



Published in final edited form as:

Cell. 2022 March 03; 185(5): 794–814.e30. doi:10.1016/j.cell.2022.01.021.

## Transcription Factor Protein Interactomes Reveal Genetic Determinants in Heart Disease

Barbara Gonzalez-Teran<sup>1,2</sup>, Maureen Pittman<sup>1</sup>, Franco Felix<sup>1,2</sup>, Reuben Thomas<sup>1</sup>, Desmond Richmond-Buccola<sup>1,2</sup>, Ruth Hüttenhain<sup>1,3,4</sup>, Krishna Choudhary<sup>1</sup>, Elisabetta Moroni<sup>5</sup>, Mauro W. Costa<sup>1,2</sup>, Yu Huang<sup>1,2</sup>, Arun Padmanabhan<sup>1,2,6</sup>, Michael Alexanian<sup>1,2</sup>, Clara Youngna Lee<sup>1,2</sup>, Bonnie E. J. Maven<sup>1,2,7</sup>, Kaitlen Samse-Knapp<sup>1,2</sup>, Sarah U. Morton<sup>8</sup>, Michael McGregor<sup>1,3,4</sup>, Casey A. Gifford<sup>1,2</sup>, J. G. Seidman<sup>9</sup>, Christine E. Seidman<sup>9,10,11</sup>, Bruce D. Gelb<sup>12,13,14</sup>, Giorgio Colombo<sup>15</sup>, Bruce R. Conklin<sup>1</sup>, Brian L. Black<sup>16</sup>, Benoit G. Bruneau<sup>1,16,17,18</sup>, Nevan J. Krogan<sup>1,3,4</sup>, Katherine S. Pollard<sup>1,18,19,\*</sup>, Deepak Srivastava<sup>1,2,16,17,20,21,\*</sup>

<sup>1</sup>Gladstone Institutes, San Francisco, CA, USA.

<sup>2</sup>Roddenberry Center for Stem Cell Biology and Medicine at Gladstone, San Francisco, CA, USA.

<sup>3</sup>Department of Cellular and Molecular Pharmacology, University of California San Francisco, San Francisco, CA, USA.

\*Address correspondence to: Deepak Srivastava and Katherine S. Pollard, Gladstone Institutes, 1650 Owens St, San Francisco, CA 94158, deepak.srivastava@gladstone.ucsf.edu, katherine.pollard@gladstone.ucsf.edu.

### AUTHOR CONTRIBUTIONS

B.G.T., K.S.P., and D.S. conceived and directed the study with input from B.G.B., B.R.C., B.L.B., and N.J.K. as part of an NIH/NHLBI-sponsored Program Project Grant and B.D.G., S.U.M., C.E.S., J.G.S. as part of the NIH/NHLBI-sponsored Bench to Bassinet Pediatric Cardiac Genomics Consortium. D.R.B. and B.G.T. generated the *GATA4-KO* hiPSC line and B. Cole and B.G.T. generated the *TBX5-KO* hiPSC line. D.R.B. and B.G.T. performed WT, *GATA4* and *TBX5* knockout CP and CM differentiations and characterization. B.C., R.H. and B.G.T. defined the appropriate affinity purification strategy for cTFs. B.G.T. performed *GATA4* and *TBX5* (GT) affinity purification and sample preparation for mass spectrometry and GT-PPI classification. M.M. prepared the AP-MS buffers and performed the desalting and lyophilization of the AP-MS samples. M.P. and K.S.P. performed the AP-MS statistical analysis and M.P., K.S.P. and B.G.T. the PPI filtering. S.U.M. and M.P. performed variant processing; M.P. and K.S.P. designed and applied GT-PPI variant enrichment analysis. M.P. and B.G.T. performed interactome gene features analysis and developed the integrative pathogenicity scoring method. M.P., R.T., B.G.T., and K.S.P. analyzed the benefit of variant filtering with the GT-PPI. C.G. analyzed the expression of GT-PPI genes with CHD variants mouse heart development cell-types. F.F. performed the *GLYR1* protein sequence alignment, *GLYR1* protein structure modeling and recombinant DNA cloning. E.M. and G.C. performed the molecular dynamics *GLYR1* simulations. F.F. and B.G.T. performed *GATA4* and *GLYR1* silencing in CPs and luciferase reporter assays. K.S. generated the cTFs ChIPseq data. B.G.T. and M.A. performed *GLYR1* and H3K36me3 ChIPseq. B.G.T. and F.F. generated the *GLYR1*<sup>P496L</sup> hiPSC line, *GLYR1*<sup>P496L</sup> cardiomyocyte differentiations, scRNAseq, *GLYR1* ChIPseq and beating analysis. R.T. and K.C. performed RNAseq and ChIPseq computational analysis. R.T. and B.G.T. performed the scRNAseq analysis. A.P. designed and generated the *Glyr1*<sup>P495L</sup> mouse line. C.Y.L. assisted with mouse colony maintenance and experimentation. Y.H. performed echocardiography. M.W.C. and B.G.T. performed *Glyr1*<sup>P495L</sup> mouse studies.

### DECLARATION OF INTERESTS

D.S. is scientific co-founder, shareholder and director of Tenaya Therapeutics. B.G.B. and B.R.C. are scientific co-founders and shareholders of Tenaya Therapeutics. K.S.P. and N.K. are shareholders of Tenaya Therapeutics. N.J.K. has received research support from Vir Biotechnology and F. Hoffmann-La Roche. N.J.K. has consulting agreements with the Icahn School of Medicine at Mount Sinai, New York, Maze Therapeutics and Interline Therapeutics, is a shareholder of Tenaya Therapeutics and has received stocks from Interline Therapeutics.

### INCLUSION AND DIVERSITY

One or more of the authors of this paper self-identifies as a member of the LGBTQ+ community. One or more of the authors of this paper self-identifies as an underrepresented ethnic minority in science. While citing references scientifically relevant for this work, we also actively worked to promote gender balance in our reference list.

**Publisher's Disclaimer:** This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

<sup>4</sup>Quantitative Biosciences Institute (QBI), University of California San Francisco, San Francisco, CA, USA.

<sup>5</sup>SCITEC-CNR, Milano, Italy.

<sup>6</sup>Division of Cardiology, Department of Medicine, University of California, San Francisco, CA, USA.

<sup>7</sup>Developmental and Stem Cell Biology PhD Program, University of California, San Francisco, CA, USA.

<sup>8</sup>Division of Newborn Medicine, Department of Medicine, Boston Children's Hospital, Boston, MA USA; and Department of Pediatrics, Harvard Medical School, Boston MA USA.

<sup>9</sup>Department of Genetics, Harvard Medical School, Boston, MA USA.

<sup>10</sup>Howard Hughes Medical Institute, Harvard Medical School, Boston, MA USA.

<sup>11</sup>Cardiovascular Division, Brigham and Women's Hospital, Boston, MA USA.

<sup>12</sup>Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA.

<sup>13</sup>Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA.

<sup>14</sup>Department of Pediatrics, Icahn School of Medicine at Mount Sinai, New York, NY, USA.

<sup>15</sup>University of Pavia, Department of Chemistry, Pavia, Italy.

<sup>16</sup>Cardiovascular Research Institute, University of California, San Francisco, CA, USA.

<sup>17</sup>Department of Pediatrics, UCSF School of Medicine, San Francisco, CA, USA.

<sup>18</sup>Chan Zuckerberg Biohub, San Francisco, CA, USA

<sup>19</sup>Department of Epidemiology & Biostatistics, Institute for Computational Health Sciences, and Institute for Human Genetics, University of California, San Francisco, CA, USA.

<sup>20</sup>Department of Biochemistry and Biophysics, University of California, San Francisco, CA, USA.

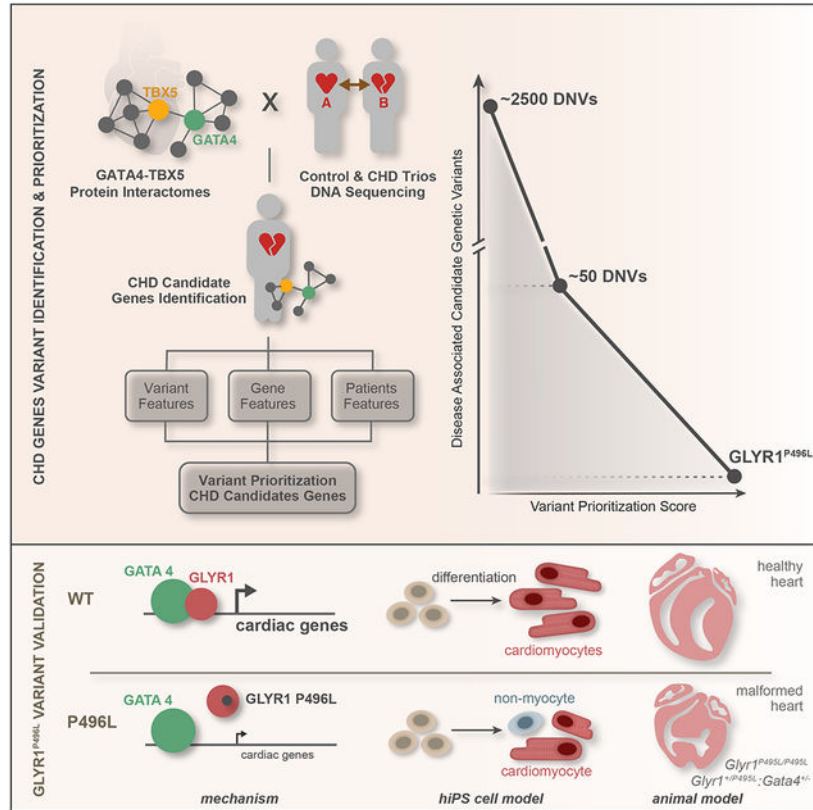
<sup>21</sup>Lead Contact.

## SUMMARY

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 mutation 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.

## Graphical Abstract



## In Brief

The integration of human protein-protein interactome networks of endogenous transcription factors known to be involved in cardiac malformations with the largest whole-exome-sequencing dataset of trios with congenital heart disease, identified an enrichment of rare *de novo* variants among interactome proteins, pointing to candidate disease genes.

## INTRODUCTION

Birth defects are complex developmental phenotypes affecting ~6% of births worldwide, yet their genetic roots are multifarious and difficult to ascertain (Christianson and Howson, 2006; 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 from our group 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 catalogue 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 display 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. (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 Genomic Consortium's 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; Molkentin et al., 1997; Mori 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; 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., 2009; 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, we leveraged an integrated proteomics and human genetics approach that dissects the protein-protein interactors of endogenous GATA4 and TBX5 in human cardiac progenitor cells to identify and prioritize potential disease genes harboring CHD-associated variants, revealing aspects of cardiac gene regulation. This approach can be leveraged to study the genetic underpinnings of many human diseases.

## RESULTS

### 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) we used antibodies against each endogenous cTF for affinity purification and mass spectrometry (AP-MS) (Figure 1A). Using CRISPR Cas9-gRNA ribonucleoproteins, we 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 (Figure S1A-E). Consistent with previous reports (Kathiriya et al., 2021; Luna-Zurita et al., 2016; Narita et al., 1997), *GATA4* and *TBX5* KO cells were able to differentiate into CMs, albeit with delayed beating and reduced differentiation efficiency (Figure S1E-G and Table S1A and 1B).

GATA4 or TBX5 mass spectrometry data were generated from three replicates of nuclei-enriched day 6 hiPSC-derived CPs from wild-type (WT) or KO samples treated with RNase and DNase to focus on nucleic acid independent interactions (Figure 1A). An initial list of GT-interactors in WT CPs was obtained by scoring the proteins identified in WT AP-MS experiments to their corresponding KO control line using the established protein-protein interaction algorithm SAINTq (Teo et al., 2016). For further stringency, additional filtering was applied for the high-scoring interactors determined by SAINTq based on nuclear localization and co-expression in the same cells as the bait protein. Proteins whose mRNA was downregulated in the KO cells compared to WT were excluded (See Methods: 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 (Enane et al., 2017; Padmanabhan et al., 2020; Waldron et al., 2016). Mutations in several of these interactors have been previously associated with human or mouse cardiac malformations, highlighting the potential of our approach for disease-gene discovery (Figure 1B and 1C, Figure S2A and S2B and Table S1C and 1D).

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 (Figure S2C). Both networks were enriched in proteins involved in similar biological processes (Figure 1B-D and Figure S2A and S2B). The top two most represented processes were transcription regulation and chromatin modification (Figure 1D), 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 (Figure 1B and 1C, Figure S2A and S2B). Chromatin modifiers (~25% or 15% of GATA4 or TBX5 interactors, respectively) predominantly belonged to ATP-dependent complexes, and we found several histone-modifying enzymes in the GATA4-PPI (Enane et al., 2017) (Figure 1B and 1C, Figure S2A and S2B). A number of RNA processing and splicing proteins, as well as members of the nuclear pore complex were also identified (Figure 1B and 1C, Figure S2A and S2B). The GT-PPIs mostly included proteins expressed ubiquitously, with a small number of tissue-enriched and cell-type enriched interactors (Figure 1E, Figure S2D).

### **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, we assessed their intersection with *de novo* variants (DNVs) and very rare (minor allele frequency (MAF) < 10<sup>-5</sup>) inherited loss-of-function variants found in CHD probands from the Pediatric Cardiac Genomics Consortium (PCGC). In addition to a previously published cohort of parent-offspring CHD trios and control trios (Jin et al., 2017), we included variant data from an additional 419 CHD probands and their parents for a total of over 3,000 trios. We used a permutation-based statistical test to analyze the frequency of variants in GT-interacting proteins among the CHD probands compared to the control group (see Methods: Permutation-based test). 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 2A). 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 odds ratio (OR) GATA4-PPI: 5.59 & adjusted p-value 0.001; TBX5-PPI: 4.34 & adjusted p-value 0.0096). By contrast, very rare inherited loss-of-function variants occurred in GT-PPI proteins with the same frequency in control and CHD groups (Figure 2B and Table S1E).

To determine whether the enrichment was predominately driven by genes previously known to be involved in cardiac development (Figure S3A), we 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 S1F). We still found an enrichment in proteins harboring protein-altering DNVs from CHD probands in both GATA4 and TBX5 interactomes (Figure 2B and Table S1E). Similar trends were observed holding out a smaller list of 144 published human CHD-genes (Table S1E) (Izarzugaza et al., 2020).

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, we generated GT-PPIs in kidney cells (HEK293) over-expressing human GATA4 or TBX5, and subjected them to the same permutation analysis with the CHD and control cohorts (Figure S3B-S3D and Table S1G and 1H). There was no significant enrichment in proteins harboring CHD-associated



protein-altering DNVs (Figure 2B and Table S1E). The GT-PPI overlap between cell types was small, with only 20 GATA4 and 13 TBX5-interactors shared (Figure S3B-D and Table S1G and H), highlighting the importance of endogenous tissue-specific protein-protein interactions in elucidating the genetic underpinnings of human diseases.

In a complementary analysis to test whether genes in the GT-PPI were enriched for protein-altering DNVs in CHD probands, we 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 we observed in interactome genes. For each gene in the GT-PPI, we identified other genes that had comparable *de novo* mutability scores (Samocha et al., 2014). We further narrowed the list of matches based on similarity of expression levels in WT CP cells. 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  $p = 0.009$ ). Conversely, there was no enrichment of non-synonymous or rare inherited loss-of-function variants among GT-PPI genes (Table S1I).

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, we 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 2C).

### **GATA4:TBX5-Interactors with Protein-Altering DNVs Unveil CHD Candidate Genes with Characteristic Features of Disease Genes**

We next asked 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 loss-of-function (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 haploinsufficiency scores than genes outside the interactome with protein-altering DNVs (Figure 3A and Figure S3E). 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 et al., 2008), 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 3B and C).

GT-interactors with protein-altering DNVs among those with CHD exhibited higher expression in the developing heart than genes with protein-altering DNVs outside the GT-PPIs (Figure 3D), but they generally displayed a broad expression pattern across most cell types (Figure 3E and Figure S3F-I) and largely involved proteins relevant to chromatin

biology (Figure 3F). Other biological processes with unexplored roles in CHD were affected, such as RNA splicing and protein folding (Figure 3F). Furthermore, although many DNVs in GT-interactors were detected in probands suffering from CHD with extracardiac abnormalities and/or neurodevelopmental defects, a sizeable number were also found in “isolated” CHD cases (Figure S3J and Table S1J).

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, we identified 20 LoF DNVs and 53 missense DNVs present in CHD cases (Table S1J). 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) (Table S1K). 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 (Table S1J and 1K). The contribution of *de novo* splice variants could not be determined due to their low counts in interactome genes from cases and controls (Table S1J and 1K).

### **An Integrative Method for Scoring Variants Identifies Specific GT-Interactors as Candidate Genes for CHD**

The GT-PPI framework combined with trio sequencing allowed us 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 et al., 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 et al., 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; Stark et al., 2017). 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 (Figure S3K), 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 4A and Figure S4A). 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 CDS length. At a residue level, a higher score indicates: (5) a variant's increased likelihood of being deleterious to protein function (CADD score); and (6) occurrence at a functional residue or protein domain. At the proband level, through the application of a weighted correction factor, a higher score indicates: (7) the background genetic variation of this individual does not include DNVs or very rare inherited LoF variants in genes known to be involved in cardiac malformations (Jin et al., 2017) 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: Variant scoring) (Figure 4A, Figure S4A). 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).

We applied this scoring method to previously identified variants implicated in CHD (Basson et al., 1999; Furtado et al., 2017; Garg et al., 2003) 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 (Gifford et al., 2019) (Figure 4B), 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 4B and Table S1L).

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 4 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. We found that variants with higher prioritization score exerted a greater effect on the encoded protein's transcriptional activity (Figure 4C).

Next, we aimed to evaluate the benefit of the GATA4 and TBX5 PPI incorporation as a filtering strategy for the identification of CHD candidate genes. To test this, we applied the variant prioritization scoring to all *de novo* missense variants from probands found in both interactome and non-interactome genes (Table S1M). We found that the variant prioritization score's 75<sup>th</sup> 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 S4B). 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 S4C). Additionally, among the missense DNVs

within the top quartile, the average prioritization score was significantly higher for GT-PPI variants than for the unfiltered missense DNVs (Figure S4D). In agreement, the proportion of missense DNVs in known CHD-causing genes that belonged to the GT-PPI network (~30% GT-PPI missense DNVs) was an order of magnitude higher than the corresponding proportion across all observed missense DNVs (~3% of all missense DNVs) observed in CHD probands. To most directly test the value of filtering the DNVs with the GT-PPI, we performed a Precision-Recall analysis that provides a performance metric, Area Under the Curve (AUC) estimated for the variant prioritization scoring with or without incorporating information from the GT-PPI. This approach demonstrated that incorporation of the GT-PPI variant-filtering strategy improved the ability of the variant prioritization scoring to predict known CHD-causing variants (GT-PPI filtered variants using VPS, AUC=0.61 vs. Random Classifier, AUC=0.3; all variants using VPS, AUC=0.18 vs Random Classifier, AUC=0.05); furthermore this outcome was not driven by a confounding effect of GT-PPI-dependent parameters included in the variant prioritization scoring re-weighting step, as a modified scoring excluding this feature performed similarly (Figure S4E and F).

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 4B). 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) (Figure 4B). 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 disperse 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 4B and Table S1L).

*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). These CHD candidate genes were detected as *GATA4* interactors, which was validated by co-immunoprecipitation (Figure S4G-J). While *GLYR1* and *SMARCC1* were previously unknown to interact with *GATA4*, we recently 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 (*Tnnt2-Cre; Brd4<sup>lox/lox</sup>*) 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.

We, therefore, investigated the genetic landscape of the *de novo* *GLYR1* variant carrier, and identified three LoF and 62 missense variants inherited from their asymptomatic parents, while no other DNVs were found in this proband (Table S1N). 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, therefore, are 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.

### The CHD-Variant in *GLYR1* Impacts Structural Dynamics and 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; Fu et al., 2006; Marabelli et al., 2019; Yu et al., 2020). The *GLYR1* missense CHD DNV we detected involved the substitution of a highly conserved proline with a leucine at amino acid (aa) 496 within the  $\beta$ -hydroxyacid dehydrogenase ( $\beta$ -HAD) domain, described to mediate the interaction between *GLYR1* monomers (Marabelli et al., 2019; Montefiori et al., 2019). Since Proline<sup>496</sup> is located within a rigid loop enriched in aromatic residues connecting two tetramerization domains (Figure 5A-5C), we hypothesized that its substitution by a leucine would impact the structural dynamics of the *GLYR1*  $\beta$ -HAD domain and therefore its ability to acquire certain functional states.

Molecular dynamics (MD) computational simulations predicted the mutant (*GLYR1*<sup>P496L</sup>)  $\beta$ -HAD to explore a narrower set of structural conformations than the WT, as shown by the time-dependent evolution of the root mean square deviation (RMSD) of frames visited during the trajectories from the reference structure (Figure S4K). This result was confirmed by the distribution of the RMSD calculated for every pair of states sampled during the simulations (Figure 5D). Furthermore, *GLYR1* structural dynamics at the local level, measured by the standard deviations of the atomic positions in the simulations (RMSF), indicated an overall lower flexibility of *GLYR1*<sup>P496L</sup> compared to the WT protein, which was evident in the Rossman-fold domain (262-437 aa) (Figure 5E). These data indicated that the P496L variant in *GLYR1* induces significant differences in the structural dynamics of the  $\beta$ -HAD domain, at the global and local levels, predicting a general increase in the structural rigidity of this region in *GLYR1*.

Increased rigidity within the  $\beta$ -HAD domain could affect *GLYR1*'s capacity to adapt to interacting partner proteins through conformational selection. Co-immunoprecipitation assays demonstrated that the *GLYR1* P496L DNV destabilized its physical interaction with *GATA4* (Figure 5F and Figure S4L) but not with previously described *GLYR1* interactors LSD2, CDK9 or Cyclin T1 (Fang et al., 2013; Yu et al., 2020) (Figure S4M). Since previous studies indicated a role for *GLYR1* in transcriptional regulation (Fei et al., 2018; Yu et al., 2020), 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 and 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 (Figure 5G and Figure S4N).

### **GATA4 & GLYR1 Co-bind a Defined Set of Heart Development Genes and Co-Regulate their Expression**

GLYR1 localizes within chromatin regions rich in histone H3 trimethylated on Lys36 (H3K36me3) at actively transcribed gene bodies to regulate transcription elongation (Fang et al., 2013; Fei et al., 2018; Marabelli et al., 2019; Yu et al., 2020). However, knowledge about how GLYR1 is recruited to specific loci or its function in homeostasis and disease is limited. Analysis of gene expression along with GLYR1 genome-wide occupancy during CM differentiation together with H3K36me3 genome-wide distribution by ChIP-sequencing (ChIPseq) in hiPSCs and CPs revealed dynamic relocalization of GLYR1 during differentiation of hiPSCs to CPs, with ~7400 differentially bound genes (FDR<0.1) between the two stages (Figure 6A and Figure S5A). K-means clustering of genes differentially bound by GLYR1 based on the three measured variables—GLYR1 ChIPseq, H3K36me3 ChIPseq, and RNA expression—highlighted GLYR1 recruitment to 4246 gene bodies upon differentiation of hiPSCs to CPs (Clusters 2 and 3). Gene ontology (GO) analysis revealed that gene programs associated with heart development were enriched in Cluster 2, which showed the highest levels of GLYR1 ChIP signal in CPs, whereas Cluster 3 was enriched for genes involved in general cellular processes. On the other hand, Cluster 1 contained ~3155 GLYR1-bound genes in hiPSCs that were lost in CPs, mainly associated with cell cycle and ribosome biogenesis terms (Figure 6A and Table S1O and 1P). Overall, GLYR1 preferentially bound to transcribed regions of active genes and co-localized with H3K36me3 (Figure 6A and Figure S5A-C). Interestingly, GLYR1 occupied ~50% of the genes up-regulated in CPs and marked with H3K36me3 (Figure S5A and B), suggesting that GLYR1 is recruited to a large set of cardiac genes during CM differentiation.

In CPs, as described in other cell types (Fei et al., 2018; Yu et al., 2020), GLYR1 broadly occupied gene bodies, from the first intron to the transcription end site (TES) on average (Figure S5C and D). On the other hand, GATA4 preferentially occupied distal regulatory elements, though some peaks were found at introns inside gene bodies, similar to GLYR1 (Figure S5C and D). To investigate GATA4-GLYR1 genomic co-occupancy in CPs, we overlapped the genes where GLYR1 was recruited in CPs (GLYR1<sup>CP</sup>: clusters 2-3, FDR<0.1 and Log<sub>2</sub>FC>0.5) with genes bound by GATA4 within the gene body window where GLYR1 typically binds (1<sup>st</sup> Intron-TES). This analysis found a statistically significant overlap between GLYR1<sup>CP</sup> and GATA4-bound gene bodies (Fisher exact p-value < 2.2E<sup>-16</sup>; OR: 2.38), identifying a defined subset of GATA4 and GLYR1-bound genes, mostly upregulated in CPs vs hiPSCs (FDR<0.05) and with greater enrichment in heart development GO terms compared to GLYR1- only and GATA4-Only occupied gene bodies (Figure 6B and C, Figure S5E and Table S1Q and 1R).

To directly evaluate if GATA4 and GLYR1 regulate the transcription of the genes they co-occupy, we analyzed the effect of silencing *GATA4* or *GLYR1* on the expression of genes in CPs by bulk RNAseq (Table S1S and 1T). *GLYR1* silencing led to reduced expression of

more than 800 genes associated with embryonic development and heart development terms compared to a control siRNA, which suggested a functional relevance for GLYR1 in the transcriptional regulation of the CM differentiation process (Figure S5F and Table S1S and 1U). Genes co-bound by GATA4 and GLYR1 were ~7 times more likely to be significantly down-regulated by both GATA4 and GLYR1 independent knockdowns compared to those not co-bound (Figure S5G). Several co-occupied and co-regulated loci (*GATA4*, *GATA6*, *TBX5*, *LRP2*, *TEMN4*, *CC2C2A*, *TTN* and *ENDRA*) are involved in human or mouse cardiac malformations (Chauveau et al., 2014; Clouthier et al., 1998; Nakamura et al., 2013; Theis et al., 2019; Pierpont et al., 2018) (Figure 6D). The observation that GATA4 mainly occupies intronic regions within GATA4:GLYR1-bound gene bodies led us to examine features characteristic of active or repressed gene regulatory elements (Akerberg et al., 2019; Kimura, 2013) within these regions. GATA4 occupancy within GATA4:GLYR1-bound gene bodies co-localized with high levels of marks associated with active regulatory elements (H3K27ac, H3K4me3, H3K4me1, MED1), as well as with TFs *TBX5*, *MEIS1*, *ISL1*, and *NKX2-5*, but with undetectable levels of the repressive mark H3K27me3 (Figure 6E and Figure S5H). Similarly, GATA4 in GATA4-Only genes occupied multi-TF intronic regions, but, on average, with lower levels of marks associated with active regulatory elements and higher levels of the repressive mark H3K27me3, in line with GATA4 acting as a repressor in about half of GATA4-Only bound neural genes (Figure 6C, Figure S5E and H). The GATA4 co-localization with multiple TFs in GATA4:GLYR1 co-bound genes raised the possibility GLYR1 interacted with other cTFs. However, GLYR1 co-immunoprecipitation assays in CPs demonstrated that while GLYR1 interacted with GATA4, it did not interact with the other cTFs tested, including *NKX2-5*, *TBX5*, *ISL1*, and *MEIS1* (Figure S5I), each of which had motif enrichment in GATA4 occupied regions (Table S1V) (Akerberg et al., 2019).

To test whether GATA4 and GLYR1 interact to positively regulate gene expression, we cloned several intronic regions with the features described above into a luciferase reporter vector under control of a minimal promoter, and tested the transactivation ability of GLYR1 and GATA4. Transfection of GATA4 alone resulted in activation of all three reporters, whereas GLYR1 alone induced luciferase activity of two out of three tested reporters, indicating that these intronic locations could function as response elements (Figure 6F). Importantly, co-transfection of GATA4 and GLYR1 increased activity of each reporter, consistent with functional co-regulation. The synergistic/additive activation induced by GLYR1 WT was strongly attenuated in the context of the GLYR1 P496L mutation (Figure 6F).

### The P496L Variant in GLYR1 Affects Cardiomyocyte Differentiation

In order to better characterize the impact of the P496L missense variant in GLYR1 protein function and gene regulation during CM differentiation, we created an hiPSC line homozygous for the missense variant P496L (*GLYR1<sup>P496L</sup>*) by CRISPR Cas9 homology-directed repair (Figure S6A-B). In parallel, we generated a *GLYR1<sup>KO</sup>* hiPSC line, and verified the reduction of RNA expression and absence of detectable protein (Figure S6C and D).



In mouse embryonic stem cells (mESCs), GLYR1 is essential for pluripotency, and depletion of GLYR1 leads to differentiation, cell cycle arrest and apoptosis (Yu et al., 2020). Similar to mESCs, GLYR1 deficiency in hiPSCs resulted in a dramatic shift in the transcriptional landscape (Figure S6E), characterized by a reduction in pluripotency and cell cycle gene expression and upregulation of tumor suppressors and apoptosis genes compared to *GLYR1<sup>WT</sup>*. However, these changes were not observed in *GLYR1<sup>P496L</sup>* hiPSCs, which were more transcriptionally similar to *GLYR1<sup>WT</sup>* cells (Figure S6E and F). These results suggested that GLYR1 is essential for the maintenance of pluripotency in hiPSCs and that the P496L variant does not lead to a complete protein loss of function.

We then subjected the *GLYR1<sup>WT</sup>* and *GLYR1<sup>P496L</sup>* lines to CM differentiation and investigated their transcriptional landscapes at the CP (day 6) and CM stages (day 18) by scRNASeq. At day 6 of differentiation, we observed three distinct subpopulations comprising both *GLYR1<sup>WT</sup>* and *GLYR1<sup>P496L</sup>* cells clustering together and expressing marker genes associated with CPs (cluster 0), endoderm (cluster 1) and vascular/endothelial cells (cluster 4), as well as a neural progenitor-like cluster mainly comprising *GLYR1<sup>P496L</sup>* cells (cluster 2) (Figure 7A and B, Figure S6G, Table S1W). *GLYR1* expression was detected in all clusters, whereas *GATA4* was only detected in the CP (cluster 0), endoderm (cluster 1) and the vascular/endothelial (cluster 4)-like clusters (Figure S6H). The intersection of the scRNAseq analysis with GLYR1 and GATA4 ChIPseq at differentiation day 6 from *GLYR1<sup>WT</sup>*, revealed that the CP-like cluster was enriched for genes co-bound by GATA4 and GLYR1 (Figure S6I). Differential expression analysis (FDR<0.05, LogFC>0.125) between *GLYR1<sup>P496L</sup>* and *GLYR1<sup>WT</sup>* CP subpopulations (cluster 0) identified 1458 downregulated genes involved in heart development, cytoskeletal organization, cell cycle, response to hypoxia and ATP metabolic processes; 1025 genes were upregulated and associated with non-cardiomyocyte developmental processes (Figure 7C, Table S1X and S1Y). Remarkably, more than 35% of the GATA4:GLYR1 co-bound genes were differentially expressed between *GLYR1<sup>P496L</sup>* and *GLYR1<sup>WT</sup>* CPs, whereas less than 15% of GATA4-Only and GLYR1-Only bound were differentially expressed (Figure 7D). ChIPseq revealed that GLYR1 occupancy was reduced in the *GLYR1<sup>P496L</sup>* CP-like cluster among GATA4:GLYR1 co-bound genes whose expression was shown to be downregulated, but not on those whose expression was upregulated or not changed (Figure 7E and F, Figure S6J and K). Importantly, GLYR1 genome-wide binding was not affected by the P496L variant (Figure S6L). These data suggest that the P496L variant affects GLYR1 occupancy and transcriptional regulation of a discrete set of target genes co-bound by GATA4, several of which have been involved in human cardiac malformations and cardiomyopathies.

At day 18 of differentiation, we observed a heterogeneous group of cellular subpopulations with significantly imbalanced contributions from *GLYR1<sup>P496L</sup>* and *GLYR1<sup>WT</sup>* cells. *GLYR1<sup>P496L</sup>* cells mainly contributed to neural-like (clusters 2, 3, 5 and 9) and hepatocyte-like (cluster 7) subpopulations, to the detriment of CM-like clusters 0 and 6 (Figure 7G and H, Figure S7A, Table S1Z and 1AA). The impaired CM-like differentiation of the *GLYR1<sup>P496L</sup>* line was also evident in the diminished cTNT levels detected by FACs and immunostaining (Figure 7I and J). Within the CM-like clusters, *GLYR1<sup>P496L</sup>* vs *GLYR1<sup>WT</sup>* cells were homogeneously distributed, and no segregation between genotypes was observed based on maturity markers (Figure S7B). However, differential expression



analysis identified 544 downregulated genes in *GLYR1*<sup>P496L</sup> CMs (FDR<0.05; logFC>0.125 *GLYR1*<sup>P496L</sup> vs *GLYR1*<sup>WT</sup>), mainly associated with precursor metabolites and energy generation, particularly ATP metabolic processes. Additionally, 803 genes were upregulated in *GLYR1*<sup>P496L</sup> CMs (FDR<0.05; logFC>0.125 *GLYR1*<sup>P496L</sup> vs *GLYR1*<sup>WT</sup>) involved in cell adhesion and migration, protein phosphorylation and cell contraction (Figure S7C, Table S1AB and 1AC). Moreover, *GLYR1*<sup>P496L</sup> CMs exhibited altered beating dynamics characterized by a significant prolongation of physical contraction and relaxation, accompanied by a decrease in the beating rate compared to control cardiomyocytes (Figure S7D and Videos S1 and S2). Overall, these data demonstrated a detrimental impact of the *GLYR1* P496L variant in CM differentiation, associated with altered *GLYR1* genomic occupancy and gene regulation at a discrete set of loci co-bound by GATA4.

### GLYR1 Mutation Disrupts Mouse Cardiac Development

In order to assess the biological importance of the *GLYR1* P496L variant *in vivo*, we generated a mouse line harboring a P495L single nucleotide variant in *GLYR1* (*Glyr1*<sup>P495L/+</sup>), homologous to human P496L, using CRISPR-Cas mediated genome editing (Figure S7E). After backcrossing four generations into the C57BL6/J background, we intercrossed *Glyr1*<sup>P495L/+</sup> mice and collected 92 pups for genotyping and heart histology at day 1 after birth (P1). Although all genotypes were born at the expected mendelian ratios ( $\chi^2 = 0.96$ ), 54% of the *Glyr1*<sup>P495L/P495L</sup> and 15.5% of the heterozygous (*Glyr1*<sup>P495L/+</sup>) mice displayed postnatal lethality between days 0 and 1, compared to only 4.4% of the WT littermates ( $\chi^2 = 0.02$  at P1) (Figure S7F). Echocardiography and histological analysis revealed ventricular septal defects (VSDs) in ~15% of the *Glyr1*<sup>P495L/P495L</sup> mice (Figure S7G and H, Videos S3 and S4). Thus, this model provides evidence for the biological importance of *GLYR1* in cardiac development, and demonstrates a deleterious effect of the P495L variant *in vivo*.

To assess whether there is a GATA4-*GLYR1* genetic interaction in mice, we crossed *Glyr1*<sup>P495L/+</sup> mice to GATA4-mutant mice (Watt et al., 2004). As expected, *Gata4*<sup>+/-</sup> mice exhibited partially penetrant cardiac defects, with a 22% incidence of VSDs, and 50% died between day 0 and day 1, consistent with previous reports for *Gata4* deletion in C57BL6/J background (Rajagopal et al., 2007), (Figure S7I, Figure 7K). *Glyr1*<sup>P495L/+</sup> mice exhibited the expected low penetrance of cardiac alterations, most frequently a persistent interatrial communication (patent foramen ovale), whereas no defects were detected in the 23 WT littermate hearts analyzed (Figure S7I, Figure 7K). We only identified eight compound heterozygote (*Glyr1*<sup>P495L/+</sup>;*Gata4*<sup>+/-</sup>) animals out of 88 pups at birth (~9.1% observed versus 25% expected,  $\chi^2 = 0.0086$ ), and five of those died in the first 24 hrs. of life (Figure S7I). Whole mount, histology and echocardiography analysis showed that all compound *Glyr1*<sup>P495L/+</sup>;*Gata4*<sup>+/-</sup> hearts were dysmorphic, with complete penetrance of cardiac septal defects, including about 80% represented by atrio-ventricular septal defects (AVSDs) (Figure 7K, Figure S7I and J, Videos S5 and S6). These data provide *in vivo* evidence for the biological relevance of the *GLYR1* P495L variant and its interaction with GATA4 in human disease.

## DISCUSSION

Here, we 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. An unbiased 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. *GLYR1* widely co-occupied cardiac regulatory elements with GATA4, and the *GLYR1* disease variant P496L disrupted the interaction with GATA4 and co-activation of cardiac developmental genes. The importance of the *GLYR1* variant and the GATA4-*GLYR1* 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.

### Integration of Tissue-specific TF-PPIs with Human Variant Data Highlights Disease Mechanisms

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 (Bryois et al., 2020; Greene et al., 2015; Izarzugaza et al., 2020; Köhler et al., 2008; Priest et al., 2016). Since most of the PPI available 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 (Barshir et al., 2014; Magger et al., 2012). 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 we 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 catalogue provided by the increasing number of large-scale sequencing studies.

### **GLYR1 Co-regulates Heart Development Genes with GATA4 and Is Mutated in CHD**

Our integrative proteomics and human genetics approach revealed GLYR1 as a GATA4 interactor in CPs that constitutes a strong candidate gene for CHD. Since the GLYR1 P496L variant impacted the protein structural dynamics and reduced its interaction with GATA4, it suggests that destabilization of tissue-specific protein-protein interactions could result in cardiac-restricted phenotypic manifestation associated with the mutation of a ubiquitously expressed chromatin reader. Evidence from human iPSC-derived CPs and in vivo mouse studies suggests the GLYR1 P496L functions as a hypomorph with discrete alterations in genome occupancy at a subset of H3K36me3-enriched regions. During CM differentiation, GATA4 physical interaction with 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. Given the overlapping functions of GATA4 and GATA6, it will be interesting to determine if the heterozygous missense variant in GATA6 in this proband functions together with the GLYR1 variant to cause disease.

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 is available to highlight disease mechanisms and provide a powerful filter for interrogating the genetic basis of disease.

### **Limitations of the study**

The variant prioritization scoring developed in this study 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 variants in *GLYR1* contribute to other cases of CHD or to other diseases in humans.

## STAR METHODS

### RESOURCE AVAILABILITY

**Lead contact**—Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Deepak Srivastava (Deepak.srivastava@gladstone.ucsf.edu).

**Materials availability**—All resources and materials reported in this paper will be shared by the lead contact upon request.

**Data and code availability**—The RNAseq, scRNAseq and ChIPseq datasets generated during this study are available at GEO [GSE159411/ <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE159411>]. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE (Perez-Riverol et al. 2019) partner repository with the dataset identifier PXD022091. Code is available at <https://github.com/mepittman/ctf-apms>. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

### EXPERIMENTAL MODEL AND SUBJECT DETAILS

**Cell Lines**—HEK293 (human [*Homo sapiens*] fetal kidney) and HeLa (human [*Homo sapiens*] cervical cancer) cells were all obtained from ATCC (<https://www.atcc.org/>). HEK-293 and HeLa cells were cultured in Dulbecco's Modified Eagle Medium (DMEM), high glucose, GlutaMAX™ Supplement (Cat. 10566016, Thermo Fisher Scientific) supplemented with 10% fetal bovine serum, 2mM sodium pyruvate, 2mM non-essential amino acids and 100 I.U./mL penicillin and 100 µg/mL streptomycin.

The WTC11 human iPSC line (Cat. GM25256, Coriell) was obtained from the Gladstone Stem Cell Core ([https://labs.gladstone.org/stem\\_cell](https://labs.gladstone.org/stem_cell)). The WTC11 line is used as a normal control by research groups all over the world. This hiPSC line was derived from a skin biopsy from a healthy adult Asian male donor in his early thirties, who showed normal function in a battery of tests. The original fibroblasts were reprogrammed using episomal methods with the following factors: LIN28A, MYC (c-MYC), POU5F1 (OCT4), and SOX2 and pluripotency state and differentiation potential characterized as previously described (Miyaoka et al. 2014).

All hiPSC clones were recurrently verified free of mycoplasma contamination and checked for normal karyotype. Karyotyping analyses were performed by Cell Line Genetics.

**Human iPSC cell line generation by CRISPR-Cas9 editing**—To generate the *GATA4*, *TBX5* and *GLYR1* knockout lines as well as the *GLYR1*<sup>P496L</sup> hiPSC line, WTC11 hiPSCs were dissociated in accutase (Cat. 07920, Stem Cell Technologies), and 250,000 cells were aliquoted per condition and nucleofected with Cas9-ribonucleoprotein complexes (Cas9-RNP) following the Primary Cell Nucleofection P3 Kit manufacturer's instructions

(V4XP-3960, Lonza). For Cas9-RNP complex preparation 180pmol of each synthetic modified sgRNA (Synthego) to target exon4 of *GATA4* (GAGGCCACUCGGCGGGAGG), exon6 *TBX5* (GCTTACCTTGTGGTTCTGGTAGG) or the *GLYR1* locus at chr16:4811270-4811248 (ATGTATTTTCAGGTAGAAATCAGG) and 20pmol of SpCas9-NLS purified protein (QB3 Macrolab, UCB) were diluted into 20µl of nucleofection buffer prepared as indicated in Primary Cell Nucleofection P3 Kit, mixed and incubated at room temperature for 10 min. After 10 min. of incubation, the aliquoted 250.000 cells were resuspended in 20µL nucleofection buffer containing the corresponding Cas9-RNP complexes, mixed 5-6 times and then transferred to the bottom of a nucleofector well (nucleofector 96 well cassette from the Primary Cell Nucleofection P3 Kit). For the generation of the *GLYR1*<sup>P496L</sup> line, the HDR template (12nM): CCTCAGATATCCTGCAAGGAACTTTAAGCTTGATTTCTACCTGAAATACATTCAG AAGGA, was added to the Cas9-RNP complexes, prior to transferring into the nucleofector wells. This was repeated for every Cas9-RNP condition. The nucleofector cassette was placed in the nucleofector instrument (Nucleofector 4D system, Lonza) and cells nucleofector using the preset program DS-138. The nucleofection cassette was brought back to sterile hood and 80µL of Essential 8 medium (E8) (Cat. A1517001, Life Technologies) with 5µM rho kinase (ROCK) inhibitor (Y-27632 2HCl, Cat. S1049, Selleckchem.com) added into each well and incubated for 10 min. at 37°C. During the incubation time, we removed the hESC-qualified LDEV-free matrigel (Cat. 354277, Corning) from 12 well plates (Corning) pre-coated for 1 hr. at 37°C with 0.5ml per well of matrigel and added 2mL of pre-warmed E8 medium plus 5µM ROCK inhibitor (Y-27632 2HCl) in each well. Then pipetted the nucleofected cells to each of the pre-coted wells with media. After ~3-5 days, wells with surviving clones were expanded to isolate gDNA for screening. A genomic fragment spanning the gRNA target sites was amplified using primers FW 5'AGAGATCTCATGCAGGGTCG3' and REV 5'TCATGATGCCTGGCCTTACT3' for *GATA4* with Titanium Taq DNA Polymerase (Cat. 639209, Takara Bio) and primers FW 5'GCAGAAACAGTTGCCAGAA3' and REV 5'CAAGGCGAATTTAGAGGGCG3' for *TBX5* or FW 5' CACCAGTGCCTCTAGCCT 3' and REV 5' TGCAGCAAATGAGGTAGGGT 3' for *GLYR1* with Phusion® High-Fidelity PCR Master Mix with GC Buffer (Cat. M0532S, New England BioLabs) and Sanger sequenced (Quintara Biosciences or MCLAB) to identify clones with frame-shift insertions/deletions. Synthego ICE analysis was run to identify clones with highest knock-out efficiency (<https://ice.synthego.com/#/>), which were subsequently subjected to the colony picking and clone sequencing until monoclonal lines were generated. The top 5 sgRNA predicted off-targets (<https://horizondiscovery.com/en/products/tools/crispr-specificity-analysis>) were verified to be intact by sanger sequencing in the final monoclonal lines and checked for normal karyotype (Cell Line Genetics).

**Human iPSC culture**—Human iPSCs were maintained on tissue culture-treated polystyrene plates (Cat. 430630, Corning) with hESC-qualified LDEV-free matrigel (Cat. 354277, Corning) in Essential 8 medium (E8) (Cat. A1517001, Life Technologies). The medium was changed daily and the hiPSCs were split every 4 – 6 days using Accutase (Cat. 07920, Stem Cell Technologies). The rho kinase (ROCK) inhibitor (Y-27632 2HCl, Cat.

S1049, [Selleckchem.com](https://www.selleckchem.com)) was included in the medium at 5 $\mu$ M final concentration on the day of passaging.

**Induced cardiomyocyte differentiation**—For human cardiac differentiation into CPs and CMs, we modified the protocols originally developed by Lian et al. and Tohyama et al. to achieve stage-specific, high yield, high-purity cardiac commitment in vitro (Lian et al. 2013; Tohyama et al. 2013). Briefly hiPSCs were detached from hESC-qualified LDEV-free matrigel (Cat. 354277, Corning) with accutase (Cat. 07920, Stem Cell Technologies) and reseeded on matrigel at 0.6–1.2x10<sup>5</sup> cells per 12well in Essential 8 medium (E8) (Cat. A1517001, Life Technologies) with 5 $\mu$ M ROCK inhibitor (Y-27632 2HCl, Cat. S1049, [Selleckchem.com](https://www.selleckchem.com)) (day-3). We optimized each hiPSC clone individually to identify the best cell seeding density that resulted in high levels of cTNT, NKX2-5, TBX5 and GATA4-positive sheet-like beating CMs at day15 of differentiation. On the next two days, media was changed daily with fresh E8 medium without ROCK inhibitor. On the day of cardiac induction (day0), 6 $\mu$ M CHIR99021 (Cat. 4423, Tocris) was added for 24 hr. in 1ml per well of B27-supplemented (without insulin) RPMI1640 media (Cat. 11875-119, Life Technologies). On day1 media was changed for freshly prepared 6 $\mu$ M CHIR99021 in B27-supplemented (without insulin) (Cat. A1895601, Life Technologies) RPMI1640 media. At day 2 and 3 IWP4 (Cat. 5214, Tocris) in B27-supplemented (without insulin) RPMI1640 media was added to activate Wnt signaling at a final concentration of 5 $\mu$ M. At day 4, IWP4 was removed, and 1ml of B27-supplemented (without insulin) RPMI1640 media per was daily added per well (days 4-9). At day 10 the media was changed to regular B27-supplemented (with insulin) (Cat. A1895601, Life Technologies) RPMI1640 hereafter. Typically, in parallel differentiations under identical conditions, WT cells started spontaneous contraction as early as days 7-8 while *TBX5-KO* and *GATA4-KO* lines tend to be delayed by 48-96 hrs., and *GLYR1<sup>P496L</sup>* line showed a delayed beating onset of 24-48hrs. To ensure consistency across the differentiations used for experimental proposes, from each differentiation cells were collected at day6 and day15-18 for cTNT, TBX5, NKX2-5 and GATA4 FACS analysis. WT hiPSC-CP samples used for AP-MS corresponded to differentiations with sheet-like beating cardiomyocytes and a minimum of 75% cTNT<sup>+</sup>, 70% TBX5<sup>+</sup>, 70% GATA4<sup>+</sup>, 70% NKX2-5<sup>+</sup> cells analyzed by FACS at day15 of differentiation. GATA4-KO and TBX5-KO CP samples corresponded to differentiations with sheet-like beating cardiomyocytes, were TBX5 or GATA4 protein absence had been confirmed, and had a minimum of 40% cTNT positive cells at day15 of differentiation. Similarly, *GLYR1<sup>P496L</sup>* samples corresponded to differentiations with sheet-like beating cardiomyocytes and had a minimum of 40% cTNT positive cells.

**Adenovirus**—Adenoviral – Human Type 5 (dE1/E3) viral particles expressing GLYR1 wild-type (Ad-GFP-EF1-h-GLYR1; PFU titer: 1x10<sup>10</sup> PFU/ml), GLYR1 P496L mutant (Ad-GFP-EF1-h-GLYR1 P496L PFU titer: 2x10<sup>9</sup> PFU/ml) or negative control (Ad-EF1a-eGFP; PFU titer: 1.2x10<sup>10</sup> PFU/ml) viral particles, were obtained from Vector Biolabs.

***Glyr1* P495L Mouse generation**—Mice were generated by blastocyst injection of ribonucleoprotein (RNP) complexes consisting of purified Cas9 protein (IDT) and guide RNAs targeting the GLYR1 locus along with a single-stranded oligonucleotide DNA



template for homology-directed repair (IDT) that leads to insertion of the P495L mutation along with introduction of a novel BsrI restriction endonuclease site. The single stranded guide RNA (ATGTATTTTCAGGTAGAAGTC) was used to target the *Glyr1* locus together with the HDR template with the *Glyr1* P495L substitution indicated below. Blastocysts were transferred to pseudo-pregnant females and pups were weaned at approximately 4 weeks of age. Founder animals were screened for introduction of the P495L mutation by PCR amplification and restriction digestion with BsrI and confirmed by sequencing. Founder animals were outcrossed to wild types and gremlin transmission of the P495L mutation was confirmed PCR/restriction digestion and sequencing. Animals were outcrossed to C57BL6 wild type mice for 4 generations. A genomic fragment spanning the gRNA target sites was amplified using primers FW 5'TTCCAGTCATTCCTTGCCCC3' and REV 5'TGATCAGAAGGGTCGGCAAG3' for *Glyr1* with Phusion® High-Fidelity PCR Master Mix with GC Buffer (Cat. M0532S, New England BioLabs) and Sanger sequenced (Quintara Biosciences or MCLAB) for genotyping.

gRNA+5: ATGTATTTTCAGGTAGAAGTC

HDR template:

```
AGGTGAGCCTGATACTCGGCGGGCAATTTTCATGTAGATCTTTTAAACTTCTAATG
AATGGC
TTTCCCTTCTCAGATATCCTACAAGGAACTTTAAACTGGACTTCTACCTGAAATA
CATTCA
GAAGGATCTCCGCCTCGCCATTGCATTGGGTGATGCAGTCAACCACCCCACTCCC
ATGGC AGCTGCAGCCAATGAG
```

## METHOD DETAILS

**Immunocytochemistry**—Media was removed from day 15 CMs and fixed on 12 well plates by adding 1ml 4% formaldehyde (Cat. 28906, ThermoScientific) followed by a 15 min. incubation at room temperature (RT). After the incubation, fixed cells were washed 3 times with 1ml of PBS (without Ca<sup>2+</sup> and Mg<sup>2+</sup>) and stored at 4°C in 1ml of PBS until processed for immunostaining. At first, cells were permeabilized for 45 min. at RT in 1ml of permeabilization/blocking buffer (5% donkey serum, 0.2% triton-X, PBS 1x). CMs were stained overnight at 4°C on gentle agitation with the following indicated primary antibodies: mouse monoclonal anti-Troponin T, Cardiac Isoform Ab-1 Clone 13-11 (REF MS-295-P, Thermo Scientific) (1:100) together with goat polyclonal C-20 anti-TBX5 (Cat. Sc-17866, Santa Cruz) (1:50) or goat polyclonal C-20 anti-GATA4 (Cat. Sc-1237, Santa Cruz) (1:50) or anti-GLYR1 (14833-1-AP, Proteintech Group) (1:50) diluted in permeabilization/blocking buffer. The next day wells were washed 3 times with 1ml of PBS per well followed by a 10 min. incubation on gentle agitation at RT. Cells were stained with Donkey Anti-Mouse Alexa Fluor 647 (Cat. A-31571, Thermo Fisher Scientific) and Donkey Anti-Goat or -Rabbit Alexa Fluor 488 (Cat. A11055 and A21206, Thermo Fisher Scientific) secondary fluorophore conjugated antibodies (1:1000) for 45 min. on gentle agitation at RT, protected from light exposure. After the incubation, cells were washed 3 times with 1ml of PBS per well followed by a 10 min. incubation on gentle agitation at RT. Immunostained samples

were counterstained with DAPI (Cat. 422801, BioLegend) and visualized in a Zeiss Z1 microscope and associated ZEN software.

**Flow Cytometry**—At CP (day6) or CM (day 15-18) stages of differentiation cells were dissociated in accutase (Cat. 07920, Stem Cell Technologies), and fixed in 1.5ml Eppendorf tubes with 1ml of 4% formaldehyde (Cat. 28906, ThermoScientific) for 15 min. on rotation at RT. After the incubation, fixed cells were washed 3 times with 1ml of PBS (without Ca<sup>2+</sup> and Mg<sup>2+</sup>) followed by centrifugation at 1700 rpm for 3 min and stored at 4°C in 1ml of PBS until processed for staining. For staining, cells were pelleted by centrifugation at 1700rpm 3min and PBS removed. Cell pellets were resuspended in 200µl of FACS buffer [1x PBS, 5% BSA, 5mM EDTA, 0.25% triton-X] and incubated at RT for 1 hr. The permeabilized cells were then pelleted resuspended in 30µl of staining-A in FACS buffer: mouse monoclonal anti-Cardiac Troponin T [1C11] (1:100) (Cat. Ab8295, Abcam) and goat polyclonal anti-NKX2-5 (1:50) (Cat. Sc-8697, Santa Cruz); or staining-B: mouse monoclonal A-6 anti-TBX5 (1:50) (Cat. Sc-515536, Santa Cruz) and goat polyclonal C-20 anti-GATA4 (1:50) (Cat. Sc-1237, Santa Cruz) and incubated at RT for 1 hr. Cells were then washed/spined down 3 times with 200µl of FACS buffer, resuspended in 30µl of FACS buffer with the secondary antibodies diluted 1:1000: Donkey Anti-Mouse Alexa Fluor 647 or 568 (Cat. A-31571 or A10037, Thermo Fisher Scientific) and Donkey Anti-Goat 488 Alexa Fluor (Cat. A11055, Thermo Fisher Scientific) and incubated for 1 hr. protected from the light at RT. The stained cells were washed/spined down 3 times with 200µl of FACS buffer and finally resuspended in 200µl of FACS buffer until analyzed. FACS stained samples were measured using FACSCalibur (BD Biosciences) or LSRII (BD Biosciences) and further analyzed using Flowjo software (<https://www.flowjo.com/>, FlowJo LLC).

**Cardiomyocyte beating characterization**—WT, *TBX5-KO*, *GATA4-KO* and *GLYR1<sup>P496L</sup>* CMs contractility beat rate parameters were measured from brightfield acquired videos with the Pulse automated analysis software (<https://www.pulsevideoanalysis.com>). Pulse applies patented computer vision algorithms to measure beating signals and their parameters from videos of beating cardiomyocytes. It uses deep learning for detection of noisy signals, enabling robust and accurate measurement of parameters of beating frequency, beat duration and amplitude. Specifically, it captures and quantifies the biomechanical beating of cardiomyocytes by performing motion analysis on the image sequence to capture changes in the image intensity due to cardiomyocyte contraction and relaxation. This platform has been previously validated in hiPSC-derived cardiomyocytes and showed high correlation with calcium transients and patch-clamp experimental outcomes (Maddah et al. 2015; Burrige et al. 2016). The onset of beating was determined by careful visual daily examination of multiple independent differentiations.

**Plasmid generation: cloning and mutagenesis**—The GFP-GATA4 plasmid (pEN563-pCAGG-eGFP-GATA4) was generated by PCR amplification of the *hGATA4* ORF from the precursor vector pAAV2.1CAG-hGATA4 (FW primer: TGGTGGATCCACCGGTATGTATCAGAGCTTGGCCATGG; REV primer: TGAGCGGCCGCGTTTAACTTACGCAGTGATTATGTCCCCGTG) and cloned following the Cold Fusion Cloning kit (Cat. MC101B-1, Systems Biosciences). Followed

manufacturer's recommendations to linearize the pEN563-CAGG-eGFP vector with PmeI and AgeI (Cat. R0560S and R0552S, New England BioLabs) enzymes to replace the previous contained ORF for the *hGATA4*. The hGLYR1-MYC and hSMARCC1-MYC plasmids (pCMV-T7-cDNA-MYC-IRES2-mCherry-pA; Cat. EX-Z0806-M73 and EX-A6386-M73) were obtained from GeneCopoeia™. The HA-hBRD4 plasmid (p6344 pcDNA4-TO-HA-Brd4FL) was a gift from Peter Howley (Addgene plasmid # 31351; <http://n2t.net/addgene:31351>; RRID: Addgene\_31351) (Rahman et al. 2011) and pCIneo-hCHD7-Kozak ATG 3' HA-bGH polyA was a gift from Peter Scacheri (Addgene plasmid # 89460; <http://n2t.net/addgene:89460>; RRID: Addgene\_89460). The NKX2-5-3xFlag plasmid was generated by modifying the pcDNA3.1 backbone, where the human *NKX2-5* ORF with a C-terminus fused 3xFLAG. The allele variants for *NKX2-5*, *CHD7*, *BRD4*, *SMARCC1* and *GLYR1* were generated by QuikChange II XL Site Directed Mutagenesis (Cat. 200521-5, Agilent Technologies) of the aforementioned plasmids. The EX-Y4729-M06-CHD7-del ATPase domain plasmid was obtained from Kai Jiao laboratory (Liu et al. 2014).

**Nuclear enriched lysis, Immunoprecipitation and Immunoblotting**—WT, GATA4-KO and TBX5-KO hiPSC derived CPs or HEK293 cells transfected with the indicated expression vectors following manufacturer instructions (FuGene HD, Promega) were lysed in Cell Lysis buffer [20mM Tris-HCl pH 8, 85mM KCl, 0.5% NP-40, freshly added protease and phosphatase inhibitors (Cat. 4693132001 Roche and Cat. 4906837001 Sigma-Aldrich)] and incubated on rotator at 4°C for 10 min. Cells were spun down at 2500 rpm for 5min. at 4°C to pellet nuclei and supernatants containing the cytosolic fraction were removed and stored for quality control proposes. Nuclei were resuspended in Nuclear Extraction Buffer (NEB) [20 mM HEPES, pH 7.4, 0.5 M NaCl, 2 mM MgCl<sub>2</sub>, 1 mM CaCl<sub>2</sub>, 0.5 % NP-40, K-Acetate 110mM, 1µM ZnCl<sub>2</sub>, and freshly added Benzoxase 2µl enzyme/ml buffer (Cat. E1014, Millipore), protease and phosphatase inhibitors] and incubated on rotation for 30 min. at 4°C. For CPs, 600µl NEB were used per 4x 12 well plates of initial CPs (~100x10<sup>6</sup>), whereas the nuclei resulting from one 10 mm confluent HEK293 plate were lysed in 300µl of NEB. After incubation, samples were centrifuged at max speed for 10 min. and the nuclear enriched lysates (supernatants) were moved to clean tubes. Nuclear lysates were then diluted 1:3 in Nuclear Dilution Buffer (NDB) [20 mM HEPES, pH 7.9, 1 mM EDTA, 0.2 % NP-40, freshly added protease and phosphatase inhibitors], 1200µl of NDB for CPs; 600µl NDB for HEK293. At this step samples can be stored and later used for immunoprecipitation or for western blotting.

The immunoprecipitations (IP) were done as described before (González-Terán et al. 2016) with modifications. Briefly, 50µl of Dynabeads™ Protein G magnetic (Cat. 10004D, Invitrogen) per IP sample were washed twice with 1ml PBS using a magnetic stand, and resuspended in an equivalent volume to the original volume of beads of PBS. Magnetic beads were then conjugated with a primary antibody: per 50µl of magnetic Dynabeads 4µg of mouse anti- GATA-4 Antibody (G-4) (Cat. Sc-25310 X, Santa Cruz), 4µg mouse anti- TBX5 Antibody (A-6) (Cat. Sc-515536 X, Santa Cruz), 2µg of anti-Myc tag antibody – ChIP Grade (Cat. Ab9132, Abcam) or 2µl of GLYR1 anti-sera 7 (James T. Kadonaga Laboratory) were added and incubated on rotation for 1 hr. at 4°C. Then, the extra non-conjugated antibody was removed by washing the incubated Dynabeads with 1ml of PBS

and twice with 1ml of Nuclear Dilution Buffer (NDB) using the magnetic stand, and resuspended in the initial volume of NDB. At this point the coated Dynabeads are ready for IP. Protein content was quantified by Quick Start™ Bradford Protein Assay Kit 1 (Cat. 5000201, BioRad) from nuclear enriched lysates and 1mg (immunoblot) or 3mg (mass spectrometry) of total protein per IP of endogenous proteins or 150µg (immunoblot) or 1mg (mass spectrometry) per IP of ectopically expressed proteins were aliquoted per condition and volumes were equalized with NDB. Prior to adding 50µl of coated beads per IP sample, 50µl of the prepared nuclear enriched lysates were set aside as “inputs”. IP samples were incubated with coated Dynabeads 4 hrs. for endogenous IP or 1 hr. for IP of overexpressed proteins, on rotator at 4°C. After incubation, samples were placed in the magnetic stand and supernatants removed and saved as “unbound-fraction” for quality control proposes. Beads were washed 2x with 1ml of NDB, and 3 times with 1ml of NDB buffer without detergent. After the last wash, beads were spun down and tubes put on the magnet to remove the remaining liquid. At this point the IP samples can be processed for mass spectrometry or subjected to immunoblotting.

For immunoblotting, samples were subjected to PAGE-SDS. Firstly, 1x of NuPAGE™ LDS Sample Buffer (4X) (Cat. NP0007, Thermo Fisher Scientific) was added to IP samples or to 25-50µg of nuclear enriched lysates and boiled at 95°C for 10 min. and 5 min. respectively. Samples were resolved in pre-cast Novex 4-12% Tris-Glycine gels (Cat. XP04122BOX, Invitrogen) and transferred over night at 35V into polyvinylidene difluoride (PVDF) membranes for endogenous proteins. For ectopically expressed proteins gels were transferred using iBlot® Transfer Stack, PVDF, mini (Cat. IB4010-32, ThermoScientific) and iBlot™ gel transfer device (Thermo Fisher Scientific) for 8 min. at 200V. Membranes were blocked with LI-COR blocking buffer (Cat. 927-40010, LI-COR Biosciences) or 10% BSA in PBS-T (PBS with 0.1% Tween) for 20 min., and incubated overnight on agitation with the primary antibody. Appropriate secondary HRP-conjugated antibody (Abcam) or fluorophore conjugated secondary antibody (LI-COR) was added for 1 hr. at a dilution of 1:5000 followed by detection with ECL Prime Western Blotting Detection Reagent (Cat. RPN2232, GELife Sciences) and exposure to autoradiography film at various time intervals or by digital imaging (LI-COR Odyssey). The following primary antibodies were used for immunoblotting: mouse anti- GATA-4 Antibody (G-4) (Cat. Sc-25310 X, Santa Cruz), mouse anti- TBX5 Antibody (A-6) (Cat. Sc-515536 X, Santa Cruz), anti-Myc tag antibody – ChIP Grade (Cat. Ab9132, Abcam), GLYR1 anti-sera 7 (James T. Kadonaga Laboratory), anti-HA tag antibody – ChIP Grade (Cat. Ab9110, Abcam), anti-MEIS1 antibody (Cat. Ab19867, Abcam), anti-ISL1 antibody (Cat. Ab20670, Abcam), anti- NKX2-5 antibody (Cat. Sc-8697, Santa Cruz), anti-LSD2 antibody (Cat. MBS4751131, MyBioSource), anti-CDK9 antibody (Cat. Sc-13130, Santa Cruz), anti-Cyclin T1 antibody (Cat. Sc-271348, Santa Cruz), and anti-Vinculin Monoclonal Antibody VLN01 (Cat. MA5-11690, Thermo Fisher Scientific).

**Mass Spectrometry**—For mass spectrometry all the IP steps described were processed in protein low-bind sterile Eppendorf tubes (Cat. 022431102, Eppendorf), aliquots specific for mass spectrometry were made for each of the IP buffers and carefully managed to avoid contaminations with filter tips. After the magnetic beads were spun down and the excess of

liquid removed, to each IP sample tube another 1ml of NDB was added, beads resuspended and transferred using wide orifice tips to new protein low-bind sterile Eppendorf tubes. IP samples were returned to the magnetic stand and supernatant carefully removed by aspiration. IP samples were spun down again to collect the remaining liquid, placed in the magnet and the remaining liquid removed. Next, we proceeded with on-bead protein digestion. One bead volume (25 $\mu$ l) of freshly prepared Alkylation Buffer [2M Urea; 50mM Tris, pH8.0; 1mM DTT; 3mM IODO; resuspended in LC-MS high grade water] was added to each IP sample and incubated at RT in the dark for 45 min. while shaking to ensure bead suspension. After incubation, an additional 3mM DTT and 750ng of trypsin per 10 $\mu$ l of bead volume were added to each tube. IP samples were on-bead digested overnight at 37°C on agitation. The next morning, beads were pelleted in a microcentrifuge at 2000 rpm for 4 min. IP samples were placed in a magnetic stand and supernatants carefully transferred to a fresh 0.5ml protein lo-bind tubes. Beads can be saved at 4°C for quality control. To each 0.5ml tube containing the digested IP supernatants, formic acid at a final concentration of 1% was added to stop the digestion process. The processed samples can be stored at this point at -80°C until desalting. Samples were subjected to desalting with OMICS tips (Agilent) and lyophilization in a speed-vacuum concentrator for 30 min. Lyophilized samples were stored at -20°C until proceeding with mass spec analysis. When ready, lyophilized samples were resuspended in 10 $\mu$ l of 0.2% formic acid/ 2%acetonitrile immediately before loading into the mass spectrometry instrument. All APMS samples were measured using the Q Exactive Hybrid Quadrupole-Orbitrap™ Mass Spectrometer.

**RNA extraction, RT-PCR and real-time PCR analysis**—Cells were harvested in TRIzol™ LS reagent (Cat. 10296010, Invitrogen) and total RNA was extracted using the Direct-Zol RNA kit (Cat. R2052, Zymo Research) according to manufacturer instruction. 1000ng of RNA were converted to cDNA using SuperScript™ III First-strand Synthesis SuperMix for qRT-PCR (Cat. 18080400, Invitrogen). For Taqman real-time PCR, 1/50 cDNA was applied for quantitative PCR reaction using Taqman Universal PCR master mix (Cat. 4305719, Life technologies). The PCR was conducted in 7900HT Fast Real-Time system (Applied Biosystem). The Taqman probe for enhanced GFP (eGFP) quantification (Mr04097229\_mr, ThermoFisher Scientific). All gene expressions were normalized with human GAPDH levels (Hs99999905\_m1, ThermoFisher Scientific).

**RNaseq Assay**—Total RNA was TRIzol-extracted (Cat. 10296010, Thermo Fisher Scientific) and further purified using the Direct-Zol RNA kit (Cat. R2052, Zymo Research) with DnaseI in-column treatment according to the manufacturer's instructions and quantified with Nanodrop (Thermo scientific). After RNA quality control with bioanalyzer Agilent 2100 (Agilent Technologies), Paired-end Poly(A)-enriched RNA libraries were prepared with the Ovation RNA-seq System V2 Kit (Cat. 7102-08, NuGEN; strand specific) from the Gladstone Genomic core. The mRNA-seq libraries were analyzed by Agilent Bioanalyzer and quantified using an Illumina Library Quantification Kit (Cat. KK4824, KAPA Biosystems). Libraries were prepared by the Gladstone Genomics Core (<http://labs.gladstone.org/genomics/home>). High-throughput sequencing was done using an Illumina HiSeq 2500 instrument (<http://humangenetics.ucsf.edu/genomics-services/sample-processing/>).



**ChIPseq assay, library preparation and sequencing**—ChIPseq was performed as previously described (Alexanian et al. 2021) with minor modifications. Briefly, CP or hiPSC cells ( $30 \times 10^6$  for cTFs and  $10 \times 10^6$  for GLYR1 and histone marks) were pelleted and suspended in 10ml DMEM and cross-linked in 1% Formaldehyde solution (Cat. 28906, Thermo Fisher Scientific) by rocking in room temperature for 10 min. Then glycine (final concentration 0.125M) was added to quench the cross-link for 5 min. Samples were centrifuged at 1000 rcf for 5 min. at 4°C. Cells were washed with 10ml of cold 1x PBS supplemented with proteinase inhibitors and phosphatase inhibitors (Cat. 4693132001 Roche and Cat. 4906837001 Sigma-Aldrich) and the pellets were snap frozen in liquid nitrogen. All samples were stored at  $-80^\circ\text{C}$  until use. When ready, cell pellets were incubated in cell lysis buffer (20 mM Tris-HCl, pH 8, 85 mM KCl, 0.5% NP-40, protease/phosphatase inhibitors) for 10 min. on a rotator at 4°C. Nuclei were isolated by centrifugation ( $2,500 \times g$ , 5 min., 4°C), resuspended in nuclear lysis buffer (50 mM Tris-HCl, pH 8, 10 mM EDTA, pH 8, 1% SDS, protease/phosphatase inhibitors) and incubated on a rotator for 30 min. at 4°C. Chromatin was sheared using a Covaris S2 sonicator (Covaris Inc) for 15 min. (60 s cycles, 5% duty cycle, 200 cycles/burst, intensity = 6) until DNA was in the 200–700 base pair range. Chromatin was diluted 3-fold in ChIP dilution buffer (0.01% SDS, 1.1% Triton X-100, 1.2mMEDTA, 16.7mMTris-HCl, pH 8, 167 mM NaCl, protease/phosphatase inhibitors) and incubated with the corresponding primary antibody at 4°C overnight under rotation. Antibody-protein complexes were immunoprecipitated using 50 $\mu\text{l}$  of Dynabeads™ Protein A/Protein G (Cat. 10015D, Invitrogen) per sample at 4°C for 2 h under rotation. After incubation, beads were washed five times (2 min./wash under rotation) with cold RIPA buffer [50 mM HEPES-KOH, pH 7.5, 500 mM LiCl, 1 mM EDTA, 1% NP-40, 0.7% Na-deoxycholate], followed by one wash in cold final wash buffer [1xTE, 50 mM NaCl]. Immunoprecipitated chromatin was eluted at 65°C with agitation for 30 min. in elution buffer [50mMTris-HCl pH 8.0, 10mMEDTA, 1% SDS]. High-salt buffer [250mM Tris-HCl, pH 7.5, 32.5 mM EDTA, pH 8, 1.25M NaCl] and Proteinase K (Cat. P8107s, New England Biolabs Inc (NEB)) were added and crosslinks were reversed overnight at 65°C. Samples were treated with Rnase A, and DNA was purified with AMPure XP beads (Cat. A63881, Beckman Coulter). Fragmented ChIP and input DNA were end-repaired, 5'-phosphorylated and dA-tailed with NEBNext Ultra II DNA Library Prep Kit for Illumina (Cat. E7645, New England BioLabs). Samples were ligated to adaptor oligos for multiplex sequencing (Cat. E7335, New England BioLabs), PCR amplified, and sequenced on an Illumina NextSeq 500 at the Gladstone Institutes. Primary antibodies used for ChIP were: GATA4 (Cat. Sc-1237 X, Santa Cruz), TBX5 (Cat. Sc-17866 X, Santa Cruz), H3K36me3 (Cat. Ab9050, Abcam), GLYR1 (anti-serum 7; provided by James T. Kadonaga's laboratory), NKX2-5 (Cat. Sc-8697 X, Santa Cruz), MEIS1 (Cat. Ab19867, Abcam), ISL1 (Cat. AF1837, R&D Systems). The specificity of the antibodies was validated in previous publications (Luna-Zurita et al. 2016; Dupays et al. 2015; Fei et al. 2018).

**siRNA transfection on CPs**—For siRNA knockdown experiments during in cardiac progenitor cells, at day 4 of differentiation cells were detached from the plates with 1ml of accutase (Cat. 07920, Stem Cell Technologies) and quenched with 1ml of B27 (minus insulin) RPMI1640 media (Cat. 11875-119, Life Technologies) per well. All cells were



combined and centrifuge at 300xg, supernatant removed and the pellet resuspended in a volume of B27 (minus insulin) RPMI with 5uM ROCK inhibitor (Y-27632 2HCl, Cat. S1049, [Selleckchem.com](https://www.selleckchem.com)) necessary for 2x the number of wells initially collected. Cells were then seeded in twice the number of 12 well plates originally collected that were pre-coated with fibronectin bovine plasma solution (Cat. F1141, Sigma-Aldrich). Cells were immediately transfected in solution, prior attachment to the well surface using lipofectamine RNAiMax (Cat. 13778075, Invitrogen). For one well of a 12 well plate, mix A (75µl Opti-MEM (Cat. 31985070, Thermo Fisher Scientific) were combined with 3µl of a 10µM siRNA stock) and mix B (75µl Opti-MEM) with 7µl of lipofectamine RNAiMax were prepared. Mix A and B were combined and incubated at RT for 5-10 min. 160µl of lipofectamine siRNA complexes were added dropwise to each well. At day 7 of differentiation, ~72 hrs. after transfection cell were collected, washed, supernatants removed and pellets snap frozen and stored at -80°C until processed. The following siRNA were used: *GATA4* Silencer Select Pre-designed SiRNA (Cat. 4392420, ID s535120, Lot. AS02F2E2, Thermo Fisher Scientific), si *GLYR1* (siRNA ID: SASI\_Hs01\_00116796, Millipore-Sigma) and Silencer Select Negative control #1 siRNA (Cat. 4390843, Thermo Fisher Scientific).

**Luciferase assay**—Transcriptional activity/synergy reporter assays for variant alleles within the reference gene *NKX2-5* and the *GATA4* interactors: *CHD7*, *SMARCC1* and *BRD4* were performed employing the previously described *Ppargc1a* promoter PGL4.23 vector (Padmanabhan et al. 2020). *GATA4*–*GLYR1* transcriptional synergy reporter assay was performed using the pANF638L vector (Knowlton et al. 1991) or pGL4.23[luc2/minP] (Cat. E8411, Promega) modified reporters in which putative intronic RE co-bound by *GATA4* & *GLYR1* were cloned. Briefly, HeLa cells were cultured in 24-well plates at 10<sup>5</sup> cells per well and transfected within 24 hrs. of seeding. Cells were co-transfected with 200 ng of luciferase reporter vector and 20 ng of Renilla luciferase control vector pRL-TK or SV40 (Cat. E2241 or E2231, Promega) in 2.4 µl FuGENE HD (Cat. E2311, Promega) and 43 µl Opti-MEM (Cat. 31985070, Thermo Fisher Scientific). The transfection mix was aliquoted in 5 tubes (1 per condition) and the following conditions were prepared for the luciferase assay using the pANF638L or *Ppargc1a* promoter PGL4.23 vector: 1.) Control: 600ng of empty vector (EV); 2.) *GATA4*: 200ng of GFP-*GATA4* vector plus 400ng EV; 3.) co-factor (*GLYR1*, *CHD7*, *SMARCC1*, *BRD4*): 400ng co-factor vector plus 200ng EV and 4.) *GATA4*+ co-factor: 200ng of GFP-*GATA4* vector with 400ng co-factor vector and 5.) *GATA4*+ mutant co-factor vector: 200ng of GFP-*GATA4* vector with 400ng mutant co-factor vector. Cells were collected at 24hrs, or 48 hrs. following transfection. Samples were processed with the Dual Luciferase Assay System (Cat. E1960, Promega) following manufacturer's instructions and measured with a luminometer (SpectraMax i3).

For analyzing the *GATA4*–*GLYR1* transcriptional synergy within putative intronic REs, *GATA4* - bound intronic regions within *GATA4* & *GLYR1* -bound cardiac development genes that co-localized with H3K27ac, H3K4me1 or H3K4me3, *MED1* and at least co-occupied by two cTFs, were cloned with Cold Phusion (Systems Biosciences) by designing gBlocks (IDT) for each of the selected putative REs (*GATA6*: chr18:19,773,958-19,774,419; *MYL4*: chr17:45,296,131-45,296,972; *TTN*: chr2:179,493,152-179,494,179 ) flanked by homology arms complementary to the pGL4.23[luc2/minP] luciferase reporter vector (Cat.

E8411, Promega). HeLa cells were plated as indicated for the pANF638L reporter and transfection mixed prepared as indicated above. The transfection mix was aliquoted in 5 tubes (1 per condition) and the following conditions were prepared for each of the cloned pGL4.23 luciferase reporter vectors: 1.) Control: 200ng of empty vector (EV); 2.) GATA4: 200ng of GFP-GATA4 vector plus Adenovirus control (Ad-EF1a-eGFP, MOI 25); 3.) GLYR1: Adenovirus GLYR1 WT (Ad-eGFP-EF1-h-GLYR1, MOI 25) and 4.) GATA4+GLYR1: 200ng of GFP-GATA4 vector and Adenovirus GLYR1 WT (MOI 25); 5.) GATA4+GLYR1 P496L: 200ng of GFP-GATA4 vector Adenovirus GLYR1 P496L (Ad-eGFP-EF1-h-GLYR1 P496L, MOI 15). The MOIs were determined by HeLa cell infection followed by quantitative PCR amplification with the eGFP Taqman probe (Mr04097229\_mr, ThermoFisher Scientific). MOIs rendering comparable eGFP expression levels were chosen for the luciferase experiments (Relative to GAPDH Avg levels for Ad GLYR1 WT MOI 25: 50.134 and Ad GLYR1 P496L MOI 15: 57.333; n=3). HeLa cells were collected 48h after transfection/infection and processed as indicated for the pANF638L vector.

**Single-cell RNAseq Cell Preparation**—Cells were wash with PBS and detached with 0.25% Trypsin (CMs) or accutase (hiPSCs and CPs) incubated for 5 minutes. Quenched with 1 mL of PBS + 1% FBS and spin down at 800-1000 rpms for 3 min. Cells were washed twice with PBS + 1% FBS, resuspend in cold PBS + 1% FBS (use 0.5mL PBS per well collected from a 12-well-plate) and put on ice. Cells were filtered to avoid clumps and manually count with hemocytometer. Both live and dead cells were included in the count and cells were diluted to 1e6 cells/ml. A total of 60,000 cells were used to proceed with the scRNAseq library construction protocol.

**Single-cell transcriptome library preparation and sequencing**—Single-cell droplet libraries from the hiPSC (1x *GLYR1<sup>KO</sup>*, 2x *GLYR1<sup>WT</sup>*, 2x *GLYR1<sup>P496L</sup>*), CM-differentiation day 6 (3x *GLYR1<sup>WT</sup>*, 3x *GLYR1<sup>P496L</sup>*) and day 18 (3x *GLYR1<sup>WT</sup>*, 3x *GLYR1<sup>P496L</sup>*) cell suspensions (Figure 7 and S6) were generated in the 10X Genomics Chromium controller according to the manufacturer's instructions in the Chromium Single Cell 3' Reagent Kit v.3 User Guide. Additional components used for library preparation include the Chromium Next GEM Single Cell 3' GEM, Library & Gel Bead Kit v3.1, (PN-1000121, 10X Genomics) and the Chromium Next GEM Chip G Single Cell Kit (PN-1000120, 10X Genomics). Libraries were prepared according to the manufacturer's instructions using the Chromium Single Cell 3' Library and Gel Bead Kit v.3.1 (PN-1000121, 10X Genomics) and 3'v3.1 Single Index Kit (PN-1000213, 10X Genomics). Final libraries were sequenced on the NovaSeq (Illumina, software v1.5). Sequencing parameters were selected according to the Chromium Single Cell v.3.1 specifications. All libraries were sequenced to a mean read depth of at least 50,000 total aligned reads per cell.

**Heart Histology**—To examine hearts at postnatal day 1, hearts were dissected from the animals and fixed overnight at 4°C in 4% paraformaldehyde. They were then washed twice in PBS and stored in 70% ethanol until processing. Hearts were paraffin embedded and sectioned to obtained a four-chamber view. Heart sections were stained with hematoxylin and eosin and imaged with a slide scanner.

**Echocardiography**—For echocardiography, newborn mice were imaged using the Vevo 3100 High Resolution Imaging System (FujiFilm VisualSonics Inc.), an ultra-high frequency linear array transducer (MX700). All the echocardiogram analyses were performed blinded (mice assigned to an alphanumeric code) until statistical analysis.

## COMPUTATIONAL ANALYSIS

**Mass spectrometry analysis of affinity purifications.**—Peptides from affinity purifications were analyzed on a Q-Exactive Plus (Thermo Fisher) mass spectrometer. The Q-Exactive Plus system was equipped with an Easy1200 nLC system (Thermo Fisher) and an analytical column (25 cm x 75  $\mu$ m I.D. packed with ReproSil Pur C18 1.9  $\mu$ m, 120Å particles, Dr. Maisch). A gradient was delivered from 2% to 30% acetonitrile over 53 minutes at a flow rate of 300 nl/min. All MS spectra were collected with Orbitrap detection, while the 20 most abundant ions were fragmented by HCD and detected in the Orbitrap. Peptide and protein identification searches, as well as label-free quantitation were performed using the MaxQuant data analysis algorithm (version 1.5.8.0) (Cox and Mann 2008). Data were searched against a database containing SwissProt Human sequences (downloaded 02/2017) concatenated to a decoy database where each sequence was randomized in order to estimate the false discovery rate (FDR).

Variable modifications were allowed for methionine oxidation and protein N-terminus acetylation. A fixed modification was indicated for cysteine carbamidomethylation. Full trypsin specificity was required. The first search was performed with a mass accuracy of  $\pm$  20 parts per million (ppm) and the main search was performed with a mass accuracy of  $\pm$  4.5 parts per million. A maximum of 5 modifications were allowed per peptide. A maximum of 2 missed cleavages were allowed. The maximum charge allowed was 7+. Individual peptide mass tolerances were allowed. For MS/MS matching, a mass tolerance of  $\pm$  20 ppm was allowed and the top 12 peaks per 100 Da were analyzed. MS/MS matching was allowed for higher charge states, water and ammonia loss events. The data were filtered to obtain a peptide, protein, and site-level false discovery rate of 0.01. The minimum peptide length was 7 amino acids.

**Selection of Interactome Proteins**—APMS data was analyzed 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, we analyze at the protein level and select proteins that interact at a BFDR cutoff of  $\leq$  0.001; to create the TBX5 interactome, we analyze at the peptide level and select those that interact at a BFDR cutoff of  $\leq$  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, we additionally filter 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 (Maere et al. 2005; Shannon et al. 2003) 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, mesoderm and neural crest cells in the developing heart were used to identify seven cell type populations (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 is 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 ( $FDR \leq 0.05$ ) with less than a 0.5 log-fold change drop in intensity were not considered to be interactors.

**Gene expression tissue distribution and specificity**—The categories for gene expression tissue distribution and tissue specificity defined by the *Tissue Atlas* within the *Human Protein Atlas* were used to classify the specified 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 subcategory 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 have, or do not have, detectable levels ( $N \geq 1$ ) of transcribed mRNA molecules. All 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

**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. Protein-coding mutations were filtered based on a Mapping Quality score  $> 59$  and Genotype Quality  $> 90$ , then annotated 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  $4 \times 10^{-4}$ , 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%).

### Permutation-based test

**Case-Control Permutation:** We tested the adjusted odds ratio of observing a *de novo* mutation in an interactome gene in CHD probands relative to controls. We 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. P-values were adjusted for multiple testing using the Bonferroni method. 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, we adjusted the odds ratios of the synonymous and protein-altering variant data by a factor that restricts the synonymous odds ratio to 1 (the 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, we repeated the analysis after removing *de novo* variants occurring in known known Human/Mouse CHD genes (sourced from Jin et al. 2017, Supplementary Data Set 2: 253 Curated) (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, we identify 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. We further filter the list of matches based on similarity of expression levels in wildtype cardiac progenitor cells, such that the measured transcripts per million (tpm) is equivalent to the order of one one-hundredth. For genes with fewer than 100 possible matches, we relax these requirements by  $\pm 0.5 \times 10^N$  (where N is 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.

### **Interactome CHD Candidate Variant and Harboring Genes Characterization**

—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. 2020). Similarly, haploinsufficiency predicts the deleteriousness of having only a single functional copy of a gene. We use the predicted haploinsufficiency values from Huang et al. 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 et al. 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. We use 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.

**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 PGC, 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 et al. 2008) (see Methods: Interactome CHD Candidate Variant Characterization). 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.

Firstly, 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 these conditions:

Factor	Conditions
0.75	Proband has another rare (MAF $10^{-5}$ ) inherited loss-of-function OR missense <i>de novo</i> 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 <i>de novo</i> mutation in an interactome gene or rare inherited loss-of-function mutation in a previously-identified CHD gene
0.25	Proband has a <i>de novo</i> missense mutation in a previously-identified CHD gene
0.10	Proband had a <i>de novo</i> 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.

Secondly, 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.

**Precision- Recall Analysis**—The ability of the Variant Prioritization Score to predict variants known to result in CHD was performed under two scenarios using a Precision-Recall (PR) Analysis. In the first scenario, prediction was performed over all missense DNVs, and in the second one, prediction was restricted to the GT-PPI missense DNVs. The PR analysis was performed using the `pr.curve` function that is part of the PRROC (Grau et al. 2015) package in R. The performance of the score was quantified in terms of the Area-Under the Curve (AUC) for the PR curves generated under the two scenarios, relative to the expected AUC from a random classifier under the corresponding scenarios. The expected AUC for the PR curve from a random classifier is equal to the fraction of variants known to cause CHD among all variants used in the prediction.

**GLYR1 Model Organisms Alignment**—Sequences of several vertebrate model organisms containing the rigid loop (bridging the two tetramerization alpha helice bundles) of the NPAC proteins' dehydrogenase domain were aligned using CLC Sequence Viewer 8.0. Amino acids 490-529 were aligned, partially spanning exons 14 and 15 (490-495, 496-5229 respectively) in the *H. sapiens* sequence. Alignments were created with the "Alignment" function with a gap open cost of 10.0, gap extension cost of 1.0, end gap cost as any other, and the very accurate (slow parameter).

Used mRNA (NM) and predicted mRNA (XM) Sequences

Chimp (*Pan Traglodyte*): XM\_016929357

Gorilla (*Gorilla gorilla*): XM\_019012358

Human (*Homo sapien*): NM\_032569

Mouse (*Mus musculus*): NM\_001359747.1

Rat (*Rattus norvegicus*): NM\_001007800

Chicken (*Gallus gallus*): NM\_001006572

Frog (*Xenopus laevis*): NM\_001030494

Zebrafish (*Danio rerio*): XM\_005164104

**GLYR1 Structural Model**—The wildtype GLYR1 structure was imported from the RCSB Protein Data Bank, entry 2UYY, the structure being elucidated through X-ray diffraction (Tickle, J. et al. 2007). The structure and domains of the NPAC monomer were edited using the PYMOL Molecular Graphics System Version 2.3.5. Domains of the NPAC dehydrogenase domain are defined as by Zhang et al 2014 (Zhang et al. 2014). The protein is shown through the cartoon function, displaying the general tertiary structure of the protein. Amino acid proline 496 and its mutant proline 496 leucine are shown in gold and through the stick function displaying the secondary structure of the amino acids to delineate the significance in change of structure. In the focused images of the rigid loop of the alpha helices tetramerization bundle, the amino acids are again shown through the stick function to delineate the secondary structure interactions of the amino acids.

**Molecular structural dynamics methods**—The initial protein structure for all-atom MD simulations in explicit water of NPAC was downloaded from the Protein Data Bank, code 2uyy.pdb.

Missing atoms and side chains were added using the Protein Preparation Wizard of the Maestro Suite of Programs (v. 2019–4). Proline to Leucine mutation was also performed using Maestro (Maestro Schrödinger, LLC 2019; Sastry et al. 2013). The simulations were run using the same protocol for both the WT and mutated monomer (subunit A).

All systems were allowed to relax with 2000 steps of steepest descent followed by another 2000 steps of conjugate gradient energy minimization. The temperature of the systems was gradually raised to 300 K in the NVT ensemble in 1.2 ns at 1 fs time-step, using the Langevin thermostat. In particular, six runs of 200 ps were performed increasing the temperature of 50 K at each step ( $T = 50, 100, 150, 200, 250,$  and  $300$  K, respectively). At 300 K, the density of the system was adjusted with 1 ns at 2 fs time-step under NPT conditions by weak coupling to a bath of constant pressure ( $P_0 = 1$  bar, coupling time  $\tau_p = 0.5$  ps). The production runs were thus carried out in the NVT ensemble. Bonds involving hydrogen atoms were constrained with the SHAKE algorithm (Miyamoto and Kollman 1992), allowing a time step of 2 fs. Electrostatic forces were computed using the particle mesh Ewald algorithm with a truncation cut-off of  $10\text{\AA}$  (Darden et al. 1993). The initial velocity of all atoms was obtained from a Maxwellian distribution at the initial temperature of 300 K.

MD simulations were run in 3 independent replicas of 500 ns each (1.5  $\mu\text{s}$  in total per system), Specifically, MD simulations were performed using Amber18 pmemd.CUDA with the all atom ff14SB force field under periodic boundary conditions (Case et al. 2017). The triclinic simulative box, filled with TIP3P (Jorgensen et al. 1983) water molecules and rendered electroneutral by addition of  $\text{Na}^+$  counterions consists of a final number of atoms of about 41 300 (monomer WT and P496L mutant), particles for each system.

The atomic positions were saved every 10 ps. The equilibrated parts of the trajectories were used for subsequent analyses. Equilibration of the trajectories was checked by monitoring the equilibration of the RMSD with respect to the initial structure and of the internal protein energy. The equilibrated parts of each trajectory for the two systems were next combined into a meta-trajectory, which was subsequently used for all the reported characterizations. Classical structural analyses were carried out with the tools in the Amber18 and Gromacs 4.5.5 package (Bekker et al. 1993) or with code written in-house.

The *root mean square deviation* (RMSD) of the backbone of the protein with respect to first frame of the trajectory along the simulation time has been calculated by least-square fitting the structure to the reference structure ( $t_2 = 0$ ) and subsequently calculating the RMSD

$$\text{RMSD}(t_1, t_2) = \sqrt{\frac{1}{M} \sum_{i=1}^N m_i \| r_i(t_1) - r_i(t_2) \|^2}$$

where  $M = \sum_{i=1}^N m_i$  and  $r_i(t)$  is the position of atom  $I$  at time  $t$ .

The *RMSF* which is a measure of the displacement of each residue averaged over the number of atoms considered, has been calculated relative to the average structure, in the equilibrated part of the simulation.

**Differential gene expression**—In order to identify genes differentially expressed between WT CP versus GATA4-KO or TBX5 CPs ( $n=5$ ); siControl vs siGATA4 CPs ( $n=3$ ) and siControl.2 vs siGLYR1 CPs ( $n=2$ ), the analyses start with raw reads/sequences in FASTQ format. Trimming of known adapters and low-quality regions of reads was

performed using Fastq-mcf (Aronesty 2013). Sequence quality control was assessed using the program FastQC (Andrews 2007) and rSeQC (Wang et al. 2012). Alignment of the provided samples to the reference genome was performed using STAR 2.5.2a (Dobin et al. 2013). Reads were aligned to the human hg19 reference assembly indicated in the header of the differential expression file. Reads were assigned to genes using *featureCounts* (Liao et al. 2014), part of the Subread suite (<http://subread.sourceforge.net/>). Gene-level counts were arrived at using Ensembl gene annotation, in GTF format. Differential expression was assessed using edgeR (Robinson et al. 2010), an R package available through Bioconductor. Genes where there were not at least two samples with at least 5 (raw) reads were filtered out from further analyses. The reads counts of remaining ones are normalized for sample-to-sample variation using *calcNormFactors* in edgeR (Robinson et al. 2010). The mean gene expression was modeled as a function of siRNA status (siRNA treatment vs scramble control) and sample id. Genes whose expression is associated with siRNA status were determined by the likelihood ratio test (Smyth 1996; Robinson and Smyth 2007; Robinson and Smyth 2008) implemented in edgeR using a FDR < 0.05 and LogFC < -0.25 threshold.

**Pathway enrichment analysis**—Functional enrichment gene-set analysis for GO (Gene Ontology) terms was performed using ToppGene Suite (<https://toppgene.cchmc.org/enrichment.jsp>) using all *Homo sapiens* genes as background. Statistically significant (Bonferroni q-value < 0.05) categories within the GO:Biological Process section were extracted and replotted.

**ChIPseq analysis**—For the ChIPseq analysis, trimming of known adapters and low-quality regions of reads was performed using Fastq-mcf. Sequence quality control was assessed using FastQC (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>). Alignment to the hg19 reference genome was performed using Bowtie 2.2.4 (Langmead and Salzberg 2012). Peaks were called using GEM (Guo et al. 2012) for TFs and BCP (Xing et al. 2012) for GLYR1 and H3K36me3 ChIPseq signals. Read counts per peak were generated with *featureCounts* (Liao et al. 2014) and normalized to account for differences in sequencing depth between samples using upper quartile normalization separately for the ChIP and input sample. Bound regions were determined using empirical Bayes F-tests for a quasi-likelihood negative binomial generalized log-linear model of the count data as implemented in edgeR. Specifically, we tested for a significant (i.e., non-zero at FDR < 5%) log<sub>2</sub> fold-increase in normalized peak signal for ChIP versus the corresponding input sample. 2 or 3 separate samples (and relative inputs) were ran from independent ChIP assays.

**- GATA4 and GLYR1 ChIPseq Genomic Features (Figure S5E):** We obtained the genomic features associated with GATA4 and GLYR1 ChIP-seq peaks using the *annotatr* (Cavalcante and Sartor 2017) package in R.

**- Metagene Plot Analysis (Figure S5D&J and 6D):** To generate metagene plots, BED files were generated containing regions of interest. The *computeMatrix scale-regions* module of deepTools (Ramírez et al. 2016), which shrinks or stretches all regions in the input BED file to the same length was used to summarize the ChIP signal profile for each region. The

ChIP signal was defined in terms of input subtracted tag densities. Specifically, the human genome, hg19 is divided into 20-bp bins. The tag density or normalized difference between the number of the reads in the ChIP sample and the input sample is computed as:

$$\# \text{ Bins In Genome} \times \left( \frac{\# \text{ chip Reads}}{\# \text{ total Chip Reads}} - \frac{\# \text{ input Reads}}{\# \text{ total Input Reads}} \right)$$

**- Analysis of Differential ChIPseq Signal:** The counts of reads mapping to genes for each of the replicates for each of the ChIPs (GLYR1 and H3K36me3) at the hiPSC and CPC stages under wild-type conditions and at the CPC stage under P496L mutant conditions were obtained using featureCounts (Liao et al. 2014) using their corresponding aligned reads in bam files. The counts of reads for each of replicates used for assaying gene expression at the two stages in the GSE137920 (Lau et al. 2019) data set were downloaded from GEO (Barrett et al. 2013). Genes where there were not at least two samples with at least 5 (raw) reads in the GLYR1 ChIPs were filtered out from further analyses. The read counts for the remaining genes corresponding to each of the three signals (GLYR1, H3K36m3 and Gene expression) are separately normalized using *calcNormFactors* in edgeR (Robinson and Oshlack 2010). Genes for which the mean GLYR1 signal in their bodies were significantly changed from CPC stage relative to hiPSC stage (or in the P496L mutant versus wild-type conditions) were determined the likelihood ratio test implemented in edgeR using FDR < 0.1 threshold. The row-normalized log2 transformed Counts-Per-Million (CPM) of GLYR1 signal for these significantly associated (with changing GLYR1 signal) genes were clustered using *kmeans* with 3 clusters implemented in R (R Core Team 2020). The resulting cluster definitions (using the GLYR1 signal) and order of genes were used to visualize the signals (in row-normalized log2 CPM units) in the H3K36me3 and RNA-seq data.

**- Definition of GATA4 and GLYR1 -bound gene categories (list used in most of the panel Figure 6):** GLYR1 bound genes in Figure 6B are defined as those genes in clusters 2 and 3 in Figure 6A which displayed enriched binding signal at the CPC stage relative to the hiPSC stage. Genes with Gata4 ChIP peaks from the first intron to Transcription End Site (TES) were defined as Gata4 bound genes.

**- Scatter plot analysis (ChIPseq/RNAseq) (Figure S5 A):** Differential gene expression between the hiPSC (day 0) and the CPC stage (day 7) was determined using quasi-likelihood F-test implemented in the *glmQLFTest* in edgeR (Lun et al. 2016) using the count matrix association with the GSE137920 (Lau et al. 2019) filtered for low counts (at least two samples with at least 5 (raw) reads) genes normalized using the *calcNormFactors*. Up-regulated genes were determined using thresholds of 1.5 for log2 fold-change ( $\log_2\text{FC} > 1.5$ ) and 0.05 for FDR ( $\text{FDR} < 0.05$ ), while down-regulated genes were determined using thresholds of -1.5 for log2 fold-change ( $\log_2\text{FC} < -1.5$ ) and 0.05 for FDR. The raw read counts for all replicates of GLYR1 at the two stages for all genes that were part of the differential gene expression analyses above, were normalized using *calcNormFactors* and Reads Per Kilobase of transcript, per Million mapped reads (RPKM) was calculated using the *rpkm* function along with the gene lengths based on the Ensembl gene annotation. Similarly, RPKM values were estimated for all replicates of the H3K36me3 ChIP at the two

stages. The scatter plots in Figure S5 use the mean log<sub>2</sub>-transformed RPKM values across replicates at a given stage.

**- Statistical Analysis of GATA4 & GLYR1 ChIPseq overlap (Figure 6B):** The significance of the overlap of genes bound within their bodies (first intron to Transcription End Site (TES)) by Gata4 and GLYR1 was determined using the Fisher's exact test implemented in R (R Core Team 2020) on 30,611 genes which had detectable reads (5 reads in at least two samples) at the day 7 in the GSE137920 (Lau et al. 2019) data set.

**- Motif Analyses:** Motif enrichment in the ChIP peak regions were performed using the *findMotifsGenome* function implemented in homer (version v4.11.1) (Heinz et al. 2010).

### Single-cell RNAseq analysis

**Preprocessing:** 6-8 replicate samples (from independent differentiation runs involving either wild-type or P496L mutant cells) were each processed in 3 batches. Reads from each sample were aligned to the hg38 human reference (version 2020-A from 10X Genomics website) using Cell Ranger version 5.0.1. The resulting count matrices for all samples within each batch were aggregated without depth normalization. The aggregated matrix was analyzed with the Seurat package (version 4.0.1) in R (Hao et al. 2021). The Seurat object was created keeping cells with the number of detected features between 30<sup>th</sup> and 95<sup>th</sup> quantiles of detected features across all cells in the batch and whose percent mitochondrial reads were below 30%. The remaining counts were normalized with the *SCTransform* function after setting the number of variable features to 3000, using the *glmGamPoi* method to estimate parameters of the fitted negative binomial distributions and regressing out the effects of the percent mitochondrial reads per cell. The *SCTransform* based normalized counts per sample were subject to correction for systematic differences between the three batches using the *RunHarmony* function as implemented in the harmony (Korsunsky et al. 2019) R package. Clustering was done using a shared nearest neighbor graph built with the top 30 dimensions of the harmony-based correction and the original Louvain algorithm for modularity optimization with resolution parameter set to 0.4. UMAP embedding was then generated using the top 30 dimensions of the harmony-based correction.

**Association with cell state:** The differences in the proportion of P496L mutant cells versus WT cells in each of the identified cell clusters at each of the sampled time-points in the differentiation protocol was quantified using an odds ratio estimate. The log odds ratios were estimated in the context of a generalized linear mixed effects model assuming the binomial family of probability distribution for the numbers of cells, as implemented in the *glmer* function in the lme4 (Bates et al. 2015) package in R. The sample or differentiation run of origin of each cell was modeled as a random effect in these models.

**Differential expression:** The association of gene expression with differences between wild-type and mutant cells were estimated using the *FindMarkers* function in Seurat (Hao et al. 2021) that implemented the MAST (Finak et al. 2015) method where the batch of the origin of each cell was modeled as a latent variable.



**Association with Chip-seq data:** The differences between the average (across replicates from independent differentiation runs) ChIP signal (quantified in Transcripts-Per-Million (TPM) units) within gene bodies among the wild-type cells and the P496L cells were visualized as scatter plots separately for genes that were up-regulated ( $\log_{2}FC > 0.125$  and  $FDR < 0.05$  in the output of the *FindMarkers* function) and down-regulated ( $\log_{2}FC < -0.125$  and  $FDR < 0.05$ ).

## QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical parameters including the exact value of n, precision measures (mean  $\pm$  SEM) and statistical significance are reported in the Figures and the Figure Legends. All calculations were performed using R or GraphPad Prism software. When several conditions were to compare, we performed a one-way ANOVA, followed by Tukey range test to assess the significance among pairs of conditions. The significance of the PPIN enrichment in CHD-associated DNVs was calculated with a permutation-based test as explained in the Computational Analysis Methods section. All the p-values related to the violin plots showing features typical of disease genes were obtained using a two-sided Mann-Whitney-Wilcoxon test with Bonferroni correction. The significance of the GATA4 & GLYR1 ChIPseq overlap was estimated using the Fisher.Exact function in R. The level of significance in all graphs is represented as follow: \*  $P < 0.05$ , \*\*  $P < 0.01$ , \*\*\*  $P < 0.001$ , \*\*\*\*  $P < 0.0001$ .

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## ACKNOWLEDGMENTS

We thank the Srivastava laboratory and Gladstone colleagues for critical discussions and feedback; Guadalupe Sabio and Mauro Costa for critical reading of the manuscript; B. Taylor for editorial assistance; Jim Kadonaga for kindly sharing GLYR1 anti-serum; Irfan Kathiriya and Kai Jiao for kindly sharing the *Nppa* luciferase reporter plasmid and the CHD7 ATP mutant plasmid, respectively; the Gladstone Genomics Core, Bioinformatic Core, Stem Cell Core, Microscopy & Histology Core, Mouse Transgenics Core and Flow Cytometry Core for their technical expertise and the Gladstone Animal Facility for support with mouse colony maintenance; David E. Gordon for sharing an optimized CRISPR/Cas9 RNP's hiPSCs knockout generation protocol. We thank Françoise Chanut for manuscript editorial support and Ana Catarina Silva (ana@anasilvaillustrations.com) for helping with figure editing and design.

## SOURCES OF FUNDING

B.G.T. is supported by the American Heart Association (18POST34080175) and AHA/CHF Congenital Heart Defect Research Award (#818798).

M.A. is supported by the Swiss National Science Foundation (P400PM\_186704).

K.S.P. is supported by NIH P01 HL098707, P01 HL146366, UM1 HL098179, Gladstone Institutes, and the San Simeon Fund.

D.S. is supported by NIH/NHLBI P01 HL098707, P01 HL146366, R01 HL057181, R01 HL127240, and by the Roddenberry Foundation, the L.K. Whittier Foundation, and the Younger Family Fund.

N.J.K. is supported by grants from the National Institutes of Health (P01 HL146366 and 1U01MH115747).

B.G.B. was supported by NIH/NHLBI P01 HL098707, P01 HL146366, and the Younger Family Fund.

A.P. is supported by the NIH (K08HL157700), Tobacco-Related Disease Research Program (578649), A. P. Giannini Foundation (P0527061), Michael Antonov Charitable Foundation and Sarnoff Cardiovascular Research Foundation.

S.U.M was supported by the American Heart Association (20POST35210452) and the Boston Children's Hospital Office of Faculty Development.

C.E.S and J.G.S are supported by NIH U01 HL098147 and UM1 HL098166.

This work was also supported by NIH/NCRR grant C06 RR018928 to the Gladstone Institutes.

## REFERENCES

- Akerberg BN, Gu F, VanDusen NJ, Zhang X, Dong R, Li K, Zhang B, Zhou B, Sethi I, Ma Q, et al. (2019). A reference map of murine cardiac transcription factor chromatin occupancy identifies dynamic and conserved enhancers. *Nat. Commun* 10, 4907. [PubMed: 31659164]
- Alexanian M, Przytycki PF, Micheletti R, et al. 2021. A transcriptional switch governs fibroblast activation in heart disease. *Nature* 595(7867), pp. 438–443. [PubMed: 34163071]
- Andrews S (2007). A quality control tool for high throughput sequence data. *bioinformatics*.
- Ang Y-S, Rivas RN, Ribeiro AJS, Srivas R, Rivera J, Stone NR, Pratt K, Mohamed TMA, Fu J-D, Spencer CI, et al. (2016). Disease model of GATA4 mutation reveals transcription factor cooperativity in human cardiogenesis. *Cell* 167, 1734–1749.e22. [PubMed: 27984724]
- Aronesty E 2013. Comparison of Sequencing Utility Programs. *The open bioinformatics journal* 7(1), pp. 1–8.
- Barrett T, Wilhite SE, Ledoux P, et al. (2013). NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Research* 41(Database issue), pp. D991–5. [PubMed: 23193258]
- Barshir R, Shwartz O, Smoly IY, and Yeger-Lotem E (2014). Comparative analysis of human tissue interactomes reveals factors leading to tissue-specific manifestation of hereditary diseases. *PLoS Comput. Biol* 10, e1003632. [PubMed: 24921629]
- Basson CT, Bachinsky DR, Lin RC, Levi T, Elkins JA, Soultz J, Grayzel D, Kroumpouzou E, Traill TA, Leblanc-Straceski J, et al. (1997). Mutations in human TBX5 [corrected] cause limb and cardiac malformation in Holt-Oram syndrome. *Nat. Genet* 15, 30–35. [PubMed: 8988165]
- Basson CT, Huang T, Lin RC, Bachinsky DR, Weremowicz S, Vaglio A, Bruzzone R, Quadrelli R, Lerone M, Romeo G, et al. (1999). Different TBX5 interactions in heart and limb defined by Holt-Oram syndrome mutations. *Proc. Natl. Acad. Sci. USA* 96, 2919–2924. [PubMed: 10077612]
- Bates D, Mächler M, Bolker B and Walker S (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1), pp. 1–48.
- Bekker H, Berendsen H, Dijkstra E, et al. (1993). Gromacs: A parallel computer for molecular dynamics simulations – ScienceOpen [Online]. Available at: <https://www.scienceopen.com/document?vid=59290415-039c-4900-8e95-d649687a2473> [Accessed: 12 September 2020].
- Bouman A, Alders M, Oostra RJ, van Leeuwen E, Thuijs N, van der Kevie-Kersemaekers A-M, and van Maarle M (2017). Oral-facial-digital syndrome type 1 in males: Congenital heart defects are included in its phenotypic spectrum. *Am. J. Med. Genet. A* 173, 1383–1389. [PubMed: 28371265]
- Bruneau BG, Logan M, Davis N, Levi T, Tabin CJ, Seidman JG, and Seidman CE (1999). Chamber-specific cardiac expression of Tbx5 and heart defects in Holt-Oram syndrome. *Dev. Biol* 211, 100–108. [PubMed: 10373308]
- Bruneau BG, Nemer G, Schmitt JP, Charron F, Robitaille L, Caron S, Conner DA, Gessler M, Nemer M, Seidman CE, 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, 709–721. [PubMed: 11572777]
- Broyis J, Skene NG, Hansen TF, Kogelman LJA, Watson HJ, Liu Z, Eating Disorders Working Group of the Psychiatric Genomics Consortium, International Headache Genetics Consortium, 23andMe Research Team, Brueggeman L, et al. (2020). Genetic identification of cell types underlying brain complex traits yields insights into the etiology of Parkinson's disease. *Nat. Genet.* 52, 482–493. [PubMed: 32341526]

- Burridge PW, Li YF, Matsa E, et al. (2016). Human induced pluripotent stem cell-derived cardiomyocytes recapitulate the predilection of breast cancer patients to doxorubicin-induced cardiotoxicity. *Nature Medicine* 22(5), pp. 547–556.
- Cao J, O’Day DR, Pliner HA, Kingsley PD, Deng M, Daza RM, Zager MA, Aldinger KA, Blecher-Gonen R, Zhang F, et al. (2020). A human cell atlas of fetal gene expression. *Science* 370.
- Case DA, Cerutti DS, Cheatham TEI, et al. (2017). Amber18. University of San Francisco.
- Castillo-Robles J, Ramírez L, Spaink HP, and Lomeli H (2018). *smarce1* mutants have a defective endocardium and an increased expression of cardiac transcription factors in zebrafish. *Sci. Rep* 8, 15369. [PubMed: 30337622]
- Cavalcante RG and Sartor MA (2017). annotatr: genomic regions in context. *Bioinformatics* 33(15), pp. 2381–2383. [PubMed: 28369316]
- Chauveau C, Bonnemann CG, Julien C, Kho AL, Marks H, Talim B, Maury P, Arne-Bes MC, Uro-Coste E, Alexandrovich A, et al. (2014). Recessive TTN truncating mutations define novel forms of core myopathy with heart disease. *Hum. Mol. Genet* 23, 980–991. [PubMed: 24105469]
- Chen J, Yuan H, Xie K, Wang X, Tan L, Zou Y, Yang Y, Pan L, Xiao J, Chen G, et al. (2020). A novel TAB2 nonsense mutation (p.S149X) causing autosomal dominant congenital heart defects: a case report of a Chinese family. *BMC Cardiovasc. Disord* 20, 27. [PubMed: 31959127]
- Christianson A, and Howson CP (2006). March of dimes. *Global Report on Birth*.
- Clouthier DE, Hosoda K, Richardson JA, Williams SC, Yanagisawa H, Kuwaki T, Kumada M, Hammer RE, and Yanagisawa M (1998). Cranial and cardiac neural crest defects in endothelin-A receptor-deficient mice. *Development* 125, 813–824. [PubMed: 9449664]
- Cox J and Mann M (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology* 26(12), pp. 1367–1372.
- Darden T, York D and Pedersen L (1993). Particle mesh Ewald: An N-log(N) method for Ewald sums in large systems. *J. Chem. Phys* 98(12), p. 10089.
- de Soysa TY, Ranade SS, Okawa S, Ravichandran S, Huang Y, Salunga HT, Schrickler A, Del Sol A, Gifford CA, and Srivastava D (2019). Single-cell analysis of cardiogenesis reveals basis for organ-level developmental defects. *Nature* 572, 120–124. [PubMed: 31341279]
- Deciphering Developmental Disorders Study (2015). Large-scale discovery of novel genetic causes of developmental disorders. *Nature* 519, 223–228. [PubMed: 25533962]
- Diets IJ, Prescott T, Champaigne NL, Mancini GMS, Krossnes B, Fri R, Kocsis K, Jongmans MCJ, and Kleefstra T (2019). A recurrent de novo missense pathogenic variant in SMARCB1 causes severe intellectual disability and choroid plexus hyperplasia with resultant hydrocephalus. *Genet. Med* 21, 572–579. [PubMed: 29907796]
- Dobin A, Davis CA, Schlesinger F, et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29(1), pp. 15–21. [PubMed: 23104886]
- Dsouza NR, Zimmermann MT, and Geddes GC (2019). A case of Coffin-Siris syndrome with severe congenital heart disease and a novel SMARCA4 variant. *Cold Spring Harb Mol Case Stud* 5.
- Dupays L, Shang C, Wilson R, et al. (2015). Sequential Binding of MEIS1 and NKX2-5 on the *Popdc2* Gene: A Mechanism for Spatiotemporal Regulation of Enhancers during Cardiogenesis. *Cell reports* 13(1), pp. 183–195. [PubMed: 26411676]
- Eilbeck K, Quinlan A, and Yandell M (2017). Settling the score: variant prioritization and Mendelian disease. *Nat. Rev. Genet* 18, 599–612. [PubMed: 28804138]
- Enane FO, Shuen WH, Gu X, Quteba E, Przychodzen B, Makishima H, Bodo J, Ng J, Chee CL, Ba R, et al. (2017). GATA4 loss of function in liver cancer impedes precursor to hepatocyte transition. *J. Clin. Invest* 127, 3527–3542. [PubMed: 28758902]
- Fang R, Chen F, Dong Z, Hu D, Barbera AJ, Clark EA, Fang J, Yang Y, Mei P, Rutenberg M, et al. (2013). LSD2/KDM1B and its cofactor NPAC/GLYR1 endow a structural and molecular model for regulation of H3K4 demethylation. *Mol. Cell* 49, 558–570. [PubMed: 23260659]
- Farwell KD, Shahmirzadi L, El-Khechen D, Powis Z, Chao EC, Tippin Davis B, Baxter RM, Zeng W, Mroske C, Parra MC, 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. *Genet. Med* 17, 578–586. [PubMed: 25356970]

- Fei J, Ishii H, Hoeksema MA, Meitinger F, Kassavetis GA, Glass CK, Ren B, and Kadonaga JT (2018). NDF, a nucleosome-destabilizing factor that facilitates transcription through nucleosomes. *Genes Dev.* 32, 682–694. [PubMed: 29759984]
- Ferrante MI, Zullo A, Barra A, Bimonte S, Messaddeq N, Studer M, Dollé P, and Franco B (2006). Oral-facial-digital type I protein is required for primary cilia formation and left-right axis specification. *Nat. Genet* 38, 112–117. [PubMed: 16311594]
- Finak G, McDavid A, Yajima M, et al. (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology* 16, p. 278. [PubMed: 26653891]
- Fu J, Yang Z, Wei J, Han J, and Gu J (2006). Nuclear protein NP60 regulates p38 MAPK activity. *J. Cell Sci* 119, 115–123. [PubMed: 16352664]
- Fuller ZL, Berg JJ, Mostafavi H, Sella G, and Przeworski M (2019). Measuring intolerance to mutation in human genetics. *Nat. Genet* 51, 772–776. [PubMed: 30962618]
- Furtado MB, Wilmanns JC, Chandran A, Perera J, Hon O, Biben C, Willow TJ, Nim HT, Kaur G, Simonds S, et al. (2017). Point mutations in murine Nkx2-5 phenocopy human congenital heart disease and induce pathogenic Wnt signaling. *JCI Insight* 2, e88271. [PubMed: 28352650]
- Garg V, Kathiriya IS, Barnes R, Schluterman MK, King IN, Butler CA, Rothrock CR, Eapen RS, Hirayama-Yamada K, Joo K, et al. (2003). GATA4 mutations cause human congenital heart defects and reveal an interaction with TBX5. *Nature* 424, 443–447. [PubMed: 12845333]
- Gifford CA, Ranade SS, Samarakoon R, Salunga HT, de Soysa TY, Huang Y, Zhou P, Elfenbein A, Wyman SK, Bui YK, et al. (2019). Oligogenic inheritance of a human heart disease involving a genetic modifier. *Science* 364, 865–870. [PubMed: 31147515]
- Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, and Barabási A-L (2007). The human disease network. *Proc. Natl. Acad. Sci. USA* 104, 8685–8690. [PubMed: 17502601]
- González-Terán B, López JA, Rodríguez E, et al. (2016). p38 $\gamma$  and  $\delta$  promote heart hypertrophy by targeting the mTOR-inhibitory protein DEPTOR for degradation. *Nature Communications* 7, p. 10477.
- Gordillo M, Vega H, and Jabs EW (1993). Roberts Syndrome. In GeneReviews®, Pagon RA, Adam MP, Ardinger HH, Wallace SE, Amemiya A, Bean LJ, Bird TD, Ledbetter N, Mefford HC, Smith RJ, et al., eds. (Seattle (WA): University of Washington, Seattle), p.
- Grau J, Grosse I and Keilwagen J (2015). PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics* 31(15), pp. 2595–2597. [PubMed: 25810428]
- Greene CS, Krishnan A, Wong AK, Ricciotti E, Zelaya RA, Himmelstein DS, Zhang R, Hartmann BM, Zaslavsky E, Sealfon SC, et al. (2015). Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet* 47, 569–576. [PubMed: 25915600]
- GTEx Consortium, Laboratory, Data Analysis & Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, et al. (2017). Genetic effects on gene expression across human tissues. *Nature* 550(7675), pp. 204–213. [PubMed: 29022597]
- Guo Y, Mahony S and Gifford DK (2012). High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Computational Biology* 8(8), p. e1002638. [PubMed: 22912568]
- Hao Y, Hao S, Andersen-Nissen E, et al. (2021). Integrated analysis of multimodal single-cell data. *Cell* 184(13), p. 3573–3587.e29. [PubMed: 34062119]
- Heinz S, Benner C, Spann N, et al. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular Cell* 38(4), pp. 576–589. [PubMed: 20513432]
- Hekselman I, and Yeger-Lotem E (2020). Mechanisms of tissue and cell-type specificity in heritable traits and diseases. *Nat. Rev. Genet* 21, 137–150. [PubMed: 31913361]
- Hinton RB, Prakash A, Romp RL, Krueger DA, Knilans TK, 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. *J. Am. Heart Assoc* 3, e001493. [PubMed: 25424575]

- Homsy J, Zaidi S, Shen Y, Ware JS, Samocha KE, Karczewski KJ, DePalma SR, McKean D, Wakimoto H, Gorham J, et al. (2015). De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. *Science* 350, 1262–1266. [PubMed: 26785492]
- Hota SK, and Bruneau BG (2016). ATP-dependent chromatin remodeling during mammalian development. *Development* 143, 2882–2897. [PubMed: 27531948]
- Hu X, Li T, Zhang C, Liu Y, Xu M, Wang W, Jia Z, Ma K, Zhang Y, and Zhou C (2011). GATA4 regulates ANF expression synergistically with Sp1 in a cardiac hypertrophy model. *J. Cell Mol. Med* 15, 1865–1877. [PubMed: 20874724]
- Huang N, Lee I, Marcotte EM and Hurles ME (2010). Characterising and predicting haploinsufficiency in the human genome. *PLoS Genetics* 6(10), p. e1001154. [PubMed: 20976243]
- Izarzugaza JMG, Ellesøe SG, Doganli C, Ehlers NS, Dalgaard MD, Audain E, Dombrowsky G, Banasik K, Sifrim A, Wilsdon A, et al. (2020). Systems genetics analysis identifies calcium-signaling defects as novel cause of congenital heart disease. *Genome Med.* 12, 76. [PubMed: 32859249]
- Ji W, Ferdman D, Copel J, Scheinost D, Shabanova V, Brueckner M, Khokha MK, and Ment LR (2020). De novo damaging variants associated with congenital heart diseases contribute to the connectome. *Sci. Rep* 10, 7046. [PubMed: 32341405]
- Jimenez-Morales D, Rosa Campos A, Von Dollen J and Swaney D (2020). artMS: Analytical R tools for Mass Spectrometry version 1.6.5 from Bioconductor. Bioconductor.
- Jin SC, Homys J, Zaidi S, Lu Q, Morton S, DePalma SR, Zeng X, Qi H, Chang W, Sierant MC, et al. (2017). Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. *Nat. Genet* 49, 1593–1601. [PubMed: 28991257]
- Jones WD, Dafou D, McEntagart M, Woollard WJ, Elmslie FV, Holder-Espinasse M, Irving M, Sagar AK, Smithson S, Trembath RC, et al. (2012). De novo mutations in MLL cause Wiedemann-Steiner syndrome. *Am. J. Hum. Genet* 91, 358–364. [PubMed: 22795537]
- Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW and Klein ML (1983). Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys* 79(2), p. 926.
- Karczewski KJ, Francioli LC, Tiao G, et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581(7809), pp. 434–443. [PubMed: 32461654]
- Kathiriyai IS, Rao KS, Iacono G, Devine WP, Blair AP, Hota SK, Lai MH, Garay BI, Thomas R, Gong HZ, et al. (2021). Modeling human TBX5 haploinsufficiency predicts regulatory networks for congenital heart disease. *Dev. Cell* 56, 292–309.e9. [PubMed: 33321106]
- Kimura H (2013). Histone modifications for human epigenome analysis. *J. Hum. Genet* 58, 439–445. [PubMed: 23739122]
- Knowlton KU, Baracchini E, Ross RS, Harris AN, Henderson SA, Evans SM, Glembotski CC, and Chien KR (1991). Co-regulation of the atrial natriuretic factor and cardiac myosin light chain-2 genes during alpha-adrenergic stimulation of neonatal rat ventricular cells. Identification of cis sequences within an embryonic and a constitutive contractile protein gene which mediate inducible expression. *J. Biol. Chem* 266, 7759–7768. [PubMed: 1850419]
- Kodo K, Nishizawa T, Furutani M, Arai S, Yamamura E, Joo K, Takahashi T, Matsuoka R, and Yamagishi H (2009). GATA6 mutations cause human cardiac outflow tract defects by disrupting semaphorin-plexin signaling. *Proc. Natl. Acad. Sci. USA* 106, 13933–13938. [PubMed: 19666519]
- Köhler S, Bauer S, Horn D, and Robinson PN (2008). Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet* 82, 949–958. [PubMed: 18371930]
- Korsunsky I, Millard N, Fan J, et al. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods* 16(12), pp. 1289–1296. [PubMed: 31740819]
- Kuo CT, Morrissey EE, Anandappa R, Sigrist K, Lu MM, Parmacek MS, Soudais C, and Leiden JM (1997). GATA4 transcription factor is required for ventral morphogenesis and heart tube formation. *Genes Dev.* 11, 1048–1060. [PubMed: 9136932]
- Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, Chen X, Taipale J, Hughes TR, and Weirauch MT (2018). The human transcription factors. *Cell* 172, 650–665. [PubMed: 29425488]



- Langmead B and Salzberg SL 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9(4), pp. 357–359. [PubMed: 22388286]
- Lau E, Han Y, Williams DR, et al. (2019). Splice-Junction-Based Mapping of Alternative Isoforms in the Human Proteome. *Cell reports* 29(11), p. 3751–3765.e5. [PubMed: 31825849]
- Lebrun N, Giurgea I, Goldenberg A, Dieux A, Afenjar A, Ghomid J, Diebold B, Mietton L, Briand-Suleau A, Billuart P, et al. (2018). Molecular and cellular issues of KMT2A variants involved in Wiedemann-Steiner syndrome. *Eur. J. Hum. Genet* 26, 107–116. [PubMed: 29203834]
- Lei I, Gao X, Sham MH, and Wang Z (2012). SWI/SNF protein component BAF250a regulates cardiac progenitor cell differentiation by modulating chromatin accessibility during second heart field development. *J. Biol. Chem* 287, 24255–24262. [PubMed: 22621927]
- Lepore JJ, Mericko PA, Cheng L, Lu MM, Morrisey EE, and Parmacek MS (2006). GATA-6 regulates semaphorin 3C and is required in cardiac neural crest for cardiovascular morphogenesis. *J. Clin. Invest* 116, 929–939. [PubMed: 16557299]
- Li QY, Newbury-Ecob RA, Terrett JA, Wilson DI, Curtis AR, Yi CH, Gebuhr T, Bullen PJ, Robson SC, Strachan T, et al. (1997). Holt-Oram syndrome is caused by mutations in TBX5, a member of the Brachyury (T) gene family. *Nat. Genet* 15, 21–29. [PubMed: 8988164]
- Li Y, Klena NT, Gabriel GC, Liu X, Kim AJ, Lemke K, Chen Y, Chatterjee B, Devine W, Damerla RR, et al. (2015). Global genetic analysis in mice unveils central role for cilia in congenital heart disease. *Nature* 521, 520–524. [PubMed: 25807483]
- Lian X, Zhang J, Azarin SM, et al. (2013). Directed cardiomyocyte differentiation from human pluripotent stem cells by modulating Wnt/ $\beta$ -catenin signaling under fully defined conditions. *Nature Protocols* 8(1), pp. 162–175. [PubMed: 23257984]
- Liao Y, Smyth GK and Shi W (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30(7), pp. 923–930. [PubMed: 24227677]
- Liu Y, Harmelink C, Peng Y, Chen Y, Wang Q, and Jiao K (2014). CHD7 interacts with BMP R-SMADs to epigenetically regulate cardiogenesis in mice. *Hum. Mol. Genet* 23, 2145–2156. [PubMed: 24293546]
- Lun ATL, Chen Y and Smyth GK (2016). It's DE-licious: A Recipe for Differential Expression Analyses of RNA-seq Experiments Using Quasi-Likelihood Methods in edgeR. *Methods in Molecular Biology* 1418, pp. 391–416. [PubMed: 27008025]
- Luna-Zurita L, Stirnimann CU, Glatt S, Kaynak BL, Thomas S, Baudin F, Samee MAH, He D, Small EM, Mileikovskiy M, et al. (2016). Complex interdependence regulates heterotypic transcription factor distribution and coordinates cardiogenesis. *Cell* 164, 999–1014. [PubMed: 26875865]
- Maddah M, Heidmann JD, Mandegar MA, et al. 2015. A non-invasive platform for functional characterization of stem-cell-derived cardiomyocytes with applications in cardiotoxicity testing. *Stem cell reports* 4(4), pp. 621–631. [PubMed: 25801505]
- Maere S, Heymans K and Kuiper M 2005. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* 21(16), pp. 3448–3449. [PubMed: 15972284]
- Maestro Schrödinger, LLC 2019. Maestro Suite of Programs (v. 2019–4). Maestro Schrödinger, LLC.
- Magger O, Waldman YY, Ruppin E, and Sharan R (2012). Enhancing the prioritization of disease-causing genes through tissue specific protein interaction networks. *PLoS Comput. Biol* 8, e1002690. [PubMed: 23028288]
- Maitra M, Koenig SN, Srivastava D, and Garg V (2010). Identification of GATA6 sequence variants in patients with congenital heart defects. *Pediatr. Res* 68, 281–285. [PubMed: 20581743]
- Maitra M, Schluterman MK, Nichols HA, Richardson JA, Lo CW, Srivastava D, and Garg V (2009). Interaction of Gata4 and Gata6 with Tbx5 is critical for normal cardiac development. *Dev. Biol* 326, 368–377. [PubMed: 19084512]
- Marabelli C, Marrocco B, Pilotto S, Chittori S, Picaud S, Marchese S, Ciossani G, Forneris F, Filippakopoulos P, Schoehn G, et al. (2019). A Tail-Based Mechanism Drives Nucleosome Demethylation by the LSD2/NPAC Multimeric Complex. *Cell Rep.* 27, 387–399.e7. [PubMed: 30970244]



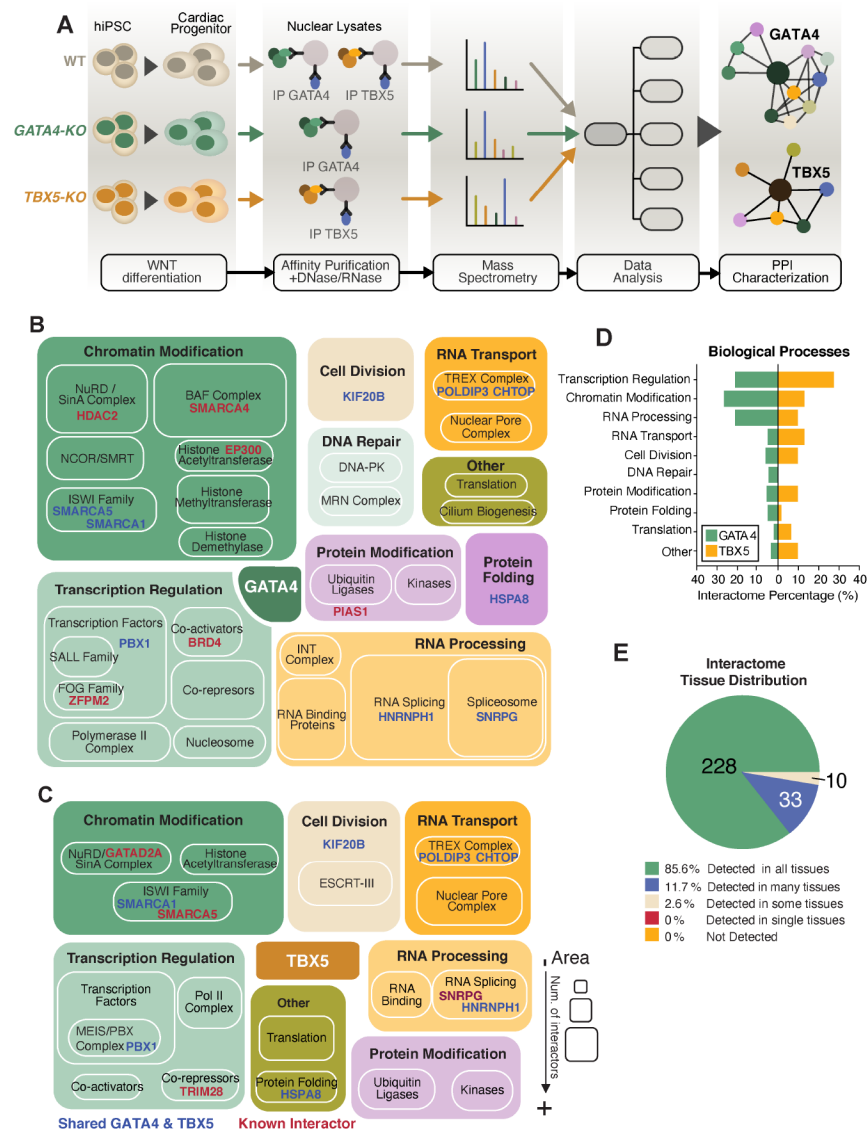
- Miyamoto S and Kollman PA 1992. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *Journal of Computational Chemistry* 13(8), pp. 952–962.
- Miyaoka Y, Chan AH, Judge LM, et al. 2014. Isolation of single-base genome-edited human iPS cells without antibiotic selection. *Nature Methods* 11(3), pp. 291–293. [PubMed: 24509632]
- Molkentin JD, Lin Q, Duncan SA, and Olson EN (1997). Requirement of the transcription factor GATA4 for heart tube formation and ventral morphogenesis. *Genes Dev.* 11, 1061–1072. [PubMed: 9136933]
- Montefiori M, Pilotto S, Marabelli C, Moroni E, Ferraro M, Serapian SA, Mattevi A, and Colombo G (2019). Impact of Mutations on NPAC Structural Dynamics: Mechanistic Insights from MD Simulations. *J. Chem. Inf. Model* 59, 3927–3937. [PubMed: 31408337]
- Mori AD, Zhu Y, Vahora I, Nieman B, Koshiba-Takeuchi K, Davidson L, Pizard A, Seidman JG, Seidman CE, Chen XJ, et al. (2006). Tbx5-dependent rheostatic control of cardiac gene expression and morphogenesis. *Dev. Biol* 297, 566–586. [PubMed: 16870172]
- Moskowitz IP, Wang J, Peterson MA, Pu WT, Mackinnon AC, Oxburgh L, Chu GC, Sarkar M, Berul C, Smoot L, et al. (2011). Transcription factor genes Smad4 and Gata4 cooperatively regulate cardiac valve development. [corrected]. *Proc. Natl. Acad. Sci. USA* 108, 4006–4011. [PubMed: 21330551]
- Nakamura H, Cook RN, and Justice MJ (2013). Mouse Tenm4 is required for mesoderm induction. *BMC Dev. Biol* 13, 9. [PubMed: 23521771]
- Narita N, Bielinska M, and Wilson DB (1997). Cardiomyocyte differentiation by GATA-4-deficient embryonic stem cells. *Development* 124, 3755–3764. [PubMed: 9367431]
- Oughtred R, Rust J, Chang C, Breitkreutz B-J, Stark C, Willems A, Boucher L, Leung G, Kolas N, Zhang F, et al. (2021). The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci.* 30, 187–200. [PubMed: 33070389]
- Padmanabhan A, Alexanian M, Linares-Saldana R, González-Terán B, Andreoletti G, Huang Y, Connolly AJ, Kim W, Hsu A, Duan Q, et al. (2020). BRD4 (Bromodomain-Containing Protein 4) Interacts with GATA4 (GATA Binding Protein 4) to Govern Mitochondrial Homeostasis in Adult Cardiomyocytes. *Circulation* 142, 2338–2355. [PubMed: 33094644]
- Parisot P, Bajolle F, Attié-Bittach T, Thomas S, Goudefroye G, Abadie V, Lyonnet S, and Bonnet D (2010). 321 Congenital heart defects in CHARGE syndrome patients with CHD7 mutations. *Archives of Cardiovascular Diseases Supplements* 2, 104–105.
- Perez-Riverol Y, Csordas A, Bai J, et al. 2019. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Research* 47(D1), pp. D442–D450. [PubMed: 30395289]
- Pierpont ME, Brueckner M, Chung WK, Garg V, Lacro RV, McGuire AL, Mital S, Priest JR, Pu WT, Roberts A, et al. (2018). Genetic basis for congenital heart disease: revisited: A scientific statement from the American Heart Association. *Circulation* 138, e653–e711. [PubMed: 30571578]
- Priest JR, Osoegawa K, Mohammed N, Nanda V, Kundu R, Schultz K, Lammer EJ, Girirajan S, Scheetz T, Waggott D, et al. (2016). De novo and rare variants at multiple loci support the oligogenic origins of atrioventricular septal heart defects. *PLoS Genet.* 12, e1005963. [PubMed: 27058611]
- R Core Team 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
- Rahman S, Sowa ME, Ottinger M, et al. 2011. The Brd4 extraterminal domain confers transcription activation independent of pTEFb by recruiting multiple proteins, including NSD3. *Molecular and Cellular Biology* 31(13), pp. 2641–2652. [PubMed: 21555454]
- Rajagopal SK, Ma Q, Obler D, Shen J, Manichaikul A, Tomita-Mitchell A, Boardman K, Briggs C, Garg V, Srivastava D, et al. (2007). Spectrum of heart disease associated with murine and human GATA4 mutation. *J. Mol. Cell Cardiol* 43, 677–685. [PubMed: 17643447]
- Ramírez F, Ryan DP, Grüning B, et al. 2016. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research* 44(W1), pp. W160–5. [PubMed: 27079975]
- Razick S, Magklaras G, and Donaldson IM (2008). iRefIndex: a consolidated protein interaction database with provenance. *BMC Bioinformatics* 9, 405. [PubMed: 18823568]

- Rentzsch P, Witten D, Cooper GM, Shendure J, and Kircher M (2019). CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* 47, D886–D894. [PubMed: 30371827]
- Richter F, Morton SU, Kim SW, Kitaygorodsky A, Wasson LK, Chen KM, Zhou J, Qi H, Patel N, DePalma SR, et al. (2020). Genomic analyses implicate noncoding de novo variants in congenital heart disease. *Nat. Genet* 52, 769–777. [PubMed: 32601476]
- Robinson MD and Oshlack A 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* 11(3), p. R25. [PubMed: 20196867]
- Robinson MD and Smyth GK 2007. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* 23(21), pp. 2881–2887. [PubMed: 17881408]
- Robinson MD and Smyth GK 2008. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics* 9(2), pp. 321–332. [PubMed: 17728317]
- Robinson MD, McCarthy DJ and Smyth GK 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26(1), pp. 139–140. [PubMed: 19910308]
- Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM, Kosmicki JA, Rehnström K, Mallick S, Kirby A, et al. (2014). A framework for the interpretation of de novo mutation in human disease. *Nat. Genet* 46, 944–950. [PubMed: 25086666]
- Sastry GM, Adzhigirey M, Day T, Annabhimoju R and Sherman W 2013. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *Journal of Computer-Aided Molecular Design* 27(3), pp. 221–234. [PubMed: 23579614]
- Sevim Bayrak C, Zhang P, Tristani-Firouzi M, Gelb BD, and Itan Y (2020). De novo variants in exomes of congenital heart disease patients identify risk genes and pathways. *Genome Med.* 12, 9. [PubMed: 31941532]
- Shannon P, Markiel A, Ozier O, et al. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research* 13(11), pp. 2498–2504. [PubMed: 14597658]
- Sifrim A, Hitz M-P, Wilsdon A, Breckpot J, Turki SHA, Thienpont B, McRae J, Fitzgerald TW, Singh T, Swaminathan GJ, et al. (2016). Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing. *Nat. Genet* 48, 1060–1065. [PubMed: 27479907]
- Smyth GK 1996. A conditional approach to residual maximum likelihood estimation in generalized linear models. *R. Stat. Soc. B*
- Stark Z, Dashnow H, Lunke S, Tan TY, Yeung A, Sadedin S, Thorne N, Macciocca I, Gaff C, Melbourne Genomics Health Alliance, et al. (2017). A clinically driven variant prioritization framework outperforms purely computational approaches for the diagnostic analysis of singleton WES data. *Eur. J. Hum. Genet* 25, 1268–1272. [PubMed: 28832562]
- Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, et al. (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 47, D607–D613. [PubMed: 30476243]
- Takeuchi JK, Lou X, Alexander JM, Sugizaki H, Delgado-Olguín P, Holloway AK, Mori AD, Wylie JN, Munson C, Zhu Y, et al. (2011). Chromatin remodelling complex dosage modulates transcription factor function in heart development. *Nat. Commun* 2, 187. [PubMed: 21304516]
- Teo G, Koh H, Fermin D, Lambert J-P, Knight JDR, Gingras A-C, and Choi H (2016). SAINTq: Scoring protein-protein interactions in affinity purification - mass spectrometry experiments with fragment or peptide intensity data. *Proteomics* 16, 2238–2245. [PubMed: 27119218]
- Theis JL, Vogler G, Missinato MA, Li X, Martinez-Fernandez A, Nielsen T, Walls SM, Kervadec A, Zeng X-XI, Kezos JN, et al. (2019). Patient-specific functional genomics and disease modeling suggest a role for LRP2 in hypoplastic left heart syndrome. *BioRxiv*.
- Thienpont B, Zhang L, Postma AV, Breckpot J, Tranchevent L-C, Van Loo P, Møllgård K, Tommerup N, Bache I, Tümer Z, et al. (2010). Haploinsufficiency of TAB2 causes congenital heart defects in humans. *Am. J. Hum. Genet* 86, 839–849. [PubMed: 20493459]

- Tickle J, Pilka ES, Bunkoczi G, et al. 2007. Structure of the cytokine-like nuclear factor n-pac [Online]. Available at: [https://www.wwpdb.org/pdb?id=pdb\\_00002uyy](https://www.wwpdb.org/pdb?id=pdb_00002uyy) [Accessed: 14 September 2020].
- Tohyama S, Hattori F, Sano M, et al. 2013. Distinct metabolic flow enables large-scale purification of mouse and human pluripotent stem cell-derived cardiomyocytes. *Cell Stem Cell* 12(1), pp. 127–137. [PubMed: 23168164]
- Tomita-Mitchell A, Maslen CL, Morris CD, Garg V, and Goldmuntz E (2007). GATA4 sequence variants in patients with congenital heart disease. *J. Med. Genet* 44, 779–783. [PubMed: 18055909]
- Uhlén M, Fagerberg L, Hallström BM, et al. 2015. Proteomics. Tissue-based map of the human proteome. *Science* 347(6220), p. 1260419. [PubMed: 25613900]
- Van Dijck A, Vulto-van Silfhout AT, Cappuyns E, van der Werf IM, Mancini GM, Tzschach A, Bernier R, Gozes I, Eichler EE, Romano C, et al. (2019). Clinical presentation of a complex neurodevelopmental disorder caused by mutations in ADNP. *Biol. Psychiatry* 85, 287–297. [PubMed: 29724491]
- Waldron L, Steimle JD, Greco TM, Gomez NC, Dorr KM, Kweon J, Temple B, Yang XH, Wilczewski CM, Davis IJ, et al. (2016). The cardiac TBX5 interactome reveals a chromatin remodeling network essential for cardiac septation. *Dev. Cell* 36, 262–275. [PubMed: 26859351]
- Wang L, Wang S and Li W 2012. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* 28(16), pp. 2184–2185. [PubMed: 22743226]
- Watt AJ, Battle MA, Li J, and Duncan SA (2004). GATA4 is essential for formation of the proepicardium and regulates cardiogenesis. *Proc. Natl. Acad. Sci. USA* 101, 12573–12578. [PubMed: 15310850]
- Wei Q, Zhan X, Zhong X, Liu Y, Han Y, Chen W, and Li B (2015). A Bayesian framework for de novo mutation calling in parents-offspring trios. *Bioinformatics* 31, 1375–1381. [PubMed: 25535243]
- Wilczewski CM, Hepperla AJ, Shimbo T, Wasson L, Robbe ZL, Davis IJ, Wade PA, and Conlon FL (2018). CHD4 and the NuRD complex directly control cardiac sarcomere formation. *Proc. Natl. Acad. Sci. USA* 115, 6727–6732. [PubMed: 29891665]
- Xin M, Davis CA, Molkenin JD, Lien C-L, Duncan SA, Richardson JA, and Olson EN (2006). A threshold of GATA4 and GATA6 expression is required for cardiovascular development. *Proc. Natl. Acad. Sci. USA* 103, 11189–11194. [PubMed: 16847256]
- Xing H, Mo Y, Liao W and Zhang MQ (2012). Genome-wide localization of protein-DNA binding and histone modification by a Bayesian change-point method with ChIP-seq data. *PLoS Computational Biology* 8(7), p. e1002613. [PubMed: 22844240]
- Yu S, Li J, Ji G, Ng ZL, Siew J, Lo WN, Ye Y, Chew YY, Long YC, Zhang W, et al. (2020). Npac Is a Co-factor of Histone H3K36me3 and Regulates Transcriptional Elongation in Mouse ES Cells. *BioRxiv*.
- Zaidi S, and Brueckner M (2017). Genetics and genomics of congenital heart disease. *Circ. Res* 120, 923–940. [PubMed: 28302740]
- Zaidi S, Choi M, Wakimoto H, Ma L, Jiang J, Overton JD, Romano-Adesman A, Bjornson RD, Breitbart RE, Brown KK, et al. (2013). De novo mutations in histone-modifying genes in congenital heart disease. *Nature* 498, 220–223. [PubMed: 23665959]
- Zhang Y, Zheng Y, Qin L, Wang S, Buchko GW and Garavito RM 2014. Structural characterization of a  $\beta$ -hydroxyacid dehydrogenase from *Geobacter sulfurreducens* and *Geobacter metallireducens* with succinic semialdehyde reductase activity. *Biochimie* 104, pp. 61–69. [PubMed: 24878278]

**HIGHLIGHTS**

- GATA4:TBX5 interactome in CPs is enriched in de novo variants associated with CHD.
- A method for scoring interactome variants identified *GLYR1* as a candidate CHD gene.
- GLYR1 and GATA4 widely co-occupied and co-activated cardiac developmental genes.
- The *GLYR1* CHD variant disrupted interaction with GATA4 and impaired cardiogenesis.



**Figure 1: Generation of GATA4 and TBX5 protein interactomes in human iPSC-derived cardiac progenitors.**

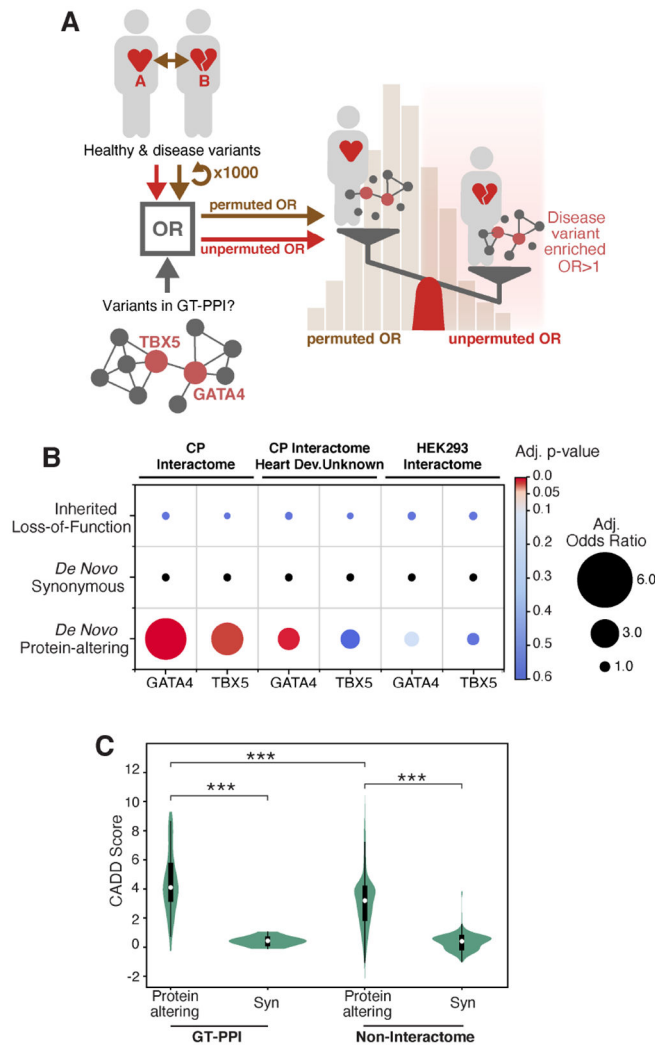
See also Figure S1, S2 and Table S1A-D.

(A) GATA4 and TBX5 AP-MS strategy from hiPSC-derived cardiac progenitors with gene knockout lines as negative controls.

(B) GATA4 and (C) TBX5 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.

(D) Distribution of GATA4 and TBX5 PPIs in biological processes, as annotated in panels B & C.

(E) Tissue expression distribution of GATA4 and TBX5 interactors across the six Human Protein Atlas categories based on transcript detection (NX = 1) in all 37 analyzed tissues (See Methods).



**Figure 2: Enrichment of *de novo* variants in CHD trios among GATA4 and TBX5 interactome proteins.**

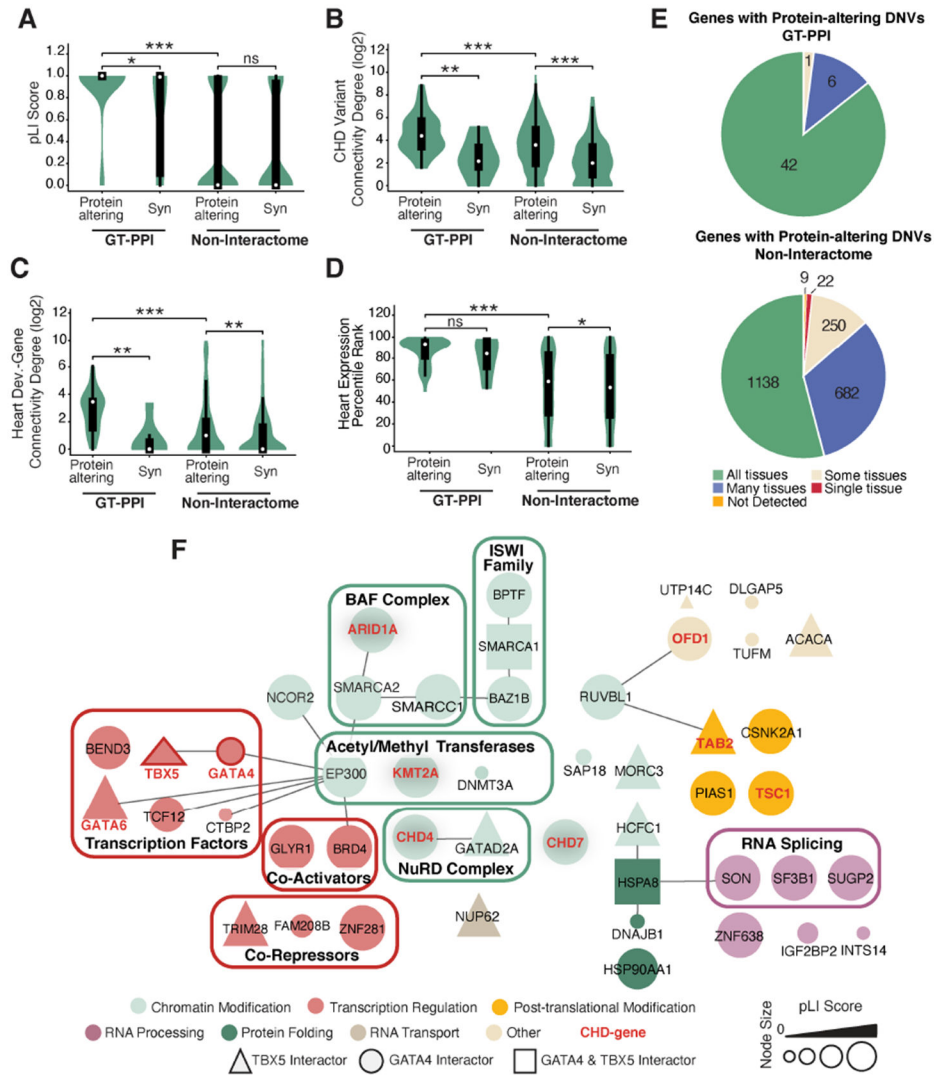
See also Figure S3 and Table S1E-I.

(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), (see STAR Methods: Permutation-based test).

(B) Results of permutation-based test in (A) for genomic variation indicated from PCGC CHD and control cohorts within the GATA4 or TBX5 interactomes in cardiac progenitors (CP Interactome), or after removing proteins involved in human or mouse cardiac malformations (CP Interactome Heart Dev. Unknown) (See Table S1F). The same analysis shown for HEK293s (HEK293 Interactome).

(C) 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 1.5x the interquartile range (thin). Two-sided Mann-Whitney-Wilcoxon test with Bonferroni correction for P-values; \*\*\*p-value<0.001.





**Figure 3: *De novo* variants in GATA4 and TBX5 interactomes exhibit features typical of disease genes.**

See also Figure S3 and Table S1J and 1K.

(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 (Jin et al., 2017); (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 1.5x 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.

(E) Pie chart of tissue expression distribution of GT-PPI or non-interactome genes harboring CHD-associated protein-altering DNVs across the six Human Protein Atlas categories (See Methods).

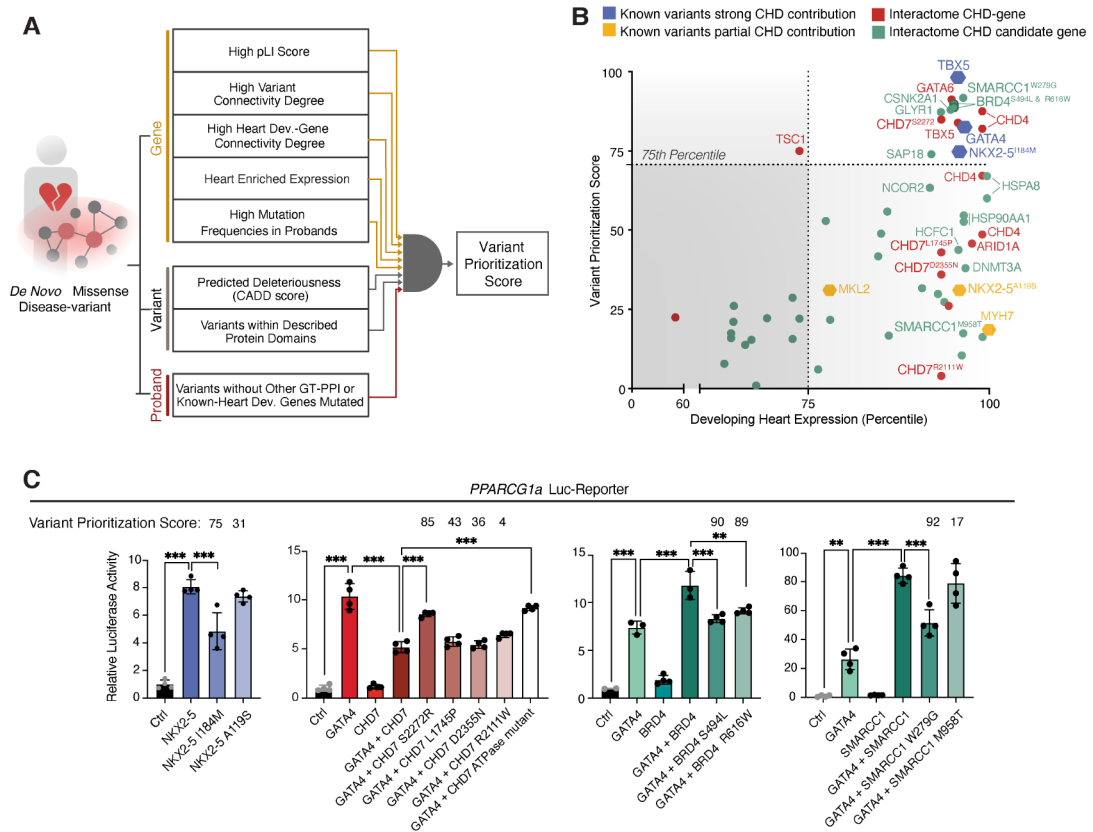
(F) Interactome CHD candidate genes represented as a network after integration with PPI information from iRefIndex database. Nodes colored based on manually annotated biological processes and protein families/complexes grouped in boxed areas. Node size reflects probability of Loss-of-function Intolerance (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).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript



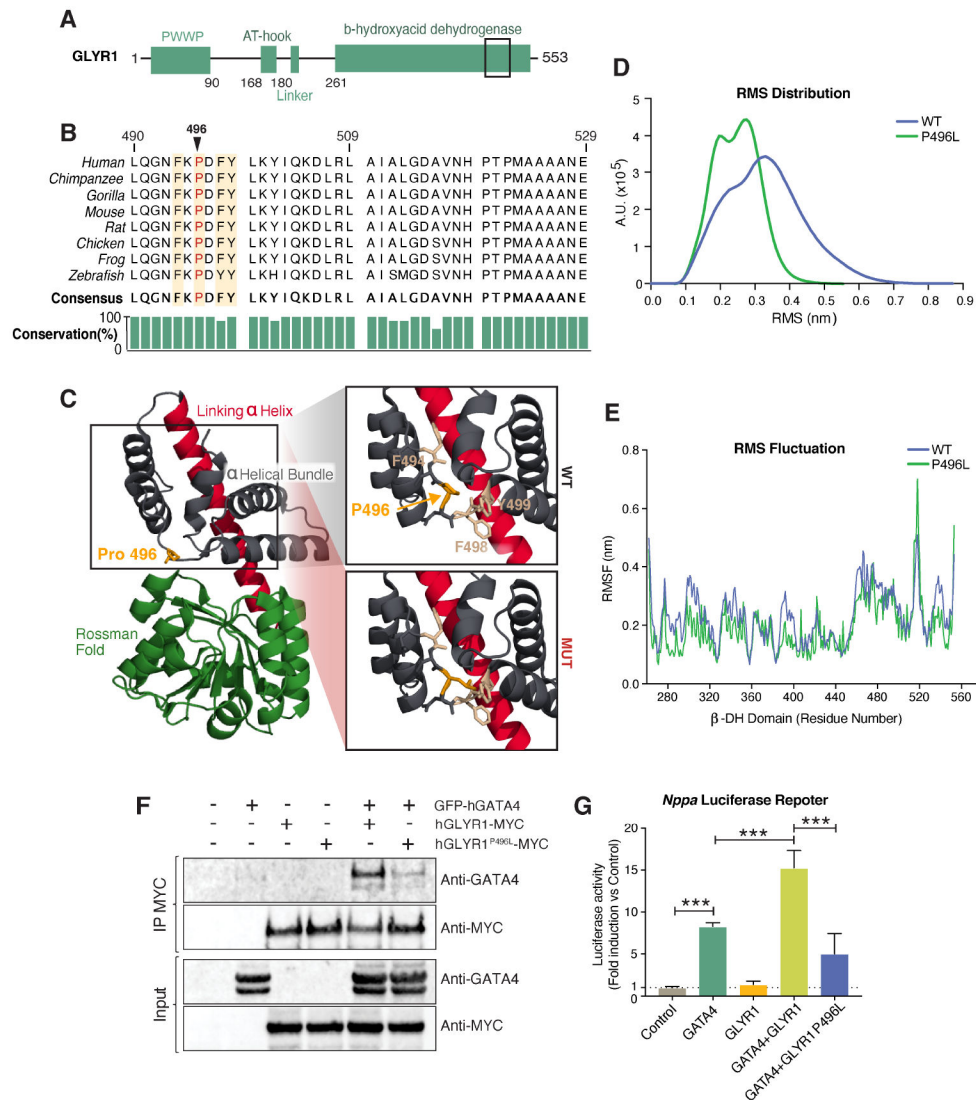
**Figure 4: Integrative variant prioritization scoring to rank CHD-associated variants.**

See also Figure S4 and Table S1L and 1M.

(A) Variant prioritization score strategy (see STAR Methods: Variant scoring and Figure S5A).

(B) 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 (blue) or partial contribution (orange) to CHD are included as references.

(C) Biochemical evaluation by luciferase assays of the functional impact for variant alleles with different prioritization scores in panel B within *NKX2-5*, *CHD7*, *BRD4* or *SMARCC1*. The *CHD7* ATPase mutant 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.



**Figure 5: Functional impact of a highly scored CHD variant in *GLYR1*.**

See also Figure S4 and Table S1N.

(A) Simplified protein schematic depicting the domain organization of human *GLYR1*.

Black rectangle indicates zoomed-in protein region in Figure 5B.

(B) Protein sequence conservation across vertebrate species for the *GLYR1* rigid loop region containing the CHD-associated P496L DNV.

(C) *GLYR1* dehydrogenase domains: Rossman-fold globular domain (green), the linking  $\alpha$ 9-helix (red), and the  $\alpha$ -helical bundle (dark blue). Right panels: zoom into the WT and mutant forms of the rigid loop with aromatic residues in beige and Proline 496 in orange.

(D) Distribution of the root mean square deviation (RMSD) of frames visited during the trajectories from the reference state represented by the starting structure of the WT (blue) and the P496L mutant (green) *GLYR1* dehydrogenase domains within the measured time.

(E) Residue flexibility analysis based on the standard deviations of the atomic positions in the simulations (RMSF) after fitting to the starting structure of the WT form (blue) and the

mutant (green) GLYR1 dehydrogenase domains. F-statistic shows lower flexibility of the mutant compared to the in the Rossmann-fold domain (residues 262-437).

(F) The ability of GLYR1 WT or P496L mutant to interact with GATA4 by immunoprecipitation (IP) of GLYR1-MYC and immunoblotting with indicated antibodies.

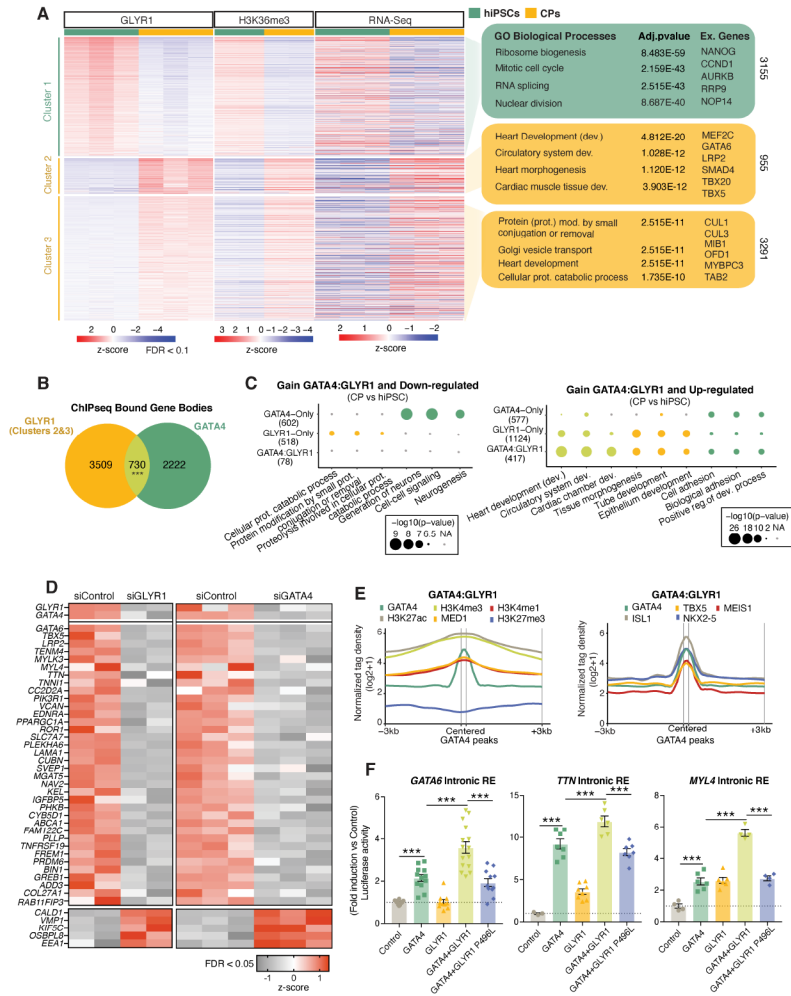
(G) Luciferase reporter assay in HeLa cells showing activation of the GATA4-dependent *Nppa* luciferase reporter upon addition of plasmids encoding indicated proteins. (n=3 independent experiments). One-way ANOVA coupled with Tukey post hoc test: \*\*\* p-value <0.001.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript



**Figure 6: GATA4-associated roles for GLYR1 in transcription regulation during cardiomyocyte differentiation.**

See also Figure S5 and Table S10-T.

(A) Heat map of genes differentially bound by GLYR1 (FDR<0.1) between hiPSCs and CPs subjected to k-means clustering based on: GLYR1 ChIPseq signal (3 representative replicates plotted, n=5), H3K36me3 ChIPseq signal (n=2) and gene expression levels (GSE137920; n=3). Statistically enriched GO Biological Process terms and example genes per cluster on the right panel.

(B) Overlap of genes bound by GLYR1 in CPs from Clusters 2 & 3 (FDR<0.1, LogFC>0.5) with genes occupied by GATA4 within the gene body (1<sup>st</sup> intron-TES). The odds of GATA4 binding to gene bodies enriched GLYR1 signal vs no GATA4 binding is 2.38 (\*\*p-value < 2.2e<sup>-16</sup>, fisher’s exact test).

(C) Gene Ontology enrichment analysis of biological process for genes up- or down-regulated in CPs compared to hiPSCs (FDR<0.05) and bound by GATA4:GLYR1, GLYR1-Only and GATA4-Only. Prot., protein; dev., development.

(D) Heat map of GATA4:GLYR1 co-bound genes differentially expressed (FDR<0.05) upon independent knockdown of GATA4 or GLYR1 at the CP stage by RNAseq.



(E) Metagene plots for GATA4:GLYR1 co-bound genes plotting the normalized ChIPseq signal for the indicated histone marks (publicly available data GSE85631 and GSM2047027) and other cardiac transcription factors centered on GATA4 peaks within the gene body (1<sup>st</sup> Intron-TES). One representative replicate plotted.

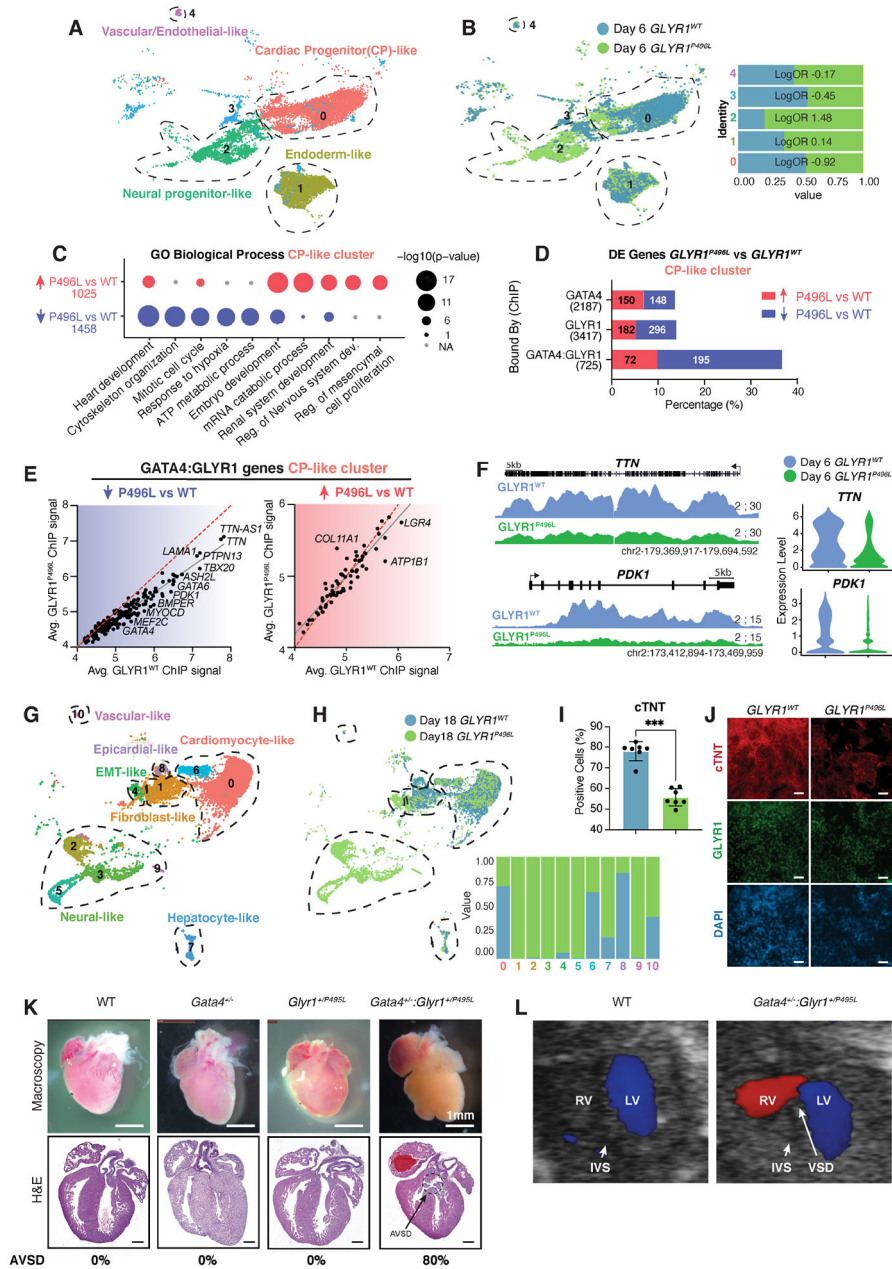
(F) Transcriptional activity of three putative intronic regulatory elements co-bound by GATA4 & GLYR1 in the presence of indicated regulatory proteins. One-way ANOVA coupled with Tukey post hoc test (n=3): \*\*\* p-value <0.001.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript



**Figure 7: Impact of the GLYR1 P496L missense variant in human iPS-derived cardiac cells and in mouse cardiogenesis.**

See also Figure S6 and S7 and Table S1U-AC.

(A-B) UMAP plot from 3 independent human iPS-CP differentiations at day 6 colored by (A) cluster identity and (B) genotype. Bar plot and natural log odds ratio (LogOR) reflect the *GLYR1*<sup>P496L</sup> cells contribution to each of the clusters compared to *GLYR1*<sup>WT</sup> cells. None reach statistical significance.

(C) Gene Ontology (GO) Biological Process enrichment analysis for genes up-regulated or down-regulated (*GLYR1*<sup>P496L</sup> vs *GLYR1*<sup>WT</sup>, FDR<0.05) within CP-like cells (cluster 0) at differentiation day 6.

(D) Percentage of GATA4:GLYR1, GLYR1-Only or GATA4-Only bound genes in CPs based on ChIPseq that were differentially expressed (*GLYR1<sup>P496L</sup>* vs *GLYR1<sup>WT</sup>*, FDR<0.05) in CP-like cells (cluster 0). Numbers within the bars: absolute numbers of genes involved.

(E) Scatter plots for GLYR1 ChIPseq signal among biological replicates from *GLYR1<sup>WT</sup>* (n=5) or *GLYR1<sup>P496L</sup>* (n= 3) CP differentiation at day 6 for GATA4:GLYR1 co-bound genes and down-regulated or up-regulated (FDR<0.05) in panel D. Dash red line: identity line; grey line: data trend line.

(F) Representative GLYR1 ChIPseq coverage tracks and expression violin plots for two representative GATA4:GLYR1 bound loci found in panel D and E to be down-regulated in CP-like cells (cluster 0) and had reduced GLYR1 occupancy in *GLYR1<sup>P496L</sup>* compared to *GLYR1<sup>WT</sup>* at differentiation day 6.

(G-H) UMAP plot from 3 independent day 18 CM differentiations colored by (G) cluster identity and (H) genotype. Bar plot and natural log odds ratio (LogOR) reflect *GLYR1<sup>P496L</sup>* cells contribution to each of the identity clusters compared to *GLYR1<sup>WT</sup>* cells (Table S1AB). The LogOR of all clusters between *GLYR1<sup>P496L</sup>* vs *GLYR1<sup>WT</sup>* are statistically significant, except for clusters 7 and 10 (FDR < 0.05).

(I) Percentage of cTNT positive cells in *GLYR1<sup>WT</sup>* and *GLYR1<sup>P496L</sup>* CM differentiation day 18 as measured by flow cytometry (n= 3). Unpaired Student's t-test: \*\*\*p-value<0.001.

(J) Representative immunostaining micrographs for cTNT (red), GLYR1 (green) or DAPI (blue) in *GLYR1<sup>WT</sup>* and *GLYR1<sup>P496L</sup>* at CM differentiation day 18. Scale (50µm).

(K) Whole mount images (scale 1 mm) and hematoxylin and eosin (H&E) images of cross-sections (scale 300 µm) from WT, *Glyr1<sup>+P495L</sup>*, *Gata4<sup>+/-</sup>* and *Glyr1<sup>+P495L</sup>:Gata4<sup>+/-</sup>* representative hearts at postnatal day 1. The AVSD incidence per genotype is indicated as a percentage of the total number of hearts analyzed by histology.

(L) Echocardiography detection of ventricular septal defects (VSD) in *Glyr1<sup>+P495L</sup>:Gata4<sup>+/-</sup>* compound heterozygous hearts at postnatal day 0. On apical 4 chamber view, in *Glyr1<sup>+P495L</sup>:Gata4<sup>+/-</sup>*, red flow in the right ventricle (RV) indicates blood that has crossed the intraventricular septum (IVS) from the left ventricle (LV) and is flowing toward the transducer.

## KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
<b>Antibodies</b>		
Mouse monoclonal anti-Troponin T Ab-1	Thermo Scientific	Clone 13-11
Mouse monoclonal anti-Cardiac Troponin T 1C11	Abcam	ab8295
Goat polyclonal C-20 anti-TBX5	Santa Cruz	sc-17866
Goat polyclonal C-20 anti-GATA4	Santa Cruz	sc-1237
Mouse monoclonal A-6 anti-TBX5	Santa Cruz	sc-515536
Mouse monoclonal G-4 anti-GATA4	Santa Cruz	sc-25310
Goat polyclonal anti-NKX2-5	Santa Cruz	sc-8697
Goat polyclonal anti-Myc tag - CHIP Grade	Abcam	ab9132
Rabbit hGLYR1 (NDF) anti-sera 7	James T. Kadonaga Laboratory	Fei et al., 2018
Rabbit polyclonal anti-HA tag - CHIP Grade	Abcam	ab9110
Rabbit polyclonal anti-H3K36me3- CHIP Grade	Abcam	ab9050
Rabbit polyclonal anti-MEIS1	Abcam	ab19867
Rabbit polyclonal anti-ISL1	Abcam	ab20670
Monoclonal anti-Vinculin VLN01	Thermo Fisher Scientific	MA5-11690
Rabbit Anti-LSD2 antibody	MyBioSource	MBS4751131
Mouse Anti-CDK9 antibody	Santa Cruz	sc-13130
Mouse Anti-Cyclin T1 antibody	Santa Cruz	sc-271348
Donkey anti-Goat Alexa Fluor 488	Thermo Fisher Scientific	A11055
Donkey anti-Mouse Alexa Fluor 647	Thermo Fisher Scientific	A-31571
Donkey Anti-Mouse Alexa Fluor 568	Thermo Fisher Scientific	A10037
Donkey anti-Rabbit Alexa Fluor 488	Thermo Fisher Scientific	A21206
<b>Bacterial and Virus Strains</b>		
Ad-GFP-EF1-h-GLYR1	Vector Biolabs	NA
Ad-GFP-EF1-h-GLYR1 P496L	Vector Biolabs	NA
Ad-EF1a-eGFP	Vector Biolabs	NA
<b>Chemicals, Peptides, and Recombinant Proteins</b>		
SpCas9-NLS purified protein	QB3 Macrolab, UCB	NA
ROCK inhibitor Y-27632 2HCl	<a href="http://Selleckchem.com">Selleckchem.com</a>	S1049
hESC-qualified LDEV-free matrigel	Corning	354277
CHIR99021	Tocris	4423
IWP4	Tocris	5214
B27-supplemented (without insulin)	Life Technologies	A1895601

REAGENT or RESOURCE	SOURCE	IDENTIFIER
B27-supplemented (with insulin)	Life Technologies	A1895601
Fibronectin bovine plasma solution	Sigma-Aldrich	F1141
Lipofectamine RNAimax	Invitrogen	13778075
FuGENE HD	Promega	E2311
Benzonase	Millipore	E1014
Dynabeads™ Protein G	Invitrogen	10004D
Dynabeads™ Protein A/Protein G	Invitrogen	10015D
AMPure XP beads	Beckman Coulter	A63881
Dulbecco's Modified Eagle Medium (DMEM), high glucose, GlutaMAX™ Supplement	Thermo Fisher Scientific	10566016
RPMI1640 media	Life Technologies	11875-119
Essential 8 medium (E8)	Life Technologies	A1517001
Novex 4-12% Tris-Glycine gels	Invitrogen	XP04122BOX
ECL Prime Western Blotting Detection Reagent	GELife Sciences	RPN2232
TRIzol™ LS reagent	Invitrogen	10296010
Opti-MEM	Thermo Fisher Scientific	31985070
<b>Critical Commercial Assays</b>		
Primary Cell Nucleofection P3 Kit	Lonza	V4XP-3960
Titanium Taq DNA Polymerase & Titanium Buffer	Takara Bio	639209
Phusion® High-Fidelity PCR Master Mix with GC Buffer	New England BioLabs	M0532S
Cold Fusion Cloning kit	Systems Biosciences	MC101B-1
iBlot® Transfer Stack PVDF mini	ThermoScientific	IB4010-32
OMIX C18 pipette tips	Agilent	A57003100
Direct-Zol RNA kit	Zymo Research	R2052
SuperScript™ III First-strand Synthesis SuperMix	Invitrogen	18080400
Taqman Universal PCR master mix	Life technologies	4305719
Ovation RNA-seq System V2 Kit	NuGEN	7102-08
Illumina Library Quantification Kit	KAPA Biosystems	KK4824
NEBNext Ultra II DNA Library Prep Kit for Illumina	New England BioLabs	E7645
NEBNext® Multiplex Oligos for Illumina® (Index Primers Set 1)	New England BioLabs	E7335
Dual Luciferase Assay System	Promega	E1960
Quick Start™ Bradford Protein Assay Kit 1	Bio-Rad	5000201
Chromium Next GEM Single Cell 3' GEM, Library & Gel Bead Kit v3.1	10X Genomics	PN-1000121
Chromium Next GEM Chip G Single Cell Kit	10X Genomics	PN-1000120
3'v3.1 Single Index Kit	10X Genomics	PN-1000213
<b>Deposited Data</b>		
RNAseq	GEO Database	GSE159411

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Single-cell RNAseq	GEO Database	GSE159411
ChIPseq	GEO Database	GSE159411
Proteomics	PRIDE Database	PXD022091
<b>Experimental Models: Cell Lines</b>		
WTC11 hiPS cell line	Gladstone Stem Cell Core (Coriell)	GM25256
TBX5-KO hiPS cell line	This paper	NA
GATA4-KO hiPS cell line	This paper	NA
GLYR1-KO hiPS cell line	This paper	NA
GLYR1-P496L hiPS cell line	This paper	NA
HeLa cell line	ATCC	ATCC® CCL-2™
HEK-293 cell line	ATCC	ATCC® CRL-1573™
<b>Experimental Models: Organisms/Strains</b>		
GLYR1 P495L mouse line	This paper	NA
<b>Oligonucleotides</b>		
GATA4 sgRNA GAGGCCACUCGCGGGAGG	Synthego	NA
TBX5 sgRNA GCTTACCTTGTGGTTCTGGTAGG	Synthego	NA
CRISPR GATA4-KO FW primer 5'AGAGATCTCATGCAGGGTTCG3'	This paper	NA
CRISPR GATA4-KO REV primer 5'TCATGATGCCTGGCCTTACT3'	This paper	NA
CRISPR TBX5-KO FW primer 5'GCAGAAACAGTTGCCAGAA3'	This paper	NA
CRISPR TBX5-KO REV primer 5'CAAGGCGAATTTAGAGGGCG3'	This paper	NA
hGATA4 Cold Fusion cloning FW primer 5'TGGTGGATCCACCGGTATGTATCAGAGCTTGCCATGG3'	This paper	NA
hGATA4 Cold Fusion cloning REV primer 5'TGAGCGGCCGCGTTTAAACTTACGCAGTGATTAIGTCCCCGTG3'	This paper	NA
hGLYR1 sgRNA ATGTATTTTCAGGTAGAAATCAGG	This paper	NA
hGLYR1-P496L HDR template: CCTCAGATATCCTGCAAGGAACTTTAAGCTTGATTCTACCTGAAATACATTCAGAAGGA	This paper	NA
CRISPR GLYR1 FW Primer 5'CACAGTGCACTTAGCCT3'	This paper	NA
CRISPR GLYR1 RV Primer 5'TGCAGCAAATGAGGTAGGGT3'	This paper	NA
P495L Mouse Genotyping FW Primer FW 5'TTCCAGTCATTCCTTGCCCC3'	This paper	NA
P495L Mouse Genotyping FW Primer RV 5'TGATCAGAAGGGTTCGGCAAG3'	This paper	NA
GATA4 Silencer Select Pre-designed siRNA ID s53120	Thermo Fisher Scientific	4392420
GLYR1 siRNA SASI_Hs01_00116796	Millipore-Sigma	NA
<b>Recombinant DNA</b>		
pCMV-T7-hGLYR1-MYC-IRES2-mCherry-pA	GeneCopoeia	EX-Z0806-M73
pCMV-T7-hGLYR1 P496L-MYC-IRES2-mCherry-pA	This paper	NA
pEN563-pCAGG-eGFP-GATA4	This paper	NA



REAGENT or RESOURCE	SOURCE	IDENTIFIER
pCMV-T7- <i>hSMARCC1</i> -MYC-IRES2-mCherry-pA	GeneCopoeia	EX-A6386-M73
pCMV-T7- <i>hSMARCC1</i> -W279G-MYC-IRES2-mCherry-pA	This paper	NA
pCMV-T7- <i>hSMARCC1</i> -M958T-MYC-IRES2-mCherry-pA	This paper	NA
p6344 pcDNA4-TO-HA- <i>Brd4FL</i>	Rahman et al., 2011	Addgene-31351
p6344 pcDNA4-TO-HA- <i>Brd4FL</i> -S494L	This paper	NA
p6344 pcDNA4-TO-HA- <i>Brd4FL</i> -R616W	This paper	NA
pCIneo- <i>hCHD7</i> -Kozak ATG 3' HA-bGH polyA	Peter Scacheri Lab	Addgene-89460
pCIneo- <i>hCHD7</i> -S2272R-Kozak ATG 3' HA-bGH polyA	This paper	NA
pCIneo- <i>hCHD7</i> -L1745P-Kozak ATG 3' HA-bGH polyA	This paper	NA
pCIneo- <i>hCHD7</i> -D2355N-Kozak ATG 3' HA-bGH polyA	This paper	NA
pCIneo- <i>hCHD7</i> -R2111W-Kozak ATG 3' HA-bGH polyA	This paper	NA
EX-Y4729-M06- <i>hCHD7</i> - del ATPase domain	Yuelong Liu et al., 2014	GeneCopoeia and mod. by Kai Jiao Lab
pcDNA3.1-CMV- <i>NKX2-5</i> -3xFLAG-bGH polyA	This paper	NA
pcDNA3.1-CMV- <i>NKX2-5</i> -I184M-3xFLAG-bGH polyA	This paper	NA
pcDNA3.1-CMV- <i>NKX2-5</i> -A119S-3xFLAG-bGH polyA	This paper	NA
pANF638 Luc vector	Knowlton et al., 1991	NA
pGL4.23[luc2/minP]	Promega	E8411
pGL4.23[luc2/minP] <i>Ppargc1a</i> Promoter	Padmanabhan et al, 2020	NA
pGL4.23[luc2/minP] <i>GATA6</i> intronic RE	This paper	NA
pGL4.23[luc2/minP] <i>TTN</i> intronic RE	This paper	NA
pGL4.23[luc2/minP] <i>MYL4</i> intronic RE	This paper	NA
pRL-TK	Promega	E2241
pRL-SV40	Promega	E2231
<b>Software and Algorithms</b>		
PRISM v10 graphing and statistical software	GraphPad Software	<a href="http://www.graphpad.com/scientific-software/prism/">http://www.graphpad.com/scientific-software/prism/</a>
Flowjo-v10	FlowJo LLC	<a href="http://www.flowjo.com">http://www.flowjo.com</a>
Pulse automated analysis software	PULSE software	<a href="https://www.pulsevideanalysis.com">https://www.pulsevideanalysis.com</a>
Cytoscape v3.7.1	Shannon et al., 2003	<a href="https://cytoscape.org">https://cytoscape.org</a>
R v.3.6.1	R Core Team, 2019	<a href="https://www.R-project.org/">https://www.R-project.org/</a>

REAGENT or RESOURCE	SOURCE	IDENTIFIER
artMS v1.4.0	Jimenez-Morales et al., 2020	<a href="https://bioconductor.org/packages/release/bioc/html/artMS.html">https://bioconductor.org/packages/release/bioc/html/artMS.html</a>
SAINTq v0.0.4	Teo et al., 2016	<a href="http://saint-apms.sourceforge.net/">http://saint-apms.sourceforge.net/</a>
fastq-mcf	ea-utils 1.1.2-537	<a href="http://code.google.com/p/ea-utils">http://code.google.com/p/ea-utils</a>
bowtie 2.2.4	Langmead and Salzberg, 2012	<a href="https://sourceforge.net/projects/bowtie-bio/files/bowtie2/2.1.0/">https://sourceforge.net/projects/bowtie-bio/files/bowtie2/2.1.0/</a>
GEM	Guo et al., 2012	<a href="http://cgs.csail.mit.edu/gem">http://cgs.csail.mit.edu/gem</a>
BCP	Xing et al., 2012	<a href="http://rulai.cshl.edu/BCP">http://rulai.cshl.edu/BCP</a>
AMBER 18	Case et al., 2017	<a href="http://ambermd.org/">http://ambermd.org/</a>
PyMOL, Molecular Graphics System, Version 1.2r3pre.	Schrödinger, LLC.	<a href="https://pymol.org/2/">https://pymol.org/2/</a>
TrioDeNovo v0.06	Wei et al., 2015	<a href="https://genome.sph.umich.edu/wiki/Triodenovo">https://genome.sph.umich.edu/wiki/Triodenovo</a>
annovar	Wang et al., 2010	<a href="https://annovar.openbioinformatics.org/en/latest/">https://annovar.openbioinformatics.org/en/latest/</a>
Seurat v4.0.1	Hao and Hao et al. 2021	<a href="https://satijalab.org/seurat/articles/install.html">https://satijalab.org/seurat/articles/install.html</a>
harmony v1.0	Korsunsky et al. 2019	<a href="https://portals.broadinstitute.org/harmony/articles/quickstart.html">https://portals.broadinstitute.org/harmony/articles/quickstart.html</a>
lme4 v1.1-26	Bates et al. 2015	<a href="https://cran.r-project.org/web/packages/lme4/index.html">https://cran.r-project.org/web/packages/lme4/index.html</a>
PRROC_1.3.1	Grau et al. 2015	<a href="https://cran.r-project.org/web/packages/PRROC/index.html">https://cran.r-project.org/web/packages/PRROC/index.html</a>

REAGENT or RESOURCE	SOURCE	IDENTIFIER
MAST v1.16	Finak et al. 2015	<a href="https://www.bioconductor.org/packages/release/bioc/html/MAST.html">https://www.bioconductor.org/packages/release/bioc/html/MAST.html</a>
homer v4.11.1	Heinz et al. 2010	<a href="http://homer.ucsd.edu/homer/download.html">http://homer.ucsd.edu/homer/download.html</a>
<b>Other</b>		
GTEX v8.1.1.9	GTEX Consortium et al., 2017	<a href="https://www.gtexportal.org/home/datasets">https://www.gtexportal.org/home/datasets</a>
iRefIndex v15.0	Razick et al., 2008	<a href="https://irefindex.vib.be/wiki/index.php/iRefIndex">https://irefindex.vib.be/wiki/index.php/iRefIndex</a>
STRING-db	Szklarczyk et al., 2019	<a href="https://string-db.org/">https://string-db.org/</a> Accessed April 22, 2021
BIOGRID	Oughtred et al., 2021	<a href="https://thebiogrid.org/">https://thebiogrid.org/</a> Accessed Apr 22, 2021
Human Protein Atlas v13	Uhlén et al., 2015	<a href="https://www.proteinatlas.org/humanproteome/tissue/tissue+specific">https://www.proteinatlas.org/humanproteome/tissue/tissue+specific</a>
Combined Annotation Dependent Depletion (CADD) v1.6	Rentzsch et al., 2019	<a href="https://cadd.gs.washington.edu/score">https://cadd.gs.washington.edu/score</a>
Haploinsufficiency prediction with imputation	Huang et al., 2010	<a href="https://doi.org/10.1371/journal.pgen.1001154.s002">https://doi.org/10.1371/journal.pgen.1001154.s002</a>
gnomAD v2.1.1	Karczewski et al., 2020	<a href="https://gnomad.broadinstitute.org/downloads#v2-constraint">https://gnomad.broadinstitute.org/downloads#v2-constraint</a>