Quantitative Biology 2021, 9(2): 216–227
https://doi.org/10.15302/J-QB-021-0248

RESEARCH ARTICLE

Integrative modeling of transmitted and *de novo* variants identifies novel risk genes for congenital heart disease

Mo Li[1,†], Xue Zeng[2,†], Chentian Jin[3,†], Sheng Chih Jin[4], Weilai Dong[2], Martina Brueckner[2,5], Richard Lifton[2,6], Qiongshi Lu[7], Hongyu Zhao[1,2,8,*]

[1] Department of Biostatistics, Yale School of Public Health, New Haven, CT 06510, USA
[2] Department of Genetics, Yale University, New Haven, CT 06510, USA
[3] Department of Molecular, Cellular & Developmental Biology, Yale University, CT 06510, USA
[4] Department of Genetics, Washington University School of Medicine, St Louis, MO 63110, USA
[5] Department of Pediatrics, Yale University, New Haven, CT 06510, USA
[6] Laboratory of Human Genetics and Genomics, Rockefeller University, New York, NY 10065, USA
[7] Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI 53792, USA
[8] Program of Computational Biology and Bioinformatics, Yale University, New Haven, CT 06510, USA
* Correspondence: hongyu.zhao@yale.edu

Received October 29, 2020; Revised February 1, 2021; Accepted March 18, 2021

***Background*: Whole-exome sequencing (WES) studies have identified multiple genes enriched for *de novo* mutations (DNMs) in congenital heart disease (CHD) probands. However, risk gene identification based on DNMs alone remains statistically challenging due to heterogenous etiology of CHD and low mutation rate in each gene.**

***Methods*: In this manuscript, we introduce a hierarchical Bayesian framework for gene-level association test which jointly analyzes *de novo* and rare transmitted variants. Through integrative modeling of multiple types of genetic variants, gene-level annotations, and reference data from large population cohorts, our method accurately characterizes the expected frequencies of both *de novo* and transmitted variants and shows improved statistical power compared to analyses based on DNMs only.**

***Results*: Applied to WES data of 2,645 CHD proband-parent trios, our method identified 15 significant genes, half of which are novel, leading to new insights into the genetic bases of CHD.**

***Conclusion*: These results showcase the power of integrative analysis of transmitted and *de novo* variants for disease gene discovery.**

Keywords: rare variants; gene-level association test; congenital heart disease; *de novo* mutation

Author summary: Whole-exome sequencing (WES) studies have successfully identified multiple risk genes for congenital heart disease (CHD). However, it remains statistically challenging due to low mutation rate of *de novo* mutations (DNMs). In this paper, we present TADA-R, an innovative statistical test for identifying trait-associated genes through jointly analyzing *de novo* and rare transmitted variants. Applied to WES data of CHD proband-parent trios, our method identified novel risk genes and provided new insights into the genetic basis of CHD. Our method may benefit future sequencing-based studies in disease trios and accelerate findings of risk genes.

INTRODUCTION

Congenital heart disease (CHD) is a common birth defect affecting 0.8% of live births. It accounts for one-third of all major congenital abnormalities [1,2]. CHD patients have structural abnormalities in the heart, including septal

[†] These authors contributed equally to this work.

defects, valve defects, and lesions affecting the outflow tract. CHD is the most important cause of death in neonates and infants before the advent of cardiac surgery, with fewer than 15% reaching adulthood. With the introduction of corrective heart surgery in the late 1950s and improvements in long term care of CHD patients, over 90% of children who survive the first year of life now live into adulthood [3]. Both genetic and environmental factors are known to play important roles in CHD, with one study performed on Danish twins estimating the genetic heritability in this population to be close to 0.5 [4]. However, the lack of comprehensive understanding of the genetic underpinnings of CHD presents a major obstacle to the reproductive counseling of CHD patients [1]. The difficulties in elucidating the genetic underpinnings of CHD stem from the fact that CHD is a genetically heterogeneous disease. Classic genetic methodologies have linked certain forms of CHD to chromosomal abnormalities, such as 22q11 deletion and trisomy 21 [5]. Targeted gene-deletion studies in mice have also identified over 300 genes which, when disrupted, lead to heart defects [6]. However, in the vast majority of CHD patients, no chromosomal abnormality or causal mutation have been identified [1].

Whole-exome sequencing (WES) has been deployed as an important tool to study the genetic contributions to disease in the past decade, partly due to substantial decrease in cost and increase in the accuracy and throughput of the technology [7]. WES studies have successfully identified novel causal genes, not just in Mendelian disorders but also for heterogeneous monogenic disorders like hearing loss and for complex disorders like cardiovascular disease [8]. Since WES often identifies tens of thousands of genetic variants in each exome, most of which are irrelevant to the disease of interest, researchers need to narrow the pool of variants being considered. For example, with sequenced exomes of healthy parents and their affected offspring, only variants that have extremely low frequencies in the general population or *de novo* mutations (DNMs) in the children are further studied, thereby greatly reducing the number of variants under consideration [1]. WES studies have achieved some success in CHD. In one notable study, Zaidi *et al.* found an excess of protein-altering DNMs in genes expressed in the developing heart, and identified disruptions to genes in the H3K4me-H3K27 pathway as contributors to CHD [1]. In a more recent study, Jin *et al*. studied rare transmitted variants in addition to DNMs, and found that DNMs accounted for 8% of cases, whereas rare transmitted mutations were implicated in 1.8% of the CHD cases [9].

Studies on DNMs often lack statistical power because of the low number of mutations. There are an estimated 1.2 DNMs per exome [10], therefore the difference between the number of DNMs in the cases and controls is generally small even in cases where particular genes contribute to a disease phenotype [11]. Statistical power is further limited since many of these studies do not incorporate the large number of inherited variants inferred through WES. The transmission and *de novo* association (TADA) framework is a hierarchical Bayesian method to identify disease genes [12] by drawing information from both inherited and *de novo* variants in the exome. Despite some success [13,14], TADA has several methodological limitations. It does not incorporate the recessive mode of inheritance nor account for factors affecting *de novo* mutability of each gene (*e.g.,* local sequence context [15]), both of which have been demonstrated to be critical for inferring risk genes for CHD [9].

In this work, we introduce TADA-R, a generalized model built upon TADA to include the recessive disease model, *i.e.,* the offspring has the recessive genotype, including homozygotes, where the affected child inherited two identical mutations from each of the two parents, and compound heterozygotes, where the affected child inherited two different mutations of the same gene, one from each parent. It is important to incorporate the recessive model when studying CHD. For example, two risk genes for human CHD (*i.e.,GDF1* and *MYH6*) were identified purely by recessive effect [9]. Besides, recessive inheritance for CHD was observed in mouse models [16]. Beyond CHD, recessive inheritance has been implicated in a few diseases, including autism and early-onset Parkinson's disease [17–19]. By taking both dominant effects and recessive effects into consideration, our model is adaptive to diverse genetic architecture. We also incorporate gene-level annotations (*e.g.*, gene length and sequence context) and data from a population reference panel (*e.g.*, gnomAD [20]) to more accurately characterize the expected frequencies of both *de novo* and transmitted rare variants, which further improved the statistical power of our method. Through simulations, we demonstrate that TADA-R consistently outperforms TADA under diverse settings. We applied our model to WES data of 2,645 CHD proband-parent trios. In total, we identified 15 significant genes, many of which are novel and were not implicated in published studies based on simple meta-analysis of *de novo* and inherited variants on the same dataset. These findings shed important light on the genetic etiology of CHD.

RESULTS

Model overview

As detailed in the Methods section, the key for the TADA-R model is the probability of rare deleterious mutations for a gene in proband-parent trios. Following TADA, we

collapse all rare mutations in a gene as a binary indicator variable for carrier status, and consider three possible genotypes for each gene: without deleterious allele (*AA*), with one deleterious allele (*Aa*), and with both alleles being deleterious (*aa*). TADA-R models four possible genotype configurations in a proband-parent trio where proband carries deleterious allele: (1) both parents are homozygous *AA* and child has a DNM (*de novo* trio); (2) one parent is heterozygous and did not transmit mutation *a* to child (non-transmitted trio); (3) one parent is heterozygous and transmitted the mutation *a* to child (transmitted trio); (4) both parent are heterozygous and child is homozygous *aa* (recessive trio). Let μ denote the *de novo* mutability of a gene, $q/2$ denote the frequency of allele *a* in the population, and $\gamma = (\gamma_d, \gamma_t, \gamma_r)$ denote the relative risks for *de novo*, transmitted, and non-transmitted variants. The probabilities for the four cases are $Pois(2\mu\gamma_d)$, $Pois(q)$, $Pois(q\gamma_t)$ and $Pois(q^2\gamma_r)$, respectively. All the statistical details are discussed in the Methods section.

If a gene is not associated with a trait (H_0), then $\gamma_d = \gamma_t = \gamma_r = 1$. Otherwise, if gene is associated with a trait (H_1), then any of $(\gamma_d, \gamma_t, \gamma_r) \neq 1$. To quantify the magnitude of gene-disease associations, we use the following Bayes factor as the test statistic:

$$B = \frac{P(X|H_1)}{P(X|H_0)} = \frac{\int (X|\gamma, q)\Pr(\gamma)\Pr(q)d\gamma dq}{\int (X|\gamma = 1, q)\Pr(q)d\gamma dq}, \quad (1)$$

where $X = \{X_d, X_n, X_t, X_r\}$ are the counts for *de novo* trios, non-transmitted trios, transmitted trios, and recessive trios, respectively. $B = 1$ = under H_0 and $B > 1$ under H_1.

Variant calling

Running our variant calling pipeline on the sequencing data and using a minor allele frequency (MAF) filter of 0.001, we found an enrichment of *de novo* deleterious missense (D-Mis) and loss-of-function (LoF) mutations in the CHD case trios as compared to the control trios. Among the 2,645 case trios, we called 369 LoF DNMs and 448 D-Mis DNMs, which was a rate of 0.14 and 0.17 mutations per trio, respectively. This is compared to 0.08 and 0.12 mutations per trio for the control trios, translating into an enrichment factor of 1.66 for *de novo* LoF mutations and 1.38 for *de novo* D-Mis mutations (Table 1).

In contrast, we observed no enrichment in tolerated missense (T-Mis) and synonymous DNMs in the cases as compared to the controls. The overall rate of DNMs per proband exome was 1.11 for CHD cases and 1.01 for controls, which is consistent with prior work [11]. The DNM enrichment in deleterious annotation categories but not in benign categories confirms that CHD probands carry more damaging, protein-altering DNMs compared to healthy controls, while the burden of non-deleterious mutations is similar between two groups. Therefore, we only consider LoF and D-Mis mutations in our following analysis.

We also observed an enrichment of transmitted LoF variants in CHD trios with one parental carrier. There were 24,311 transmitted variants in the CHD case trios, corresponding to 9.19 variants per proband, as compared to 7.85 variants per control sample. This represents an enrichment factor of 1.17. By comparison, the enrichment factor for D-Mis variants was 1.06.

We also investigated the enrichment for recessive trios, where the proband had mutations in both copies of the gene. We observed a comparable enrichment for recessive trios with de novo trios. An enrichment of 1.39 and 1.45 was seen for LoF and D-Mis mutations, respectively. The higher enrichment for D-Mis mutation may be attributed to the relatively low counts.

Simulation study

We first evaluated the type-I error of our method. We compared the statistical power of four methods, including

Table 1 Mutation counts from 2,645 CHD case trios and 1,789 autism study control trios

Genotype	Mutation type	Cases		Controls		Cases/Controls
		N	Rate	N	Rate	
De novo	D-Mis	448	0.17	220	0.12	1.38
	LoF	369	0.14	150	0.08	1.66
Non-transmitted	D-Mis	59765	22.60	38476	21.51	1.05
	LoF	24974	9.44	14227	7.95	1.19
Transmitted	D-Mis	58870	22.26	37583	21.01	1.06
	LoF	24311	9.19	14051	7.85	1.17
Recessive	D-Mis	357	0.13	166	0.09	1.45
	LoF	175	0.06	85	0.05	1.39

Enrichment observed for *de novo*, non-transmitted, transmitted, and recessive trios by mutation types. N refers to the number of mutations and rate refers to the ratio of number of mutations and sample sizes.

(1) TADA-Denovo, (2) TADA, (3) TADA-R, and (4) TADA-R with gene-specific prior (see Methods). No method showed type-I error inflation, though they were likely to be more conservative than expected (Fig.1). This is because mutations with very low frequencies had few observations. Statistical power is exactly zero when no mutation is observed. Next, we evaluated the statistical power of our method. We compared the statistical power of the four methods in the three simulation settings:

Setting 1: $\gamma_d = 20$, $\gamma_t = 4$, $\gamma_r = 1$. This is the case where there is only dominant disease mode.

Setting 2: $\gamma_d = 1$, $\gamma_t = 1$, $\gamma_r = 16$. This is the case where there is only recessive disease mode.

Setting 3: $\gamma_d = 20$, $\gamma_t = 4$, $\gamma_r = 16$. This is the case where there is both dominant and recessive disease modes.

TADA-R with gene-specific prior had the best performance in all simulation settings (Fig. 1). In particular, for a dominant trait, the latter three methods had similar statistical power. However, for a recessive trait, TADA could not effectively identify risk genes, while TADA-R was consistently more powerful. Incorporating gene-specific prior further improved statistical power. We observed a similar pattern for traits with both dominant and recessive genes, with 15.6% and 25.4% improvement in statistical power for TADA-R without/with gene-specific prior in comparison to TADA.

Real data analysis

Finally, we performed the TADA-R analysis on 2,645 CHD proband-parent trios. 15 genes reached the genome-wide level of significance (Table 2). We decomposed the Bayes factors of significant genes into contributions from dominant trios (*i.e.*, *de novo*, non-transmitted, and transmitted trios) and recessive trios (Fig. 2). 11 out of

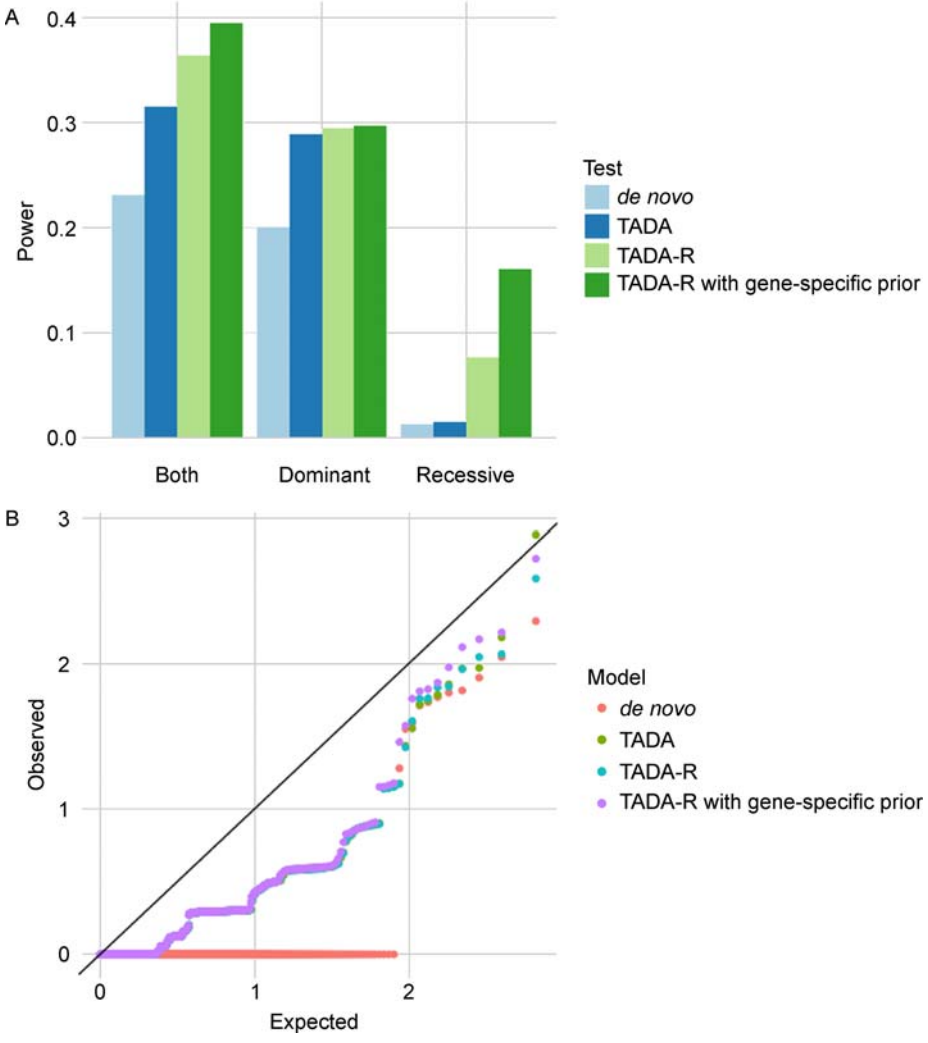

Figure 1. The statistical power and type-I error of different methods. We compared the statistical power (A) and type-I error (B) of TADA-*de novo*, TADA, TADA-R, and TADA-R with gene-specific prior with different disease architectures.

Table 2 Significant genes in TADA-R analysis of CHD case trios

Gene	BF_LoF	BF_D-Mis	FDR	Jin (2017)	Annotation
CHD7	1.76E + 16	3.25E−01	0.001	1	H-CHD
KMT2D	3.97E + 16	2.52E + 13	0.001	2	H-CHD
PTPN11	5.32E−01	1.24E + 17	0.001	3	H-CHD
RBFOX2	1.63E + 04	1.87	0.001	6	
GDF1	7.51E + 03	8.23E + 09	0.001		H-CHD
POGZ	2.72E + 02	4.39	0.020	23	
ACTB	2.66E + 01	3.80E + 01	0.024		H-CHD
CYP21A2	3.56	7.13	0.059		
ADIPOQ	5.11E-02	2.23E + 08	0.059		
NSD1	2.94E + 03	9.85E + 01	0.059	15	H-CHD
SULF1	3.01E−02	2.26E + 14	0.059		
RPL5	7.79E + 03	1.12	0.059		H-CHD
AKAP12	1.49E + 02	8.47E−01	0.059		
NOTCH1	3.39E + 06	1.01E + 26	0.059	5	H-CHD
SMAD2	1.39E + 02	9.12E + 02	0.096	19	

TADA-R analysis was performed on 2,645 CHD case trios. Bayes factor values were computed separately for LoF and D-Mis variants (shown as BF_LoF and BF_D-Mis). We annotated significant genes on whether they are among the top 25 genes identified by Jin *et al*. or known human CHD genes (H-CHD).

15 significant genes only showed dominant inheritance, including *CHD7*, *KMT2D*, *PTPN11*, *RBFOX2*, *POGZ*, *ACTB*, *CYP21A2*, *RPL5*, *AKAP12*, *NOTCH1*, and *SMAD2*. Eight have been previously reported as human CHD genes in association analysis or gene expression analysis [9]. We identified four genes showing both dominant and recessive associations, including *GDF1*, *SULF1*, *NSD1*, and *ADIPOQ*. Homozygosity of a mutation in *GDF1* was identified as a major cause of severe CHD among Ashkenazim [9]. We annotated these genes with the percentile rank of gene expression in developing mouse heart [1], and their intolerance to LoF mutation (pLI) in the gnomAD database [20]. Most of the genes are extremely intolerant to LoF mutations (pLI>0.85) and highly expressed in mouse developmental hearts (expression quantile>75%).

Majority (n = 5/7) of the genes identified by Jin *et al*. are also significant in our study, including *CHD7*, *KMT2D*, *PTPN11*, *RBFOX2*, and *NOTCH1*. We observed fewer mutations in the other two genes due to differences in the variant filtering criteria, therefore they were not significant in our analysis. All five significantly associated genes recapitulated in our study are highly intolerant to LoF and D-Mis mutations, with pLI⩾0.99 and Z⩾2.5. *CHD7* (chromodomain helicase DNA binding protein 7) is essential for the formation of multipotent migratory neural crest, which is required for the development of cardiac structures [21]. Damaging mutations in *KMT2D* (lysine methyltransferase 2D) have been strongly associated with Kabuki syndrome; CHDs are present in 70% of *KMT2D* positive Kabuki syndrome patients [22]. Mutations in *PTPN11* (protein tyrosine phosphatase non-receptor Type 11) are identified in over 50% of Noonan syndrome patients as well as patients with LEOPARD Syndrome 1; CHDs are featured in both syndromes. *NOTCH1* (notch receptor 1) is a central cardiac development factor that controls fetal cardiac development and cardiomyocyte proliferation [23]. Mutations in *NOTCH1* have been associated with a spectrum of aortic valve anomalies in human [16–19,24–26]. *RBFOX2* (RNA binding fox-1 homolog 2) is highly expressed in heart, liver, and pancreas. LoF mutations in *RBFOX2* have been associated with hypoplastic left heart syndrome (HLHS) [27]. A study in HLHS patient positive for *RBFOX2* mutations showed that Rbfox2 regulates mRNA levels of targets with 3′UTR binding sites contributing to aberrant gene expression in CHD patients [28].

Among the seven novel CHD risk genes identified in our study, *POGZ*, *ACTB*, and *SMAD2* are extremely intolerant to LoF and D-Mis mutations and are highly expressed in developing human heart. Additionally, all damaging variants identified in these three genes in our cohort have not been observed in ExAC [29] and gnomAD [20]. All identified D-Mis mutations alter highly conserved amino acid loci, supporting the pathogenicity of these variants. *POGZ* (pLI = 1.00, Z score = 3.58) encodes Pogo transposable element derived with ZNF domain. *De novo* protein-truncating mutations in *POGZ* have been reported in multiple cases of White-Sutton syndrome [30,31], among which CHD has been reported in one patient. *ACTB* (pLI = 0.99, Z score = 5.1) encodes actin beta. Heterozygous mutation in *ACTB* is

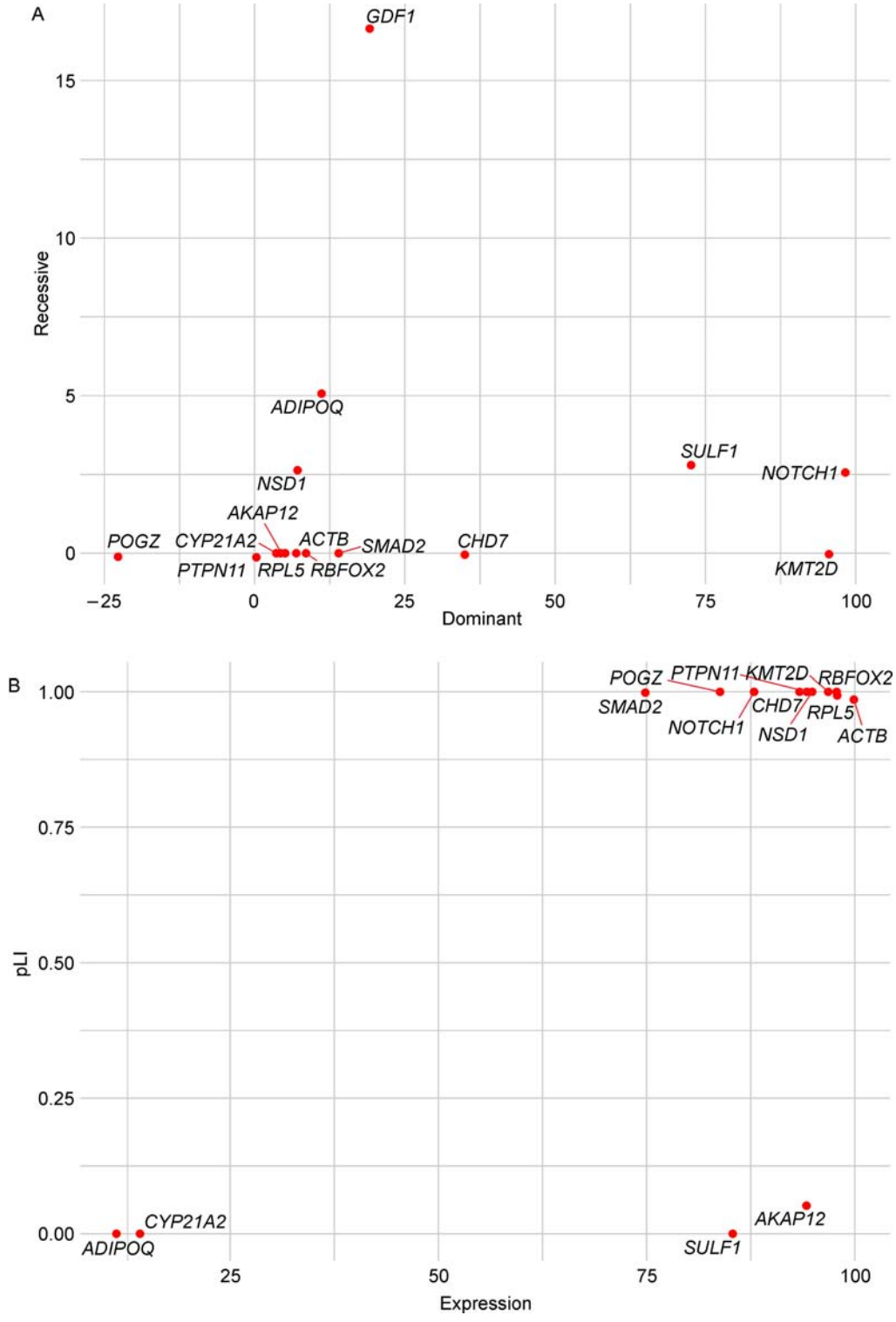

Figure 2. Bayes factor, pLI score and gene expression in developing mouse heart for significant genes in TADA-R analysis. (A) *x*/*y*-axis denotes the Bayes factor of dominant trios/recessive tios after log transformation. (B) *x*-axis denotes the percentile rank of gene expression in developing mouse heart at E14.5, and *y*-axis denotes the pLI score in the gnomAD database.

associated with Baraitser-Winter syndrome 1 [32], where CHDs have been reported in multiple unrelated *ACTB* mutation positive patients [33]. This suggests *ACTB* plays a role in cardiac development. *SMAD2* (pLI = 1.00, Z score = 3.8) encodes SMAD family member 2. Mouse knockout of *SMAD2* has abnormal dorsal aorta and heart tube morphology. Additionally, a transmitted LoF mutation in *SMAD2* has been reported as a disease-causing mutation in a familial case of CHD [1].

Additionally, we identified two *de novo* and one transmitted frameshift mutations in *AKAP12* (pLI = 0.05, Z score = 0.74), a gene encoding A-kinase anchoring protein 12. All identified *AKAP12* mutations are novel in gnomAD and ExAC. *AKAP12* is highly expressed in developing human heart and mouse knockout of *AKAP12* exhibits abnormal heart left ventricle and heart septum morphology, suggesting a role of *AKAP12* in CHDs.

Significant results were also observed in *CYP21A2* (pLI = 0, Z score = 1.62; cytochrome P450 family 21 subfamily A member 2), *ADIPOQ* (pLI = 0, Z score = –0.08; Adiponectin), and *SULF1* (pLI = 0, Z score = 1.58; Sulfatase 1). A large number of risk alleles (n = 76) that are relatively more common were observed in our cohort in these three genes, with 38% and 63% of the alleles having MAF $\geqslant$ 1E–4 and 1E–5, respectively. *SULF1* is highly expressed in developing human heart. Additionally, a microdeletion encompassing two genes, *SULF1* and *SLCO5A1*, on chromosome 8q13 has been associated with the mesomelia-synostosis syndrome (MSS); CHDs are one of the complications of MSS. These results provide potential target genes for functional validation in the future.

We tested for type I errors in the WES pipeline and TADA analysis by evaluating data from 1,789 control trios, which were the healthy siblings of autism probands in the simons simplex collection. No gene was found to be statistically significant. Besides, there is no known CHD gene among the top gene list (Table 3). This further suggests that our method does not have type I error inflation.

DISCUSSION

In this paper, we have introduced TADA-R, a novel method to perform gene-level association analysis in WES studies. This method integrates signals from DNMs and transmitted variants, and models both dominant and recessive inheritance to further improve statistical power. This is helpful for the study of CHD, where recessive genes play an important role [9]. Besides, TADA did not consider the difference of mutability among genes. We address this limitation by integrating gene-specific prior on mutation frequency learned from a large external database. Through simulations, we demonstrated that TADA-R achieves better performance under various genetic architectures.

To further demonstrate the performance of TADA-R, we conducted a case study on CHD. In total, we identified 15 significant genes for CHD and successfully replicated five significant genes as discussed in the previous paper [9]. More importantly, our model identified seven novel CHD risk genes that were not implicated in previous analysis using the same dataset. One example is *POGZ*, a gene that is intolerant to loss-of-function variants (pLI = 1.00), highly expressed in developing human heart, and has been reported in CHD cases. Our analysis provides novel insight into the disease mechanism of CHD.

Despite the success, our method has several limitations. Although including transmitted variants in analysis may improve statistical power, it may bring difficulty in interpreting the results. It may be difficult to pinpoint the driving source of genetic risk given multiple mutations scattered around the coding regions, especially for those that are unlikely to be pathogenic. In our analysis, we decomposed the contribution of mutations into dominant

Table 3 Significant genes in TADA-R analysis of autism control trios

Gene	BF_LoF	BF_D-Mis	p_value	FDR	Annotation
RAMP3	2.95	1.01E + 03	4.00E−04	1	Unknown
ANKRD55	1.86E + 02	1.24	5.00E−04	1	Unknown
BCAR3	1.46E + 01	2.92	7.00E−04	1	Unknown
KLK9	5.95E−02	2.04E + 02	9.00E−04	1	Unknown
PCCA	1.04E−01	1.91E + 06	0.001	1	Unknown
TELO2	8.37E−02	2.96E + 03	0.001	1	Unknown
FAM126A	1.23E−01	1.05E + 04	0.0011	1	Unknown
SYTL3	1.03E + 01	6.14	0.0012	1	Unknown
DOCK10	4.05E−02	1.27E + 04	0.0012	1	Unknown
MAPK8	2.63E−01	4.42E + 01	0.0012	1	Unknown

TADA-R analysis was performed on 1,789 autism control trios. Bayes factor values are computed separately for LoF and D-Mis variants (shown as BF_LoF and BF_D-Mis). We annotate significant genes on where they are known human CHD genes.

effect and recessive effect, which makes it easier to interpret the disease mechanism. However, methods for further decomposition are needed. Another limitation is that our method does not account for inbreeding when estimating the prior of parental genotype frequencies in recessive trios. Modeling the elevated rate of inbreeding in the case cohort has the potential to further improve the performance of our model.

In summary, TADA-R is a powerful and flexible framework to perform gene-level association analysis for *de novo* and transmitted variants. By integrating both dominant and recessive effects of genes, TADA-R can achieve better performance in all simulation scenarios. We have successfully applied TADA-R to analyze a CHD dataset, and shown that TADA-R is able to replicate previous findings, as well as identify novel significant genes. Further explorations suggest that these genes likely play important roles in CHD mechanism. Besides CHD, TADA-R can be applied to analyze WES data for other disease trios. As WES data continue to be generated for more traits and more individuals, we hope that TADA-R can lead to more gene identifications and biological insights.

METHODS

Whole-exome sequencing data

WES data for CHD case trios were downloaded from the database of Genotypes and Phenotypes (dbGaP) website (phs000571.v1.p1) [9]. The study included a total of 2,645 families that were recruited to the Congenital Heart Disease Network Study of the Pediatric Cardiac Genomics Consortium. Each family includes unaffected parents and one CHD offspring. Details on the inclusion criteria, sequencing protocol, and variant calling pipelines have been previously reported [9]. WES data for control trios were downloaded from (https://ndar.nih.gov/study.html?id=353) [34]. The data contain 1,789 families with two unaffected parents, one offspring with autism, and one unaffected sibling. Only the unaffected sibling and parents were analyzed in this study [9,34,35].

We considered rare homozygous and compound heterozygous variants supported by *high quality* sequence reads. We defined rare variants as those that had an allele frequency less than 0.1% across all the samples in the 1000 Genomes, EVS, and ExAC datasets [29,36,37], while high quality sequence reads are those that passed GATK Variant Score Quality Recalibration, had a minimum 8 total reads for both proband and parents, and had a genotype quality (GQ) $\geqslant$ 20. Rare variants were annotated by ANNOVAR [38]. Only LoF mutations (nonsense, canonical splice-site, frameshift indels, and start loss) and D-Mis variants were considered potentially damaging to the disease.

Probabilistic model

TADA-R models four possible genotype configurations in a proband-parent trio where proband carries deleterious allele: (1) both parents are homozygous AA and child has a *de novo* mutation (*de novo* trio); (2) one parent is heterozygous and did not transmit mutation a to child (non-transmitted trio); (3) one parent is heterozygous and transmitted the mutation a to child (transmitted trio); (4) both parent are heterozygous and child is homozygous aa (recessive trio). The probability of observing each trio conditional on unaffected parents is the multiplication of the probability of parent genotype, child genotype, and child phenotype.

If we denote the frequency of allele a in the population as $q/2$, the frequencies of parent genotype being AA, Aa, and aa genotypes are $1-q+q^2/4$, $q-(1-q/2)$, and $q^2/4$, respectively. As q is very small in the population, we ignore the possibility of aa genotype for unaffected parents and approximate the frequencies of homozygous (AA) and heterozygous (Aa) genotypes as $1-q$ and q, respectively. Let μ denote the *de novo* mutability of a gene which quantifies the expected DNM counts in the gene per individual and per chromosome. Conditioning on parental genotypes, the probabilities for observing the affected child's genotype for these four cases are $1-2\mu$, 2μ, $(1-2\mu)/2$, and $(1+2\mu)/2$, respectively. We use the tri-nucleotide sequence context approach to estimate de novo mutability μ [15]. For phenotype probability, let f denote the penetrance of AA genotype, and let $\gamma=(\gamma_d,\gamma_t,\gamma_r)$ denote the relative risk values for *de novo*, transmitted, and non-transmitted variants. The penetrance for the proband in *de novo*, non-transmitted, transmitted, and recessive trio is $\gamma_d f$, f, $\gamma_t f$, and $\gamma_r f$, respectively (Fig. 3).

As q and μ are small for rare variant, we ignore terms $1-q$, $1-2\mu$, and $1+2\mu$. The probabilities of the four types of trios can be approximated by $2\mu\gamma_d$, q, $q\gamma_t$ and $q^2\gamma_r$, and for a cohort of N trios, their counts can be approximated by the following distribution:

$$X_d \sim Pois(2\mu\gamma_d N),\ X_n \sim Pois(qN),$$
$$X_t \sim Pois(q\gamma_t N),\ X_r \sim Pois(q^2\gamma_r N), \tag{2}$$

where X_d, X_n, X_t, X_r are the counts for *de novo* trios, non-transmitted trios, transmitted trios, and recessive trios, respectively.

The full likelihood of our model is then

$$L(\Theta)=\prod_{i=1}^{M}\int (X_i|\gamma,q_i)\Pr(\gamma)\Pr(qi)d\gamma dq_i, \tag{3}$$

where M is the number of genes, $X_i=(X_{d\_i}, X_{n\_i}, X_{t\_i}, X_{r\_i})$ are the counts of different family genotype configurations for gene i, q_i is frequency of having rare

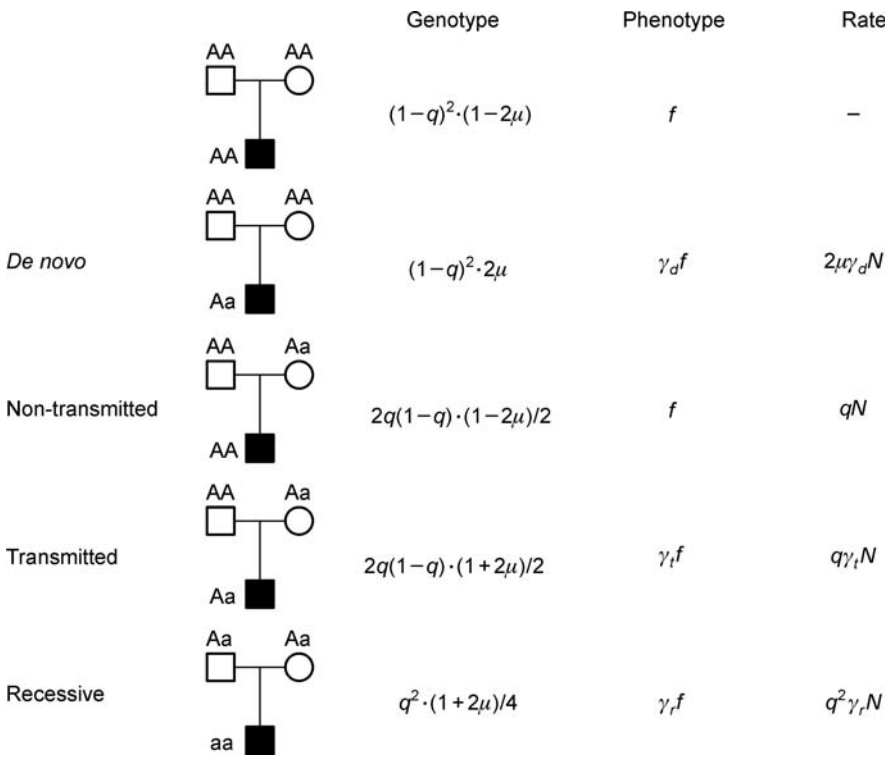

Figure 3. TADA-R probabilistic model for a family trio with an affected child. Genotype probabilities are computed as the marginal probability of parental genotypes times the conditional probability of the child, given the parents.

variants in gene i in the population, and $\gamma = (\gamma_d, \gamma_t, \gamma_r)$ are the relative risk values for *de novo*, transmitted, and non-transmitted variants across all genes. Prior selection of parameters γ_d, γ_t, γ_r and q_i will be discussed in the next session.

Prior estimation

The model involves four parameters, including cross-gene parameters γ_d, γ_t, γ_r, and gene-specific parameter q_i. We use the conjugate prior for these parameters:

$$\gamma_d \sim Gamma(\rho_d,\nu_d),\ \gamma_t \sim Gamma(\rho_t,\nu_t),$$
$$\gamma_r \sim Gamma(\rho_r,\nu_r),\ q_i \sim Gamma(\omega_i,\tau), \quad (4)$$

where ρ_d, ν_d, ρ_t, ν_t, ρ_r, ν_r are the hyperparameters for the relative risks. We use an empirical Bayes approach to estimate these hyperparameters. The details about hyper-parameter estimation for γ were discussed in the TADA model [12]. ω_i and τ are the hyperparameters for the rare allele frequency q_i, where ω_i controls the expectation of the mutation frequency for gene i, and τ controls the variance of frequency distributions across all genes. We estimated ω_i and τ from the gnomAD database. Since we want to focus on extremely rare variants, there is high variance for the observed counts of these rare variants, and it is possible that a gene may lack any rare variant. The point estimates of allele frequencies in gnomAD cannot be plugged in directly. Instead, we 'smoothen' the point estimates via an empirical Bayes prior that integrates information across all the genes in the genome.

Specifically, we consider an additional hierarchical layer on the distribution of q_i. We propose the following prior distribution for q_i

$$f(q_i) = Gamma(\alpha_i\beta,\beta), \quad (5)$$

where α_i is the expectation of this distribution and is calculated as the re-weighted frequency based on the total counts of rare variants in gnomAD. We use mutability scores m_i estimated based on sequence context by Samocha *et al*. as the weights for mutations [15]. We have

$$\alpha_i = \frac{m_i}{\sum_i m_i} \times \frac{\sum_i z_i}{2N_{gnomAD}}, \quad (6)$$

where z_i is the observed counts of rare variants for gene i in gnomAD, N_{gnomAD} is the sample size for gnomAD, and z_i follows a Poisson distribution given the allele frequency q_i

$$z_i \mid q_i \sim Poisson\ (2N_{gnomAD}\ q_i). \quad (7)$$

The other parameter β controls the variance of the prior

distribution (variance = α_i/β) and is shared between all genes. We estimate the unknown parameters β by maximizing the likelihood of the observed variant counts in gnomAD.

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}} L(z) = \underset{\beta}{\operatorname{argmax}} \prod_i \int_0^1 P(z_i \mid q_i) f(q_i) dq_i \quad (8)$$

Then, the posterior distribution of q_i given the variant counts in gnomAD is

$$P(q_i \mid z_i) = \frac{P(z_i \mid q_i) f(q_i)}{\int_0^1 P(z_i \mid q_i) f(q_i) dq_i}$$

$$= Gamma(\alpha_i \hat{\beta} + z_i, \hat{\beta} + 2N_{gnomad}) \quad (9)$$

We use this distribution as the prior distribution of q_i in our model, where $\omega_i = \alpha_i \hat{\beta} + z_i$, and $\tau = \hat{\beta} + 2N_{gnomad}$.

Simulation settings

We conducted simulations to evaluate the performance of our model. Simulation data were generated under both the null and alternative hypotheses to evaluate the type-I error rate and statistical power. We fixed the sample size of our simulation study as 5,000, and randomly selected 200 genes for simulation. For each gene, we further replicated 5 times. In each repeat, we sampled the allele frequency of each gene based on the prior distribution $P(q_i|Z_i)$ estimated from gnomAD. Then, we sampled the number of *de novo* trios, non-transmitted trios, transmitted trios, and recessive trios for each gene from the following Poisson distribution. We adjusted mutation rates by considering inbreeding, with inbreeding factor F = 0.002. This is the same as that reported in the CHD dataset [9]. After the adjustment, we have:

$$X_d \sim Pois(2\mu\gamma_d N),\ X_n \sim Pois((q - qF)N),$$

$$X_t \sim Pois((q - qF)\gamma_t N),\ X_r \sim Pois(qF\gamma_r N/2). \quad (10)$$

To generate mutation counts under the alternative hypothesis, we considered the following three simulation settings:

Setting 1: $\gamma_d = 20,\ \gamma_t = 4,\ \gamma_r = 1$. This is the case where there is only dominant disease mode.

Setting 2: $\gamma_d = 1,\ \gamma_t = 1,\ \gamma_r = 16$. This is the case where there is only recessive disease mode.

Setting 3: $\gamma_d = 20,\ \gamma_t = 4,\ \gamma_r = 16$. This is the case where there is both dominant and recessive disease modes.

We note that the values of the relative risks chosen here are consistent with previous estimations [12].

Using the same mutation counts as input, we compared the performance of four methods: (1) TADA-Denovo only takes DNMs as input; (2) TADA takes both *de novo* and transmitted dominant mutations as input; (3) TADA-R takes *de novo*, transmitted dominant, and transmitted recessive mutations as input; and (4) TADA-R with gene-specific prior further considers gene-specific prior in the TADA-R model. As shown in Eq. (1), we used a Bayes factor as test statistic to quantify the magnitude of gene-disease association. The *p*-value of this Bayes factor was then calculated by a permutation test, where samples from the null distribution were generated by setting the relative risk of variants as zero, and *p*-value is the proportion of permuted samples with values greater than the observed Bayes factor. Finally, the false discovery rate (FDR) was calculated as Benjamini-Hochberg correction of multiple testing *p*-values. The FDR is estimated for all the above methods under different simulation scenarios. Statistical power was calculated as the proportion of genes with FDRs smaller than 0.05. Under the null hypothesis, we generated mutation counts under $\gamma_d = 1, \gamma_t = 1, \gamma_r = 1$ and repeated the whole procedure. Similarly, type-I error was calculated as the proportion of genes with FDRs smaller than 0.05 under the null.

SUPPORTING INFORMATION

TADA software is available as an R package at https://github.com/limo936/TADA-R. Statistical tests for genes can run independently and parallelly to speed up the algorithm. On average, each gene takes 40 seconds for computation.

ACKNOWLEDGEMENTS

This study was supported in part by the National Institutes of Health (NIH) grants R01 GM134005, and the National Science Foundation (NSF) grants DMS 1902903. Dr. Sheng Chih Jin's effort was supported by the Pathway to Independence Award (K99/R00) program, grants K99HL143036-01A1 and R00HL143036-02.

COMPLIANCE WITH ETHICS GUIDELINES

The authors Mo Li, Xue Zeng, Chentian Jin , Sheng Chih Jin, Weilai Dong, Martina Brueckner, Richard Lifton, Qiongshi Lu, and Hongyu Zhao declare that they have no conflict of interests.

The article does not contain any human or animal subjects performed by any of the authors.

OPEN ACCESS

REFERENCES

1. Zaidi, S., Choi, M., Wakimoto, H., Ma, L., Jiang, J., Overton, J. D., Romano-Adesman, A., Bjornson, R. D., Breitbart, R. E., Brown, K. K., *et al.* (2013) *De novo* mutations in histone-modifying genes in congenital heart disease. Nature, 498, 220–223
2. Postma, A. V., Bezzina, C. R. and Christoffels, V. M. (2016) Genetics of congenital heart disease: the contribution of the noncoding regulatory genome. J. Hum. Genet., 61, 13–19
3. Gilboa, S. M., Salemi, J. L., Nembhard, W. N., Fixler, D. E. and Correa, A. (2010) Mortality resulting from congenital heart disease among children and adults in the United States, 1999 to 2006. Circulation, 122, 2254–2263
4. Wienke, A., Herskind, A. M., Christensen, K., Skytthe, A. and Yashin, A. I. (2005) The heritability of CHD mortality in danish twins after controlling for smoking and BMI. Twin Res. Hum. Genet., 8, 53–59
5. Lalani, S. R. and Belmont, J. W. (2014) Genetic basis of congenital cardiovascular malformations. Eur. J. Med. Genet., 57, 402–413
6. Bentham, J. and Bhattacharya, S. (2008) Genetic mechanisms controlling cardiovascular development. Ann. N. Y. Acad. Sci., 1123, 10–19
7. Teer, J. K. and Mullikin, J. C. (2010) Exome sequencing: the sweet spot before whole genomes. Hum. Mol. Genet., 19, R145–R151
8. Rabbani, B., Tekin, M. and Mahdieh, N. (2014) The promise of whole-exome sequencing in medical genetics. J. Hum. Genet., 59, 5–15
9. Jin, S. C., Homsy, J., Zaidi, S., Lu, Q., Morton, S., DePalma, S. R., Zeng, X., Qi, H., Chang, W., Sierant, M. C., *et al.* (2017) Contribution of rare inherited and *de novo* variants in 2,871 congenital heart disease probands. Nat. Genet., 49, 1593–1601
10. Kong, A., Frigge, M. L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S. A., Sigurdsson, A., Jonasdottir, A., Jonasdottir, A., *et al.* (2012) Rate of *de novo* mutations and the importance of father's age to disease risk. Nature, 488, 471–475
11. Sanders, S. J., Murtha, M. T., Gupta, A. R., Murdoch, J. D., Raubeson, M. J., Willsey, A. J., Ercan-Sencicek, A. G., DiLullo, N. M., Parikshak, N. N., Stein, J. L., *et al.* (2012) *De novo* mutations revealed by whole-exome sequencing are strongly associated with autism. Nature, 485, 237–241
12. He, X., Sanders, S. J., Liu, L., De Rubeis, S., Lim, E. T., Sutcliffe, J. S., Schellenberg, G. D., Gibbs, R. A., Daly, M. J., Buxbaum, J. D., *et al.* (2013) Integrated model of *de novo* and inherited genetic variants yields greater power to identify risk genes. PLoS Genet., 9, e1003671
13. De Rubeis , S., He, X., Goldberg, A. P., Poultney, C. S., Samocha, K., Cicek, A. E., Kou, Y., Liu, L., Fromer, M., Walker, S., *et al.* (2014) Synaptic, transcriptional and chromatin genes disrupted in autism. Nature, 515, 209–215
14. Singh, T., Kurki, M. I., Curtis, D., Purcell, S. M., Crooks, L., McRae, J., Suvisaari, J., Chheda, H., Blackwood, D., Breen, G., *et al.* (2016) Rare loss-of-function variants in SETD1A are associated with schizophrenia and developmental disorders. Nat. Neurosci., 19, 571–577
15. Samocha, K. E., Robinson, E. B., Sanders, S. J., Stevens, C., Sabo, A., McGrath, L. M., Kosmicki, J. A., 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
16. Bruneau, B. G. (2008) The developmental genetics of congenital heart disease. Nature, 451, 943–948
17. Shanks, M. E., Downes, S. M., Copley, R. R., Lise, S., Broxholme, J., Hudspith, K. A., Kwasniewska, A., Davies, W. I., Hankins, M. W., Packham, E. R., *et al.* (2013) Next-generation sequencing (NGS) as a diagnostic tool for retinal degeneration reveals a much higher detection rate in early-onset disease. Eur. J. Hum. Genet., 21, 274–280
18. Chahrour, M. H., Yu, T. W., Lim, E. T., Ataman, B., Coulter, M. E., Hill, R. S., Stevens, C. R., Schubert, C. R., Greenberg, M. E., Gabriel, S. B., *et al.* (2012) Whole-exome sequencing and homozygosity analysis implicate depolarization-regulated neuronal genes in autism. PLoS Genet., 8, e1002635
19. Schormair, B., Kemlink, D., Mollenhauer, B., Fiala, O., Machetanz, G., Roth, J., Berutti, R., Strom, T. M., Haslinger, B., Trenkwalder, C., *et al.* (2018) Diagnostic exome sequencing in early-onset Parkinson's disease confirms VPS13C as a rare cause of autosomal-recessive Parkinson's disease. Clin. Genet., 93, 603–612
20. Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., *et al.* (2020) The mutational constraint spectrum quantified from variation in 141,456 humans. Nature, 581, 434–443
21. Bajpai, R., Chen, D. A., Rada-Iglesias, A., Zhang, J., Xiong, Y., Helms, J., Chang, C. P., Zhao, Y., Swigut, T. and Wysocka, J. (2010) CHD7 cooperates with PBAF to control multipotent neural crest formation. Nature, 463, 958–962
22. Digilio, M. C., Gnazzo, M., Lepri, F., Dentici, M. L., Pisaneschi, E., Baban, A., Passarelli, C., Capolino, R., Angioni, A., Novelli, A., *et al.* (2017) Congenital heart defects in molecularly proven Kabuki syndrome patients. Am. J. Med. Genet. A., 173, 2912–2922
23. Kasahara, A., Cipolat, S., Chen, Y., Dorn, G. W. 2nd and Scorrano, L. (2013) Mitochondrial fusion directs cardiomyocyte differentiation via calcineurin and Notch signaling. Science, 342, 734–737
24. Garg, V., Muth, A. N., Ransom, J. F., Schluterman, M. K., Barnes, R., King, I. N., Grossfeld, P. D. and Srivastava, D. (2005) Mutations in NOTCH1 cause aortic valve disease. Nature, 437, 270–274
25. Mohamed, S. A., Aherrahrou, Z., Liptau, H., Erasmi, A. W., Hagemann, C., Wrobel, S., Borzym, K., Schunkert, H., Sievers, H. H. and Erdmann, J. (2006) Novel missense mutations (p.T596M and p.P1797H) in NOTCH1 in patients with bicuspid aortic valve. Biochem. Biophys. Res. Commun., 345, 1460–1465

26. McBride, K. L., Riley, M. F., Zender, G. A., Fitzgerald-Butt, S. M., Towbin, J. A., Belmont, J . W. and Cole, S. E. (2008) NOTCH1 mutations in individuals with left ventricular outflow tract malformations reduce ligand-induced signaling. Hum. Mol. Genet., 17, 2886–2893
27. Homsy, J., Zaidi, S., Shen, Y., Ware, J. S., Samocha, K. E., Karczewski, K. J., DePalma, S. R., 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
28. Verma, S. K., Deshmukh, V., Nutter, C. A., Jaworski, E., Jin, W., Wadhwa, L., Abata, J., Ricci, M., Lincoln, J., Martin, J. F., *et al.* (2016) Rbfox2 function in RNA metabolism is impaired in hypoplastic left heart syndrome patient hearts. Sci. Rep., 6, 30896
29. Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., *et al.* (2016) Analysis of protein-coding genetic variation in 60,706 humans. Nature, 536, 285–291
30. Stessman, H. A. F., Willemsen, M. H., Fenckova, M., Penn, O., Hoischen, A., Xiong, B., Wang, T., Hoekzema, K., Vives, L., Vogel, I., *et al.* (2016) Disruption of POGZ Is Associated with Intellectual Disability and Autism Spectrum Disorders. Am. J. Hum. Genet., 98, 541–552
31. The Deciphering Developmental Disorders Study (2015) Large-scale discovery of novel genetic causes of developmental disorders. Nature, 519, 223–228
32. Di Donato, N., Rump, A., Koenig, R., Der Kaloustian, V. M., Halal, F., Sonntag, K., Krause, C., Hackmann, K., Hahn, G., Schrock, E., *et al.* (2014) Severe forms of Baraitser-Winter syndrome are caused by ACTB mutations rather than ACTG1 mutations. Eur. J. Hum. Genet., 22, 179–183
33. Cuvertino, S., Stuart, H. M., Chandler, K. E., Roberts, N. A., Armstrong, R., Bernardini, L., Bhaskar, S., Callewaert, B., Clayton-Smith, J., Davalillo, C. H., *et al.* (2017) ACTB loss-of-function mutations result in a pleiotropic developmental disorder. Am. J. Hum. Genet., 101, 1021–1033
34. Krumm, N., Turner, T. N., Baker, C., Vives, L., Mohajeri, K., Witherspoon, K., Raja, A., Coe, B. P., Stessman, H. A., He, Z. X., *et al.* (2015) Excess of rare, inherited truncating mutations in autism. Nat. Genet., 47, 582–588
35. Fischbach, G. D. and Lord, C. (2010) The Simons Simplex Collection: a resource for identification of autism genetic risk factors. Neuron, 68, 192–195
36. 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. Nature, 526, 68–74
37. NHLBI Exome Sequencing Project (ESP) website. http://evs.gs.washington.edu/EVS/. Accessed: Sep 1, 2020
38. Wang, K., Li, M. and Hakonarson, H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res., 38, e164

