PLOS GENETICS

RESEARCH ARTICLE

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**[1,2,3]*

1 Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut, United States of America, **2** Department of Genetics, Yale School of Medicine, New Haven, Connecticut, United States of America, **3** Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut, United States of America

☯ These authors contributed equally to this work.
* hongyu.zhao@yale.edu

Citation: Xie Y, Li M, Dong W, Jiang W, Zhao H (2021) M-DATA: A statistical approach to jointly analyzing *de novo* mutations for multiple traits. PLoS Genet 17(11): e1009849. https://doi.org/10.1371/journal.pgen.1009849

Editor: Mingyao Li, University of Pennsylvania, UNITED STATES

Received: June 16, 2021

Accepted: September 29, 2021

Published: November 4, 2021

Data Availability Statement: 2,645 CHD data can be downloaded from the supplement of PMID28991257 (https://static-content.springer.com/esm/art%3A10.1038%2Fng.3970/MediaObjects/41588_2017_BFng3970_MOESM3_ESM.xlsx), and autism data can be acquired from denovo-db (https://denovo-db.gs.washington.edu/denovo-db/Download.jsp). The use of autism data from Simons Simplex Collection (SSC) and Simons VIP data sets is limited to projects related to advancing the field of autism and related developmental disorder research. Questions on SSC/VIP consents should be directed to

Abstract

Recent studies have demonstrated that multiple early-onset diseases have shared risk genes, based on findings from *de novo* mutations (DNMs). 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 for CHD from joint analysis, including 12 novel genes, which is substantially more than single-trait analysis, leading to novel insights into CHD disease etiology.

Author summary

With the development of new generation sequencing technology, germline mutations such as *de novo* mutations (DNMs) with deleterious effects can be identified to aid in discovering the genetic causes for early on-set diseases such as congenital heart disease (CHD). However, the statistical power is still limited by the small sample size of DNM studies due to the high cost of recruiting and sequencing samples, and the low occurrence of DNMs given its rarity. Compared to DNM analyses for other diseases, it is even more challenging for CHD given its genetic heterogeneity. Recent research has suggested shared disease mechanisms between early-onset neurodevelopmental diseases and CHD based on findings from DNMs. Currently, there are few methods that can jointly analyze DNM data on multiple traits. Therefore, we develop a framework to identify risk genes for multiple traits simultaneously for DNM data. The new method is applied to CHD and autism

collections@sfari.org. Summary statistics of real data application can be downloaded from https://github.com/JustinaXie/MDATA/tree/main/data.

Funding: This work was supported in part by NIH grant R03HD100883-01A1 (Y.X. and H.Z.) and R01GM134005-01A1 (W.J. and H.Z.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

as a case study to demonstrate its improved power in identifying risk genes compared with single-trait analyses. Our results lead to new insights on the disease etiology of CHD, and the shared etiological mechanisms between CHD and autism.

Introduction

The development of sequencing technologies such as Whole Exome Sequencing (WES) has led to the identifications of the genetic causes of many diseases in the past decades. Studies based on WES have successfully identified novel causal genes in both Mendelian disorders and complex disorders [1,2]. Because WES may generate a large number of genetic variants in studied individuals, a strategy to narrow down the pool of candidate variants is to scan for *de novo* mutations (DNMs) by comparing the WES data between the healthy parents and their affected offspring (proband). As DNMs with deleterious effects have not been through natural selection, they have proved very informative in identifying risk genes for early on-set diseases such as congenital heart disease (CHD) [3–7]. For instance, Homsy et al. identified an excess of protein-damaging DNMs in 1,213 exome-sequenced CHD parent-offspring trios, especially in genes highly expressed in the developing heart and brain [4]. In a recent study, Jin et al. found that DNMs accounted for 8% of CHD cases and identified striking overlap between genes with damaging DNMs in probands with CHD and autism [5]. These studies showed that DNM analyses can play an important role in exploring the genetic etiology of CHD. However, the statistical power for identifying risk genes is still hampered by the limited sample size of DNM studies 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 [8, 9]. 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 [4,10]. 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 ***De 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 and has improved power of identifying risk genes over single-trait analyses. We applied M-DATA 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.

Methods

Ethics statement

This study is approved by Yale Human Research Protection Program Institutional Review Boards (IRB protocol ID 2000028735).

Probabilistic model

First, 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, μ_i is the mutability of gene i estimated using the framework in Samocha et al. [27] and γ_i is the relative risk of the DNMs in the risk gene and is assumed to be larger than 1. The derivation of the parameter of the Poisson distribution is the same as that in TADA [28,29]. We define this model as the single-trait model without annotation in our main text.

To leverage information from functional annotations, we use an exponential link between γ_i and X_i,

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

where X_i^T is the transpose of the functional annotation vector of gene i, and β is the effect size vector of the functional annotations. Under the assumption that risk genes have higher burden than non-risk genes, we expect the estimated value of γ_i to be larger than 1.

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:

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

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

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

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

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

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

where π 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 the latent variables $Z_{10} + Z_{11}$ and $Z_{01} + Z_{11}$ are independent, $\pi_{11} = (\pi_{10} + \pi_{11})(\pi_{01} + \pi_{11})$. The difference between π_{11} and $(\pi_{10} + \pi_{11})(\pi_{01} + \pi_{11})$ reflects the magnitude of global pleiotropy between the two traits. μ_i is the same as our one-trait model. N_1, γ_{i1} and X_{i1} are the case cohort size, relative risk and annotation vector of gene i for the first trait. N_2, γ_{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

$$L(\Theta) = \prod_{i=1}^{M}\sum_{l\in\{00,10,01,11\}}[\pi_l \Pr(Y_{i1}, Y_{i2}|Z_{il} = 1;\Theta)]^{Z_{il}}$$

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

$$l(\Theta) = \sum_{i=1}^{M}\log\sum_{l\in\{00,10,01,11\}}[\pi_l \Pr(Y_{i1}, Y_{i2}|Z_{il} = 1;\Theta)]^{Z_{il}}.$$

Estimation

Parameters of our models can be estimated using the Expectation-Maximization (EM) algorithm [30]. It is very computationally efficient for our model without annotation 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

$$l(\Theta) \geq Q(\Theta) = \sum_{i=1}^{M}\sum_{l\in\{00,10,01,11\}}[Z_{il}[\log(\pi_l) + \log(\Pr(Y_{i1}, Y_{i2}|Z_{il} = 1;\Theta))]].$$

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,

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

In the M-step, we update the parameters in Θ based on the estimation of Z_{il} in the E-step by maximizing $Q(\Theta)$. For π, 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. Taking the first order derivative of $Q(\Theta)$ with respect to β_1 as 0, we have

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

If we do not add any functional annotations to our model (X_{i1} degenerates to 1 and β_1 degenerates to a scalar), there exists an analytical solution for β_1.

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

However, there is no explicit solution for β_1, so we adopt the Newton-Raphson method for estimation after adding functional annotations into our model. 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 β_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 and feature selection

As we have discussed, there are multiple sources of functional annotations for DNMs. For gene-level annotations, we can directly plug into our gene-based model. For variant-level annotations, 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 γ_{it} and mutability μ_{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

$$\begin{aligned}\frac{P(Y_i|H_1)}{P(Y_i|H_0)} &= \frac{\prod_t \exp(\beta_0 + \beta_1 X_{it} I\{Y_{it} = 1\})\exp(-2N\mu_{it}\exp(\beta_0 + \beta_1 X_{it}))}{\prod_t \exp(-2N\mu_{it})} \\ &= \exp\left(\sum\nolimits_t (\beta_0 + \beta_1 X_{it} I\{Y_{it} = 1\})\right)\exp\left(\sum\nolimits_t - 2N\mu_{it}[\exp(\beta_0 + \beta_1 X_{it}) - 1]\right)\end{aligned}$$

Assuming the variant-level effect size β_1 is small, we can apply Taylor expansion to the second term of the above equation,

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

If we center the collapsed variant-level annotations, we can apply $\Sigma t\ X_{it} = 0$ to the above equation and further simplify it as

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

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

We used variant-level annotations from ANNOVAR [31] in our analysis. We define loss-of-function (LoF) as frameshift insertion/deletion, splice site alteration, stopgain and stoploss predicated by ANNOVAR, and define deleterious missense variants (Dmis) predicted by MetaSVM [32]. Specifically, we included four categories of features including variant-level deleteriousness (PolyPhen (D), PolyPhen(P) [33], MPC [34], CADD [35], REVEL [36], and LoF), variant-level allele frequencies (gnomAD_exome and gnomAD_genome [37]), variant-level splicing scores (dbscSNV_ADA_score, dbscSNV_RF_score [38] and dpsi_zscore [39]) and gene conservation scores (pLI and mis_z) downloaded from gnomAD v2.1.1 [37] in real data analysis. To construct gene-level annotation scores, variant-level annotations were collapsed by summing up values calculated from the mutation information for each gene. All continuous gene-level features were normalized before model fitting.

Before performing multi-trait analysis, features were selected separately for each trait by single-trait analysis. For each trait, all gene-level features were evaluated by Pearson's correlation. If the Pearson's correlation between two annotations was larger than 0.7, only one annotation was kept. After model fitting, we kept annotations with the absolute values of effect sizes larger than 0.01 and refit the model with the selected annotations. For multi-trait analyses, we constructed the annotation matrices using the features selected from each trait (see more details in S1 Text).

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) [40]. For joint analysis of two traits, the Jlfdr of whether gene i is associated with the first trait is

$$\begin{aligned}\mathrm{Jlfdr}_1(Y_{i1}, Y_{i2}) &= Pr(Z_{i00} + Z_{i01} = 1|Y_{i1}, Y_{i2}) \\ &= \frac{\pi_{00} Pr(Y_{i1}, Y_{i2}|Z_{i00} = 1;\Theta) + \pi_{01} Pr(Y_{i1}, Y_{i2}|Z_{i01} = 1;\Theta)}{\sum_{l' \in \{00,01,10,11\}}[\pi_{l'} Pr(Y_{i1}, Y_{i2}|Z_{il'} = 1;\Theta)]} \\ &= \frac{\pi_{00}\mathrm{Poisson}(Y_{i1}, 2N_1\mu_i)\mathrm{Poisson}(Y_{i2}, 2N_2\mu_i) + \pi_{01}\mathrm{Poisson}(Y_{i1}, 2N_1\mu_i)\mathrm{Poisson}(Y_{i2}, 2N_2\mu_i\gamma_{i2})}{\sum_{l' \in \{00,01,10,11\}}[\pi_{l'} Pr(Y_{i1}, Y_{i2}|Z_{il'} = 1;\Theta)]},\end{aligned}$$

where $\gamma_{i1} = \exp(X_{i1}^T\beta_1)$ and $\gamma_{i2} = \exp\left(X_{i2}^T\beta_2\right)$. When there is no annotation, both β_1 and β_2 degrade from vectors to single intercept values. Then γ_{i1} and γ_{i2} share the same values $\exp(\beta_1)$ and $\exp(\beta_2)$ across all genes. Same formula can be used to compute the Jlfdr of each gene. The definition of the Jlfdr is the posterior probability of a null hypothesis being true, given the

observed DNM count vector (Y_1, Y_2). If we consider the first trait, the corresponding null hypothesis is the gene i associates with neither trait or only associates with the second trait, i.e., $Z_{i00} + Z_{i01} = 1$. And the corresponding Jlfdr is $\text{jlfdr}_1(Y_{i1}, Y_{i2}) = \Pr(Z_{i00} + Z_{i01} = 1 | Y_{i1}, Y_{i2})$. In comparison, the p-value is defined as the probability of observing more extreme results given the null hypothesis being true, i.e., p-value = Pr(More extreme than $(Y_{i1}, Y_{i2}) | Z_{i00} + Z_{i01} = 1$). To compute it, we need to firstly define a partial order for comparing two-dimensional vector (Y_1, Y_2), with which the genes associate with the first trait can stand out. One way to define the partial order is to summarize the vector into a one-dimensional test statistic. Since this is not our focus, we will not discuss how to derive a new test statistic in the article. Although the Jlfdr_1 already informs the probability of whether the gene is associated with the first trait, we should not directly use it as the p-value to infer the association status due to their different definitions and properties. In the simulation studies and real data application, we used Jlfdr as our inference method for risk gene identification.

The following relationship between Jlfdr and false discovery rates (Fdr) was shown in Jiang and Yu [40],

$$\text{Fdr}_1(\mathcal{R}) = E(\text{Jlfdr}_1(Y_1, Y_2) | (Y_1, Y_2) \in \mathcal{R}) \approx \frac{1}{|\{(Y_{i1}, Y_{i2}) \in \mathcal{R}\}|} \sum_{(Y_{i1}, Y_{i2}) \in \mathcal{R}} \text{Jlfdr}_1(Y_{i1}, Y_{i2}),$$

where the rejection region is the set of two-dimensional vector (Y_1, Y_2) such that the null hypothesis can be rejected based on a specific rejection criterion. For example, we can specify a rejection criterion to select genes with large values of the weighted average DNM counts: $0.9Y_1 + 0.1Y_2 \geq 5$, then the corresponding rejection region is the upper right region above the line of $0.9Y_1 + 0.1Y_2 = 5$. Here we omit the gene indicator i since the rejection region is defined on DNM count pairs of two traits regardless of the exact gene labels. Jiang and Yu [40] showed that the most powerful rejection region for a given Fdr level q is $\{\text{jldr}_1(Y_1, Y_2) \leq t(q)\}$. To determine the threshold $t(q)$, we sort the calculated jldr_1 value of each gene in an ascending order first. Denote the a-th jlfdr_1 value as Jlfdr_1^a. We can approximate the Fdr of the region $\mathcal{R}_a = \{(Y_1, Y_2) | \text{Jlfdr}_1(Y_{i1}, Y_{i2}) \leq \text{Jlfdr}_1^a\}$ as

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

Denote $c = \max\{a | \text{Fdr}(\mathcal{R}_a) \leq q\}$, and then the threshold $t(q)$ for Jlfdr_1 is Jlfdr_1^c. For testing association with the first trait, we reject all genes with $\text{Jlfdr}_1(\text{Y}_{i1}, \text{Y}_{i2}) \leq t(q)$. For both simulation and real data analyses, the global Fdr is controlled at $q = 0.05$. The global Fdr is abbreviated as FDR in the following text.

Implementation of mTADA

We used extTADA [11] to estimate the hyperpriors input for mTADA. For simulation and real data application, we applied 2 MCMC chains and 10,000 iterations as recommended by the authors [23]. We applied posterior probability>0.8 as the threshold for risk gene inference. We benchmarked the computational time of mTADA and M-DATA on Intel Xon Gold 6240 processors (2.6GHZ).

Misspecified model

We tested if M-DATA have proper power when functional annotations affect the latent variables Z_{il}, $l \in \{00, 01, 10, 11\}$ rather than the relative risk parameters γ_{i1} and γ_{i2}. Further, we assumed that the latent variable Z_{i10} is associated with the functional annotation vector X_{i1}, which is the functional annotation vector for gene i of the first trait, Z_{i01} is associated with X_{i2},

which is the functional annotation vector for gene i of the second trait, and Z_{i11} is associated with both X_{i1} and X_{i2} through the following forms:

$$P(Z_{i00}) = \frac{1}{1+\exp(X_{i1}^T\beta_1)+\exp(X_{i2}^T\beta_2)+\exp(X_{i1}^T\beta_1+X_{i2}^T\beta_2)}$$

$$P(Z_{i10}) = \frac{\exp(X_{i1}^T\beta_1)}{1+\exp(X_{i1}^T\beta_1)+\exp(X_{i2}^T\beta_2)+\exp(X_{i1}^T\beta_1+X_{i2}^T\beta_2)}$$

$$P(Z_{i01}) = \frac{\exp(X_{i2}^T\beta_2)}{1+\exp(X_{i1}^T\beta_1)+\exp(X_{i2}^T\beta_2)+\exp(X_{i1}^T\beta_1+X_{i2}^T\beta_2)}$$

$$P(Z_{i11}) = \frac{\exp(X_{i1}^T\beta_1+X_{i2}^T\beta_2)}{1+\exp(X_{i1}^T\beta_1)+\exp(X_{i2}^T\beta_2)+\exp(X_{i1}^T\beta_1+X_{i2}^T\beta_2)}$$

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

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

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

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

where π is the corresponding risk proportion of genes belonging to each class, with $\sum_{l\in\{00,\ 10,\ 01,\ 11\}}\pi_l = 1$. Here, μ_i is the mutability of gene i. N_1, γ_{i1} and X_{i1} are the case cohort size, relative risk and annotation vector of gene i for the first trait. Similarly, N_2, γ_{i2} and X_{i2} are defined for the second trait.

Verification and comparison

Estimation evaluation

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 v2.1.1 [37]. We set the size of the case cohort at 2000, 5000 and 10000, corresponding to a small, medium and large WES study. We assumed the proportion of risk genes to be 0.1 for each trait (i.e., $\pi_{10} + \pi_{11} = \pi_{01} + \pi_{11} = 0.1$), and varied the shared risk proportion π_{11} at 0.01, 0.03, 0.05, 0.07 and 0.09. When $\pi_{11} = 0.01$, it corresponds to the independence of latent variables $Z_{10} + Z_{11}$ and $Z_{01} + Z_{11}$ 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 varied the effect sizes of annotations $(\beta_{j0}, \beta_{j1}, \beta_{j2})$, $j = 1, 2$ from (3, 0.1, 0.1) (3, 0.1, 0) and (3, 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 π_{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

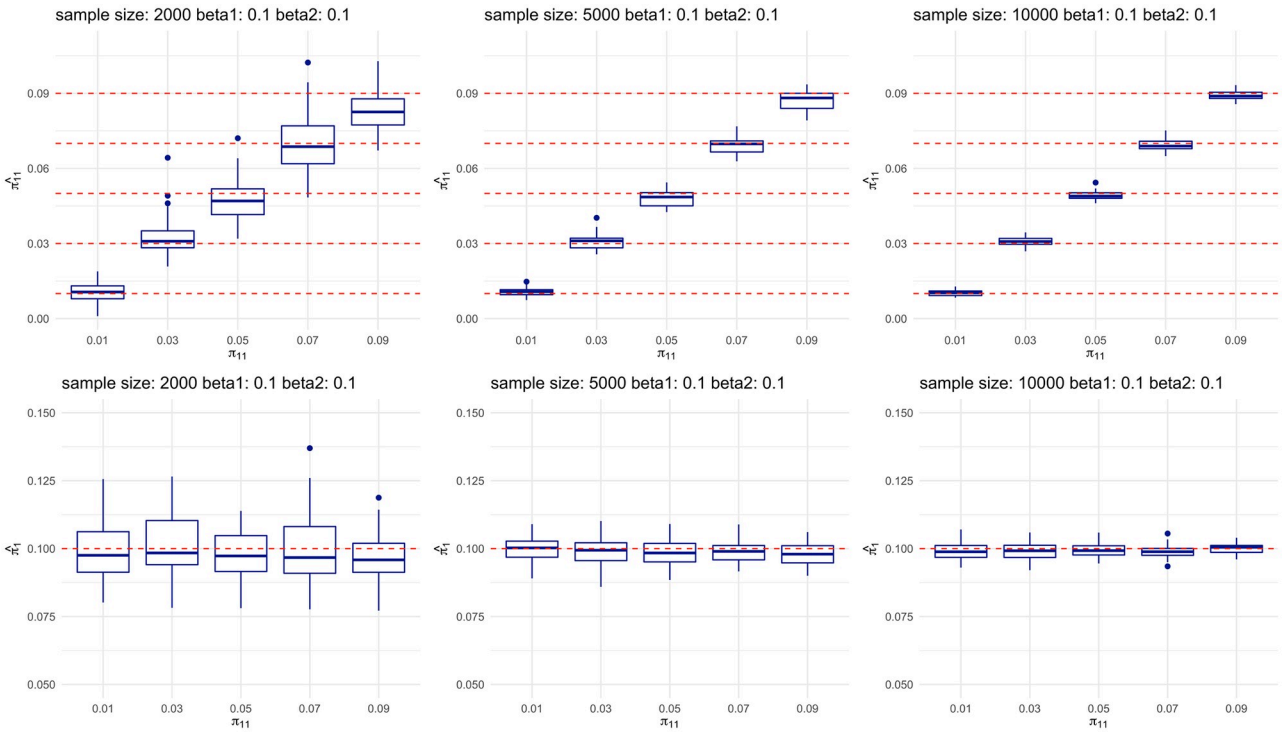

Fig 1. Multi-trait analysis can accurately estimate the proportion of shared risk genes and single-trait risk genes. Top panels show the estimation of shared risk proportion, and bottom panels show the estimation of a single trait. For each panel, each plot from left to right represents study sample 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. Each scenario is replicated for 50 times in our simulations. True values are shown in red dashed lines.

https://doi.org/10.1371/journal.pgen.1009849.g001

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 Fig 1. The rest of scenarios are shown in Fig A in S1 Text.

Power evaluation

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 varied the effect size of annotations (β_{j0}, β_{j1}, β_{j2}, β_{j3}), $j = 1, 2$ from (3, 0.1, 0.1, 0.1) (3, 0.3, 0.3, 0.3) and (3, 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 Jlfdr (see Methods) can control FDR (Fig B in S1 Text) under theses settings and then evaluated power (Fig 2), type I error (Fig C in S1 Text), and AUC (Fig D in S1 Text) for our single-trait models and multi-trait models. There are in total 45 simulation settings. Under each setting, the data were simulated based on our multi-trait model with annotations (Methods).

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 stronger, the power performance of models with annotations are better than models without annotations (Figs E and F in S1 Text). With the increase of shared risk proportion, the power performance of multi-trait models become better than single-trait models.

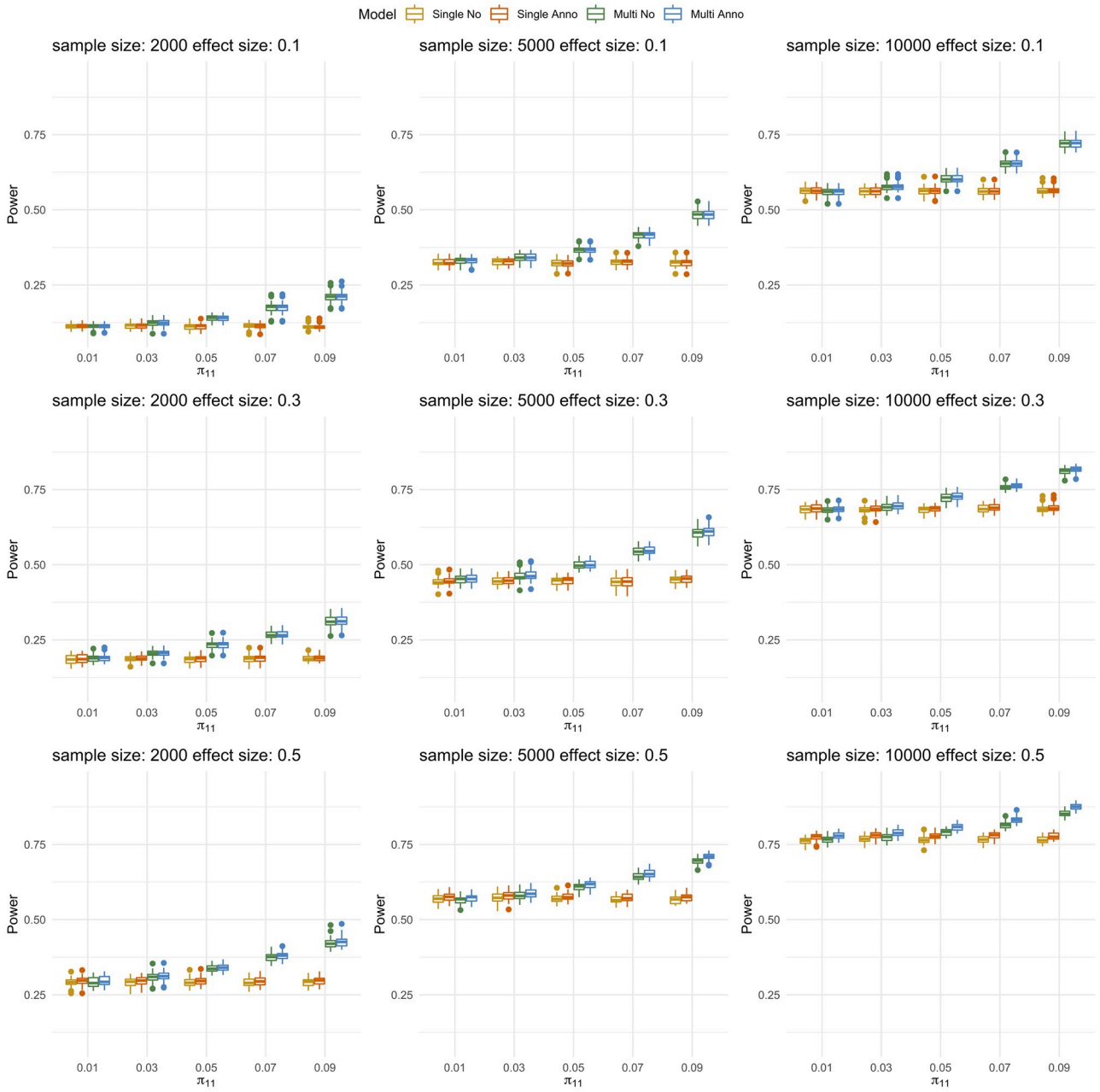

Fig 2. Power performance under different strengths of annotations. The panels from top to bottom show the power performance under weak, moderate and strong annotations, respectively. For each panel, each plot from left to right represents study sample size of 2000, 5000, and 10000, respectively. Within 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. Each scenario is replicated for 50 times in our simulations.

https://doi.org/10.1371/journal.pgen.1009849.g002

Comparison with mTADA

Under the same settings in the previous section, we compared the power performance of mTADA and M-DATA. In the simulation, we observed that both methods could control FDR, while mTADA was more conservative than M-DATA for FDR control (Fig G in S1 Text). M-DATA has higher power than mTADA when the effect size of annotations is larger (Fig 3). The result is consistent with our observation in the real data (Application). In the time

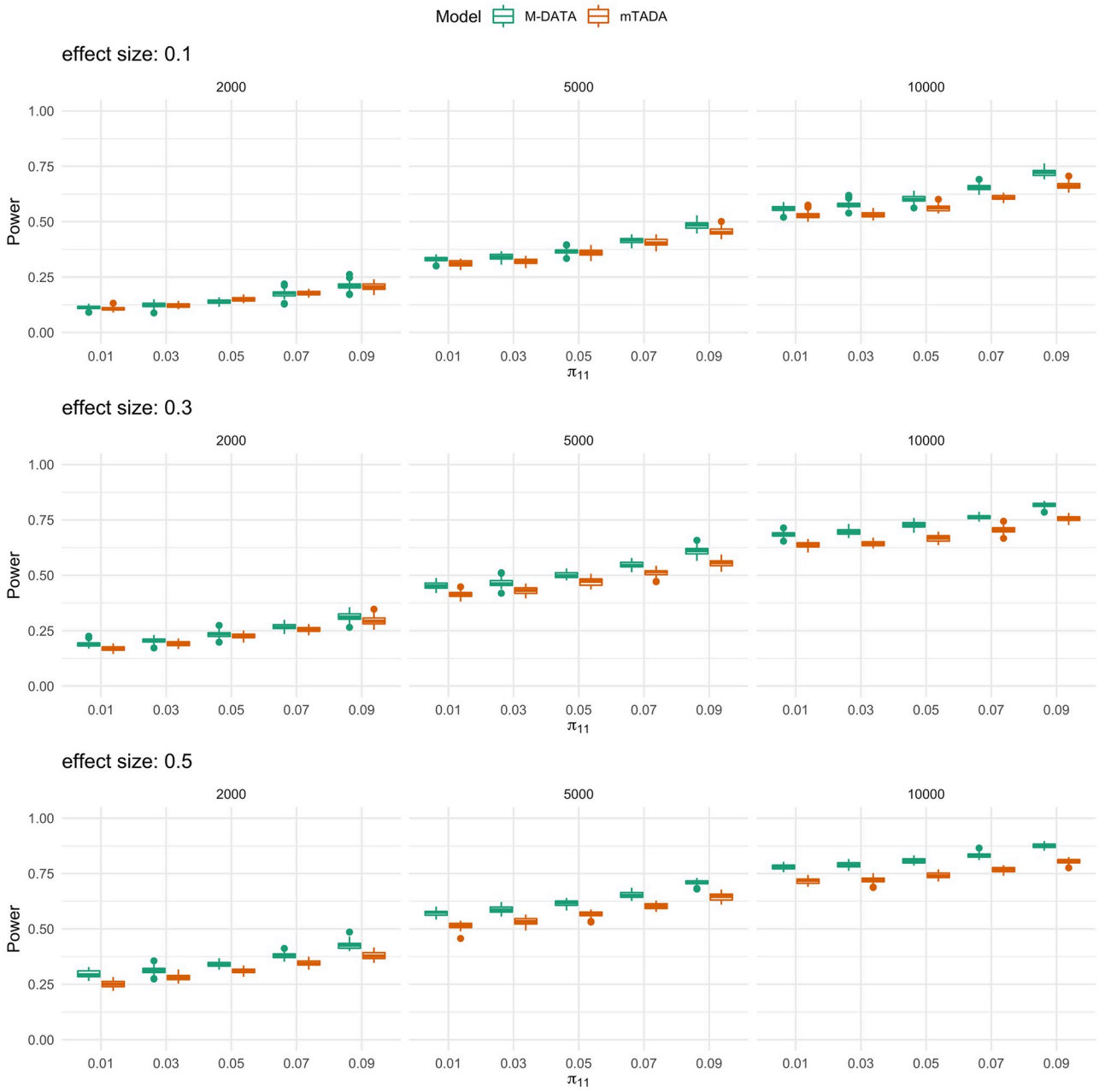

Fig 3. Comparisons of M-DATA and mTADA under different strengths of annotations. The panels from top to bottom show the power performance under weak, moderate and strong annotations, respectively. For each panel, each plot from left to right represents study sample size of 2000, 5000, and 10000, respectively. 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. Each scenario is replicated for 50 times in our simulations.

https://doi.org/10.1371/journal.pgen.1009849.g003

comparison, we observed that our method converged faster than the MCMC method adopted by mTADA (Table D in S1 Text).

Robustness to model misspecification

We also evaluated the power performance of M-DATA under misspecified models (Methods), where we simulated two Bernoulli annotations that affect the latent variables Z_{il}, $l \in \{00, 01,$

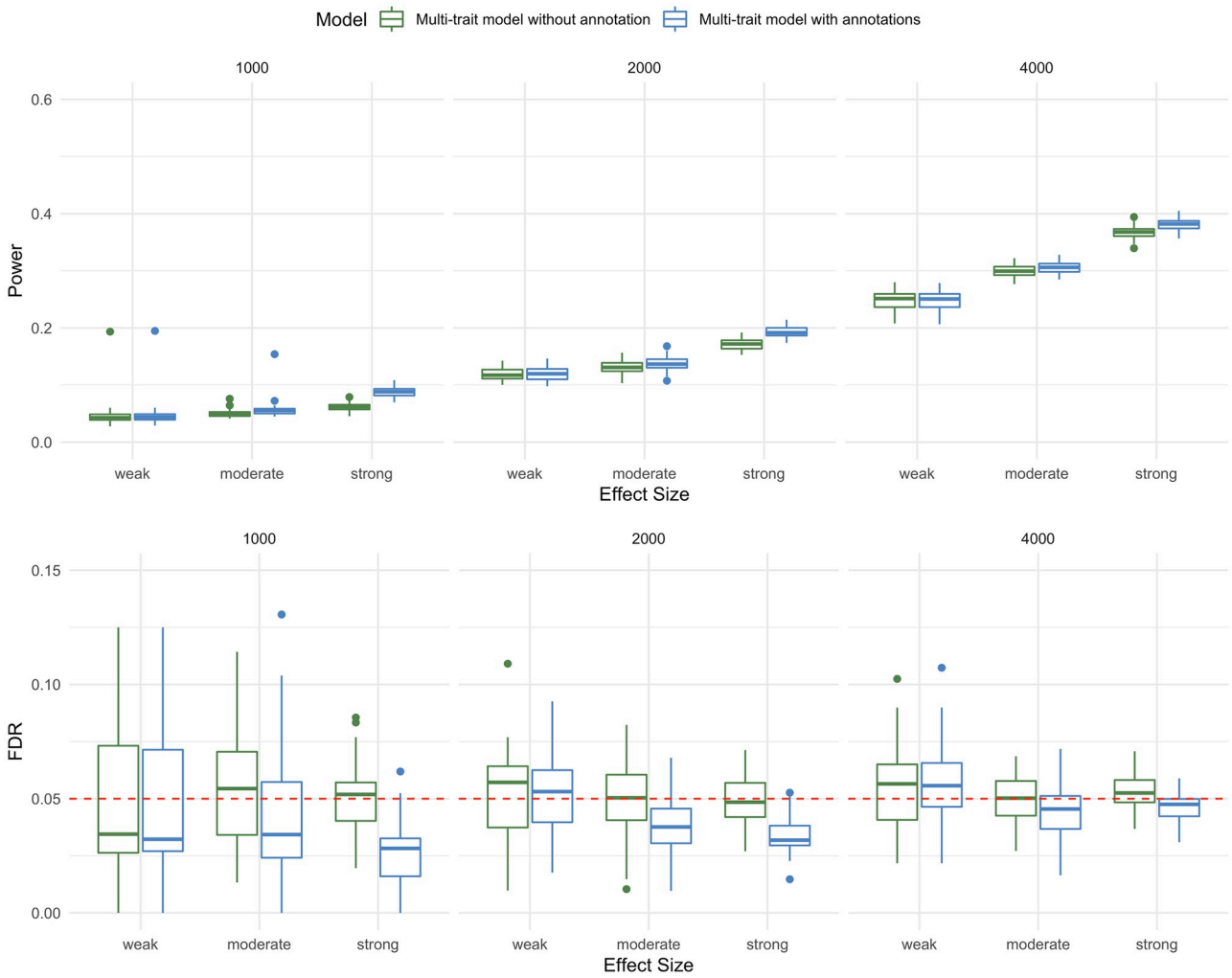

Fig 4. Power and FDR of M-DATA under model misspecification. The top panel and bottom panel show the power and FDR under weak, moderate and strong annotations on the latent variables $\boldsymbol{Z_{il}}$, $\boldsymbol{l} \in \{\mathbf{00, 01, 10, 11}\}$ respectively. For each panel, each plot from left to right represents study sample size of 1000, 2000, and 4000, respectively. Each scenario is replicated for 50 times in our simulations.

https://doi.org/10.1371/journal.pgen.1009849.g004

10, 11}, and set the parameter of the Bernoulli distributions to 0.5 for both traits. We varied the effect sizes of annotations on the latent variables (β_{j0}, β_{j1}, β_{j2}), j = 1, 2 at (-3, 0.5, 0.5), (-3, 1, 1) and (-3, 1.5, 1.5), which corresponds to the case when the effect of annotations is weak, moderate, and strong, respectively. The relative risk parameters γ_{i1} and γ_{i2} were set at 25. We simulated DNM counts under this misspecified model and evaluated the performance of M-DATA multi-trait models for different sizes of DNM cohort (1000, 2000, and 4000). We observed that M-DATA can control FDR under all settings and the multi-trait model with annotations had better power than the multi-trait model without annotation with the increase of the effect size of annotations (Fig 4).

Application

We applied M-DATA to real DNM data from 2,645 CHD probands reported in Jin et al. [5] and 5,623 autism probands acquired from denovo-db [41]. We only considered damaging mutations (LoF and Dmis) in our analysis as the number of non-deleterious mutations is not

Table 1. Results for M-DATA single-trait and multi-trait models.

Model	FDR<0.05	FDR<0.01
Single no Anno: CHD/Autism	15/28	8/17
Single with Anno: CHD/Autism	19/35	10/22
Multi no Anno: CHD/Autism	18/28	11/19
Multi with Anno: CHD/Autism	23/37	14/23

https://doi.org/10.1371/journal.pgen.1009849.t001

expected to provide information to differentiate cases from controls biologically [42]. Details of functional annotation and feature selection are included in Methods and S1 Text. In total, there were 18,856 genes tested by M-DATA.

We performed single-trait analysis on CHD and autism data separately, followed by joint analysis both CHD and autism data with the 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 (Table 1). It demonstrates that M-DATA is able to identify more genes by jointly analyzing multiple traits and incorporating information from functional annotations. We visualized the identified genes with Venn diagrams (Fig 5 and Fig H in S1 Text).

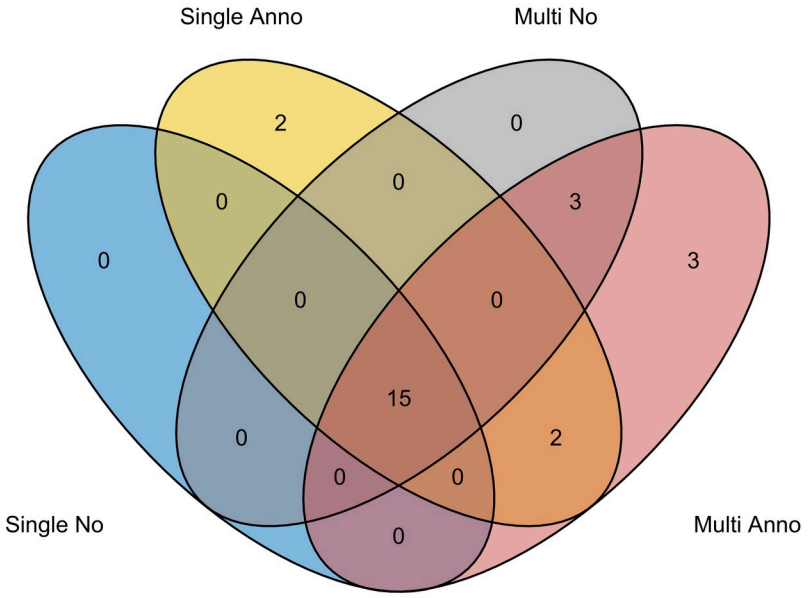

Fig 5. Venn diagram of identified genes in different models. Compared to the single-trait model without annotation, the single-trait model with annotations identified 4 additional genes. Compared to the multi-trait model without annotation, the multi-trait model with annotations identified 5 additional genes. In total, the multi-trait models identified 6 different genes compared to the single-trait models, including 4 novel human CHD genes (*CDK13*, *FRYL*, *LZTR1* and *NAA15*).

https://doi.org/10.1371/journal.pgen.1009849.g005

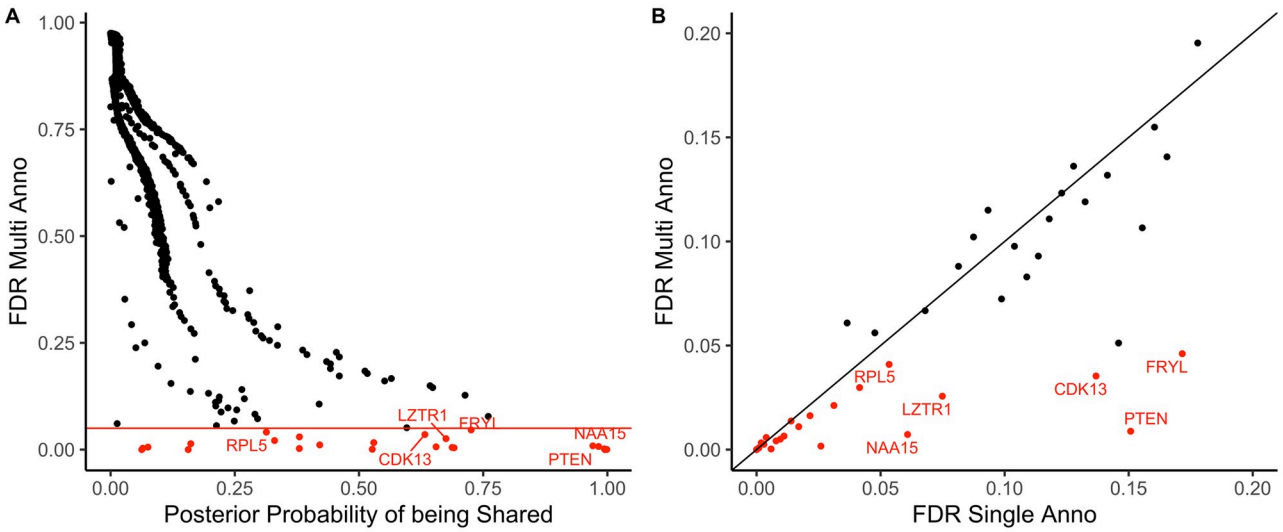

Fig 6. Multi-trait analyses prioritized additional genes with high posterior probability of being shared risk genes for CHD. The 23 genes identified by the multi-trait model with annotations are marked in red on the plot and the additional 6 genes that were identified by the multi-trait models are annotated with gene symbols. (A) shows that the 6 additional genes identified by the multi-trait models had high posterior probability of being shared. The x-axis represents the posterior probability of being shared calculated from the multi-trait model with annotations. The y-axis represents the FDR of genes calculated from the multi-trait model with annotations. (B) shows that the top genes in the multi-trait model with annotations also had low FDR (<0.2) in the single-trait model with annotations. The x-axis represents the FDR of genes calculated from the single-trait model with annotations. The y-axis represents the FDR of genes calculated from the multi-trait model with annotations.

https://doi.org/10.1371/journal.pgen.1009849.g006

We further demonstrate the results by taking CHD as an example. Compared with the single-trait model without annotation, the multi-trait model without annotation identified 3 additional genes, which are *FRYL*, *NAA15* and *PTEN*. Compared with the single-trait model with annotations, the multi-trait model with annotations identified 6 additional genes, including *CDK13*, *FRYL*, *LZTR1*, *NAA15*, *PTEN* and *RPL5*. There are two additional genes identified by the single-trait model with annotations, but not the multi-trait models. Both of these two genes did not have DNMs in autism and are around the margin of FDR threshold (0.05) for the multi-trait model with annotations (*AHNAK* 0.056, *MYH6* 0.061).

To further illustrate the gain of power from multi-trait analysis, we visualized the posterior probability of being shared risk gene for CHD and autism of identified genes in the multi-trait model with annotations in Fig 6A (CHD) and Fig I in S1 Text (autism). In the main text, we further illustrate the results with the 23 significant CHD genes. The 23 significant genes are colored red, and the 6 additional genes identified by multi-trait analyses are annotated with gene symbols. From this figure, we can see that most genes (5/6) have high posterior probability of being shared. *RPL5* is at the margin of FDR threshold in the single-trait models and may be prioritized in the multi-trait models by chance (Fig 6B). In addition, we checked the correlation between the FDR of top genes identified by the multi-trait model with annotations in the single-trait model with annotations (Fig 6B). All 6 genes have low FDR (<0.2) in the single-trait model with annotations, which indicates multi-trait analysis can prioritize marginal signals in single-trait analysis.

We take the 5 CHD genes identified by the multi-trait models, but not the single-trait models as examples to demonstrate the pleiotropic effect. We selected the DNM counts of CHD and autism, FDR of the single-trait model with annotations and FDR of the multi-trait model with annotations model from the results (Table 2). From this table, we can see *CDK13*, *FRYL*, *LZTR1*, *NAA15* and *PTEN* have 2 DNM counts for CHD and at least 1 shared DNM count

Table 2. Pleiotropic effect boosts power for M-DATA multi-trait models.

Gene	*CHD Counts*	*Autism Counts*	*FDR Single Anno*	*FDR Multi Anno*
CDK13	2	1	0.137	0.0354
FRYL	2	2	0.172	0.0461
LZTR1	2	1	0.0749	0.0257
NAA15	2	3	0.0609	0.00726
PTEN	2	4	0.151	0.00882

https://doi.org/10.1371/journal.pgen.1009849.t002

with autism. For *PTEN*, it has 4 autism DNM counts, and we can see a substantial increase of significance in terms of FDR. Thus, the insight is that genes with shared counts with autism are more likely to be prioritized for CHD in multi-trait analyses by leveraging the pleiotropic effect.

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 253 known CHD genes [5]. They are involved in essential developmental pathways or biological processes, such as Notch signaling (*NOTCH1*), RAS signaling (*PTPN11*, *RAF1*, *SOS1*), PI3K/AKT signaling (*PTEN*), chromatin modeling (*CHD7*, *KMT2D*, *NSD1*), transcriptional regulations (*GATA6*), and cell structural support (*ACTB*, *RPL5*) [43, 44].

Among the 12 novel genes, *RBFOX2*, *SMAD2*, *CDK13* are three emerging CHD risk genes that have been recently reported to cause hypoplastic left heart syndrome [4,45,46], laterality defect [3,47], and septal defects and pulmonary valve abnormalities [48], 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 [49,50]. Interestingly, one of the reported mutation carriers also presented cardiac defect [51].

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 [52,53]. The deletion of *KDM5B* in mice is neonatal lethal with respiratory failure and neurodevelopmental defects [54]. 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 [55].

POGZ, *KDM5B*, and *NAA15* are all highly expressed in developmental heart at mice embryonic day E14.5 [5]. *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.

Table 3. Comparison of M-DATA multi-trait models with mTADA.

	M-DATA No	M-DATA Anno	mTADA
CHD	18	23	20
Autism	28	37	28

https://doi.org/10.1371/journal.pgen.1009849.t003

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 [4].

Further, we compared the performance of M-DATA with mTADA [23] using the same real data of CHD and autism. We fitted both methods with damaging mutations (LoF and Dmis mutations). mTADA identified all 18 genes identified by our no annotation model, and missed 3 genes (*CDK13*, *SAMD11*, and *RPL5*) identified by our annotation model for CHD (Table 3). We visualized the results with Venn diagrams (Fig 7 and Fig J in S1 Text). We also compared our results with the results of CHD-ASD pair reported by mTADA using CHD data [4], autism data [11], and mutability data downloaded from the github webpage of mTADA (Table E in S1 Text).

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. The implemented model is available at https://github.com/JustinaXie/MDATA. 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 variant-level or gene-level, which can further improve the statistical power. Through simulation study, we demonstrated that our multi-trait model with annotations could not only gain accurate estimates on the proportion of shared risk genes between two traits and the proportion of risk genes for a single trait under various settings, but also gained statistical power compared to the 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 our simulation study, we found that the algorithm converges faster than methods that use MCMC for estimation (Table D in S1 Text).

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. Currently, we did not consider the influence of admixed population in M-DATA. In a recent study, Kessler et al. studied DNM across 1,465 diverse genomes and discovered mutation rates may be affected by the environment more significantly than previously known [56]. Confounding from the environment on mutation rates could be further explored through cross-ancestry rare variant studies.

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

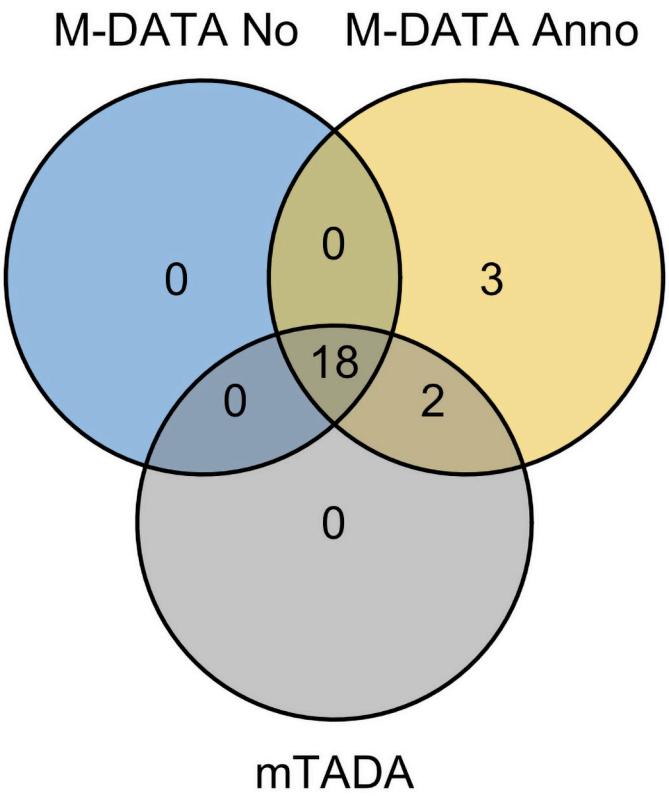

Fig 7. Venn diagram of genes identified by M-DATA and mTADA for CHD. M-DATA multi-trait model with annotations identified 3 additional genes *(CDK13*, *SAMD11* and *RPL5)*.

https://doi.org/10.1371/journal.pgen.1009849.g007

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 significant 23 genes for our multi-trait model with annotations. 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 [29], analyzing DNMs from whole-genome sequencing (WGS) data [25], and incorporating pathway information [57]. Extension of these methods to multi-trait analysis is a potential future direction.

Supporting information

S1 Table. Simulation of estimation and power evaluation.
(XLSX)

S2 Table. Simulation of comparison with mTADA.
(XLSX)

S3 Table. Results of real data application.
(XLSX)

S1 Text. Supplementary notes on methods and results. Fig A. Multi-trait analysis can accurately estimate the proportion of shared risk genes and single-trait risk genes when one annotation is effective or no annotation is effective. Fig B. M-DATA can control FDR under different simulation settings. Fig C. Type 1 errors under different simulation settings. Fig D. AUCs under different simulation settings. Fig E. Multi-trait model with annotations has more increase in power than multi-trait model without annotation when the effect size of annotations is stronger. Fig F. Multi-trait models can control FDR when the effect size of annotations is stronger. Fig G. mTADA is more conservative than M-DATA for FDR control. Fig H. Venn diagram of identified genes in different models for autism. Fig I. Multi-trait analyses prioritized additional genes with high posterior probability of being shared risk genes for autism. Fig J. Venn diagram of genes identified by M-DATA and mTADA for autism. Table A. Summary for Functional Annotations. Table B. Significant genes (FDR<0.05) identified by M-DATA. Table C. Parameter estimates of M-DATA multi-trait analysis of CHD and autism. Table D. Time comparison of M-DATA and mTADA under multiple settings in the simulation study. Table E. Comparison of M-DATA multi-trait models with the results reported in mTADA. (DOCX)

Acknowledgments

We thank denovo-db for providing publicly accessible *de novo* mutation information and Jin et al. [5] for providing publicly accessible *de novo* mutation data for congenital heart disease. We thank Dr. Sheng Chih (Peter) Jin, Geyu Zhou and Hanmin Guo for helpful discussions.

Author Contributions

Conceptualization: Yuhan Xie, Mo Li, Hongyu Zhao.

Data curation: Yuhan Xie, Mo Li, Weilai Dong.

Formal analysis: Yuhan Xie, Mo Li.

Funding acquisition: Hongyu Zhao.

Methodology: Yuhan Xie, Mo Li, Wei Jiang.

Project administration: Hongyu Zhao.

Resources: Weilai Dong, Hongyu Zhao.

Software: Yuhan Xie.

Supervision: Hongyu Zhao.

Validation: Yuhan Xie.

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

Writing – review & editing: Yuhan Xie, Mo Li, Weilai Dong, Wei Jiang, Hongyu Zhao.

References

1. Teer JK, Mullikin JC. Exome sequencing: the sweet spot before whole genomes. Human Molecular Genetics. 2010; 19(R2):R145–R51. https://doi.org/10.1093/hmg/ddq333 PMID: 20705737
2. Rabbani B, Tekin M, Mahdieh N. The promise of whole-exome sequencing in medical genetics. Journal of Human Genetics. 2014; 59(1):5–15. https://doi.org/10.1038/jhg.2013.114 PMID: 24196381
3. Zaidi S, Choi M, Wakimoto H, Ma L, Jiang J, Overton JD, et al. De novo mutations in histone-modifying genes in congenital heart disease. Nature. 2013; 498(7453):220–3. Epub 2013/05/12. https://doi.org/10.1038/nature12141 PMID: 23665959.

4. Homsy J, Zaidi S, Shen Y, Ware JS, Samocha KE, Karczewski KJ, et al. De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. Science. 2015; 350 (6265):1262–6. Epub 2016/01/20. https://doi.org/10.1126/science.aac9396 PMID: 26785492

5. Jin SC, Homsy J, Zaidi S, Lu Q, Morton S, DePalma SR, et al. Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. Nat Genet. 2017; 49(11):1593–601. Epub 2017/10/11. https://doi.org/10.1038/ng.3970 PMID: 28991257

6. Richter F, Morton SU, Kim SW, Kitaygorodsky A, Wasson LK, Chen KM, et al. Genomic analyses implicate noncoding de novo variants in congenital heart disease. Nature genetics. 2020; 52(8):769–77. https://doi.org/10.1038/s41588-020-0652-z PMID: 32601476

7. Watkins WS, Hernandez EJ, Wesolowski S, Bisgrove BW, Sunderland RT, Lin E, et al. De novo and recessive forms of congenital heart disease have distinct genetic and phenotypic landscapes. Nature communications. 2019; 10(1):1–12. https://doi.org/10.1038/s41467-018-07882-8 PMID: 30602773

8. Coe BP, Stessman HAF, Sulovari A, Geisheker MR, Bakken TE, Lake AM, et al. Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity. Nature Genetics. 2019; 51(1):106–16. https://doi.org/10.1038/s41588-018-0288-4 PMID: 30559488

9. Zhernakova A, Stahl EA, Trynka G, Raychaudhuri S, Festen EA, Franke L, et al. Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci. PLoS Genet. 2011; 7(2):e1002004. Epub 2011/03/09. https://doi.org/10.1371/journal.pgen.1002004 PMID: 21383967

10. Willsey AJ, Morris MT, Wang S, Willsey HR, Sun N, Teerikorpi N, et al. The Psychiatric Cell Map Initiative: A Convergent Systems Biological Approach to Illuminating Key Molecular Pathways in Neuropsychiatric Disorders. Cell. 2018; 174(3):505–20. Epub 2018/07/28. https://doi.org/10.1016/j.cell.2018.06.016 PMID: 30053424

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

12. Li J, Cai T, Jiang Y, Chen H, He X, Chen C, et al. Genes with de novo mutations are shared by four neuropsychiatric disorders discovered from NPdenovo database. Mol Psychiatry. 2016; 21(2):290–7. Epub 2015/04/08. https://doi.org/10.1038/mp.2015.40 PMID: 25849321

13. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. 2014; 10(5):e1004383. Epub 2014/05/17. https://doi.org/10.1371/journal.pgen.1004383 PMID: 24830394

14. Solovieff N, Cotsapas C, Lee PH, Purcell SM, Smoller JW. Pleiotropy in complex traits: challenges and strategies. Nat Rev Genet. 2013; 14(7):483–95. Epub 2013/06/12. https://doi.org/10.1038/nrg3461 PMID: 23752797

15. Chung D, Yang C, Li C, Gelernter J, Zhao H. GPA: a statistical approach to prioritizing GWAS results by integrating pleiotropy and annotation. PLoS Genet. 2014; 10(11):e1004787. Epub 2014/11/14. https://doi.org/10.1371/journal.pgen.1004787 PMID: 25393678

16. Flutre T, Wen X, Pritchard J, Stephens M. A statistical framework for joint eQTL analysis in multiple tissues. PLoS Genet. 2013; 9(5):e1003486. Epub 2013/05/15. https://doi.org/10.1371/journal.pgen.1003486 PMID: 23671422

17. Sul JH, Han B, Ye C, Choi T, Eskin E. Effectively identifying eQTLs from multiple tissues by combining mixed model and meta-analytic approaches. PLoS Genet. 2013; 9(6):e1003491. Epub 2013/06/21. https://doi.org/10.1371/journal.pgen.1003491 PMID: 23785294

18. Duong D, Gai L, Snir S, Kang EY, Han B, Sul JH, et al. Applying meta-analysis to genotype-tissue expression data from multiple tissues to identify eQTLs and increase the number of eGenes. Bioinformatics. 2017; 33(14):i67–i74. Epub 2017/09/09. https://doi.org/10.1093/bioinformatics/btx227 PMID: 28881962

19. Li G, Jima D, Wright FA, Nobel AB. HT-eQTL: integrative expression quantitative trait loci analysis in a large number of human tissues. BMC Bioinformatics. 2018; 19(1):95. Epub 2018/03/11. https://doi.org/10.1186/s12859-018-2088-3 PMID: 29523079

20. Li C, Yang C, Gelernter J, Zhao H. Improving genetic risk prediction by leveraging pleiotropy. Hum Genet. 2014; 133(5):639–50. Epub 2013/12/18. https://doi.org/10.1007/s00439-013-1401-5 PMID: 24337655

21. Maier R, Moser G, Chen GB, Ripke S, Coryell W, Potash JB, et al. Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. Am J Hum Genet. 2015; 96(2):283–94. Epub 2015/02/03. https://doi.org/10.1016/j.ajhg.2014.12.006 PMID: 25640677

22. Hu Y, Lu Q, Liu W, Zhang Y, Li M, Zhao H. Joint modeling of genetically correlated diseases and functional annotations increases accuracy of polygenic risk prediction. PLoS Genet. 2017; 13(6):e1006836. Epub 2017/06/10. https://doi.org/10.1371/journal.pgen.1006836 PMID: 28598966

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

24. Lu Q, Yao X, Hu Y, Zhao H. GenoWAP: GWAS signal prioritization through integrated analysis of genomic functional annotation. Bioinformatics. 2016; 32(4):542–8. Epub 2015/10/28. https://doi.org/10.1093/bioinformatics/btv610 PMID: 26504140

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

26. Butkiewicz M, Blue EE, Leung YY, Jian X, Marcora E, Renton AE, et al. Functional annotation of genomic variants in studies of late-onset Alzheimer's disease. Bioinformatics. 2018; 34(16):2724–31. https://doi.org/10.1093/bioinformatics/bty177 PMID: 29590295

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

28. He X, Sanders SJ, Liu L, De Rubeis S, Lim ET, Sutcliffe JS, et al. Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. PLoS Genet. 2013; 9(8):e1003671. Epub 2013/08/24. https://doi.org/10.1371/journal.pgen.1003671 PMID: 23966865

29. Mo Li XZ, Chentian Jin, Sheng Chih Jin, Weilai Dong, Martina Brueckner, Richard Lifton, Qiongshi Lu, Hongyu Zhao. Integrative modeling of transmitted and de novo variants identifies novel risk genes for congenital heart disease. Quant Biol.0-${article.jieShuYe}.

30. Moon TK. The expectation-maximization algorithm. IEEE Signal Processing Magazine. 1996; 13 (6):47–60. https://doi.org/10.1109/79.543975

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

32. Kim S, Jhong J-H, Lee J, Koo J-Y. Meta-analytic support vector machine for integrating multiple omics data. BioData Mining. 2017; 10(1):2. https://doi.org/10.1186/s13040-017-0126-8 PMID: 28149325

33. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. 2013; Chapter 7:Unit7.20. Epub 2013/01/15. https://doi.org/10.1002/0471142905.hg0720s76 PMID: 23315928

34. Samocha KE, Kosmicki JA, Karczewski KJ, O'Donnell-Luria AH, Pierce-Hoffman E, MacArthur DG, et al. Regional missense constraint improves variant deleteriousness prediction. BioRxiv. 2017:148353.

35. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014; 46(3):310–5. Epub 2014/02/04. https://doi.org/10.1038/ng.2892 PMID: 24487276

36. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet. 2016; 99 (4):877–85. Epub 2016/09/27. https://doi.org/10.1016/j.ajhg.2016.08.016 PMID: 27666373

37. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020; 581(7809):434–43. https://doi.org/10.1038/s41586-020-2308-7 PMID: 32461654

38. Jian X, Boerwinkle E, Liu X. In silico prediction of splice-altering single nucleotide variants in the human genome. Nucleic Acids Research. 2014; 42(22):13534–44. https://doi.org/10.1093/nar/gku1206 PMID: 25416802

39. Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RKC, et al. The human splicing code reveals new insights into the genetic determinants of disease. Science. 2015; 347(6218):1254806. https://doi.org/10.1126/science.1254806 PMID: 25525159

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

41. Turner TN, Yi Q, Krumm N, Huddleston J, Hoekzema K, HA FS, et al. denovo-db: a compendium of human de novo variants. Nucleic Acids Res. 2017; 45(D1):D804–d11. Epub 2016/12/03. https://doi.org/10.1093/nar/gkw865 PMID: 27907889

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

43. Zaidi S, Brueckner M. Genetics and Genomics of Congenital Heart Disease. Circ Res. 2017; 120 (6):923–40. Epub 2017/03/18. https://doi.org/10.1161/CIRCRESAHA.116.309140 PMID: 28302740

44. Pierpont ME, Brueckner M, Chung WK, Garg V, Lacro RV, McGuire AL, et al. Genetic Basis for Congenital Heart Disease: Revisited: A Scientific Statement From the American Heart Association. Circulation. 2018; 138(21):e653–e711. Epub 2018/12/21. https://doi.org/10.1161/CIR.0000000000000606 PMID: 30571578

45. McKean DM, Homsy J, Wakimoto H, Patel N, Gorham J, DePalma SR, et al. Loss of RNA expression and allele-specific expression associated with congenital heart disease. Nat Commun. 2016; 7:12824. Epub 2016/09/28. https://doi.org/10.1038/ncomms12824 PMID: 27670201

46. Verma SK, Deshmukh V, Nutter CA, Jaworski E, Jin W, Wadhwa L, et al. Rbfox2 function in RNA metabolism is impaired in hypoplastic left heart syndrome patient hearts. Sci Rep. 2016; 6:30896. Epub 2016/08/04. https://doi.org/10.1038/srep30896 PMID: 27485310

47. Granadillo JL, Chung WK, Hecht L, Corsten-Janssen N, Wegner D, Nij Bijvank SWA, et al. Variable cardiovascular phenotypes associated with SMAD2 pathogenic variants. Hum Mutat. 2018; 39(12):1875–84. Epub 2018/08/30. https://doi.org/10.1002/humu.23627 PMID: 30157302.

48. Sifrim A, Hitz MP, Wilsdon A, Breckpot J, Turki SH, Thienpont B, et al. Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing. Nat Genet. 2016; 48(9):1060–5. Epub 2016/08/02. https://doi.org/10.1038/ng.3627 PMID: 27479907

49. Stessman HAF, Willemsen MH, Fenckova M, Penn O, Hoischen A, Xiong B, et al. Disruption of POGZ Is Associated with Intellectual Disability and Autism Spectrum Disorders. Am J Hum Genet. 2016; 98 (3):541–52. Epub 2016/03/05. https://doi.org/10.1016/j.ajhg.2016.02.004 PMID: 26942287

50. Matsumura K, Seiriki K, Okada S, Nagase M, Ayabe S, Yamada I, et al. Pathogenic POGZ mutation causes impaired cortical development and reversible autism-like phenotypes. Nat Commun. 2020; 11 (1):859. Epub 2020/02/28. https://doi.org/10.1038/s41467-020-14697-z PMID: 32103003

51. White J, Beck CR, Harel T, Posey JE, Jhangiani SN, Tang S, et al. POGZ truncating alleles cause syndromic intellectual disability. Genome Med. 2016; 8(1):3. Epub 2016/01/08. https://doi.org/10.1186/s13073-015-0253-0 PMID: 26739615

52. Kidder BL, Hu G, Zhao K. KDM5B focuses H3K4 methylation near promoters and enhancers during embryonic stem cell self-renewal and differentiation. Genome Biol. 2014; 15(2):R32. Epub 2014/02/06. https://doi.org/10.1186/gb-2014-15-2-r32 PMID: 24495580

53. Kurup JT, Campeanu IJ, Kidder BL. Contribution of H3K4 demethylase KDM5B to nucleosome organization in embryonic stem cells revealed by micrococcal nuclease sequencing. Epigenetics Chromatin. 2019; 12(1):20. Epub 2019/04/04. https://doi.org/10.1186/s13072-019-0266-9 PMID: 30940185

54. Albert M, Schmitz SU, Kooistra SM, Malatesta M, Morales Torres C, Rekling JC, et al. The histone demethylase Jarid1b ensures faithful mouse development by protecting developmental genes from aberrant H3K4me3. PLoS Genet. 2013; 9(4):e1003461. Epub 2013/05/03. https://doi.org/10.1371/journal.pgen.1003461 PMID: 23637629

55. Cheng H, Dharmadhikari AV, Varland S, Ma N, Domingo D, Kleyner R, et al. Truncating Variants in NAA15 Are Associated with Variable Levels of Intellectual Disability, Autism Spectrum Disorder, and Congenital Anomalies. Am J Hum Genet. 2018; 102(5):985–94. Epub 2018/04/17. https://doi.org/10.1016/j.ajhg.2018.03.004 PMID: 29656860

56. Kessler MD, Loesch DP, Perry JA, Heard-Costa NL, Taliun D, Cade BE, et al. De novo mutations across 1,465 diverse genomes reveal mutational insights and reductions in the Amish founder population. Proceedings of the National Academy of Sciences. 2020; 117(5):2560–9. https://doi.org/10.1073/pnas.1902766117 PMID: 31964835

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