# M-DATA: A Statistical Approach to Jointly Analyzing *De Novo* Mutations for Multiple Traits

Yuhan Xie[1#], Mo Li[1#], Weilai Dong[2], Wei Jiang[1], Hongyu Zhao[1,3*]

[1] Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA 06510

[2] Department of Genetics, Yale School of Medicine, New Haven, CT, USA 06510

[3] Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA 06511

[#] These authors contributed to this work equally

* To whom correspondence should be addressed:

Prof Hongyu Zhao

Department of Biostatistics, Yale School of Public Health, 60 College Street, New Haven, CT, 06520, USA

Email: hongyu.zhao@yale.edu

## Abstract

Recent studies have demonstrated that multiple early-onset diseases have shared risk genes, based on findings from *de novo* mutations (DMNs). Therefore, we may leverage information from one trait to improve statistical power to identify genes for another trait. However, there are few methods that can jointly analyze DNMs from multiple traits. In this study, we develop a framework called M-DATA (**M**ulti-trait framework for ***De novo*** mutation **A**ssociation **T**est with **A**nnotations) to increase the statistical power of association analysis by integrating data from multiple correlated traits and their functional annotations. Using the number of DNMs from multiple diseases, we develop a method based on an Expectation-Maximization algorithm to both infer the degree of association between two diseases as well as to estimate the gene association probability for each disease. We apply our method to a case study of jointly analyzing data from congenital heart disease (CHD) and autism. Our method was able to identify 23 genes from joint analysis, including 12 novel genes, which is substantially more than single-trait analysis, leading to novel insights into CHD disease etiology.

## Introduction

Congenital heart disease (CHD) is the most common birth defect. It affects 0.8% of live birth and accounts for one-third of all major congenital abnormalities [1,2]. CHD is associated with both genetic and environmental factors [2]. It is genetically heterogenous and the estimated heritability in a Danish twin study is close to 0.5 [3].

Studies on *de novo* mutations (DNMs) have been successful in identifying risk genes for early on-set diseases as DNMs with deleterious effects have not been through natural selection. By conducting whole-exome sequencing (WES) studies for parent-offspring trios, there are cumulative findings of potential risk genes for CHD and neurodevelopmental disorders by identifying genes with more DNMs than expected by chance [4-6]. However, the statistical power for identifying risk genes is still hampered by the limited sample size of WES due to its relatively high cost in recruiting and sequencing samples, as well as the low occurrence of DNMs given its rarity.

Meta-analysis and joint analysis are two major approaches to improve the statistical power by integrating information from different studies. Meta-analysis studies on WES DNMs and Genome-wide Association Studies (GWAS) for multiple traits have been conducted [7,8]. However, these approaches may overlook the heterogeneity among traits, thus hinder the ability to interpret finding for each single trait. By identifying the intersection of top genes from multiple traits, some recent studies have shown that there are shared risk genes between CHD and autism [9,10]. Shared disease mechanism for early-onset neurodevelopmental diseases has also been reported [11,12]. Based on these findings, joint analysis methods have been proposed and gained success in GWAS and expression quantitative trait loci (eQTL) studies. Studies have shown that multi-trait analysis can improve statistical power [13-19] and accuracy of genetic risk

prediction [20-22]. Currently, there lacks joint analysis methods to analyze DNM data on multiple traits globally, with the exception of mTADA[23].

In addition to joint analysis, integrating functional annotations has also been shown to improve statistical power in GWAS [15,24] and facilitate the analysis of sequencing studies [25 26]. There is a growing number of publicly available tools to annotate mutations in multiple categories, such as the genomic conservation, epigenetic marks, protein functions and human health. With these resources, there is a need to develop a statistical framework for jointly analyzing traits with shared genetic architectures and integrating functional annotations for DNM data.

In this article, we propose a **M**ulti-trait ***D****e novo* mutation **A**ssociation **T**est with **A**nnotations, named M-DATA, to identify risk genes for multiple traits simultaneously based on pleiotropy and functional annotations. We demonstrate the performance of M-DATA through extensive simulation studies and real data examples. Through simulations, we illustrate that M-DATA is able to accurately estimate the proportion of disease-causing genes between two traits under various genetic architectures. M-DATA outperformed single-trait approaches and methods even if annotation information was not used. Annotations can further boost the power of M-DATA. We applied M-TADA to identify risk genes for CHD and autism. There are 23 genes discovered to be significant for CHD, including 12 novel genes, bringing novel insight to the disease etiology of CHD.

## Materials and Methods

### *Probabilistic model*

Firstly, we consider the simplest case with only one trait, and then we extend our model to multiple traits. We denote $Y_i$ as the DNM count for gene $i$ in a case cohort, and assume $Y_i$ come

from the mixture of null ($H_0$), and non-null ($H_1$), with proportion $\pi_0 = 1 - \pi$ and $\pi_1 = \pi$ respectively. Let $Z_i$ be the latent binary variable indicating whether this gene is associated with the trait of interest, where $Z_i = 0$ means gene $i$ is unassociated ($H_0$), and $Z_i = 1$ means gene $i$ is associated ($H_1$). Then, we have the following model:

$$Z_i \sim Bernoulli(\pi)$$

$$Y_i | Z_i = 0 \sim Poisson(2N\mu_i)$$

$$Y_i | Z_i = 1 \sim Poisson(2N\mu_i\gamma_i)$$

where N is the sample size of the case cohort, $\mu_i$ is the mutability of gene $i$ estimated using the framework in Samocha et al. [27], and $\gamma_i$ is the relative risk of the DNMs in the risk gene and is assumed to be larger than 1. The derivation of the model is the same as that in TADA [6]. We define this model as the single-trait without annotation model in the following text.

To leverage information from functional annotations, we use an exponential link between $\gamma_i$ and $X_i$ , the functional annotation vector of gene $i$ to make sure $\gamma_i$ is greater than or equal to 1 as follows:

$$\gamma_i = exp(X_i^T \beta),$$

where $\beta$ is the effect size vector of the functional annotations. We use the variant-level annotation from ANNOVAR[28]. We define loss-of-function (LoF) as frameshift insertion/deletion, splice site alteration, stopgain and stoploss predicated by ANNOVAR, and define deleterious variants (Dmis) predicted by MetaSVM[29]. Variant-level annotations are then collapsed into gene-level annotation, which is discussed in details in real data analysis of section Results. We also consider using pLI and mis-z scores directly as gene-level annotations from gnomAD v2.1.1[28].

Now we extend our model to consider multiple traits simultaneously. To unclutter our notations, we present the model for the two-trait case. Suppose we have gene counts $Y_{i1}$ and $Y_{i2}$ for gene $i$ from two cohorts with different traits. Similarly, we introduce latent variables $Z_i = [Z_{i00}, Z_{i10}, Z_{i01}, Z_{i11}]$ to indicate whether gene $i$ is associated with the traits. Specifically, $Z_{i00} = 1$ means the gene $i$ is associated with neither trait, $Z_{i10} = 1$ means that it is only associated with the first trait, $Z_{i01} = 1$ means that it is only associated with the second trait, and $Z_{i11} = 1$ means that it is associated with both traits. Then, we have:

$$Z_i \sim Multinomial(1, \pi), \text{ with } \pi = (\pi_{00}, \pi_{10}, \pi_{01}, \pi_{11})$$

$$\pi_{00} = \Pr(Z_{i00} = 1),\ Y_{i1}|Z_{i00} \sim Poisson(2N_1\mu_i),\ Y_{i2}|Z_{i00} \sim Poisson(2N_2\mu_i)$$

$$\pi_{10} = \Pr(Z_{i10} = 1),\ Y_{i1}|Z_{i10} \sim Poisson\left(2N_1\mu_i\gamma_{i1}\right),\ Y_{i2}|Z_{i10} \sim Poisson(2N_2\mu_i)$$

$$\pi_{01} = \Pr(Z_{i01} = 1),\ Y_{i1}|Z_{i01} \sim Poisson(2N_1\mu_i),\ Y_{i2}|Z_{i01} \sim Poisson(2N_2\mu_i\gamma_{i2})$$

$$\pi_{11} = \Pr(Z_{i11} = 1),\ Y_{i1}|Z_{i11} \sim Poisson(2N_1\mu_i\gamma_{i1}),\ Y_{i2}|Z_{i11} \sim Poisson(2N_2\mu_i\gamma_{i2})$$

$$\gamma_{i1} = \exp(X_{i1}^T\beta_1),\ \gamma_{i2} = \exp\left(X_{i2}^T\beta_2\right)$$

where $\pi$ is the corresponding risk proportion of genes belonging to each class, with $\sum_{l\in\{00,10,01,11\}} \pi_l = 1$. Then, the risk proportion of the first trait and second trait is $\pi_{10} + \pi_{11}$ and $\pi_{01} + \pi_{11}$, respectively. When there is no pleiotropy of the two traits, $\pi_{11} = (\pi_{10} + \pi_{11})(\pi_{01} + \pi_{11})$. The difference between $\pi_{11}$ and $(\pi_{10} + \pi_{11})(\pi_{01} + \pi_{11})$ reflects the magnitude of global pleiotropy between the two traits. $\mu_i$ is the same as our one-trait model. $N_1, \gamma_{i1}$ and $X_{i1}$ are the case cohort size, relative risk and annotation vector of gene $i$ for the first trait. $N_2, \gamma_{i2}$ and $X_{i2}$ are similarly defined for the second trait.

Denote $\Theta = (\pi, \beta_1, \beta_2)$ the parameters to be estimated in our model. As we only consider *de novo* mutations, they can be treated as independent as they occur with very low frequency. The full likelihood function can be written as

$$L(\Theta) = \prod_{i=1}^{M} \sum_{l \in \{00,10,01,11\}} [\pi_l \Pr(Y_i | Z_{il} = 1; \Theta)]^{Z_{il}}$$

where M is the number of genes. The log-likelihood funciton is

$$l(\Theta) = \sum_{i=1}^{M} \log \sum_{l \in \{00,10,01,11\}} [\pi_l \Pr(Y_i | Z_{il} = 1; \Theta)]^{Z_{il}}.$$

***Estimation***

Parameters of our models can be estimated using the Expectation-Maximization (EM) algorithm [30]. It is very computationally efficient for our model without annotations because we have explicit solutions for the estimation of all parameters in the M-step.

By Jenson's inequality, the lower bound $Q(\Theta)$ of the log-likelihood function is

$$l(\Theta) \geq Q(\Theta) = \sum_{i=1}^{M} \sum_{l \in \{00,10,01,11\}} \left[ Z_{il} [\log(\pi_l) + \log(\Pr(Y_i | Z_{il} = 1; \Theta))] \right].$$

The algorithm has two steps. In the E-step, we update the estimation of latent variables $Z_{il}, l \in \{00,01,10,11\}$ by its posterior probability under the current parameter estimates in round s. That is,

$$Z_{il}^{(s)} = \Pr\left(Z_{il} = 1 \middle| Y_i; \Theta^{(s)}\right) = \frac{\Pr\left(Z_{il} = 1, Y_i \middle| \Theta^{(s)}\right)}{\Pr\left(Y_i \middle| \Theta^{(s)}\right)} = \frac{\pi_l^{(s)} \Pr\left(Y_{i1}, Y_{i2} \middle| Z_{il} = 1; \Theta^{(s)}\right)}{\sum_{l' \in \{00,01,10,11\}} \left[\pi_{l'}^{(s)} \Pr(Y_{i1}, Y_{i2} | Z_{il'} = 1; \Theta^{(s)}\right]}.$$

In the M-step, we update the parameters in $\Theta$ based on the estimation of $Z_{il}$ in the E-step by maximizing $Q(\Theta)$. For $\pi$, there is an analytical solution, which is

$$\pi_l^{(s+1)} = \frac{\sum_{i=1}^{M} Z_{il}^{(s)}}{M}$$

For the rest of derivation, we take the estimation process for the first trait as an example. If we do not add any functional annotations to our model, there exists an analytical solution for $\beta_1$ by taking the first order derivative of $Q(\Theta)$ with respect to $\beta_1$ as 0.

$$d_{\beta_1} Q(\Theta)^{(s)} = \sum_{i=1}^{M} (Z_{i10} + Z_{i11}) \left( Y_{i1} X_{i1} - 2N_1 \mu_i \exp\left(X_{i1}^T \beta_1\right) X_{i1} \right) = 0.$$

Then, the solutions for $\beta_1$ is

$$\beta_1^{(s+1)} = \log \frac{\sum_{i=1}^{M} Y_{i1} (Z_{i10} + Z_{i11})}{\sum_{i=1}^{M} 2N_1 \mu_i (Z_{i10} + Z_{i11})}$$

However, by adding functional annotations into our model, there is no explicit solution for $\beta_1$, so we adopt the Newton-Raphson method for estimation. The second-order derivatives for $Q(\Theta)$ is

$$d_{\beta_1}^2 Q(\Theta) = -\sum_{i=1}^{M} (Z_{i10} + Z_{i11}) \left( 2N_1 \mu_i \exp\left(X_{i1}^T \beta_1\right) X_{i1} X_{i1}^T \right),$$

Then, the estimate of $\beta_1$ can be obtained as

$$\beta_1^{(s+1)} = \beta_1^{(s)} - \left[ d_{\beta_1}^2 Q(\Theta)^{(s)} \right]^{-1} d_{\beta_1} Q(\Theta)^{(s)},$$

***Functional Annotation***

As we have discussed, there are multiple sources of functional annotations for DNMs. For gene-level annotations such as pLI and mis-z scores, we can directly plug into our gene-based model. For mutation-level annotations, such as prediction of pathogenicity, allele frequency, and splicing, it is important to collapse the variant-level information into gene-level without diluting useful information. Simply pulling over variant-level annotations of all base pairs within a gene may not be the best approach. To better understand the relationship, we calculate the likelihood ratio of the DNM counts under $H_1$ and $H_0$. Under $H_1$, for all positions t within a gene $i$, the DNM count $Y_{it}$ follows the Poisson distribution with relative risk $\gamma_{it}$ and mutability $\mu_{it}$ , then we have

$$\frac{P(Y_i|H_1)}{P(Y_i|H_0)} = \frac{\prod_t P(Y_{it}|H_1)}{\prod_t P(Y_{it}|H_0)} = \frac{\prod_t Poisson(2N\mu_{it}\gamma_{it})}{\prod_t Poisson(2N\mu_{it})},$$

where $\gamma_{it} = \exp(\beta_0 + \beta_1 X_{it})$. There is likely to be at most one mutation at each position *t* due to the low frequency of DNM. We can further simplify the above equation to

$$\frac{P(Y_i|H_1)}{P(Y_i|H_0)} = \frac{\prod_t \exp(\beta_0 + \beta_1 X_{it} I\{Y_{it} = 1\})\exp(-2N\mu_{it}\exp(\beta_0 + \beta_1 X_{it}))}{\prod_t \exp(-2N\mu_{it})}$$

$$= \exp\left(\sum_t (\beta_0 + \beta_1 X_{it} I\{Y_{it} = 1\}\right) \exp\left(\sum_t -2N\mu_{it}[\exp(\beta_0 + \beta_1 X_{it}) - 1]\right)$$

Without loss of generality, we assume the functional annotations are centered at 0. Then, by applying $\sum_t X_{it} = 0$, we obtain

$$\frac{P(Y_i|H_1)}{P(Y_i|H_0)} \approx \exp\left(\sum_t (\beta_0 + \beta_1 X_{it} I\{Y_{it} = 1\}\right) \exp\left(\sum_t -2N\mu_{it}[\exp(\beta_0) - 1]\right)$$

$$= \exp\left(\beta_0' + \beta_1' \sum_t (X_{it} I\{Y_{it} = 1\}\right).$$

The above approximation motivates us to aggregate variant-level annotations to gene-level annotations by summing up all annotation values of mutations within a gene after preprocessing each mutaiton-level annotation.

***Hypothesis Testing***

Without loss of generality, we take the first trait as an example to illustrate our testing procedure. After we estimate the parameters, genes can be prioritized based on their joint local false discovery rate (Jlfdr) [31]. For joint analysis of two traits, the Jlfdr of whether gene $i$ is associated with the first trait is

$$\mathrm{Jlfdr}_1(\mathrm{Y}_{i1}, \mathrm{Y}_{i2}) = \Pr\left(\mathrm{Z}_{i00} + \mathrm{Z}_{i01} = 1|\mathrm{Y}_{i1}, \mathrm{Y}_{i2}\right)$$

The following relationship between Jlfdr and false discovery rates (Fdr) was shown in Jiang and Yu [31],

$$\mathrm{Fdr}_1(\mathcal{R}) = E(\mathrm{Jlfdr}_1(\mathrm{Y}_{i1}, \mathrm{Y}_{i2}) | Z_{i00} + Z_{i01} \in \mathcal{R}) \approx \frac{1}{|\{\mathrm{Z}_{i00}+\mathrm{Z}_{i01}\in\mathcal{R}\}|} \sum_{\mathrm{Z}_{i00}+\mathrm{Z}_{i01}\in\mathcal{R}} \mathrm{Jlfdr}_1(\mathrm{Y}_{i1}, \mathrm{Y}_{i2}),$$

where $\mathcal{R}$ is the rejection region. Following their procedure, we determine the optimal rejection region by controlling $\mathrm{Jlfdr}_1$ smaller than a threshold $t(q)$. To determine the threshold t(q), we sort the calculated $\mathrm{Jlfdr}_1$value of each gene in an ascending order first. Denote the a-th $\mathrm{Jlfdr}_1$ value as $\mathrm{Jlfdr}_1^a$. We can approximate the Fdr of the region $\mathcal{R}_a = \{\mathrm{Z}_{i00} + \mathrm{Z}_{i01} | \mathrm{Jlfdr}_1(\mathrm{Y}_{i1}, \mathrm{Y}_{i2}) \leq \mathrm{Jlfdr}_1^a\}$ as

$$\mathrm{Fdr}\ (\mathcal{R}_a) = \frac{1}{a} \sum_{b=1}^{a} \mathrm{Jldfr}_1^b$$

Denote $c = \max\ \{a | \mathrm{Fdr}\ (\mathcal{R}_a) \leq q\}$ , and then the threshold t(q) for $\mathrm{Jldfr}_1$ is $\mathrm{Jldfr}_1^c$. For testing association with the first trait, we reject all genes with $\mathrm{Jlfdr}_1(\mathrm{Y}_{i1}, \mathrm{Y}_{i2}) \leq t(q)$. For both simulation and real data analyses, the global Fdr is controlled at $q = 0.05$.

## Results

### *Simulation studies*

We conducted comprehensive simulation studies to evaluate the estimation and power performance of M-DATA. We set the total number of genes M to 10,000, where genes were randomly selected from gnomAD v2.1 [32]. We set the size of the case cohort at 2000, 5000 and 10000, corresponding to a small, medium and large WES study. We fixed the risk proportion to 0.1 for each trait, and varied the shared risk proportion $\pi_{11}$ at 0.01, 0.03, 0.05, 0.07 and 0.09. When $\pi_{11} = 0.01$, it corresponds to the absence of pleiotropy between two traits, and we expect our multi-trait models to perform similarly as our single-trait models.

We first evaluated the performance of estimation for our models, and then we conducted power analysis for our single-trait models and multi-trait models. To evaluate the estimation performance for multi-trait models, we simulated the true model with two Bernoulli annotations,

and set the parameter of the Bernoulli distributions to 0.5 for both traits. We fixed the log of baseline relative risk $\beta_{j0}$ at 3, and varied the effect size $(\beta_{j1}, \beta_{j2}), j = 1,2$ from $(0.1, 0.1)$ $(0.1,0)$ and $(0,0)$, which corresponds to the cases when both annotations are effective, only the first annotation is effective and no annotation is effective. We evaluated the estimates of shared proportion of risk genes $\pi_{11}$ and the risk gene proportion for a single trait. There are in total 27 simulation settings for estimation evaluation. To obtain an empirical distribution of our estimated parameters, we replicated the process for 50 times for each setting. We simulated the two traits in a symmetrical way, so we only present the results of the first trait. The performance of estimation under the scenario that both annotations are effective$((\beta_{j1}, \beta_{j2}) = (0.1,0.1), j = 1,2)$ are shown in Figure 1. The rest of scenarios are shown in the supplemental data (Figure S1).

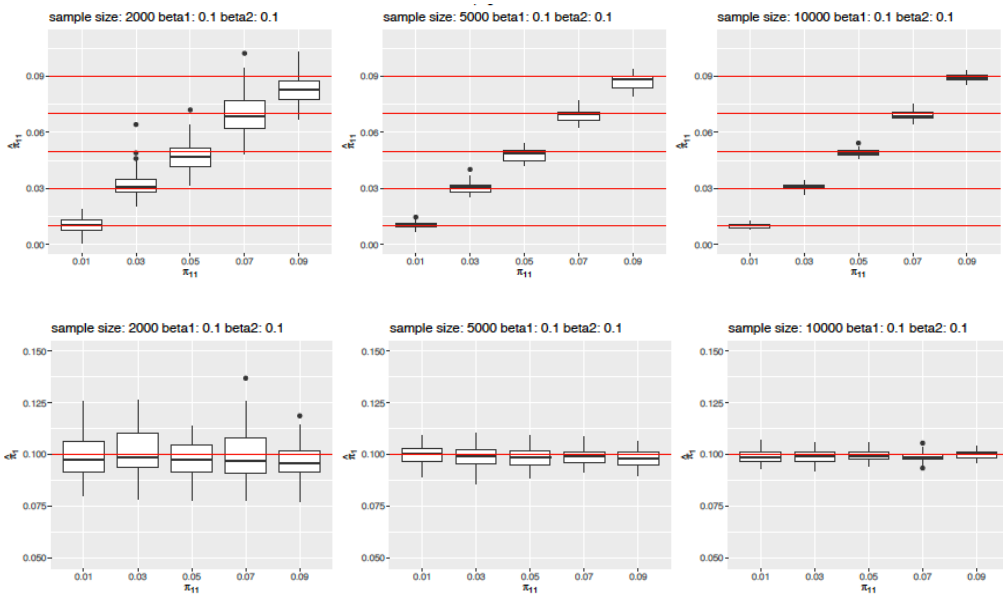
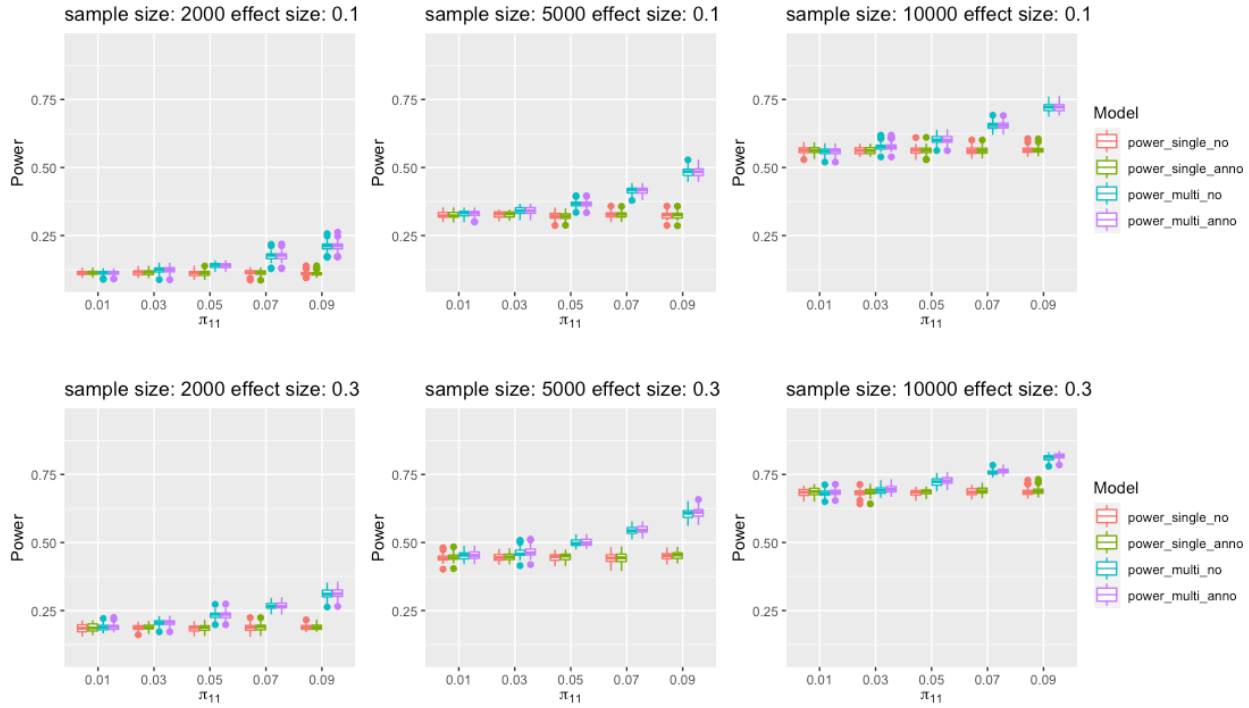



**Figure 1. Multi-trait analysis can accurately estimate the proportion of shared risk genes and single-trait risk genes.** Top panels show the estimation of shared risk proportion, and bottom panels show the estimation of a single trait. For each panel, each plot from left to right represents study sample size of 2000, 5000, and 10000, respectively. Within each plot, boxes from left to right represent the proportion of shared risk genes being 0.01, 0.03, 0.05, 0.07 and 0.09, respectively. True values are shown in red lines.

Given that the effective number of functional annotations for DNM data in real world is unknown, we explored the power performance of single-trait and multi-trait models when annotations are only partially observed. We fixed the log of baseline relative risk $\beta_{j0}$ at 3, and varied the effect size $(\beta_{j1}, \beta_{j2}, \beta_{j3}), j = 1,2$ from $(0.1, 0.1, 0.1)$ $(0.3, 0.3, 0.3)$ and $(0.5, 0.5, 0.5)$, which corresponds to the cases when effect of annotations is weak, moderate, and strong. We assumed that only the first two annotations can be observed. We first demonstrated our model can control global Fdr (Figure S2) under theses settings and then evaluated power (Figure 2), type I error (Figure S3), and AUC (Figure S4) for our single-trait model and multi-trait model. There are in total 45 simulation settings. Under each setting, the data were simulated based on our multi-trait model with annotations in the method section.



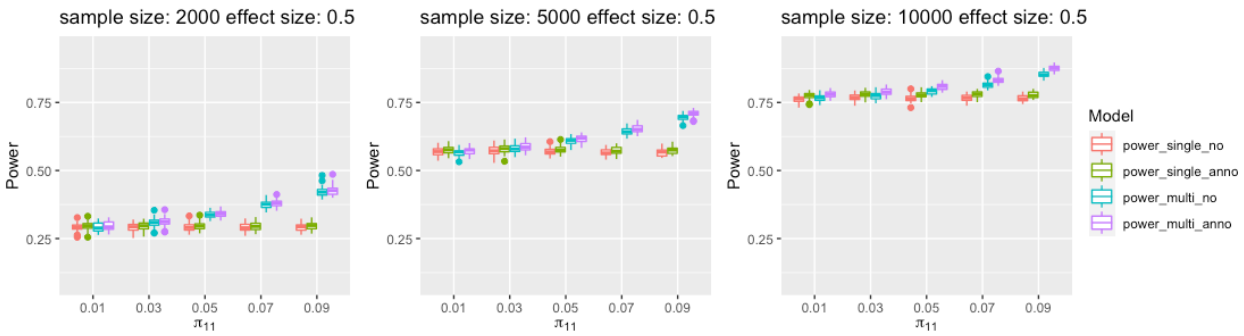



**Figure 2. Power performance under different strengths of annotations.** The panels from top to bottom show the power performance under weak, moderate and strong annotations. For each panel, each plot from left to right represents study sample size of 2000, 5000, and 10000, respectively. Within each plot, boxes from left to right represent the proportion of shared risk genes being 0.01, 0.03, 0.05, 0.07 and 0.09, respectively.

With the increase of the sample size, the performance of all four models becomes better. Under weak annotations, the power performance of models with annotations and without annotations are comparable. However, when annotations are strong, the power performance of models with annotations are better than models without annotations. With the increase of shared risk proportion, the power performance of multi-trait models become better than single-trait models.

### *Real Data Analysis*

We applied M-DATA to real DNM data from 2,645 CHD samples reported in Jin et al. [4] and 5,623 autism samples acquired from denovo-db [33]. We only considered LoF and D-Mis mutations in our analysis as the number of non-deleterious mutations is not expected to provide information to differentiate cases from controls biologically [34]. We used the output from ANNOVAR [28] as our annotation data. Specifically, we selected Polyphen2 [35], MPC [36], CADD [37] and REVEL [38] as the deleteriousness annotations, exome/genome allele frequency from gnomAD [32] as allele frequency annotation, ADA and RF score [39] as splicing annotations from ANNOVAR [28] and obtained pLI and misZ scores as gene conservation annotations from

gnomAD v2.1 [32]. The preprocessing steps of each annotation are provided in Table S1. All preprocessed annotations are centered and scaled before model fitting.

M-DATA can either use the preselected annotations in real data analysis, or consider annotations provided by the users and conduct feature selections if required. For feature selection, M-DATA firstly evaluates the correlation between all annotations and keeps only one annotation from a correlated cluster when the Pearson's correlation between two annotations is larger than a preset threshold. And then it keeps the top annotations with large effect sizes and refits the model with the selected annotations.

We performed single-trait analysis on CHD and autism data separately, followed by joint analysis both CHD and autism data with our multi-trait models. We compared the performance of single-trait models and multi-trait models for CHD under different significance thresholds. With a stringent significance threshold (i.e. FDR < 0.01), single-trait model without annotation identified 8 significant genes, single-trait model with annotation identified 10 significant genes, multi-trait model without annotation identified 11 significant genes, and multi-trait model with annotation identified 14 genes. With FDR < 0.05, single-trait model without annotation identified 15 significant genes, single-trait model with annotation identified 19 significant genes, multi-trait model without annotation identified 18 significant genes, and multi-trait model with annotation identified 23 significant genes. It demonstrates that M-DATA is able to identify more genes by jointly analyzing multiple traits and incorporating information from functional annotations.

| Model | FDR<0.05 | FDR<0.01 |
|---|---|---|
| Single no Anno: CHD/Autism | 15/28 | 8/17 |
| Single with Anno: CHD/Autism | 19/34 | 10/22 |
| Multi no Anno: CHD/Autism | 18/28 | 11/19 |
| Multi with Anno: CHD/Autism | 23/36 | 13/22 |

**Table 1. Results for M-DATA Single-Trait and Multi-Trait Models**

Among the 23 identified genes from joint model with annotations, 11 were well established known CHD genes based on a previously compiled gene list with 254 known CHD genes [4]. They are involved in essential developmental pathways or biological processes, such as Notch signaling (*NOTCH1*), RAS signaling (*PTPN11*, *RAF1*, *SOS1*, *BRAF*), PI3K/AKT signaling (*PTEN*), chromatin modeling (*CHD7*, *KMT2D*, *NSD1*), transcriptional regulations (*GATA6*), and cell structural support (*ACTB*) [40,41].

Among the 12 novel genes, *FRYL* and *NAA15* are identified by the multi-trait models, but not identified by the single-trait models. Compared to the single-trait with annotation model, the multi-trait with annotation model can identify 2 additional genes, which are *CDK13* and *LZTR1*.

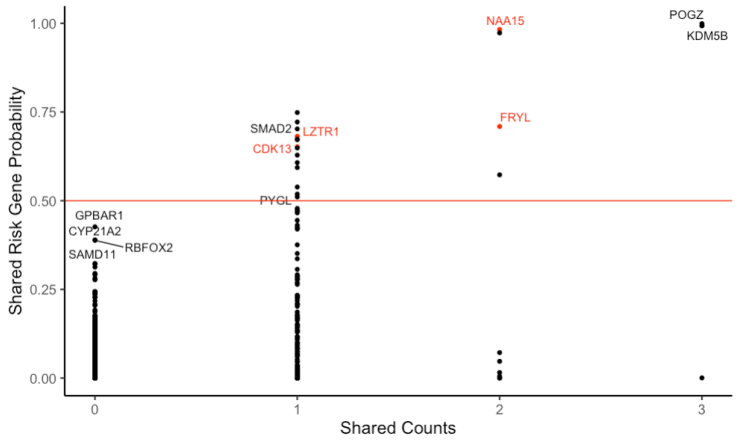



**Figure 3. Correlation between estimated shared risk probability and shared mutation counts.** The x-axis shows the shared mutation counts (i.e., minimum mutation counts between two traits). The y-axis shows the estimated shared risk gene probability of genes. Gene names of the 12 novel genes are annotated on the plot and the additional 4 genes that can be identified by the multi-trait models are marked in red.

*RBFOX2*, *SMAD2, CDK13* are three emerging CHD risk genes that have been recently reported to cause hypoplastic left heart syndrome [9,42,43], laterality defect [1,44], and septal defects and pulmonary valve abnormalities [45], respectively.

Additionally, 4 novel genes, *POGZ, KDM5B*, *NAA15*, and *FRYL,* harbored at least two *de novo* mutations in both CHD and autism cohorts.

*POGZ*, encoding a heterochromatin protein 1 alpha-binding protein, participates in chromatin modeling and gene regulations. It binds to chromatin and facilitates the packaging of DNA onto chromosomes. *POGZ* damaging *de novo* mutations were strongly linked with autism spectrum disorders and other neurodevelopmental disorders [46,47]. Interestingly, one of the reported mutation carriers also presented cardiac defect [48].

*KDM5B* is a lysine-specific histone demethylase. Studies have shown that it regulates H3K4 methylation near promoter and enhancer regions in embryonic stem cells and controls the cell pluripotency [49,50]. The deletion of *KDM5B* in mice is neonatal lethal with respiratory failure and neurodevelopmental defects [51]. Recessive mutations in the gene were associated with mental retardation (OMIM: 618109) and one reported patient presented atrial septal defect.

*NAA15* encodes the auxiliary subunit of N-Alpha-Acetyltransferase 15, which catalyzes one of the most common post-translational modification essential for normal cell functions. Protein-truncating mutations in *NAA15* were reported in intellectual disability and autism patients, some of whom also presented a variety of cardiac abnormalities including ventricular septal defect, heterotaxy, pulmonary stenosis and tetralogy of Fallot [52].

*POGZ*, *KDM5B*, and *NAA15* are all highly expressed in developmental heart at mice embryonic day E14.5 [4]. *POGZ* and *NAA15* are intolerant for both LoF and missense mutations, given that they have a pLI score > 0.9 and a missense z-score > 3. *KDM5B* is intolerant for missense mutations with a missense z-score of 1.78. Considering their intolerance of protein-altering variants, the identification of damaging *de novo* mutations in them is highly unlikely. Therefore, our analyses suggest that *POGZ*, *KDM5B* and *NAA15* may be considered as new candidate CHD genes.

Furthermore, among the 17 genes with at least one *de novo* mutation in CHD and autism cohorts, 5 genes (*KMT2D*, *NSD1*, *POGZ*, *SMAD2*, *KDM5B*) play a role in chromatin modeling. Such high proportion is consistent with previous studies that chromatin modeling-related transcriptional regulations are essential for both cardiac and neuro-development, and genes with critical regulatory roles in the process may be pleotropic [9].

We also compared our results with mTADA [23] for multi-trait analysis. We downloaded the CHD data [53], autism data [11], and mutability data from the github webpage of mTADA. We fitted our multi-trait model by plugging in the sum of LoF and Dmis mutation count of a gene as the mutation count for a gene in our model and the sum of LoF and Dmis mutability of a gene as the mutability for our model. This model is defined as our multi-trait model without annotation in the comparison. For our multi-trait model with annotation, we used the LoF mutation count of each gene as the annotation. Following their risk gene identification method, we also define genes with posterior probability (PP) larger than 0.8 as risk genes. For CHD data, among 19,358 genes tested, mTADA identified 11 risk genes, with 9 of 11 are unknown human CHD genes, our model without adding annotations identified 17 risk genes, with 13 are unknown human CHD genes, and our model with annotations identified 19 genes, with 15 unknown human CHD genes. All 11 risk genes identified by mTADA were also identified by our two models. The 6 additional unknown human CHD genes identified by our models are *BRD4, CPD, GANAB, NCKAP1, RABGAP1L* and *SMAD2*. If we relax the threshold of mTADA to PP>0.5, all 6 genes can also be found by mTADA.

| Number of Genes | No Anno Model | Anno Model | mTADA |
|---|---|---|---|
| M-DATA CHD | 17 | 19 | 11 |
| Known Human CHD[4] | 4 | 4 | 2 |
| Overlap CHD with mTADA | 11/11 | 11/11 | - |
| M-DATA Autism | 33 | 75 | 45 |
| Overlap Autism with Satterstrom et al.[54] | 25 | 40 | 27 |
| Overlap Autism with mTADA | 30/45 | 34/45 | - |

**Table 2. Comparison of M-DATA multi-trait models with mTADA**

## Discussion

In this paper, we have introduced M-DATA, a method to jointly analyze *de novo* mutations from multiple traits by integrating shared genetic information across traits. This approach can increase the effective sample size for all traits, especially for those with small sample size. M-DATA also provides a flexible framework to incorporate external functional annotations, either mutation-level or gene-level, which can further improve the statistical power. Through simulation study, we demonstrated that our multi-trait with annotation model could not only gain accurate estimates on the proportion of shared risk genes between two traits and the proportion of risk genes for a trait under various settings, but also gained statistical power compared to single-trait models. In addition, M-DATA adopts the Expectation-Maximization (EM) algorithm in estimation, which does not require prior parameter specification or pre-estimation. In practice, we found that the algorithm converges much faster than methods that use MCMC for estimation.

Despite the success, there are some limitations in the current M-DATA model. In our real data analysis, we used two different data sources for CHD and autism. Samples with both diseases in our multi-trait analysis may bring bias because of the violation of independence assumption in our multi-trait models. The autism DNM data in our analysis are from different studies, and different filtering criteria across studies may also bring bias and dilute our signals. In addition, we only considered two traits simultaneously. Though it is straightforward to extend our model to more than two traits, the number of groups (i.e., the dimension of latent variables $Z_i$) increases exponentially with the number of traits ($2^N$ for N traits) [23]. This might bring difficulty in estimation and have more computational cost. Model performance with more than two traits need further exploration.

In conclusion, M-DATA is a novel and powerful approach to performing gene-based association analysis for DNMs across multiple traits. Not only does M-DATA have better statistical power than single-trait methods, it also provides reasonable estimation of shared proportion of risk genes between two traits, which gives novel insights in the understanding of disease mechanism. We have successfully applied M-DATA to study CHD, which identified 25 significant genes for our multi-trait annotation model. Moreover, our method provides a general framework in extending single-trait method to multi-trait method which can also incorporate information from functional annotations. Recently, there are several advancements in the association analysis for rare variants, such as jointly analyzing DNMs and transmitted variants [34], analyzing DNMs from whole-genome sequencing (WGS) data [25], and incorporating pathway information [55]. Extension of these methods to multi-trait analysis is a potential future direction.

**Supplemental Data**

Supplemental Data includes Table S1-S3 and Figure S1- S4.

**Acknowledgement**

Supported in part by NIH grant R03HD100883-01A1.

**Web Resources**

M-DATA, https://github.com/JustinaXie/MDATA

denovo-db, https://denovo-db.gs.washington.edu/denovo-db/

ANNOVAR, https://annovar.openbioinformatics.org/en/latest/

OMIM, http://www.omim.org

## 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. 10.1038/nature12141.
2. Postma, A.V., Bezzina, C.R., and Christoffels, V.M. (2016). Genetics of congenital heart disease: the contribution of the noncoding regulatory genome. Journal of Human Genetics *61*, 13-19. 10.1038/jhg.2015.98.
3. 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. 10.1375/1832427053435328.
4. 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. 10.1038/ng.3970.
5. 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. 10.1038/nature10945.
6. 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. 10.1371/journal.pgen.1003671.
7. Coe, B.P., Stessman, H.A.F., Sulovari, A., Geisheker, M.R., Bakken, T.E., Lake, A.M., Dougherty, J.D., Lein, E.S., Hormozdiari, F., Bernier, R.A., and Eichler, E.E. (2019). Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity. Nature Genetics *51*, 106-116. 10.1038/s41588-018-0288-4.
8. Zhernakova, A., Stahl, E.A., Trynka, G., Raychaudhuri, S., Festen, E.A., Franke, L., Westra, H.J., Fehrmann, R.S., Kurreeman, F.A., Thomson, B., et al. (2011). Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci. PLoS Genet *7*, e1002004. 10.1371/journal.pgen.1002004.
9. 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. 10.1126/science.aac9396.
10. Willsey, A.J., Morris, M.T., Wang, S., Willsey, H.R., Sun, N., Teerikorpi, N., Baum, T.B., Cagney, G., Bender, K.J., Desai, T.A., et al. (2018). The Psychiatric Cell Map Initiative: A Convergent Systems Biological Approach to Illuminating Key Molecular Pathways in Neuropsychiatric Disorders. Cell *174*, 505-520. 10.1016/j.cell.2018.06.016.
11. Nguyen, H.T., Bryois, J., Kim, A., Dobbyn, A., Huckins, L.M., Munoz-Manchado, A.B., Ruderfer, D.M., Genovese, G., Fromer, M., Xu, X., et al. (2017). Integrated Bayesian

analysis of rare exonic variants to identify risk genes for schizophrenia and neurodevelopmental disorders. Genome Med *9*, 114. 10.1186/s13073-017-0497-y.
12. Li, J., Cai, T., Jiang, Y., Chen, H., He, X., Chen, C., Li, X., Shao, Q., Ran, X., Li, Z., et al. (2016). Genes with de novo mutations are shared by four neuropsychiatric disorders discovered from NPdenovo database. Mol Psychiatry *21*, 290-297. 10.1038/mp.2015.40.
13. Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C., and Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet *10*, e1004383. 10.1371/journal.pgen.1004383.
14. Solovieff, N., Cotsapas, C., Lee, P.H., Purcell, S.M., and Smoller, J.W. (2013). Pleiotropy in complex traits: challenges and strategies. Nat Rev Genet *14*, 483-495. 10.1038/nrg3461.
15. Chung, D., Yang, C., Li, C., Gelernter, J., and Zhao, H. (2014). GPA: a statistical approach to prioritizing GWAS results by integrating pleiotropy and annotation. PLoS Genet *10*, e1004787. 10.1371/journal.pgen.1004787.
16. Flutre, T., Wen, X., Pritchard, J., and Stephens, M. (2013). A statistical framework for joint eQTL analysis in multiple tissues. PLoS Genet *9*, e1003486. 10.1371/journal.pgen.1003486.
17. Sul, J.H., Han, B., Ye, C., Choi, T., and Eskin, E. (2013). Effectively identifying eQTLs from multiple tissues by combining mixed model and meta-analytic approaches. PLoS Genet *9*, e1003491. 10.1371/journal.pgen.1003491.
18. Duong, D., Gai, L., Snir, S., Kang, E.Y., Han, B., Sul, J.H., and Eskin, E. (2017). Applying meta-analysis to genotype-tissue expression data from multiple tissues to identify eQTLs and increase the number of eGenes. Bioinformatics *33*, i67-i74. 10.1093/bioinformatics/btx227.
19. Li, G., Jima, D., Wright, F.A., and Nobel, A.B. (2018). HT-eQTL: integrative expression quantitative trait loci analysis in a large number of human tissues. BMC Bioinformatics *19*, 95. 10.1186/s12859-018-2088-3.
20. Li, C., Yang, C., Gelernter, J., and Zhao, H. (2014). Improving genetic risk prediction by leveraging pleiotropy. Hum Genet *133*, 639-650. 10.1007/s00439-013-1401-5.
21. Maier, R., Moser, G., Chen, G.B., Ripke, S., Coryell, W., Potash, J.B., Scheftner, W.A., Shi, J., Weissman, M.M., Hultman, C.M., et al. (2015). Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. Am J Hum Genet *96*, 283-294. 10.1016/j.ajhg.2014.12.006.
22. Hu, Y., Lu, Q., Liu, W., Zhang, Y., Li, M., and Zhao, H. (2017). Joint modeling of genetically correlated diseases and functional annotations increases accuracy of polygenic risk prediction. PLoS Genet *13*, e1006836. 10.1371/journal.pgen.1006836.
23. Nguyen, T.-H., Dobbyn, A., Brown, R.C., Riley, B.P., Buxbaum, J.D., Pinto, D., Purcell, S.M., Sullivan, P.F., He, X., and Stahl, E.A. (2020). mTADA is a framework for identifying risk genes from de novo mutations in multiple traits. Nature Communications *11*, 2929. 10.1038/s41467-020-16487-z.
24. Lu, Q., Yao, X., Hu, Y., and Zhao, H. (2016). GenoWAP: GWAS signal prioritization through integrated analysis of genomic functional annotation. Bioinformatics *32*, 542-548. 10.1093/bioinformatics/btv610.

25. Liu, Y., Liang, Y., Cicek, A.E., Li, Z., Li, J., Muhle, R.A., Krenzer, M., Mei, Y., Wang, Y., Knoblauch, N., et al. (2018). A Statistical Framework for Mapping Risk Genes from De Novo Mutations in Whole-Genome-Sequencing Studies. Am J Hum Genet *102*, 1031-1047. 10.1016/j.ajhg.2018.03.023.
26. Butkiewicz, M., Blue, E.E., Leung, Y.Y., Jian, X., Marcora, E., Renton, A.E., Kuzma, A., Wang, L.-S., Koboldt, D.C., Haines, J.L., and Bush, W.S. (2018). Functional annotation of genomic variants in studies of late-onset Alzheimer's disease. Bioinformatics *34*, 2724-2731. 10.1093/bioinformatics/bty177.
27. 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. 10.1038/ng.3050.
28. Yang, H., and Wang, K. (2015). Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. Nat Protoc *10*, 1556-1566. 10.1038/nprot.2015.105.
29. Kim, S., Jhong, J.-H., Lee, J., and Koo, J.-Y. (2017). Meta-analytic support vector machine for integrating multiple omics data. BioData Mining *10*, 2. 10.1186/s13040-017-0126-8.
30. Moon, T.K. (1996). The expectation-maximization algorithm. IEEE Signal Processing Magazine *13*, 47-60. 10.1109/79.543975.
31. Jiang, W., and Yu, W. (2016). 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 *33*, 500-507. 10.1093/bioinformatics/btw690.
32. 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. 10.1038/s41586-020-2308-7.
33. Turner, T.N., Yi, Q., Krumm, N., Huddleston, J., Hoekzema, K., HA, F.S., Doebley, A.L., Bernier, R.A., Nickerson, D.A., and Eichler, E.E. (2017). denovo-db: a compendium of human de novo variants. Nucleic Acids Res *45*, D804-d811. 10.1093/nar/gkw865.
34. Li, M. (2020). Gene-based Association Analysis for Genome-wide Association and Whole-exome Sequencing Studies. Doctor of Philosophy (Yale University).
35. Adzhubei, I., Jordan, D.M., and Sunyaev, S.R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet *Chapter 7*, Unit7.20. 10.1002/0471142905.hg0720s76.
36. Samocha, K.E., Kosmicki, J.A., Karczewski, K.J., O'Donnell-Luria, A.H., Pierce-Hoffman, E., MacArthur, D.G., Neale, B.M., and Daly, M.J. (2017). Regional missense constraint improves variant deleteriousness prediction. BioRxiv, 148353.
37. Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet *46*, 310-315. 10.1038/ng.2892.
38. Ioannidis, N.M., Rothstein, J.H., Pejaver, V., Middha, S., McDonnell, S.K., Baheti, S., Musolf, A., Li, Q., Holzinger, E., Karyadi, D., et al. (2016). REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet *99*, 877-885. 10.1016/j.ajhg.2016.08.016.

39. Jian, X., Boerwinkle, E., and Liu, X. (2014). In silico prediction of splice-altering single nucleotide variants in the human genome. Nucleic Acids Research *42*, 13534-13544. 10.1093/nar/gku1206.
40. Zaidi, S., and Brueckner, M. (2017). Genetics and Genomics of Congenital Heart Disease. Circ Res *120*, 923-940. 10.1161/circresaha.116.309140.
41. Pierpont, M.E., Brueckner, M., Chung, W.K., Garg, V., Lacro, R.V., McGuire, A.L., Mital, S., Priest, J.R., Pu, W.T., Roberts, A., et al. (2018). Genetic Basis for Congenital Heart Disease: Revisited: A Scientific Statement From the American Heart Association. Circulation *138*, e653-e711. 10.1161/cir.0000000000000606.
42. McKean, D.M., Homsy, J., Wakimoto, H., Patel, N., Gorham, J., DePalma, S.R., Ware, J.S., Zaidi, S., Ma, W., Patel, N., et al. (2016). Loss of RNA expression and allele-specific expression associated with congenital heart disease. Nat Commun *7*, 12824. 10.1038/ncomms12824.
43. 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. 10.1038/srep30896.
44. Granadillo, J.L., Chung, W.K., Hecht, L., Corsten-Janssen, N., Wegner, D., Nij Bijvank, S.W.A., Toler, T.L., Pineda-Alvarez, D.E., Douglas, G., Murphy, J.J., et al. (2018). Variable cardiovascular phenotypes associated with SMAD2 pathogenic variants. Hum Mutat *39*, 1875-1884. 10.1002/humu.23627.
45. Sifrim, A., Hitz, M.P., Wilsdon, A., Breckpot, J., Turki, S.H., Thienpont, B., McRae, J., Fitzgerald, T.W., Singh, T., Swaminathan, G.J., et al. (2016). Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing. Nat Genet *48*, 1060-1065. 10.1038/ng.3627.
46. 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. 10.1016/j.ajhg.2016.02.004.
47. Matsumura, K., Seiriki, K., Okada, S., Nagase, M., Ayabe, S., Yamada, I., Furuse, T., Shibuya, H., Yasuda, Y., Yamamori, H., et al. (2020). Pathogenic POGZ mutation causes impaired cortical development and reversible autism-like phenotypes. Nat Commun *11*, 859. 10.1038/s41467-020-14697-z.
48. White, J., Beck, C.R., Harel, T., Posey, J.E., Jhangiani, S.N., Tang, S., Farwell, K.D., Powis, Z., Mendelsohn, N.J., Baker, J.A., et al. (2016). POGZ truncating alleles cause syndromic intellectual disability. Genome Med *8*, 3. 10.1186/s13073-015-0253-0.
49. Kidder, B.L., Hu, G., and Zhao, K. (2014). KDM5B focuses H3K4 methylation near promoters and enhancers during embryonic stem cell self-renewal and differentiation. Genome Biol *15*, R32. 10.1186/gb-2014-15-2-r32.
50. Kurup, J.T., Campeanu, I.J., and Kidder, B.L. (2019). Contribution of H3K4 demethylase KDM5B to nucleosome organization in embryonic stem cells revealed by micrococcal nuclease sequencing. Epigenetics Chromatin *12*, 20. 10.1186/s13072-019-0266-9.
51. Albert, M., Schmitz, S.U., Kooistra, S.M., Malatesta, M., Morales Torres, C., Rekling, J.C., Johansen, J.V., Abarrategui, I., and Helin, K. (2013). The histone demethylase Jarid1b

ensures faithful mouse development by protecting developmental genes from aberrant H3K4me3. PLoS Genet *9*, e1003461. 10.1371/journal.pgen.1003461.

52. Cheng, H., Dharmadhikari, A.V., Varland, S., Ma, N., Domingo, D., Kleyner, R., Rope, A.F., Yoon, M., Stray-Pedersen, A., Posey, J.E., et al. (2018). Truncating Variants in NAA15 Are Associated with Variable Levels of Intellectual Disability, Autism Spectrum Disorder, and Congenital Anomalies. Am J Hum Genet *102*, 985-994. 10.1016/j.ajhg.2018.03.004.
53. 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 (New York, N.Y.) *350*, 1262-1266. 10.1126/science.aac9396.
54. Satterstrom, F.K., Kosmicki, J.A., Wang, J., Breen, M.S., De Rubeis, S., An, J.-Y., Peng, M., Collins, R., Grove, J., Klei, L., et al. (2020). Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. Cell *180*, 568-584.e523. https://doi.org/10.1016/j.cell.2019.12.036.
55. Nguyen, T.H., He, X., Brown, R.C., Webb, B.T., Kendler, K.S., Vladimirov, V.I., Riley, B.P., and Bacanu, S.A. (2021). DECO: a framework for jointly analyzing de novo and rare case/control variants, and biological pathways. Brief Bioinform. 10.1093/bib/bbab067.