Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Genetics & Development

Statistical models of the genetic etiology of congenital heart disease

Guojie Zhong[1,2] and Yufeng Shen[1,3,4,*]

Congenital heart disease (CHD) is a collection of anatomically and clinically heterogeneous structure anomalies of heart at birth. Finding genetic causes of CHD can not only shed light on developmental biology of heart, but also provide basis for improving clinical care and interventions. The optimal study design and analytical approaches to identify genetic causes depend on the underlying genetic architecture. A few well-known syndromes with CHD as core conditions, such as Noonan and CHARGE, have known monogenic causes. The genetic causes of most of CHD patients, however, are unknown and likely to be complex. In this review, we highlight recent studies that assume a complex genetic architecture of CHD with two main approaches. One is genomic sequencing studies aiming for identifying rare or *de novo* risk variants with large genetic effect. The other is genome-wide association studies optimized for common variants with moderate genetic effect.

Addresses
[1] Department of Systems Biology, Columbia University Irving Medical Center, New York, NY, USA
[2] Integrated Program in Cellular, Molecular, and Biological Studies, Columbia University Irving Medical Center, New York, NY, USA
[3] Department of Biomedical Informatics, Columbia University Irving Medical Center, New York, NY, USA
[4] JP Sulzberger Columbia Genome Center, Columbia University Irving Medical Center, New York, NY, USA

Corresponding author: Yufeng Shen (ys2411@cumc.columbia.edu)
* Twitter account: @joshuas

Current Opinion in Genetics & Development 2022, **76**:101967

This review comes from a themed issue on **Molecular and Genetic Basis of Disease**

Edited by **Neil Hanchard** and **Heather C Mefford**

https://doi.org/10.1016/j.gde.2022.101967

Introduction

Congenital heart disease (CHD) is an anatomically heterogeneous condition. Historically, CHD is a severe condition with high mortality and morbidity. Recent advance in medicine has improved survival, but many CHD patients still have lingering medical problems later in life. As a result, CHD still causes reduced productive fitness. A study on Tetralogy of Fallot (TOF) reported that the average number of offspring per CHD individual is about 30% smaller than age-matched controls [1]. Therefore, genetic factors with large effect size must be under strong negative selection and be rare in the population. In fact, the genes of known syndromes with CHD as a core condition, such as *PTPN11* in Noonan [2–6] and *KMT2D* in Kabuki [7,8], often harbor *de novo* mutations in individuals with CHD. Consistently, identification of new risk genes by *de novo* variants (DNVs) is the main study design in several recent prominent genetic studies in CHD [9••–11]. Overall, the population-attributable risk percentage (PAR%) of CHD explained by *de novo* coding variants is about 20–30% in syndromic cases that have additional congenital anomalies or neurodevelopmental disorders, and likely less than 10% in isolated cases [9••–11].

$$PAR = \frac{Incidence\ Rate\ in\ total\ population - Incidence\ Rate\ unexposed}{Incidence\ Rate\ in\ total\ population}$$

$$h^2 = \frac{Phenotypic\ Variation\ Due\ to\ Genetic\ Factors}{Total\ Phenotypic\ Variation}$$

The bulk of CHD risk remains unknown. A range of CHD subtypes has high heritability (h^2) [12–16]. A common hypothesis is that CHD in most of affected individuals is caused by multiple environmental and genetic factors, and most of these genetic factors may have moderate or small effect. Genome-wide association studies (GWAS) are designed to identify such genetic risk by searching for association of common variants with moderate effect size [17••]. Figure 1 summarizes two genetic models for CHD. Under the "Denali" model, which is essentially monogenic, the genetic risk is dominated by highly penetrant mutations ("peaks"). Classical Mendelian CHD genes are under this model. Under the "Everest" model, the aggregated genetic risk from common inherited variants forms a high plateau of the disease liability that may be sufficient to cause CHD in most of the patients. Rare or DNVs with larger effect ("peaks") may still play a role in these patients, especially in those with severe or syndromic conditions.

Overall, recent human genetic studies on CHD are driven by genomic technological advances, meticulously enrolled large cohorts, and advanced statistical models. In this review, we will summarize the statistical methods used in recent studies.

Figure 1

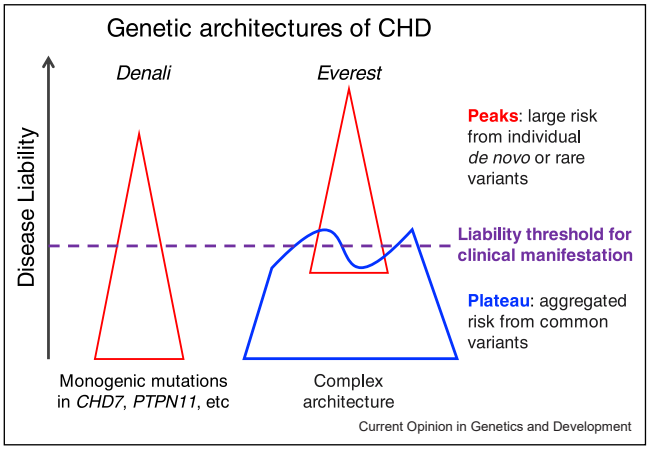

Illustration of genetic architectures of CHD by two types of mountains. The height of the mountain represents disease liability. Left, the *Denali* model, disease liability is dominated by highly penetrant monogenic mutations. Right, the *Everest* model, disease liability is contributed by multiple genetic factors, including the high plateau (aggregated risk from common variants with small effect) and peaks (rare or DNVs with large effect). Horizontal dashed line indicates liability threshold for clinical manifestation of CHD. The aggregated risk from common variants in some individuals may be large enough to reach liability threshold. The illustration is adapted from lastfrontier.org.

Main statistical models of recent genetic association studies

The basic model of statistical analysis of rare and common variants is similar. Under the null model, which states that a variant or a gene is not associated with CHD risk, the allele frequency in cases follows a statistical distribution with parameters determined by theoretical models or empirical data of a population without CHD (Table 1). In a frequentist approach, one can assess the null model by estimating type-I error, that is, p-value, as the probability of data as extreme as what was observed in the cases based on the null model. A candidate risk variants or genes are identified if one can reject the null hypothesis when the p-value is smaller than a type-I error threshold. In a Bayesian approach, one must define an alternative model, which states that the variant or gene is associated with CHD risk and the effect size follows a specified distribution. The model has two key parameters, the prior probability of being associated with the risk and the effect size, that is, the magnitude of the association, often defined as odds ratio or relative risk (Table 1).

A crucial issue in genome-wide analysis is multiple test adjustment. The standard approach in GWAS is to use Bonferroni correction of p-values to minimize false discovery (Table 1), with a commonly used threshold of 5e-8, which reflects the effective number of independent common variants in human genome, no matter how many SNPs are tested. The same approach can be applied to gene-based tests used in analysis of rare variants, adjusting for about 20 000 protein-coding genes. This approach is straightforward and effective in preventing false discoveries. However, it errs on being conservative. More powerful methods, based on false-discovery rate (FDR) or q-values, have been used in recent studies. We will review the complexity of this approach in the section on *de novo*-coding variant analysis.

Analysis of *de novo*-coding variants

DNVs are new mutations that are present in offspring but not in parents. DNVs, especially the ones located in protein-coding regions, are a group of variants with largest effect size a priori. In Zaidi et al., 2013 [18], the Pediatric Cardiac Genomics Consortium team analyzed exome sequencing data of 362 CHD case–parent trios and 264 control trios. Compared with controls, CHD cases showed an excess of deleterious DNVs, especially in chromatin modifiers required for H3K4/H3K27 methylation and highly expressed genes during embryonic heart development. Only one gene, *NAA15*, harbors more than one *de novo* protein-altering variants in the data. This indicates that the pool of genes that contribute to CHD through *de novo* coding variants is very large, similar to the genetic architecture of autism-spectrum disorder based on early exome sequencing studies [19–23]. As the statistical power for identification of individual risk genes is severely limited by the small sample size, they introduced a heuristic method to improve signal-to-noise ratio of putative disease-causing DNVs based on filters at variant and gene levels. First, select likely deleterious DNVs by filtering based on functional annotations and conservation; second, select genes that are likely to play a role in CHD based on high-expression rank in developing mouse heart (embryonic day 14.5). With this method, they showed CHD cases were significantly enriched of deleterious DNVs in biologically plausible genes compared with controls. This paradigm has been used since then in later studies in CHD and other birth defects [24–28].

In Homsy et al., 2015 [9••], the same trend was observed in a larger cohort of 1213 trios, where they identified an excess of protein-damaging DNVs, especially in genes highly expressed in the developing heart and brain that are involved in morphogenesis, chromatin modification, and transcriptional regulation. The study introduced a new method to improve power by comparing observed data in cases with what is expected by chance based on a background mutation model [29]. The background mutation model, first introduced in Samocha et al., 2014 [30], estimates context-specific mutation rate based on large-scale human genome sequencing data. In a transcript (of a gene), the aggregated mutation rate of each functional class of variants (e.g.,

Table 1

Summary of all statistical methods mentioned in this paper.

Type of variants	Method	Likelihood model	Criteria for identify candidate risk genes or variants	Publication	Data size	Description
De novo coding variants	Frequentist (Poisson exact test)	Null: $m \sim Poisson(\lambda)$	P-value threshold determined by Bonferroni correction	Homsy et al, 2015	1213 trios	Test the probability of data as extreme as what was observed in the cases
	Bayesian (TADA)	Null: $m \sim Poisson(\lambda)$ Alternative: $m \sim Poisson(\gamma^*\lambda)$	Bayesian FDR	Sifrim et al, 2021	1891 trios	Test the probability of observed data belongs to alternative distribution against null distribution
Rare inherited variants	Binomial exact test	$Expected\ number\ of\ RGs = \beta_0 + \beta_1{}^*mutability + \beta_2{}^*mutability^2$	P-value threshold determined by Bonferroni correction	Jin et al, 2017 [10••,38]	2645 trios	The null distribution is estimated as a polynomial regression function of gene mutability
	VAAST	$\lambda = ln\left(\frac{L_{null}}{L_{alt}}\right) \sim \chi^2 distribution$	P-value threshold determined by Bonferroni correction	Watkins et al, 2019 [40••]	2391 trios	A gene-burden test that ranks the probability being a risk gene or gene set based on empirically calibrated rarity of genotypes
CNVs	Binomial exact test	$m_{case} \sim Binomial\left(n_{case}, \frac{m_{control}}{n_{control}}\right)$	P-value threshold determined by Bonferroni correction	Audain et al., 2021 [44]	7958 cases and 14 082 controls	Test the probability of data as extreme as what was observed in the cases
De novo noncoding variants	Binomial test HeartENN	$m_{case} \sim Binomial\left(n_{case}, \frac{m_{control}}{n_{control}}\right)$	P-value threshold determined by Bonferroni correction	Richter et al, 2020 [45••]	749 probands and 1611 control trios	Variants of cases and controls were scored and filtered with HeartENN with optimal threshold, followed by one-tailed binomial test
Common variants	GWAS	–	P-value threshold (5×10^{-8})	Lahm et al, 2021 [17••]	4034 cases and 8486 controls	Test the probability of data as extreme as what was observed in the cases

protein-truncating variants, missense variants, and synonymous variants) can be estimated by adding up all possible point mutation rates of that class. Missense variants can be further filtered as damage missense ("D-mis") variants by various prediction tools (REVEL [31], CADD [32], MPC [33], metaSVM [34], etc). Using background mutation rate to calculate expectation under the null is equivalent to having infinitely large number of controls. With the background model, the number of DNVs in each gene follows a Poisson distribution

$$H_0: m \sim Poisson(\lambda) \quad (1)$$

where m is the observed number of DNVs of a certain type (e.g., missense variants) in N individuals, λ is the mean of the distribution, estimated by the number of individuals (N) multiplied by the background mutation rate of the type of the variants in the gene. Under the null, one can test the significance of DNV burden of a gene by a *Poisson exact test*. It is applicable to both single genes and gene sets. A limitation of such method comes from the uncertainty of estimated background mutation rate because the model cannot completely account for the variation of mutation rate along the genome [30,35], especially short insertions and deletions. This issue can be partly improved by modeling the mutation rate as a random variable with mean and variance. While a Poisson test only accounts for the mean of this random variable, a Gamma-Poisson test (equivalent to negative binomial) can account for both.

Homsy et al., 2015 [36] reported three genes (*PTPN11*, *KMT2D*, and *RBFOX2*) with genome-wide significance (Bonferroni correction) and additional 18 genes with ≥2 damaging DNVs. While it is challenging to establish the association of these 18 genes individually by DNV data, simulation analysis showed that conditioned on the total number of damaging DNVs, the number of genes with multiple damaging DNVs by chance is far smaller, indicating that most of these 18 genes are likely to be true risk genes. This important insight is the foundation of formal methods to estimate FDR through Bayesian approaches.

Indeed, in a more recent study of 1891 CHD cases by Sifrim et al., 2016 [11], the authors used a hierarchical Bayesian model, TADA [21], to estimate FDR by explicitly modeling the alternative hypothesis and the priors. Just like Poisson tests, TADA assumes that the observed number of DNVs in each gene follows a Poisson distribution (Equation (1)).

Under the null hypothesis, λ can be estimated by background mutation rate using the same method based on Samocha 2014 model. Under alternative hypothesis, TADA specifies two parameters: the prior probability of being a risk gene (π) and the relative risk (γ) of a class of variants. Given γ, the observed number of DNVs follows Poisson distribution with mean of $\gamma^*\lambda$:

$$H_1 \mid \gamma : m \sim Poisson(\gamma^*\lambda) \quad (2)$$

γ is treated as a random variable with prior distribution. The Bayes factor is calculated as

$$BF = \int \frac{p(m|\gamma^*\lambda)}{p(m|\lambda)} d\gamma \quad (3)$$

The posterior odds (PO) of two hypotheses can be calculated as

$$PO = \frac{P(m \mid H_1)}{P(m \mid H_0)} = \frac{\pi}{1-\pi} BF \quad (4)$$

TADA methods estimate the hyperparameters (π, γ) by fitting the overall model with overall enrichment of DNVs and how DNVs are distributed across genes [21,37].

Finally, one can estimate posterior probability of association (PPA) from posterior odds, and then a Bayesian FDR by false-discovery proportion implied by PPA. A candidate gene can be nominated if its FDR is below a threshold. From this paradigm, they identified 16 genes with FDR below 0.01. This approach has been used since then in many other studies. This study also introduced a new method for analyzing the missense variants by clustering of mutations within genes, where they computed the geometric mean of distances between each pair of mutations along the protein sequence and compare with random simulations. This method can increase the power when integrated with the mutation-burden test.

Analysis of rare inherited variants

Rare inherited variants are variants in offsprings that inherited from parents with low minor allele frequency in the population, for example, less than 1%. As DNVs only account for ~30% of PAR of syndromic cases and < 10% of isolated cases, it is important to investigate rare variants in both dominant and recessive models to search for additional genetic factors.

Jin et al., 2017 [38] developed a rigorous method to quantify the contribution of recessive genotypes of rare variants. As in previous analysis with implied dominant model, they only considered damaging coding (LoF, D-Mis, indels) variants. Under outbred population, the observed recessive genotypes should be proportional to the square of the cumulative frequency of damaging alleles, while under inbred population, it instead increases linearly with this number. This could be represented as a polynomial regression function of gene mutability

$$Number\ of\ RGs = \beta_0 + \beta_1 * mutability + \beta_2 * mutability^2 + \epsilon \quad (5)$$

They tested the enrichment of observed recessive genotypes with a one-tailed binomial test in a specific gene or gene set in cases. This frequentist approach provides additional statistical significance in identifying CHD risk genes. Through this approach, they found enrichment of a single *GDF1* founder mutation in Ashkenazim population. They also identified *FLT4* with significantly enriched rare heterozygous LoF variants in the patients with TOF, indicating its dominant genetic role to this subtype of CHD. *FLT4*, together with *NOTCH1*, was further confirmed to be the most frequent site of genetic variants that predisposed to TOF [39].

Watkins et al., 2019 [40••], analyzed the same exome sequencing dataset with a different approach, VAAST [41,42]. VAAST is a gene-burden test that ranks the probability being a risk gene or gene set based on empirically calibrated rarity of genotypes. VAAST approach is especially relevant for assessing the impact of compound heterozygous genotypes. Using VAAST, they find that cilia and cilia-related genes are enriched for rare, damaging recessive variants, suggesting their recessive homozygous and compound heterozygous leading to CHD.

Analysis of copy number variants

Pathogenic copy number variants (CNVs) were estimated to be presented in 4–14% of patients with isolated CHD and 15–20% of patients with CHD and extracardiac anomalies [43]. The basic model of analyzing CNVs is similar to small rare variants, that is, to test whether CNVs are enriched in cases compared with controls in genomic intervals such as genes. Audain et al., 2021 [44] reported an integrated analysis of CNVs and DNVs. The study assembled CNV data of 7958 cases and 14 082 controls, mostly from public repositories, and coding DNVs of 2489 cases from two recent publications [11,38]. They performed enrichment test on rare CNVs and DNVs separately and combined p-values using Fisher's method. The study identified 21 genes with significant association with loss-of-function point mutations or deletions.

Analysis of *de novo* noncoding variants

Most of CHD cases, even syndromic ones, do not carry damaging DNVs in coding regions or pathogenic CNVs. This motivated the investigation of *de novo* noncoding variants in Richter et al., 2020 [45••]. Unlike protein-altering DNVs, there is no significant overall enrichment of *de novo* noncoding variants in cases. To quantify how much *de novo* noncoding variants contribute to CHD risk, it is crucial to preselect variants that may have a functional impact. This requires a systematic integration of functional genomics data. In Richter et al., 2020 [45••], they introduced HeartENN, a deep learning-based epigenomic-effect model based on DeepSEA [46]. HeartENN uses convolutional neural network architecture to predict the genome-wide features for human and mouse based on the heart-specific chromatin profile, including histone markers in promoter and enhancer regions. Noncoding variants were scored and filtered with HeartENN with optimal threshold, followed by case–control binomial tests. They observed a significant enrichment of noncoding variants with HeartENN score ≥0.1 in 749 CHD trios compared with 1611 unaffected trios. In addition to transcriptional regulatory disruption, they also tested the enrichment of noncoding variants that may disrupt post-transcriptional regulation, through a combination of RNA-binding protein-binding sites and histone markers of transcribing gene body. To account for testing of many hypotheses among which many are correlated, they used a method described in Werling et al., 2018 [47] to estimate the effect number of independent tests by the number of eigen vectors that explains ≥99% of the variance of the correlations between features. This number is used to perform Bonferroni correction of p-values of individual tests. Those tests provided additional evidence of disturbed post-transcriptional regulation machinery that may contribute to CHD.

Analysis of common variants

GWAS are designed to test the association of individual genetic variants across the genome with phenotypes. It has been widely used [48] in complex diseases and traits with variants that are common in the population, for example, allele frequency above 1. A recent GWAS study on CHD was performed on 4034 cases with CHD and 8486 healthy controls [17••]. They identified 20 genome-wide significant SNPs with a wide range of effect size (odds ratio from 1.57 to 6.11). As expected, the effect size is negatively correlated with allele frequency. All the significant SNPs with odds ratio above 2 have allele frequency below 5%. Interestingly, all but one SNP are associated with specific CHD subtypes. Since the sample size of each subtype is still modest in the study, the statistical power of GWAS in CHD is limited. How to effectively identify genetic risk factors that are shared across subtypes and subtype-specific using the same dataset without incurring heavy multiple testing adjustment is an open statistical problem.

Conclusions

Recent studies on rare and common variants in CHD have significantly improved the understanding of the etiology and genetic architecture of CHD. Several advances in the statistical analysis and variant annotation tools, including applications of background mutation model (1213 trios, [9]••), Bayesian association methods

(1891 trios, [11]), recent missense pathogenicity prediction tools, VAAST (2391 trios, Watkins et al., 2019 [40••]), and HeartENN (749 probands and 1611 control trios, Richter et al., 2020 [45••]) have significantly increased the statistical power.

There are significant gaps in our understanding of human CHD genetics. The identified genetic risk factors only account for a minor fraction of population-attributable risk, especially isolated CHD without additional congenital anomalies or neurodevelopmental disorders. Some risk genes are shared in across CHD subtypes, while others are specific to certain subtypes. Comprehensive identification of new risk genes by both rare and common variants is still a key to advance the field and provide the basis for improving understanding of developmental biology of CHD and long-term clinical outcomes. To improve statistical power, we need more international collaboration and effective data sharing. Furthermore, we need new statistical and computational methods that can take advantage of recent advances in machine learning, protein structure, and single-cell technologies. For example, modeling of single-cell expression and functional genomics data to infer regulatory networks during fetal heart development would be of great help for understanding the genetic architectures of CHD. Finally, accurate prediction of the pathogenicity and mode of action of missense variants is critically important to improve power of new gene discovery and clinical interpretation. Recent advances in machine-learning modeling of protein sequence and structure can improve our ability to make predictions of the functional and genetic impact of genetic variants.

Conflict of interest

None.

Acknowledgments

This work was supported by National Institutes of Health (NIH) grants R01GM120609 and U01HL153009.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest.

1. Chin-Yee NJ, Costain G, Swaby JA, Silversides CK, Bassett AS: **Reproductive fitness and genetic transmission of tetralogy of Fallot in the molecular age**. *Circ Cardiovasc Genet* 2014, **7**:102-109.

2. Tartaglia M, Mehler EL, Goldberg R, Zampino G, Brunner HG, Kremer H, van der Burgt I, Crosby AH, Ion A, Jeffery S, *et al.*: **Mutations in PTPN11, encoding the protein tyrosine phosphatase SHP-2, cause Noonan syndrome**. *Nat Genet* 2001, **29**:465-468.

3. Tartaglia M, Kalidas K, Shaw A, Song X, Musat DL, van der Burgt I, Brunner HG, Bertola DR, Crosby A, Ion A, *et al.*: **PTPN11 mutations in Noonan syndrome: molecular spectrum, genotype-phenotype correlation, and phenotypic heterogeneity**. *Am J Hum Genet* 2002, **70**:1555-1563.

4. Becker K, Hughes H, Howard K, Armstrong M, Roberts D, Lazda EJ, Short JP, Shaw A, Patton MA, Tartaglia M: **Early fetal death associated with compound heterozygosity for Noonan syndrome-causative PTPN11 mutations**. *Am J Med Genet A* 2007, **143A**:1249-1252.

5. Bertola DR, Pereira AC, de Oliveira PS, Kim CA, Krieger JE: **Clinical variability in a Noonan syndrome family with a new PTPN11 gene mutation**. *Am J Med Genet A* 2004, **130A**:378-383.

6. Binder G, Neuer K, Ranke MB, Wittekindt NE: **PTPN11 mutations are associated with mild growth hormone resistance in individuals with Noonan syndrome**. *J Clin Endocrinol Metab* 2005, **90**:5377-5381.

7. Micale L, Augello B, Maffeo C, Selicorni A, Zucchetti F, Fusco C, De Nittis P, Pellico MT, Mandriani B, Fischetto R, *et al.*: **Molecular analysis, pathogenic mechanisms, and readthrough therapy on a large cohort of Kabuki syndrome patients**. *Hum Mutat* 2014, **35**:841-850.

8. Van Laarhoven PM, Neitzel LR, Quintana AM, Geiger EA, Zackai EH, Clouthier DE, Artinger KB, Ming JE, Shaikh TH: **Kabuki syndrome genes KMT2D and KDM6A: functional analyses demonstrate critical roles in craniofacial, heart and brain development**. *Hum Mol Genet* 2015, **24**:4443-4453.

9. •• Homsy J, Zaidi S, Shen Y, Ware JS, Samocha KE, Karczewski KJ, DePalma SR, McKean D, Wakimoto H, Gorham J, *et al.*: ***De novo* mutations in congenital heart disease with neurodevelopmental and other congenital anomalies**. *Science* 2015, **350**:1262-1266.

In this study, the authors showed that protein-damaging DNVs are enriched in CHD cases, especially in genes highly expressed during embroynic heart development that involved in morphogenesis, chromatin modification, and transcriptional regulation. They identified three genes, PTPN11, KMT2D and RBFOX2, that are associated with CHD risk by DNVs with genome-wide significance. Furthermore, using permuations, they showed that most of the other genes with multiple protein-damaging DNVs are also CHD risk genes, even though there is a lack of statistical evidence to identify risk genes individually among these genes.

10. •• Jin SC, Homsy J, Zaidi S, Lu Q, Morton S, DePalma SR, Zeng X, Qi H, Chang W, Sierant MC, *et al.*: **Contribution of rare inherited and *de novo* variants in 2871 congenital heart disease probands**. *Nat Genet* 2017, **49**:1593-1601.

The authors systematic analyzed the impact of rare inherited recessive and dominant variants and of DNMs via exome sequencing on 2871 CHD cases. They identified GDF1, MYH6, FLT4 that are associated with severe CHD in Ashkenazim, Shone complex and TOF, respectively. They developed a method to model rare recessive genotypes and quantified the contribution of recessive genotypes to CHD.

11. • Sifrim A, Hitz MP, Wilsdon A, Breckpot J, Turki SH, Thienpont B, McRae J, Fitzgerald TW, Singh T, Swaminathan GJ, *et al.*: **Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing**. *Nat Genet* 2016, **48**:1060-1065.

In this study, the authors exome sequenced 1891 probands, including 610 syndromic CHD and 1281 nonsyndromic CHD. They confirmed a significant enrichment of de novo PTVs in syndromic cases, while significant enrichment of PTVs inherited from unaffected parents in nonsyndromic cases. They identified three genome-wide significant genes that contribute to syndromic CHD by de novo mutations, CHD4, CDK13 and PRKD1. Additionally, they applied a Bayesian association method, TADA, to identify additional candidate risk genes with calibrated FDR.

12. • McBride KL, Pignatelli R, Lewin M, Ho T, Fernbach S, Menesses A, Lam W, Leal SM, Kaplan N, Schliekelman P, *et al.*: **Inheritance analysis of congenital left ventricular outflow tract obstruction malformations: segregation, multiplex relative risk, and heritability**. *Am J Med Genet A* 2005, **134A**:180-186.

The authors analyzed data from 124 families with left ventricular outflow tract (LVOTO) malformations. They estimated a heritability of 0.71-0.90 and a relative risk of 36.9 for first-degree relatives of affected individuals. The results support a complex pattern of inheritance, likely oligogenic, of LVOTO.

13. • Pierpont ME, Brueckner M, Chung WK, Garg V, Lacro RV, McGuire AL, Mital S, Priest JR, Pu WT, Roberts A, *et al.*: **Genetic basis for**

congenital heart disease: revisited: a scientific statement from the American Heart Association. *Circulation* 2018, **138**:e653-e711.

The review provided an overview of epidemiology and genetic basis of CHD. It summarized several studies of CHD subtypes with estimated heritability in the range of 70% to 90%, supporting strong genetic contribution to those subtypes.

14. Cripe L, Andelfinger G, Martin LJ, Shooner K, Benson DW: **Bicuspid aortic valve is heritable**. *J Am Coll Cardiol* 2004, **44**:138-143.

15. Hinton RB Jr., Martin LJ, Tabangin ME, Mazwi ML, Cripe LH, Benson DW: **Hypoplastic left heart syndrome is heritable**. *J Am Coll Cardiol* 2007, **50**:1590-1595.

16. Nogee JM, Jay PY: **The heritable basis of congenital heart disease: past, present, and future**. *Circ Cardiovasc Genet* 2016, **9**:315-317.

17. •• Lahm H, Jia M, Dressen M, Wirth F, Puluca N, Gilsbach R, Keavney BD, Cleuziou J, Beck N, Bondareva O, *et al.*: **Congenital heart disease risk loci identified by genome-wide association study in European patients**. *J Clin Invest* (2) 2021, **131**:e141837.

The authors performed a genome-wide association study (GWAS) of 4034 patients with CHD and 8486 healthy controls. They identified 20 SNPs reached genome-wide significance in various subtypes of CHD. These SNPs are close to MACROD2, GOSR2, WNT3, and MSX1. Further single cell RNA-Seq analyses provided strong functional evidence that those genes play important roles during embryonic development of the human heart.

18. Zaidi S, Choi M, Wakimoto H, Ma L, Jiang J, Overton JD, Romano-Adesman A, Bjornson RD, Breitbart RE, Brown KK, *et al.*: ***De novo* mutations in histone-modifying genes in congenital heart disease**. *Nature* 2013, **498**:220-223.

19. O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, Coe BP, Levy R, Ko A, Lee C, Smith JD, *et al.*: **Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations**. *Nature* 2012, **485**:246-250.

20. Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, Sabo A, Lin CF, Stevens C, Wang LS, Makarov V, *et al.*: **Patterns and rates of exonic *de novo* mutations in autism spectrum disorders**. *Nature* 2012, **485**:242-245.

21. He X, Sanders SJ, Liu L, De Rubeis S, Lim ET, Sutcliffe JS, Schellenberg GD, Gibbs RA, Daly MJ, Buxbaum JD, *et al.*: **Integrated model of *de novo* and inherited genetic variants yields greater power to identify risk genes**. *PLoS Genet* 2013, **9**:e1003671.

22. Iossifov I, O'Roak BJ, Sanders SJ, Ronemus M, Krumm N, Levy D, Stessman HA, Witherspoon KT, Vives L, Patterson KE, *et al.*: **The contribution of *de novo* coding mutations to autism spectrum disorder**. *Nature* 2014, **515**:216-221.

23. De Rubeis S, He X, Goldberg AP, Poultney CS, Samocha K, Cicek AE, Kou Y, Liu L, Fromer M, Walker S, *et al.*: **Synaptic, transcriptional and chromatin genes disrupted in autism**. *Nature* 2014, **515**:209-215.

24. Qi H, Yu L, Zhou X, Wynn J, Zhao H, Guo Y, Zhu N, Kitaygorodsky A, Hernan R, Aspelund G, *et al.*: ***De novo* variants in congenital diaphragmatic hernia identify MYRF as a new syndrome and reveal genetic overlaps with other developmental disorders**. *PLoS Genet* 2018, **14**:e1007822.

25. Wang J, Ahimaz PR, Hashemifar S, Khlevner J, Picoraro JA, Middlesworth W, Elfiky MM, Que J, Shen Y, Chung WK: **Novel candidate genes in esophageal atresia/tracheoesophageal fistula identified by exome sequencing**. *Eur J Hum Genet* 2021, **29**:122-130.

26. Qiao L, Xu L, Yu L, Wynn J, Hernan R, Zhou X, Farkouh-Karoleski C, Krishnan US, Khlevner J, De A, *et al.*: **Rare and *de novo* variants in 827 congenital diaphragmatic hernia probands implicate LONP1 as candidate risk gene**. *Am J Hum Genet* 2021, **108**:1964-1980.

27. Zhong G, Ahimaz P, Edwards NA, Hagen JJ, Faure C, Kingma P, Middlesworth W, Khlevner J, Fiky ME, Schindel D, *et al.*: **Identification and validation of novel candidate risk genes in endocytic vesicular trafficking associated with esophageal atresia and tracheoesophageal fistulas**. *HGG Adv.* (3) 2022, **3**:100107.

28. Bishop MR, Diaz Perez KK, Sun M, Ho S, Chopra P, Mukhopadhyay N, Hetmanski JB, Taub MA, Moreno-Uribe LM, Valencia-Ramirez LC, *et al.*: **Genome-wide enrichment of *de novo* coding mutations in orofacial cleft trios**. *Am J Hum Genet* 2020, **107**:124-136.

29. Ware JS, Samocha KE, Homsy J, Daly MJ: **Interpreting *de novo* variation in human disease using denovolyzeR**. *Curr Protoc Hum Genet* 2015, **87**:7.25.1-7.25.15.

30. Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM, Kosmicki JA, Rehnstrom K, Mallick S, Kirby A, *et al.*: **A framework for the interpretation of *de novo* mutation in human disease**. *Nat Genet* 2014, **46**:944-950.

31. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, Musolf A, Li Q, Holzinger E, Karyadi D, *et al.*: **REVEL: an ensemble method for predicting the pathogenicity of rare missense variants**. *Am J Hum Genet* 2016, **99**:877-885.

32. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M: **CADD: predicting the deleteriousness of variants throughout the human genome**. *Nucleic Acids Res* 2019, **47**:D886-D894.

33. Evans P, Wu C, Lindy A, McKnight DA, Lebo M, Sarmady M, Abou, Tayoun AN: **Genetic variant pathogenicity prediction trained using disease-specific clinical sequencing data sets**. *Genome Res* 2019, **29**:1144-1151.

34. Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, Liu X: **Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies**. *Hum Mol Genet* 2015, **24**:2125-2137.

35. Carlson J, Locke AE, Flickinger M, Zawistowski M, Levy S, Myers RM, Boehnke M, Kang HM, Scott LJ, Li JZ, *et al.*: **Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans**. *Nat Commun* 2018, **9**:3753.

36. Homsy J, Zaidi S, Shen Y, Ware JS, Samocha KE, Karczewski KJ, DePalma SR, McKean D, Wakimoto H, Gorham J, *et al.*: ***De novo* mutations in congenital heart disease with neurodevelopmental and other congenital anomalies**. *Science* 2015, **350**:1262-1266.

37. Nguyen HT, Bryois J, Kim A, Dobbyn A, Huckins LM, Munoz-Manchado AB, Ruderfer DM, Genovese G, Fromer M, Xu X, *et al.*: **Integrated Bayesian analysis of rare exonic variants to identify risk genes for schizophrenia and neurodevelopmental disorders**. *Genome Med* 2017, **9**:114.

38. Jin SC, Homsy J, Zaidi S, Lu Q, Morton S, DePalma SR, Zeng X, Qi H, Chang W, Sierant MC, *et al.*: **Contribution of rare inherited and *de novo* variants in 2871 congenital heart disease probands**. *Nat Genet* 2017, **49**:1593-1601.

39. Page DJ, Miossec MJ, Williams SG, Monaghan RM, Fotiou E, Cordell HJ, Sutcliffe L, Topf A, Bourgey M, Bourque G, *et al.*: **Whole exome sequencing reveals the major genetic contributors to nonsyndromic tetralogy of fallot**. *Circ Res* 2019, **124**:553-563.

40. •• Watkins WS, Hernandez EJ, Wesolowski S, Bisgrove BW, Sunderland RT, Lin E, Lemmon G, Demarest BL, Miller TA, Bernstein D, *et al.*: ***De novo* and recessive forms of congenital heart disease have distinct genetic and phenotypic landscapes**. *Nat Commun* 2019, **10**:4722.

The authors analyzed whole exome sequencing (WES) data of 2391 trios. They used VAAST to prioritize disease associated genotypes. They identified 229 damaged cilia-related genes with enriched recessive genotypes while depleted DNVs. They observed opposite trend for chromatin-modifying genes. Their analysis revealed that dominant and recessive CHD genes are associated with different functions.

41. Yandell M, Huff C, Hu H, Singleton M, Moore B, Xing J, Jorde LB, Reese MG: **A probabilistic disease-gene finder for personal genomes**. *Genome Res* 2011, **21**:1529-1542.

42. Hu H, Huff CD, Moore B, Flygare S, Reese MG, Yandell M: **VAAST 2.0: improved variant classification and disease-gene identification using a conservation-controlled amino acid substitution matrix**. *Genet Epidemiol* 2013, **37**:622-634.

43. Andersen TA, Troelsen Kde L, Larsen LA: **Of mice and men: molecular genetics of congenital heart disease**. *Cell Mol Life Sci* 2014, **71**:1327-1352.

44. Audain E, Wilsdon A, Breckpot J, Izarzugaza JMG, Fitzgerald TW, Kahlert AK, Sifrim A, Wunnemann F, Perez-Riverol Y, Abdul-Khaliq H, *et al.*: **Integrative analysis of genomic variants reveals new associations of candidate haploinsufficient genes with congenital heart disease**. *PLoS Genet* 2021, **17**:e1009679.

45. •• Richter F, Morton SU, Kim SW, Kitaygorodsky A, Wasson LK, Chen KM, Zhou J, Qi H, Patel N, DePalma SR, *et al.*: **Genomic analyses implicate noncoding *de novo* variants in congenital heart disease**. *Nat Genet* 2020, **52**:769-777.
The authors compared genome sequences from 749 CHD probands and their parents with those from 1611 unaffected trios. They develped a neural network based non-coding variant transcriptional impact prediction method and identified enrichment of damaing non-coding DNVs in cases compared to controls. Overall, their findings implicated enrichment of potentially disruptive regulatory noncoding DNVs in a fraction of CHD, and highlighted the potential of WGS to more fully elucidate the genetic architecture of CHD.

46. Zhou J, Troyanskaya OG: **Predicting effects of noncoding variants with deep learning-based sequence model**. *Nat Methods* 2015, **12**:931-934.

47. Werling DM, Brand H, An JY, Stone MR, Zhu L, Glessner JT, Collins RL, Dong S, Layer RM, Markenscoff-Papadimitriou E, *et al.*: **An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder**. *Nat Genet* 2018, **50**:727-736.

48. Tam V, Patel N, Turcotte M, Bosse Y, Pare G, Meyre D: **Benefits and limitations of genome-wide association studies**. *Nat Rev Genet* 2019, **20**:467-484.