PLOS GENETICS

RESEARCH ARTICLE

Network assisted analysis of *de novo* variants using protein-protein interaction information identified 46 candidate genes for congenital heart disease

Yuhan Xie[1], Wei Jiang[1], Weilai Dong[2], Hongyu Li[1], Sheng Chih Jin[3], Martina Brueckner[2,4], Hongyu Zhao[1,2,5]*

1 Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut, United States of America, **2** Department of Genetics, Yale School of Medicine, New Haven, Connecticut, United States of America, **3** Department of Genetics, Washington University School of Medicine, St Louis, Missouri, United States of America, **4** Department of Pediatrics, Yale University, New Haven, Connecticut, United States of America, **5** Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut, United States of America

* hongyu.zhao@yale.edu

Check for updates

OPEN ACCESS

Citation: Xie Y, Jiang W, Dong W, Li H, Jin SC, Brueckner M, et al. (2022) Network assisted analysis of *de novo* variants using protein-protein interaction information identified 46 candidate genes for congenital heart disease. PLoS Genet 18(6): e1010252. https://doi.org/10.1371/journal.pgen.1010252

Editor: Xiaofeng Zhu, Case Western Reserve University, UNITED STATES

Received: December 13, 2021

Accepted: May 12, 2022

Published: June 7, 2022

Copyright: © 2022 Xie et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Summary statistics of real data application can be downloaded from https://github.com/JustinaXie/NDATA. Results of simulation and real data can be downloaded from S1 Table and S2 Table in the Supporting Information.

Funding: This work was supported in part by NIH grant R03HD100883-01A1 (Y.X. and H.Z.) and R01GM134005-01A1 (W.J., H.L., and H.Z.). The funders had no role in study design, data collection

Abstract

De novo variants (DNVs) with deleterious effects have proved informative in identifying risk genes for early-onset diseases such as congenital heart disease (CHD). A number of statistical methods have been proposed for family-based studies or case/control studies to identify risk genes by screening genes with more DNVs than expected by chance in Whole Exome Sequencing (WES) studies. However, the statistical power is still limited for cohorts with thousands of subjects. Under the hypothesis that connected genes in protein-protein interaction (PPI) networks are more likely to share similar disease association status, we developed a Markov Random Field model that can leverage information from publicly available PPI databases to increase power in identifying risk genes. We identified 46 candidate genes with at least 1 DNV in the CHD study cohort, including 18 known human CHD genes and 35 highly expressed genes in mouse developing heart. Our results may shed new insight on the shared protein functionality among risk genes for CHD.

Author summary

The topologic information in a pathway may be informative to identify functionally interrelated genes and help improve statistical power in DNV studies. Under the hypothesis that connected genes in PPI networks are more likely to share similar disease association status, we developed a novel statistical model that can leverage information from publicly available PPI databases. Through simulation studies under multiple settings, we proved our method can increase statistical power in identifying additional risk genes compared to methods without using the PPI network information. We then applied our method to a

and analysis, decision to publish, or preparation of the manuscript.

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

real example for CHD DNV data, and then visualized the subnetwork of candidate genes to find potential functional gene clusters for CHD.

Introduction

Congenital heart disease (CHD) is the most common birth defect affecting ~ 1% of live births and accounts for one-third of all major congenital abnormalities [1–3]. There is substantial evidence that CHD has a strong genetic component [4]. Although it is estimated that aneuploidies and copy number variations account for about 23% of CHD cases, few individual disease-causing genes have been identified in published studies [5–8]. Therefore, the limited knowledge of the underlying genetic causes poses an obstacle to the reproductive counseling of CHD patients [9].

Whole Exome Sequencing (WES) studies have successfully boosted novel causal gene identification for both Mendelian and complex disorders [10,11]. To narrow down the pool of candidate variants from WES, family-based studies have been conducted to scan for *de novo* variants (DNVs) from parent-offspring trios. DNV studies have been shown to play an important role in risk gene identification for CHD [1,3,5,6,12–15]. From the analysis of 1,213 CHD parent-offspring trios, Homsy et al. identified a greater burden of damaging DNVs, especially in genes with likely functional roles in heart and brain development [12]. Recently, Jin et al. inferred that DNVs in ~440 genes were likely contributors to CHD [5]. Despite these advances, it remains challenging to capture the causal genes with only DNV data as CHD is very genetically heterogeneous [6].

Several statistical methods have been proposed to identify risk genes by integrating DNVs with other genetic variants and additional biological data. He et al. developed a Bayesian framework, namely the Transmission And De novo Association (TADA), to increase statistical power of inferring risk genes by incorporating both DNMs and rare inherited variants [16]. A hierarchical Bayes strategy was adopted for parameter estimation in TADA. Following this idea, a number of methods have been proposed to improve TADA, with some focusing on leveraging the shared genetic information in multiple correlated phenotypes, such as neurodevelopmental disorders and CHD [17,18], whereas others extend the method by integrating DNMs with other types of genetic variants and functional annotations [19–22]. Please note that, except for DECO [22], all these methods treat each gene individually and do not consider the interaction effects of genes. Thus, there is a pressing need for developing network-based frameworks to consider the functional connectivities among genes.

Network-based approaches have been successful in prioritizing risk genes for downstream genomic and transcriptomic studies [23–26]. Chen et al. [24] proposed a Markov Random Field (MRF) model to incorporate pathway topology structure for Genome-Wide Association Studies (GWAS). They showed that their method is more powerful than single gene-based methods through both simulation and real data analyses. In 2015, Liu et al. adopted a similar idea as Chen et al. to analyze DNV data from WES studies [27]. Their framework, namely DAWN, combines TADA p-values with the estimated network from gene co-expression data. In their real data analysis for autism, 333 genes were prioritized by integrating DNV summary statistics and expression data from brain tissue. However, the above methods require summary statistics (Z scores or p-values) from genetic association analysis as their input, which may not be provided from results of DNV analysis [17,19].

More recently, Bayrak et al. developed a priority score to quantify the proximity of genes to the known CHD risk genes using DNV data [3]. Utilizing canonical pathways and human

gene networks, their analyses identified 23 novel genes that are likely to contribute to CHD pathogenesis. Their results further support the potential to improve power by integrating network information with DNV data. Then, the question becomes how to choose an informative gene network for CHD. As there is a limited number of co-expression data sets for human developmental heart, a natural choice for network information would be human PPI databases. There are multiple primary PPI network databases such as BioGRID [28], IntAct [29], DIP [30], MINT [31], and HPRD [32]. Most network-based studies apply their real data on two or more of databases to obtain their results. Nonetheless, it is hard to check the overlapping information between two PPI databases and interpret the divergent results. Multiple integrative databases such as STRING [33], HINT [34], UniHI [35], hPRINT [36] and GPS-Prot [37] provide a platform to resolve the above problems [38]. Among them, STRING is a popular PPI resource that imports protein association knowledge from physical interaction and curated knowledge from the primary PPI databases and other pathway information knowledge such as KEGG [39–41] and GO [42,43]. In addition, it provides a score to measure the likelihood of interactions. Some studies have used STRING in their post-association analysis for gene-based DNV studies and showed significant enrichment of candidate CHD risk genes in the STRING PPI network [44,45]. These results suggest that incorporating PPI network information from STRING may identify additional risk genes with more biological interpretability.

As an illustrative example, we applied TADA *de novo* test [16] with the CHD DNV data curated in our previous work [18], and conducted a post-association analysis on the p-values returned from the test. After false discovery rate (FDR) adjustment of p-values, we identified 21 genes with FDR<0.1 among 18,856 genes tested, and found that the number of edges formed by the 21 genes (20 edges, blue line in Fig 1A) is much larger than the upper tail of the empirical distribution sampled from 21 randomly selected genes in the STRING V11.0 database (score threshold: 400) for 10,000 times (Fig 1A). This suggests that the candidate CHD genes are highly enriched in terms of their interactions in the STRING database. To further illustrate that PPI information may contribute to CHD gene discovery, we showed the number of edges formed by the top genes ranked by adjusted p-values for CHD and compared it with

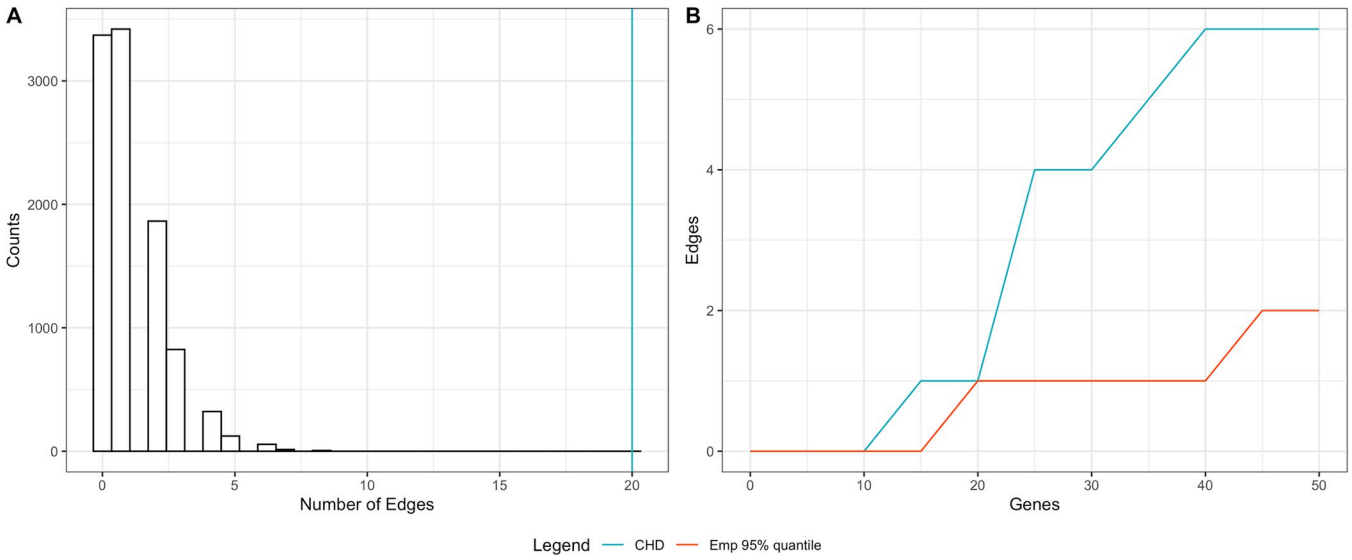

Fig 1. CHD top genes are more connected than randomly selected genes in the STRING PPI network. (A) Empirical distribution of the number of edges formed by 21 randomly selected genes. Blue line represents the number of edges formed by the 21 CHD top genes from TADA *de novo* analysis. (B) Blue line represents the number of edges formed by CHD top genes and red line represents randomly selected genes.

https://doi.org/10.1371/journal.pgen.1010252.g001

the number of edges formed by randomly selected genes with a more stringent selection of PPI edges in the STRING database (score threshold: 950) in Fig 1B. We considered 95th percentile of the empirical distribution derived from 10,000 sets of random genes in the PPI network as a baseline. When more than 20 top CHD genes are selected, the number of edges formed by these genes is significantly more than that from randomly selected genes. This suggests top genes in CHD tend to be neighbors in the STRING PPI network.

Motivated by the observation from Fig 1, we develop a **N**etwork assisted model for **D**e novo **A**ssociation **T**est using protein-protein inter**A**ction information, named N-DATA, to leverage prior information of interactions among genes from the PPI network to boost statistical power in identifying risk genes for CHD based on the 'guilt by association' principle [46, 47]. In the following, we first introduce the inference procedure for our model, and then demonstrate the performance of our method through simulation studies and real data applications.

Methods

In this section, we introduce the statistical model for the proposed framework. The network information in the PPI database is represented by an undirected graph $G = (V,E)$, where $V = \{1,..,n\}$ is a set of n genes in the network, and $E = \{<i,j>: i \text{ and } j \text{ are genes connected by the edges}\}$. The degree of a gene i is defined as the number of direct neighbors (N_i) for gene i in the network and denoted as d_i. We denote the latent association status of gene i with a disease of interest, e.g., CHD, as S_i, where $S_i = 1$ if gene i is associated with the disease, $S_i = -1$ if gene i is not associated with the disease. $S = \{S_1,..,S_n\}$ are the corresponding latent states for genes in $V = \{1,\ldots,n\}$. The DNV count of each gene i is defined as Y_i. We propose a simple discrete Markov random field model [48, 49] with a nearest neighbor Gibbs measure [50] to model the following joint probability function $S = \{S_1,..,S_n\}$:

$$P(S|\theta_0) \propto exp\left\{h\sum_{i\in V} I_1(S_i) + \tau_0 \sum_{<i,j>\in E} (w_i + w_j) I_{-1}(S_i) I_{-1}(S_j) + \tau_1 \sum_{<i,j>\in E} (w_i + w_j) I_1(S_i) I_1(S_j)\right\},$$

where w_i is the weight for gene i and will be chosen based on the characteristics of the network. In real data analysis, we set w_i as the square root of the degree of gene i in the network ($w_i = \sqrt{d_i}$) following Chen et al. [24]. $\theta_0 = (h, \tau_0, \tau_1)$ are hyperparameters related to the network. Specifically, h determines the marginal distribution of S_i when all genes are independent i.e., $P(S_i = 1|h, \tau_0 = \tau_1 = 0) = \frac{\exp(h)}{1+\exp(h)}$. τ_0 and τ_1 characterize the prior weights of edges between non-associated genes and associated genes, respectively. We further assume that, given the latent state S_i, the DNV count Y_i follows a Poisson distribution. The mutability of gene i (μ_i) can be estimated using the framework in Samocha et al. [12, 51]. Based on the derivation in TADA [16], the probability of observing DNVs for gene i in each trio can be approximated by $2\mu_i\gamma$, where γ is the relative risk of the DNVs. Further, the expected count of DNVs for gene i in N trios is $2N\mu_i\gamma$. When gene i is not a risk gene, γ is equal to 1. Then, we have the following model for DNV counts:

$$Y_i|S_i = -1 \sim \text{Poisson}(2N\mu_i),$$

$$Y_i|S_i = 1 \sim \text{Poisson}(2N\mu_i\gamma),$$

$$\theta_0 = (h, \tau_0, \tau_1); \theta_1 = \gamma.$$

To reduce the computational burden from a fully Bayesian solution for maximizing the marginal likelihood, we propose an empirical Bayes method to estimate the parameters θ_0 and

θ_1, and the latent association status S by maximizing the pseudo conditional likelihood (PCLK) for n genes as follows

$$\text{PCLK} = \prod_{i=1}^{n} \text{f}(Y_i|S_i, \theta_1)\text{Pr}(S_i|S_{N_i}, \theta_0),$$

where S_{N_i} represents the latent association status for neighbors of gene i. It has been shown that the estimator from the PCLK in a general Markov random field setting is consistent under mild regularity conditions [24,49]. When maximizing the PCLK, we can estimate the hyper-parameters θ_0, θ_1 and latent status S iteratively.

We can obtain an empirical estimate for θ_0 by maximizing $\prod_{i=1}^{n} \text{Pr}(S_i|S_{N_i}, \theta_0)$, which is equivalent to maximizing the parameters in the following logistic regression model:

$$\text{logit Pr}(S_i|S_{N_i}, \theta_0) = h + \tau_1 X_{i1} - \tau_0 X_{i0},$$

where $X_{i1} = w_i \sum_{k\in N_i} I_1(S_k) + \sum_{k\in N_i} w_k I_1(S_k)$ and $X_{i0} = w_i \sum_{k\in N_i} I_{-1}(S_k) + \sum_{k\in N_i} w_k I_{-1}(S_k), i = 1, \ldots, n$. To make sure the estimated θ_0 is finite, we can add a ridge penalty term $\lambda(h^2 + \tau_0^2 + \tau_1^2)$ to the likelihood function to solve the maximization problem by the Newton-Raphson's method [52].

We then update the latent status S by maximizing the PCLK using the iterative conditional mode method [49]. After we obtain the updated values θ_0 and S, we can estimate the hyper-parameter θ_0 by maximizing $\prod_{i=1}^{n} f(Y_i|S_i, \theta_1)$ by using the following closed-form expression:

$$\log L(\theta_1|Y) \propto \log \prod_{S_i=1} \exp(-2\mu N\gamma + Y_i \log\gamma)$$

$$\frac{\partial \log L(\theta_1|Y)}{\partial \gamma} = -\sum_{S_i=1} 2\mu N + \frac{\sum_{S_i=1} Y_i}{\gamma}$$

$$\hat{\gamma} = \frac{\sum_{S_i=1} Y_i}{\sum_{S_i=1} 2\mu N}$$

Algorithm 1: Procedure for Parameter Estimation

1. Set initial configuration $\boldsymbol{s}^0$
2. In the jth iteration, for given $\boldsymbol{s}^{(j-1)}$, obtain $\hat{\theta}_0^j$ from

$$\text{logit Pr}\,(S_i^{(j-1)}|S_{N_i}^{(j-1)}, \theta_0^{(j-1)}) = h + \tau_1 X_{i1} - \tau_0 X_{i0},\ i = 1, \cdots, n$$

3. Sequentially update the labels of nodes to obtain $S^{(j)}$ (ICM)

$$S_i^{(j)} = \arg\max_{s_i} f\,(Y_i|S_i, \hat{\theta}_1^{(j-1)})\text{Pr}\,(S_i|S_{N_i}^{(j-1)}, \hat{\theta}_0^{(j)}) \prod_{k\in S_N} \text{Pr}\,(S_k^{(j-1)}|S_i, S_{N_k-i}^{(j-1)}, \hat{\theta}_0^{(j)})$$

4. Obtain $\hat{\theta}_1^i(\hat{\gamma}^{(j)})$ from

$$\hat{\theta}_1^{(j)} = \text{argmax}_{\theta_1} \log L(\theta_1|\theta_0^{(j)}, \text{S}^{(j)}, \text{Y})$$

5. Repeat steps 2, 3, and 4 until convergence

Finally, after we obtain the estimated $\hat{\theta}_0$ and $\hat{\theta}_1$, we use Gibbs sampling based on the conditional distribution $P(S_i|S_{N_i}, \hat{\theta}_0, \hat{\theta}_1)$. This method has been proved to be valid for multiple testing under dependence in a compound decision theoretic framework [53,54]. Then, we can estimate the marginal posterior probability $q_i = P(S_i = -1|Y)$. Let $q_{(i)}$ be the sorted values of q_i in descending order. For each gene i, the null hypothesis and alternative hypothesis are

$$H_{i0}: \text{ Gene } i \text{ is not associated with the trait of interest}$$

$$H_{i1}: \text{ Gene } i \text{ is associated with the trait of interest}$$

As shown by Jiang and Yu [55], the relationship between global FDR and local FDR (lfdr) is FDR = $E(\text{lfdr}|Y \in \mathcal{R})$, where the rejection region $\mathcal{R}$ is the set of Y such that the null hypothesis can be rejected based on a specific rejection criterion. To control the expected global FDR less than α, we propose the following procedure: let $m = \max\left\{s : \frac{1}{s}\sum_{i=1}^{s} q_{(i)}\right\}$, we reject all the null hypotheses corresponding to $H_{(1)},\ldots,H_{(m)}$.

Verification and comparison

We used network information from the STRING PPI database and simulated DNV count data to study the performance of our method. First, we randomly selected 2,000 genes, retrieved their mutability from the real data, and extracted the corresponding PPI network formed by these 2,000 genes. Then, we simulated the latent status of genes with Gibbs sampling under the given network information, and the count of DNVs for each gene with the Poisson distribution given the latent status of the gene. We evaluated FDR and power under various settings of sample size N and relative risk parameter γ.

We fixed true network parameter h as -4 and varied τ_1 from 0.1 to 0.9 to make the total number of risk genes in the network of 2,000 random genes vary from 57 to 353. We varied the sample size N at 2,000, 5,000 and 10,000 to evaluate the performance of N-DATA in small, medium, and large WES cohorts, respectively. In addition, we varied β (log relative risk parameter γ) at 3, 3.5, and 4 to investigate the performance of N-DATA around the burden estimated results from real data (In real data analysis, $\hat{\beta} = 3.60$). Each simulation setting was replicated 100 times. For Gibbs sampling-based inference, we used 5,000 MCMC iterations, and set the first 2,000 iterations as burn-ins. These numbers were chosen empirically based on the diagnostic plots for convergence.

First, we compared the performance of N-DATA model with and without the PPI network as input. For N-DATA model without the PPI network, we assigned the weight of gene i $w_i = 0$ for inference. We present the power and FDR performance of N-DATA models in Fig A and Fig B in S1 Text. Then, we compared the power of TADA *de novo* test (TADA-*De novo*), DAWN, and N-DATA using the same simulation settings. Hyperprior of TADA-*De novo* was estimated from the function *denovo.MOM* based on the recommendation from the authors [16]. Power of TADA was calculated based on TADA p-values under FDR adjustment. DAWN v1.0 was downloaded from http://www.compgen.pitt.edu/DAWN/DAWN_homepage.htm. We adapted the code of DAWN by substituting the adjacency matrix inferred from its Partial Neighborhood Selection algorithm to the adjacency matrix from network. We used TADA-*De novo* p-values and PPI network as the input of DAWN. We applied default settings for parameters in DAWN.

We compared the performance of TADA-*De novo*, TADA-*De novo* p-values + DAWN and N-DATA under different simulation settings. We reported the power performance under FDR threshold 0.05 in the main text (Fig 2). We first checked if all three methods could control the

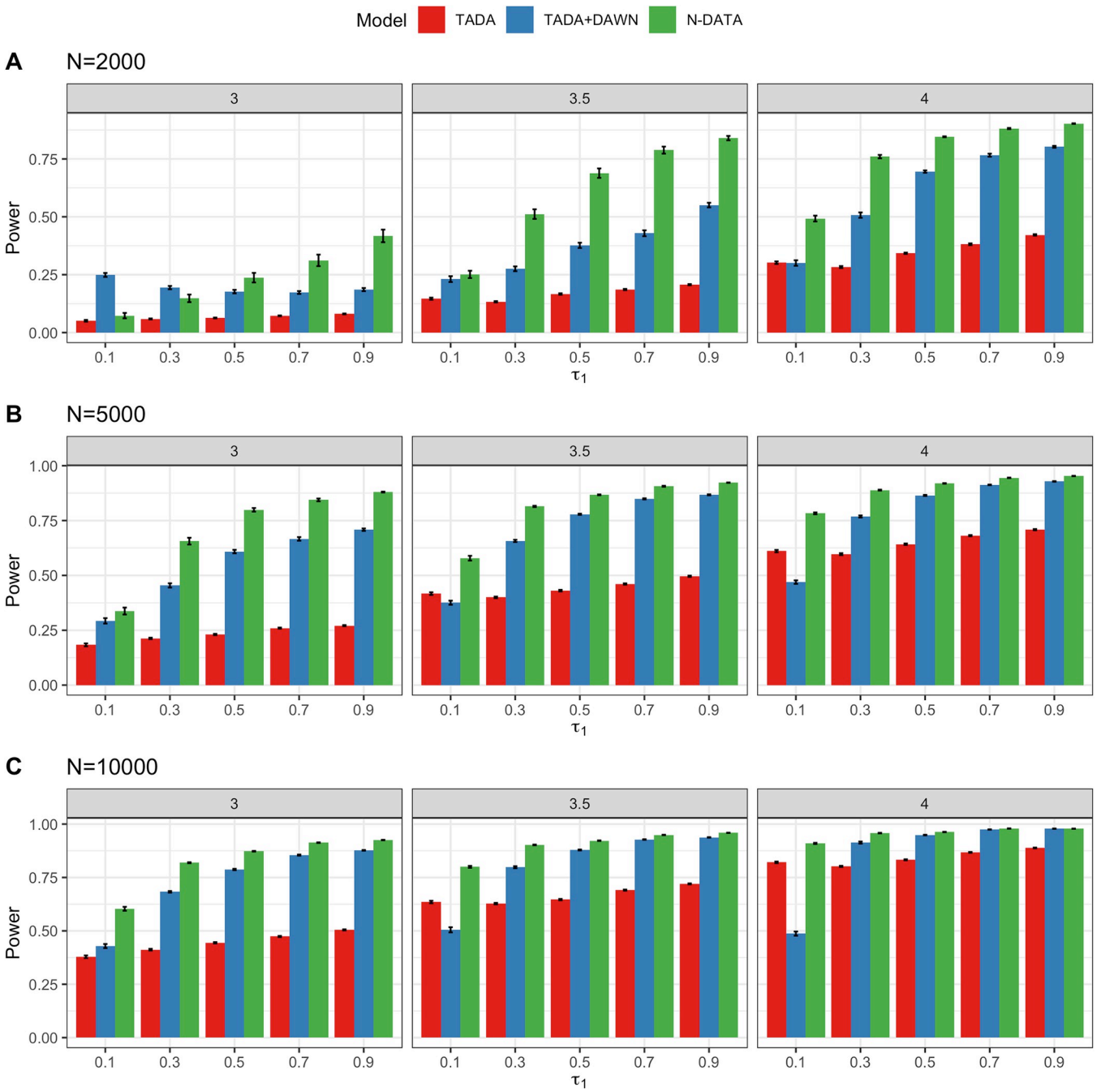

Fig 2. Power comparison of TADA-*De novo*, TADA-*De novo* p-values + DAWN and N-DATA. Error bars represent standard errors estimated from 100 replications of simulation. Three panels in each sub-figure from left to right represent $\beta = 3$, $\beta = 3.5$, and $\beta = 4$, respectively. Each panel shows the change of power when τ_1 varies from 0.1 to 0.9. (A) Power comparison between the two models when the sample size is small ($N = 2{,}000$). (B) Power comparison between the three models when the sample size is medium ($N = 5{,}000$). (C) Power comparison between the three models when the sample size is large ($N = 10{,}000$).

https://doi.org/10.1371/journal.pgen.1010252.g002

global FDR when the threshold is 0.05 (Fig C in S1 Text). Overall, N-DATA controlled the FDR well and had the best power under all scenarios. We observed that when τ_1, N, and β are all small, DAWN had FDR inflations for some runs. We suspect that this may be due to the discreteness of p-values, resulting in the violation of the normal distribution assumption for corresponding z-scores used in the input of DAWN. When the number of risk genes is small,

DAWN may have lower power than TADA and N-DATA. When τ_1, N, and β became larger, the power of DAWN was comparable with N-DATA. Time comparisons for the three models are presented in Fig D in S1 Text.

Application

We applied N-DATA to DNV data from 2,645 CHD trios reported in Jin et al [5], and annotated the CHD variants by ANNOVAR [56]. We denoted loss of function (LoF) as frameshift insertion/deletion, splice site alteration, stopgain and stoploss predicted by ANNOVAR, and deleterious missense (Dmis) predicted by the MetaSVM [57] algorithm. We only consider damaging variants (LoF and Dmis) in our analysis as the number of non-deleterious variants is not expected to provide the information to differentiate cases from controls biologically [58].

For network information, we first downloaded STRING v11.0 with medium edge likelihood via interface from STRINGdb package in R and call this original network from STRING $\mathcal{G}_0$. We obtained the curated list of known human CHD genes from Jin el al [5] and expanded the gene list by including additional candidate genes (FDR<0.1) from the single-trait analysis in our previous work [18]. This gene list (258 genes) was set as seed genes for our network. Then, we extracted the subnetwork including the seed genes and the direct neighbors with likelihood score larger than 950 of those genes and call this subnetwork $\mathcal{G}_1$. We only kept overlapping genes with our DNV data in $\mathcal{G}_1$ and called the final network used in our real application as $\mathcal{G}_2$. There were in total 1,814 genes and 21,468 edges in $\mathcal{G}_2$.

To show that our method can leverage network information to boost risk gene identification, we applied our algorithm without using the network as an input. When there was no prior information from the network, we identified 18 significant genes with FDR<0.05. To include the network information from $\mathcal{G}_2$we denote the degree of gene i in network $\mathcal{G}_2$ as d_i, and let the weight in the prior as $w_i = \sqrt{d_i}$. After adding the network information from $\mathcal{G}_2$, we identified 46 genes with at least 1 DNV, and 26 genes harboring at least 2 DNVs with FDR<0.05 in the CHD cohort.

We also compared the results of N-DATA with TADA-*De novo* test [16]. As in the simulation study, we observed that DAWN may not control the FDR under the preset threshold under our network and cohort settings. Thus, we did not include the results of DAWN in the comparison. TADA-*De novo* test (p-values with FDR adjustment) identified 28 significant genes. Without integrating the network information, N-DATA can identify 18 significant genes with FDR<0.05. After integrating the $\mathcal{G}_2$ network, N-DATA identified 323 genes with FDR<0.05. As some of the genes may be prioritized due to network characteristics, but did not have DNV count in the study cohort (more details in S1 Text), we further filtered out genes without DNV and considered the 46 genes identified with FDR<0.05 and at least 1 DNV as the candidate genes. (Table 1)

We visualized the overlap of 258 seed genes, genes that were identified by TADA-*De novo* p-values, N-DATA w/o network model, and N-DATA in Fig 3. Fig 3A shows the 323 genes

Table 1. Comparison of TADA and N-DATA models.

Method	Criteria	Number of Identified Genes
TADA-*De novo* p-values	FDR<0.05	28
N-DATA w/o Network	FDR<0.05	18
N-DATA	FDR<0.05	323
(Network $\mathcal{G}_2$ network + DNV counts)	DNVs≥1	46

https://doi.org/10.1371/journal.pgen.1010252.t001

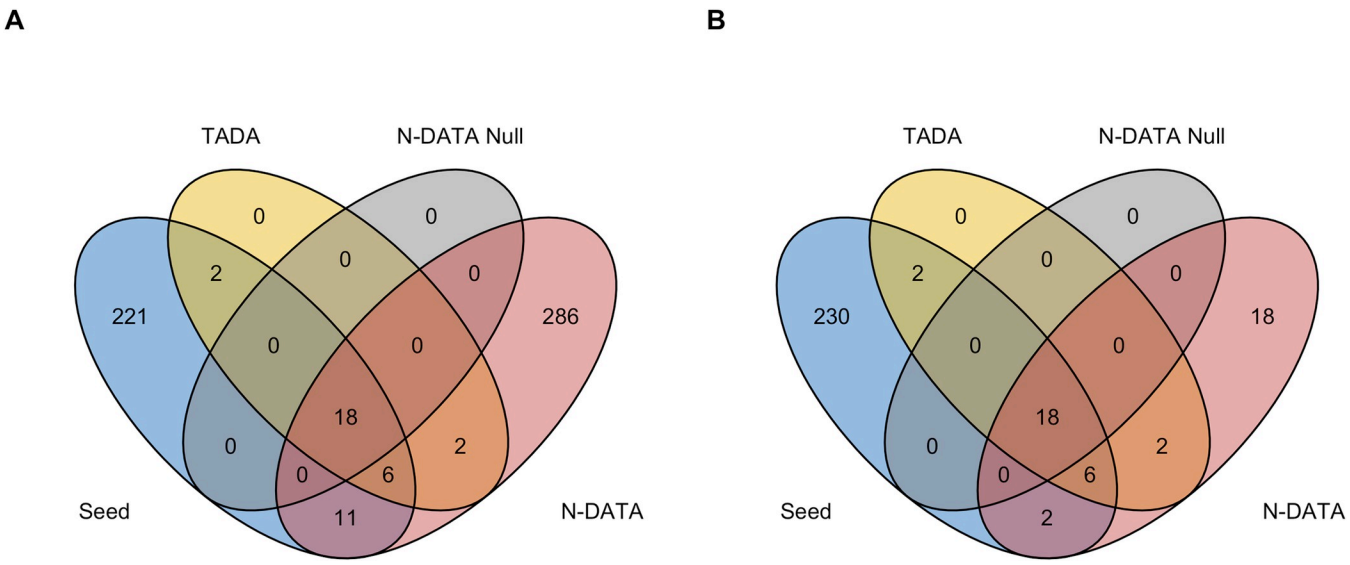

Fig 3. Venn diagram of 258 seed genes, TADA genes, N-DATA w/o network genes and N-DATA genes. (A) Overlapping genes between 258 seed genes, TADA genes, N-DATA w/o network (N-DATA Null) genes and 323 N-DATA genes. (B) Overlapping genes between 258 seed genes, TADA genes, N-DATA w/o network (N-DATA Null) genes and 46 N-DATA candidate genes.

https://doi.org/10.1371/journal.pgen.1010252.g003

identified by N-DATA, while Fig 3B shows the 46 genes with at least 1 DNV. From Fig 3B, N-DATA found most of the genes that can be identified by TADA (26 out of 28).

Further, we calculated the overlap of the significant genes identified by N-DATA and TADA, and 872 genes that are highly expressed (top 25%) in mouse developing heart at E14.5 [12] and in the 1,814 gene network (HHE genes) (Fig 4). Among the 323 N-DATA identified genes, 27 are known human CHD genes and 213 genes are HHE genes. Among the 46 genes, 18 are known human CHD genes and 35 are HHE genes.

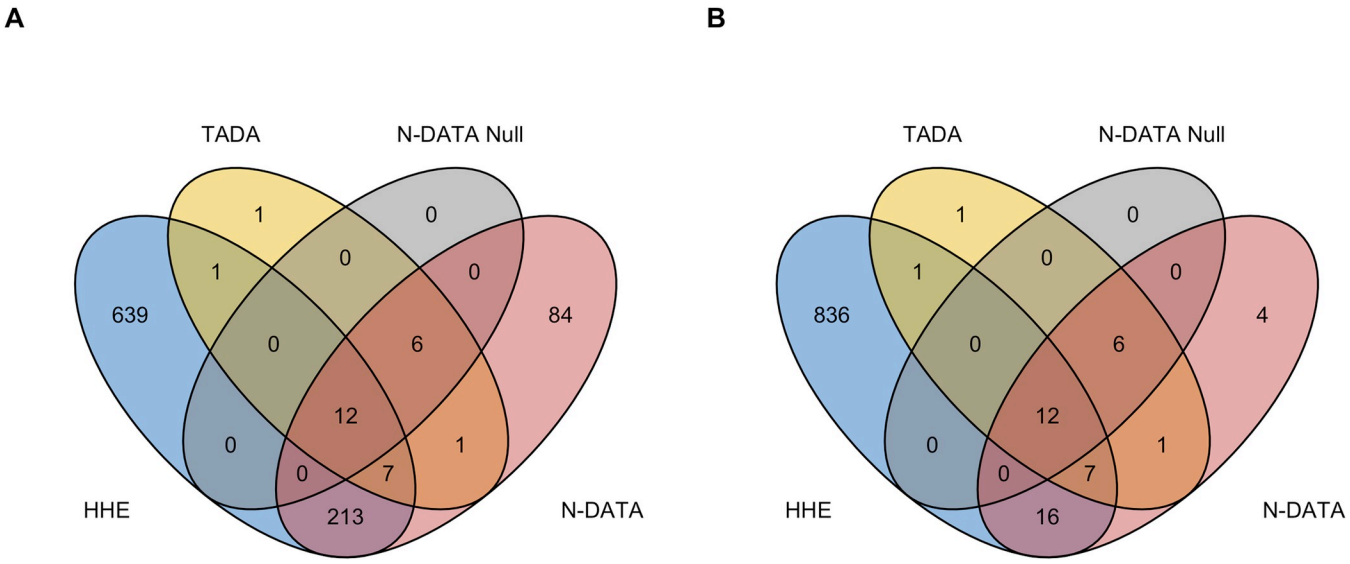

Fig 4. Venn diagram of HHE genes, TADA genes, N-DATA w/o network genes, and the N-DATA genes. (A) Overlapping genes between 872 HHE genes, TADA genes, N-DATA w/o network (N-DATA Null) genes, and 323 N-DATA genes. (B) Overlapping genes between 872 HHE genes, TADA genes, N-DATA w/o network (N-DATA Null) genes, and 46 N-DATA candidate genes.

https://doi.org/10.1371/journal.pgen.1010252.g004

We visualized the 323 genes identified in the $\mathcal{G}_2$ network (S1 Fig). The 323 genes formed two major clusters. The bigger cluster (right) is an extended cluster for protein synthesis genes, including ribosome protein genes (RPL-, RPS-), peptide chain elongation genes (EEF-, EIF-, SPR-, GSPT-), rRNA processing genes (UTP-, WDR-, RIOK-, NO-, IMP-), etc. Though without finding DNVs in the current cohort, ribosome genes *RPL11*, *RPL35A*, *RPS10*, *RPS19*, *RPS24*, *RPS26*, and *RPS7* are known CHD genes. Ribosome dysfunctions have been implicated in a variety of developmental disorders, including CHD [59]. For instance, multiple genes encoding ribosome subunits are known to cause Diamond-Blackfan anemia and 30% of the patients also presented CHD [60]. Functional studies showed that the deficiency in ribosomes can impact cell growth which might be a potential mechanism to cause CHD [61,62].

The other cluster (left) is the extended cluster for mRNA splicing genes, which encode various components of spliceosome and associated factors, such as snRNP (LSM-, SNRNP-, SNRP-), pre-mRNA processing factors (PRPF-), RNA helicases (DDX-, DHX-), hnRNPs (HNRNP-), and splicing factors (SF3-, SRS-, CWC-) [63]. Heart development involves many alternative splicing events. Mutations in splicing associated factor genes such as *RBM24*, *RBFOX2* and *SF3B1*, have been shown to cause cardiac malformation in mouse and human [64]. A specific type of snRNP called snoRNA and its targets showed reduced expression in myocardium of infants with Tetralogy of Fallot and impacted heart development through impairing spliceosome functions [65].

Thus, genes in the two clusters may be associated with CHD via disruption of protein synthesis or mRNA splicing events.

Further, we zoomed in on the 46 genes with at least 1 DNV in the $\mathcal{G}_2$ network to demonstrate that the PPI network information can help boost statistical power and provide biological interpretation for the current CHD cohort. (Fig 5)

Among the 46 candidate genes, *PTPN11*, *RAF1* and *RIT1* had 2 recurrent DNVs, and *CHD7*, *NOTCH1*, *NSD1* and *PYGL* also had recessive genotypes in the CHD cohort [5]. The 46 candidate genes form 4 clusters in the $\mathcal{G}_2$ network (Fig 5). The biggest cluster includes seven known CHD genes *TBX5*, *KMT2D*, *PTPN11*, *SOS1*, *ACTB*, *NOTCH1*, and *PTEN*, which are involved in transcriptional regulation and early cell growth or differentiation processes. The six new genes *SMAD2*, *KLF4*, *CTNNB1*, *CDC42*, *ITSN2*, and *WWTR1* also function in similar pathways and have varied implications for cardiac development. For instance, *KLF4* and *CTNNB1* have been implicated in cardiac cell differentiation [66]. *Cdc42* cardiomyocyte knock-out mice presented heart defects such as ventricular septum defects and thin ventricular walls [67]. *WWTR1* encodes a transcription regulator, which serves as an effector of Hippo pathway and regulates cardiac wall maturation in zebrafish [68].

The second biggest cluster is constituted of 7 new genes, all of which are involved in mRNA splicing. Specifically, *SART1*, *SRRM2*, *PRPF38A*, *PRPF8*, and *SF3B1* are associated factors or components of spliceosome; *HNRNPK* encodes a pre-mRNA-binding protein; *DHX9* encodes an RNA helicase which promotes R-loop formation while RNA splicing is perturbed [69]. Alternative splicing plays an essential role in heart development, homeostasis, and disease pathogenesis. Mouse knockouts of multiple splice factors had impaired cardiogenesis [70]. *SF3B1*, specifically, has been shown to upregulate to induce heart disease in both human and mice [64]. Thus, though not fully investigated, DNVs in those mRNA splicing-related genes may contribute to CHD pathogenesis.

The third cluster contains genes involved in protein synthesis, including the known gene *RPL5* and genes not previously associated with CHD (*EIF4*, *EIF5*, *EEF2*, and *RPL10)*. *RPL5* and *RPL10* encode the ribosome subunits. Mutations in *RPL5* and other ribosomal genes can lead to multiple congenital anomalies, including CHD [71]. *EIF4* and *EIF5* encode translation initiation factors while *EEF2* encodes the elongation factor that regulate peptide chain

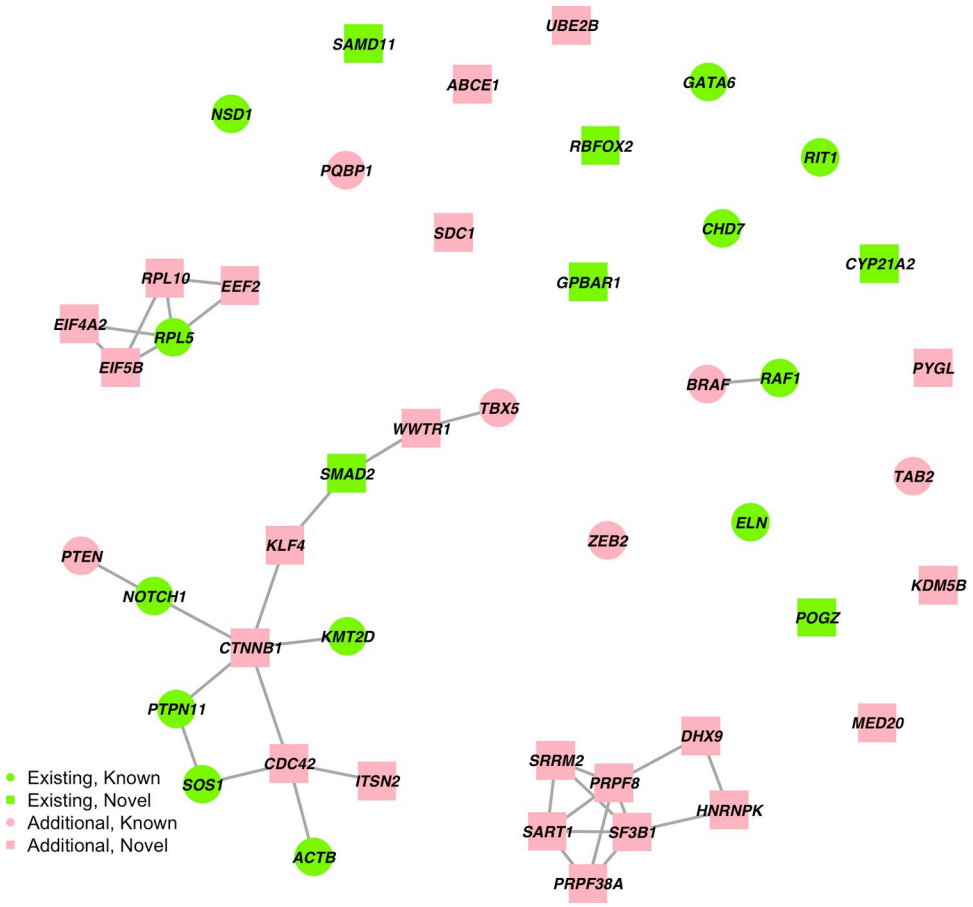

Fig 5. N-DATA model identified 46 candidate genes with at least 1 DNV. Green labels indicate the 18 genes identified when no network information was provided for N-DATA, and red labels indicate the additional 28 genes identified when the $\mathcal{G}_2$ network was integrated. Circles indicate the 18 known human CHD genes, and squares indicate the 28 novel genes identified by N-DATA.

https://doi.org/10.1371/journal.pgen.1010252.g005

elongation during protein synthesis. A recent study reported that the deficiency in ribosome associated NatA complex reduces ribosomal protein and subsequently impact cell development as a mechanism to cause CHD [62]. Thus, DNVs in the above genes may lead to CHD via impairment of protein synthesis.

The last cluster contains the known CHD genes *BRAF* and *RAF1*, both of which encode key kinases in Ras signaling and are related to Noonan syndrome with CHD as a common feature.

Among the un-clustered genes, six are identified after using the network information: *ABCE1*, *UBE2B*, *SDC1*, *PYGL*, *KDM5B*, *MED20*. *UBE2B* and *KDM5B*, encoding epigenetic modifiers, have shown suggestive evidence in cardiac development or CHD [72,73] and might be potential CHD genes.

Discussion

In this article, we have introduced a Bayesian framework to integrate PPI network information as the prior knowledge into DNV analysis for CHD. The implemented model is available at https://github.com/JustinaXie/NDATA. This approach adopts MRF to model the interactions among genes. We apply an empirical Bayes strategy to estimate parameters in the model and conduct statistical inference based on the posterior distribution sampled from a Gibbs

sampler. The simulation studies and real data analysis on CHD suggest that the proposed method has improved power to identify risk genes over methods without integrating network information.

Our proposed framework is innovative in the following aspects. First, it does not need to estimate hyperprior based on other sources compared to the existing pathway-based test for DNV data [22,45]. Second, it does not require external expression data for the DNV cohort and uses the publicly available PPI database instead, which makes it more applicable to different diseases. This method not only increases power in risk gene identification, but also assists in biological interpretation by visualizing clusters of risk genes with functional relevance in the network.

However, there are some limitations in the current N-DATA model. First, our model is dependent on the choice of network. Using different PPI networks and different filtering criteria could result in a different set of significant genes. Currently, we did not provide a way to prioritize existing networks. We have provided details on a comparison of using the HINT network versus the STRING network (more details in S1 Text). For the two networks compared in our study, HINT has the advantage of leveraging additional information from PDB [74], and being manually curated to filter out erroneous and low-quality interactions; while STRING has the advantage of providing a score to measure the likelihood of interactions, and including information from multiple pathway databases. We also found that the risk genes identified from the two databases had a significant overlap ($p = 3.14\times10^{-12}$). The overlapping risk genes were highly enriched for Human Phenotype Ontology (HPO) [75] terms related to CHD from g:Profiler [76] analysis, and p-values of overlapped pathway outputs from Ingenuity Pathways Analysis (IPA, QIAGEN Inc.) had a significant Pearson's correlation ($R = 0.48$; $p<2.2\times10^{-16}$).

Second, our model may be only used for early on-set disorders with a strong DNV signal. For diseases with small relative risks or small sample sizes, our model may suffer from convergency issues (more details in S1 Text). In real applications, it is important to conduct an initial analysis on the enrichment of top genes identified from *de novo* association tests in the network like our motivating example.

Third, we applied an empirical Bayes strategy to obtain point estimates of hyperparameters instead of using a fully Bayesian approach considering the computation burden. A fully Bayesian model that can account for intrinsic uncertainties would be a potential future direction. Fourth, likelihood-based inference may suffer from local maxima [24]. Although we didn't identify significant differences of different initiation points from our simulation study (more details in S1 Text), we recommend initiating the labels of genes from a known risk gene set or running with multiple starts for real data application (more details in S1 Text). Also, we observe the Gibbs sampler tends to move around local maxima for some time before convergence. Empirically, we suggest running at least 2,000 times of iterations and discarding the first 1,000 iterations as burn-ins. Fifth, we only considered the simulation verification under the ground truth model based on our assumptions, the generalizability to other alternative models is unexplored.

Sixth, to apply our model to other diseases, practitioners should be cautious if they would like to use the mutability of genes from a public dataset. 1) For WES data, the target region for each study could be different, which further results in differences in the calculation of mutability for the coding region 2) Mutability may be calculated based on a specific functional annotation of variants. Studies that use divergent classification criteria for variants should not share the same mutability. 3) Publicly available mutability may have been adjusted for cohort-specific parameters, such as sequencing depth, which may also affect the results if adapted to another cohort.

In addition, we only considered damaging DNVs and assumed the relative risk parameter γ is the same across all genes in N-DATA, which may cause our model to lose power if it varies across variants with different functions (e.g., LoF and Dmis). Future studies may explore adding functional annotation of variants as a layer in the model to further improve statistical power.

Supporting information

S1 Text. Supplementary notes on methods and results. Fig A in S1 Text: Power comparison of N-DATA w/o and w/ PPI network models. Fig B in S1 Text: FDR comparison of N-DATA w/o and w/ PPI network models. Fig C in S1 Text: FDR comparison of TADA-De novo, TADA-De novo p-values + DAWN and N-DATA. Fig D in S1 Text: Time comparison of TADA-De novo, TADA-De novo p-values + DAWN and N-DATA.
(DOCX)

S1 Fig. N-DATA identified 323 genes in network $\mathcal{G}_2$. Green labels indicate the 18 genes identified when no network information was provided for N-DATA, and red labels indicate the additional 305 genes identified when the $\mathcal{G}_2$ network was integrated. Circles indicate the 27 known human CHD genes, and squares indicate the 296 novel genes.
(SVG)

S1 Table. Simulation results.
(XLSX)

S2 Table. Results of real data application.
(XLSX)

Acknowledgments

We thank Jin et al. [5] for sharing the *de novo* variant data of CHD. We thank Andrew Xu and Dr. Min Chen for discussions on coding, Ziyu Jiang for discussions on diagnostic for Bayesian inference, and Dr, Peifeng Ruan for discussions on PPI networks.

Author Contributions

Conceptualization: Yuhan Xie, Wei Jiang.

Data curation: Yuhan Xie, Hongyu Li.

Formal analysis: Yuhan Xie.

Funding acquisition: Hongyu Zhao.

Investigation: Hongyu Zhao.

Methodology: Yuhan Xie, Wei Jiang, Hongyu Li.

Project administration: Yuhan Xie, Hongyu Zhao.

Resources: Weilai Dong, Sheng Chih Jin, Martina Brueckner, Hongyu Zhao.

Software: Yuhan Xie.

Supervision: Wei Jiang, Hongyu Zhao.

Validation: Weilai Dong, Sheng Chih Jin, Martina Brueckner.

Visualization: Yuhan Xie.

Writing – original draft: Yuhan Xie, Wei Jiang, Weilai Dong, Hongyu Zhao.

Writing – review & editing: Yuhan Xie, Wei Jiang, Weilai Dong, Hongyu Li, Sheng Chih Jin, Martina Brueckner, Hongyu Zhao.

References

1. Zaidi S, Choi M, Wakimoto H, Ma L, Jiang J, Overton JD, et al. De novo mutations in histone-modifying genes in congenital heart disease. Nature. 2013; 498(7453):220–3. Epub 2013/05/12. https://doi.org/10.1038/nature12141 PMID: 23665959.
2. Postma AV, Bezzina CR, Christoffels VM. Genetics of congenital heart disease: the contribution of the noncoding regulatory genome. Journal of Human Genetics. 2016; 61(1):13–9. https://doi.org/10.1038/jhg.2015.98 PMID: 26223183
3. Sevim Bayrak C, Zhang P, Tristani-Firouzi M, Gelb BD, Itan Y. De novo variants in exomes of congenital heart disease patients identify risk genes and pathways. Genome Med. 2020; 12(1):9. Epub 2020/01/17. https://doi.org/10.1186/s13073-019-0709-8 PMID: 31941532; PubMed Central PMCID: PMC6961332.
4. Diab NS, Barish S, Dong W, Zhao S, Allington G, Yu X, et al. Molecular Genetics and Complex Inheritance of Congenital Heart Disease. Genes (Basel). 2021; 12(7). Epub 2021/07/03. https://doi.org/10.3390/genes12071020 PMID: 34209044; PubMed Central PMCID: PMC8307500.
5. Jin SC, Homsy J, Zaidi S, Lu Q, Morton S, DePalma SR, et al. Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. Nat Genet. 2017; 49(11):1593–601. Epub 2017/10/11. https://doi.org/10.1038/ng.3970 PMID: 28991257; PubMed Central PMCID: PMC5675000.
6. Zaidi S, Brueckner M. Genetics and Genomics of Congenital Heart Disease. Circ Res. 2017; 120(6):923–40. Epub 2017/03/18. https://doi.org/10.1161/CIRCRESAHA.116.309140 PMID: 28302740; PubMed Central PMCID: PMC5557504.
7. Glessner JT, Bick AG, Ito K, Homsy J, Rodriguez-Murillo L, Fromer M, et al. Increased frequency of de novo copy number variants in congenital heart disease by integrative analysis of single nucleotide polymorphism array and exome sequence data. Circ Res. 2014; 115(10):884–96. Epub 2014/09/11. https://doi.org/10.1161/CIRCRESAHA.115.304458 PMID: 25205790; PubMed Central PMCID: PMC4209190.
8. Soemedi R, Wilson IJ, Bentham J, Darlay R, Töpf A, Zelenika D, et al. Contribution of global rare copy-number variants to the risk of sporadic congenital heart disease. Am J Hum Genet. 2012; 91(3):489–501. Epub 2012/09/04. https://doi.org/10.1016/j.ajhg.2012.08.003 PMID: 22939634; PubMed Central PMCID: PMC3511986.
9. Pierpont ME, Brueckner M, Chung WK, Garg V, Lacro RV, McGuire AL, et al. Genetic Basis for Congenital Heart Disease: Revisited: A Scientific Statement From the American Heart Association. Circulation. 2018; 138(21):e653–e711. Epub 2018/12/21. https://doi.org/10.1161/CIR.0000000000000606 PMID: 30571578; PubMed Central PMCID: PMC6555769.
10. Teer JK, Mullikin JC. Exome sequencing: the sweet spot before whole genomes. Human Molecular Genetics. 2010; 19(R2):R145–R51. https://doi.org/10.1093/hmg/ddq333 PMID: 20705737
11. Rabbani B, Tekin M, Mahdieh N. The promise of whole-exome sequencing in medical genetics. Journal of Human Genetics. 2014; 59(1):5–15. https://doi.org/10.1038/jhg.2013.114 PMID: 24196381
12. Homsy J, Zaidi S, Shen Y, Ware JS, Samocha KE, Karczewski KJ, et al. De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. Science. 2015; 350(6265):1262–6. Epub 2016/01/20. https://doi.org/10.1126/science.aac9396 PMID: 26785492; PubMed Central PMCID: PMC4890146.
13. Richter F, Morton SU, Kim SW, Kitaygorodsky A, Wasson LK, Chen KM, et al. Genomic analyses implicate noncoding de novo variants in congenital heart disease. Nature genetics. 2020; 52(8):769–77. https://doi.org/10.1038/s41588-020-0652-z PMID: 32601476
14. Watkins WS, Hernandez EJ, Wesolowski S, Bisgrove BW, Sunderland RT, Lin E, et al. De novo and recessive forms of congenital heart disease have distinct genetic and phenotypic landscapes. Nature communications. 2019; 10(1):1–12.
15. Sifrim A, Hitz M-P, Wilsdon A, Breckpot J, Turki SHA, Thienpont B, et al. Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing. Nature Genetics. 2016; 48(9):1060–5. https://doi.org/10.1038/ng.3627 PMID: 27479907
16. He X, Sanders SJ, Liu L, De Rubeis S, Lim ET, Sutcliffe JS, et al. Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. PLoS Genet. 2013; 9(8):e1003671. Epub 2013/08/24. https://doi.org/10.1371/journal.pgen.1003671 PMID: 23966865; PubMed Central PMCID: PMC3744441.

17. Nguyen T-H, Dobbyn A, Brown RC, Riley BP, Buxbaum JD, Pinto D, et al. mTADA is a framework for identifying risk genes from de novo mutations in multiple traits. Nature Communications. 2020; 11 (1):2929. https://doi.org/10.1038/s41467-020-16487-z PMID: 32522981

18. Xie Y, Li M, Dong W, Jiang W, Zhao H. M-DATA: A statistical approach to jointly analyzing de novo mutations for multiple traits. PLoS Genet. 2021; 17(11):e1009849. Epub 2021/11/05. https://doi.org/10.1371/journal.pgen.1009849 PMID: 34735430; PubMed Central PMCID: PMC8568192.

19. Nguyen HT, Bryois J, Kim A, Dobbyn A, Huckins LM, Munoz-Manchado AB, et al. Integrated Bayesian analysis of rare exonic variants to identify risk genes for schizophrenia and neurodevelopmental disorders. Genome Med. 2017; 9(1):114. Epub 2017/12/22. https://doi.org/10.1186/s13073-017-0497-y PMID: 29262854; PubMed Central PMCID: PMC5738153.

20. Liu Y, Liang Y, Cicek AE, Li Z, Li J, Muhle RA, et al. A Statistical Framework for Mapping Risk Genes from De Novo Mutations in Whole-Genome-Sequencing Studies. Am J Hum Genet. 2018; 102 (6):1031–47. Epub 2018/05/15. https://doi.org/10.1016/j.ajhg.2018.03.023 PMID: 29754769; PubMed Central PMCID: PMC5992125.

21. Mo Li XZ, Chentian Jin, Sheng Chih Jin, Weilai Dong, Martina Brueckner, Richard Lifton, Qiongshi Lu, Hongyu Zhao. Integrative modeling of transmitted and de novo variants identifies novel risk genes for congenital heart disease. Quant Biol.0-${article.jieShuYe}. https://doi.org/10.15302/j-qb-021-0248 PMID: 35414959

22. Nguyen TH, He X, Brown RC, Webb BT, Kendler KS, Vladimirov VI, et al. DECO: a framework for jointly analyzing de novo and rare case/control variants, and biological pathways. Brief Bioinform. 2021. Epub 2021/04/02. https://doi.org/10.1093/bib/bbab067 PMID: 33791774.

23. Lee I, Blom UM, Wang PI, Shim JE, Marcotte EM. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome Res. 2011; 21(7):1109–21. Epub 2011/05/04. https://doi.org/10.1101/gr.118992.110 PMID: 21536720; PubMed Central PMCID: PMC3129253.

24. Chen M, Cho J, Zhao H. Incorporating biological pathways via a Markov random field model in genome-wide association studies. PLoS Genet. 2011; 7(4):e1001353. Epub 2011/04/15. https://doi.org/10.1371/journal.pgen.1001353 PMID: 21490723; PubMed Central PMCID: PMC3072362.

25. Hou L, Chen M, Zhang CK, Cho J, Zhao H. Guilt by rewiring: gene prioritization through network rewiring in genome wide association studies. Hum Mol Genet. 2014; 23(10):2780–90. Epub 2014/01/02. https://doi.org/10.1093/hmg/ddt668 PMID: 24381306; PubMed Central PMCID: PMC3990172.

26. Li H, Zhu B, Xu Z, Adams T, Kaminski N, Zhao H. A Markov random field model for network-based differential expression analysis of single-cell RNA-seq data. BMC Bioinformatics. 2021; 22(1):524. Epub 20211026. https://doi.org/10.1186/s12859-021-04412-0 PMID: 34702190; PubMed Central PMCID: PMC8549347.

27. Liu L, Lei J, Roeder K. Network assisted analysis to reveal the genetic basis of autism. The Annals of Applied Statistics. 2015; 9(3):1571–600, 30. https://doi.org/10.1214/15-AOAS844 PMID: 27134692

28. Oughtred R, Stark C, Breitkreutz BJ, Rust J, Boucher L, Chang C, et al. The BioGRID interaction database: 2019 update. Nucleic Acids Res. 2019; 47(D1):D529–d41. Epub 2018/11/27. https://doi.org/10.1093/nar/gky1079 PMID: 30476227; PubMed Central PMCID: PMC6324058.

29. Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, et al. The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res. 2014; 42(Database issue):D358–63. Epub 2013/11/16. https://doi.org/10.1093/nar/gkt1115 PMID: 24234451; PubMed Central PMCID: PMC3965093.

30. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. The Database of Interacting Proteins: 2004 update. Nucleic Acids Res. 2004; 32(Database issue):D449–51. Epub 2003/12/19. https://doi.org/10.1093/nar/gkh086 PMID: 14681454; PubMed Central PMCID: PMC308820.

31. Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, Galeota E, et al. MINT, the molecular interaction database: 2012 update. Nucleic Acids Res. 2012; 40(Database issue):D857–61. Epub 2011/11/19. https://doi.org/10.1093/nar/gkr930 PMID: 22096227; PubMed Central PMCID: PMC3244991.

32. Mishra GR, Suresh M, Kumaran K, Kannabiran N, Suresh S, Bala P, et al. Human protein reference database—2006 update. Nucleic Acids Res. 2006; 34(Database issue):D411–4. Epub 2005/12/31. https://doi.org/10.1093/nar/gkj141 PMID: 16381900; PubMed Central PMCID: PMC1347503.

33. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019; 47(D1):D607–d13. Epub 2018/11/27. https://doi.org/10.1093/nar/gky1131 PMID: 30476243; PubMed Central PMCID: PMC6323986.

34. Das J, Yu H. HINT: High-quality protein interactomes and their applications in understanding human disease. BMC Systems Biology. 2012; 6(1):92. https://doi.org/10.1186/1752-0509-6-92 PMID: 22846459

35. Kalathur RKR, Pinto JP, Hernández-Prieto MA, Machado RSR, Almeida D, Chaurasia G, et al. UniHI 7: an enhanced database for retrieval and interactive analysis of human molecular interaction networks.

Nucleic acids research. 2014; 42(Database issue):D408–D14. Epub 2013/11/08. https://doi.org/10.1093/nar/gkt1100 PMID: 24214987.

36. Elefsinioti A, Saraç Ö S, Hegele A, Plake C, Hubner NC, Poser I, et al. Large-scale de novo prediction of physical protein-protein association. Mol Cell Proteomics. 2011; 10(11):M111.010629. Epub 20110811. https://doi.org/10.1074/mcp.M111.010629 PMID: 21836163; PubMed Central PMCID: PMC3226409.

37. Fahey ME, Bennett MJ, Mahon C, Jäger S, Pache L, Kumar D, et al. GPS-Prot: A web-based visualization platform for integrating host-pathogen interaction data. BMC Bioinformatics. 2011; 12(1):298. https://doi.org/10.1186/1471-2105-12-298 PMID: 21777475

38. Bajpai AK, Davuluri S, Tiwary K, Narayanan S, Oguru S, Basavaraju K, et al. Systematic comparison of the protein-protein interaction databases from a user's perspective. Journal of Biomedical Informatics. 2020; 103:103380. https://doi.org/10.1016/j.jbi.2020.103380 PMID: 32001390

39. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research. 2000; 28(1):27–30. https://doi.org/10.1093/nar/28.1.27 PMID: 10592173.

40. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Research. 2015; 44(D1):D457–D62. https://doi.org/10.1093/nar/gkv1070 PMID: 26476454

41. Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. KEGG: integrating viruses and cellular organisms. Nucleic Acids Res. 2021; 49(D1):D545–d51. https://doi.org/10.1093/nar/gkaa970 PMID: 33125081; PubMed Central PMCID: PMC7779016.

42. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics. 2000; 25(1):25–9. https://doi.org/10.1038/75556 PMID: 10802651.

43. Gene Ontology C. The Gene Ontology resource: enriching a GOld mine. Nucleic acids research. 2021; 49(D1):D325–D34. https://doi.org/10.1093/nar/gkaa1113 PMID: 33290552.

44. Nguyen TH, Dobbyn A, Brown RC, Riley BP, Buxbaum JD, Pinto D, et al. mTADA is a framework for identifying risk genes from de novo mutations in multiple traits. Nat Commun. 2020; 11(1):2929. Epub 2020/06/12. https://doi.org/10.1038/s41467-020-16487-z PMID: 32522981; PubMed Central PMCID: PMC7287090.

45. Nguyen HT, Dobbyn A, Charney AW, Bryois J, Kim A, Mcfadden W, et al. Integrative analysis of rare variants and pathway information shows convergent results between immune pathways, drug targets and epilepsy genes. bioRxiv. 2018:410100. https://doi.org/10.1101/410100

46. Oti M, Brunner H. The modular nature of genetic diseases. Clinical Genetics. 2007; 71(1):1–11. https://doi.org/10.1111/j.1399-0004.2006.00708.x PMID: 17204041

47. Moreau Y, Tranchevent L-C. Computational tools for prioritizing candidate genes: boosting disease gene discovery. Nature Reviews Genetics. 2012; 13(8):523–36. https://doi.org/10.1038/nrg3253 PMID: 22751426

48. Besag J. Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society: Series B (Methodological). 1974; 36(2):192–225.

49. Besag J. On the statistical analysis of dirty pictures. Journal of the Royal Statistical Society: Series B (Methodological). 1986; 48(3):259–79.

50. Kindermann R. Markov random fields and their applications. American mathematical society. 1980.

51. Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM, et al. A framework for the interpretation of de novo mutation in human disease. Nat Genet. 2014; 46(9):944–50. Epub 2014/08/05. https://doi.org/10.1038/ng.3050 PMID: 25086666; PubMed Central PMCID: PMC4222185.

52. Le Cessie S, Van Houwelingen JC. Ridge estimators in logistic regression. Journal of the Royal Statistical Society: Series C (Applied Statistics). 1992; 41(1):191–201.

53. Sun W, Tony Cai T. Large-scale multiple testing under dependence. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2009; 71(2):393–424.

54. Li H, Wei Z, Maris J. A hidden Markov random field model for genome-wide association studies. Biostatistics. 2010; 11(1):139–50. Epub 2009/10/14. https://doi.org/10.1093/biostatistics/kxp043 PMID: 19822692; PubMed Central PMCID: PMC2800164.

55. Jiang W, Yu W. Controlling the joint local false discovery rate is more powerful than meta-analysis methods in joint analysis of summary statistics from multiple genome-wide association studies. Bioinformatics. 2016; 33(4):500–7. https://doi.org/10.1093/bioinformatics/btw690 PMID: 28011772

56. Yang H, Wang K. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. Nat Protoc. 2015; 10(10):1556–66. Epub 2015/09/18. https://doi.org/10.1038/nprot.2015.105 PMID: 26379229; PubMed Central PMCID: PMC4718734.

57. Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Human Molecular Genetics. 2014; 24(8):2125–37. https://doi.org/10.1093/hmg/ddu733 PMID: 25552646

58. Li M. Gene-based Association Analysis for Genome-wide Association and Whole-exome Sequencing Studies: Yale University; 2020.

59. Narla A, Ebert BL. Ribosomopathies: human disorders of ribosome dysfunction. Blood. 2010; 115 (16):3196–205. Epub 20100301. https://doi.org/10.1182/blood-2009-10-178129 PMID: 20194897; PubMed Central PMCID: PMC2858486.

60. Vlachos A, Osorio DS, Atsidaftos E, Kang J, Lababidi ML, Seiden HS, et al. Increased Prevalence of Congenital Heart Disease in Children With Diamond Blackfan Anemia Suggests Unrecognized Diamond Blackfan Anemia as a Cause of Congenital Heart Disease in the General Population: A Report of the Diamond Blackfan Anemia Registry. Circ Genom Precis Med. 2018; 11(5):e002044. https://doi.org/10.1161/CIRCGENETICS.117.002044 PMID: 29748317; PubMed Central PMCID: PMC5951415.

61. Cheng Z, Mugler CF, Keskin A, Hodapp S, Chan LY, Weis K, et al. Small and Large Ribosomal Subunit Deficiencies Lead to Distinct Gene Expression Signatures that Reflect Cellular Growth Rate. Mol Cell. 2019; 73(1):36–47.e10. Epub 20181129. https://doi.org/10.1016/j.molcel.2018.10.032 PMID: 30503772; PubMed Central PMCID: PMC6382079.

62. Ward T, Tai W, Morton S, Impens F, Van Damme P, Van Haver D, et al. Mechanisms of Congenital Heart Disease Caused by NAA15 Haploinsufficiency. Circ Res. 2021; 128(8):1156–69. Epub 2021/02/10. https://doi.org/10.1161/CIRCRESAHA.120.316966 PMID: 33557580; PubMed Central PMCID: PMC8048381.

63. Shi Y. Mechanistic insights into precursor messenger RNA splicing by the spliceosome. Nat Rev Mol Cell Biol. 2017; 18(11):655–70. Epub 20170927. https://doi.org/10.1038/nrm.2017.86 PMID: 28951565.

64. van den Hoogenhof MM, Pinto YM, Creemers EE. RNA Splicing: Regulation and Dysregulation in the Heart. Circ Res. 2016; 118(3):454–68. Epub 2016/02/06. https://doi.org/10.1161/CIRCRESAHA.115.307872 PMID: 26846640.

65. Nagasawa C, Ogren A, Kibiryeva N, Marshall J, O'Brien JE, Kenmochi N, et al. The Role of scaRNAs in Adjusting Alternative mRNA Splicing in Heart Development. J Cardiovasc Dev Dis. 2018; 5(2). Epub 20180508. https://doi.org/10.3390/jcdd5020026 PMID: 29738469; PubMed Central PMCID: PMC6023535.

66. Kami D, Kitani T, Kawasaki T, Gojo S. Cardiac mesenchymal progenitors differentiate into adipocytes via Klf4 and c-Myc. Cell Death Dis. 2016; 7(4):e2190–e. https://doi.org/10.1038/cddis.2016.31 PMID: 27077806.

67. Liu Y, Wang J, Li J, Wang R, Tharakan B, Zhang SL, et al. Deletion of Cdc42 in embryonic cardiomyocytes results in right ventricle hypoplasia. Clin Transl Med. 2017; 6(1):40–. https://doi.org/10.1186/s40169-017-0171-4 PMID: 29101495.

68. Lai JKH, Collins MM, Uribe V, Jiménez-Amilburu V, Günther S, Maischein HM, et al. The Hippo pathway effector Wwtr1 regulates cardiac wall maturation in zebrafish. Development. 2018; 145(10). Epub 2018/05/19. https://doi.org/10.1242/dev.159210 PMID: 29773645.

69. Chakraborty P, Huang JTJ, Hiom K. DHX9 helicase promotes R-loop formation in cells with impaired RNA splicing. Nat Commun. 2018; 9(1):4346. Epub 2018/10/21. https://doi.org/10.1038/s41467-018-06677-1 PMID: 30341290; PubMed Central PMCID: PMC6195550.

70. Zahr HC, Jaalouk DE. Exploring the Crosstalk Between LMNA and Splicing Machinery Gene Mutations in Dilated Cardiomyopathy. Front Genet. 2018; 9:231. Epub 2018/07/28. https://doi.org/10.3389/fgene.2018.00231 PMID: 30050558; PubMed Central PMCID: PMC6052891.

71. Gazda HT, Sheen MR, Vlachos A, Choesmel V, O'Donohue MF, Schneider H, et al. Ribosomal protein L5 and L11 mutations are associated with cleft palate and abnormal thumbs in Diamond-Blackfan anemia patients. Am J Hum Genet. 2008; 83(6):769–80. Epub 2008/12/09. https://doi.org/10.1016/j.ajhg.2008.11.004 PMID: 19061985; PubMed Central PMCID: PMC2668101.

72. Robson A, Makova SZ, Barish S, Zaidi S, Mehta S, Drozd J, et al. Histone H2B monoubiquitination regulates heart development via epigenetic control of cilia motility. Proc Natl Acad Sci U S A. 2019; 116 (28):14049–54. Epub 2019/06/27. https://doi.org/10.1073/pnas.1808341116 PMID: 31235600; PubMed Central PMCID: PMC6628794.

73. Audain E, Wilsdon A, Breckpot J, Izarzugaza JMG, Fitzgerald TW, Kahlert AK, et al. Integrative analysis of genomic variants reveals new associations of candidate haploinsufficient genes with congenital heart disease. PLoS Genet. 2021; 17(7):e1009679. Epub 2021/07/30. https://doi.org/10.1371/journal.pgen.1009679 PMID: 34324492; PubMed Central PMCID: PMC8354477.

74. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic acids research. 2000; 28(1):235–42. https://doi.org/10.1093/nar/28.1.235 PMID: 10592235.

75. Köhler S, Gargano M, Matentzoglu N, Carmody LC, Lewis-Smith D, Vasilevsky NA, et al. The Human Phenotype Ontology in 2021. Nucleic Acids Res. 2021; 49(D1):D1207–d17. Epub 2020/12/03. https://doi.org/10.1093/nar/gkaa1043 PMID: 33264411; PubMed Central PMCID: PMC7778952.

76. Reimand J, Arak T, Adler P, Kolberg L, Reisberg S, Peterson H, et al. g:Profiler-a web server for functional interpretation of gene lists (2016 update). Nucleic acids research. 2016; 44(W1):W83–W9. Epub 2016/04/20. https://doi.org/10.1093/nar/gkw199 PMID: 27098042.