.” *American Journal of Human Genetics* 108 (2): 269–83.

Palstra, Robert-Jan, and Frank Grosveld. 2012. “Transcription Factor Binding at Enhancers: Shaping a Genomic Regulatory Landscape in Flux.” *Frontiers in Genetics* 3 (September): 195.

Priest, James R., Kazutoyo Osoegawa, Nebil Mohammed, Vivek Nanda, Ramendra Kundu, Kathleen Schultz, Edward J. Lammer, et al. 2016. “De Novo and Rare Variants at Multiple Loci Support the Oligogenic Origins of Atrioventricular Septal Heart Defects.” *PLoS Genetics* 12 (4): e1005963.

Rajderkar, Sudha, Iros Barozzi, Yiwen Zhu, Rong Hu, Yanxiao Zhang, Bin Li, Yoko

Fukuda-Yuzawa, et al. 2021. “Topologically Associating Domain Boundaries Are Commonly Required for Normal Genome Function.” *bioRxiv*. https://doi.org/10.1101/2021.05.06.443037.

Rao, Suhas S. P., Miriam H. Huntley, Neva C. Durand, Elena K. Stamenova, Ivan D. Bochkov, James T. Robinson, Adrian L. Sanborn, et al. 2014. “A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping.” *Cell* 159 (7): 1665–80.

Smith, Emily M., Bryan R. Lajoie, Gaurav Jain, and Job Dekker. 2016. “Invariant TAD Boundaries Constrain Cell-Type-Specific Looping Interactions between Promoters and Distal Elements around the CFTR Locus.” *American Journal of Human Genetics* 98 (1): 185–201.

Tachiwana, Hiroaki, Tatsuro Yamamoto, and Noriko Saitoh. 2020. “Gene Regulation by Non-Coding RNAs in the 3D Genome Architecture.” *Current Opinion in Genetics & Development* 61 (April): 69–74.

Xu, Mingjie, Jie Yao, Yingchao Shi, Huijuan Yi, Wukui Zhao, Xinhua Lin, and Zhongzhou Yang. 2021. “The SRCAP Chromatin Remodeling Complex Promotes Oxidative Metabolism during Prenatal Heart Development.” *Development* 148 (8). https://doi.org/10.1242/dev.199026.

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:
Maureen Pittman
7CFEC616571A4B0... Author Signature

3/22/2023
Date