









mTADA is a framework for identifying risk genes from de novo mutations in multiple traits

Tan-Hoang Nguyen ^{1,2}, Amanda Dobbyn^{1,3}, Ruth C. Brown², Brien P. Riley ², Joseph D. Buxbaum⁴, Dalila Pinto ^{4,5,6}, Shaun M. Purcell⁷, Patrick F. Sullivan⁸, Xin He ^{9,10} & Eli A. Stahl ^{1,11}

Joint analysis of multiple traits can result in the identification of associations not found through the analysis of each trait in isolation. Studies of neuropsychiatric disorders and congenital heart disease (CHD) which use de novo mutations (DNMs) from parent-offspring trios have reported multiple putatively causal genes. However, a joint analysis method designed to integrate DNMs from multiple studies has yet to be implemented. We here introduce multiple-trait TADA (mTADA) which jointly analyzes two traits using DNMs from non-overlapping family samples. We first demonstrate that mTADA is able to leverage genetic overlaps to increase the statistical power of risk-gene identification. We then apply mTADA to large datasets of >13,000 trios for five neuropsychiatric disorders and CHD. We report additional risk genes for schizophrenia, epileptic encephalopathies and CHD. We outline some shared and specific biological information of intellectual disability and CHD by conducting systems biology analyses of genes prioritized by mTADA.

¹Division of Psychiatric Genomics, Department of Genetics and Genomic Sciences, Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, USA. ²Virginia Institute for Psychiatric and Behavioral Genetics, Department of Psychiatry, Virginia Commonwealth University, Richmond, VA, USA. ³Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. ⁴Seaver Autism Center, Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA. ⁵The Mindich Child Health & Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. ⁶Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. ⁷Sleep Center, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. ⁸Departments of Genetics and Psychiatry, University of North Carolina, Chapel Hill, NC, USA. ⁹Department of Human Genetics, University of Chicago, Chicago, IL, USA. ¹⁰Grossman Institute for Neuroscience, Quantitative Biology and Human Behavior, University of Chicago, Chicago, IL, USA. ¹¹Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA. ✉email: tanhoang.nguyen@vcuhealth.org; xinhe@uchicago.edu; eli.stahl@mssm.edu

The analysis of multiple traits can help characterize the genetic architectures of complex disorders¹. One approach is to meta-analyze results derived from separate single-trait studies². However, joint analysis of multiple traits can better accommodate heterogeneity of genetic effects of the same variants or genes across traits^{3,4}. Numerous studies have jointly analyzed two or more traits and successfully identified shared common-variant associations^{5–8}. In addition, additional risk loci have been identified using these approaches^{7,9}. However, none of these studies has examined rare variation from case-control data, or de novo variants for which mutation rates should be taken into account. For these rare variants, gene based tests have identified several genes associated with different disorders^{10–13}. Some recent studies have shown that there are multiple risk genes that are shared between neurodevelopmental disorders^{10,14,15}, and also with congenital heart disease (CHD)^{16,17}. These results are based on the intersection among the top prioritized genes from each disorder; therefore, reported numbers of genes shared by multiple disorders remain low^{10,17}. Development of multi-trait rare-variant methods for neuropsychiatric disorders (NPDs) and related disorders will facilitate the understanding of this important aspect of genetic architecture for these phenotypes.

Currently, there is still a limitation in the risk gene identification for a single trait of NPDs and relevant disorders from parent-offspring trio studies. One reason is that risk gene discovery is underpowered when sample sizes are limited, as well as when relative risks are not large^{10,11}. Multiple risk genes have been reported for undiagnosed developmental disorders (DD), intellectual disability (ID) and autism spectrum disorder (ASD)^{12,18,19} thanks to large sample sizes and/or relative risks¹⁰. However, there are a few risk genes identified for schizophrenia (SCZ), epileptic encephalopathies (EE) and other disorders because of small gene-level relative risks or small sample sizes^{10,20,21}. Increasing sample sizes will increase power to identify additional risk genes, but this is an expensive solution and may not be feasible for some studies. If there are genetic overlaps, methods that can leverage the information from one trait to increase power for risk-gene identification of another trait could help in obtaining additional genes for these disorders.

Here, we have developed a new statistical model, mTADA (multi-trait transmission and de novo association test), that jointly analyzes de novo mutations (DNMs) of two traits in order to estimate the gene-level genetic overlap of the two traits, and to identify additional risk genes for each analyzed trait as well as shared and specific risk genes. First, we utilize simulation data and demonstrate that, compared with a single-trait method, mTADA substantially improves the power of risk-gene identification when genetic overlaps increase, especially for traits with smaller sample sizes or smaller relative risks. For example, mTADA is able to statistically increase evidence for multiple

genes in a tested trait which shows 1) marginally statistical evidence in that trait, and 2) strong evidence in the other trait if the two traits have a high genetic overlap. To illustrate the advantage of the new pipeline over its previous single-trait version, we apply the method to large data sets of different NPDs and CHD (>13,000 parent-offspring trios) and identify shared genes between each pair of these disorders. mTADA identifies additional risk genes for each disorder by borrowing the information of other traits. We validate these results in an independent cohort of 1,241 trios with CHD, 197 trios with EE, and 4,877 SCZ cases and 6,203 controls. In addition, we demonstrate that mTADA's results could be used to better understand the shared and specific biological information for two tested disorders by using multiple systems biology approaches to test the top prioritized risk genes of the CHD-ID pair. CHD-specific genes are specific to certain biological pathways.

Results

The mTADA framework. The mTADA method is gene-based and requires input data of the number of DNMs and mutation rate per gene. If the DNMs are stratified on the basis of predicted effect (e.g., 'missense', 'nonsense', etc.), then each gene-annotation category should have its own mutation rate that reflects the predicted effects of the mutations within. In summary, for each gene, we consider four models M_j ($j = 0..3$) reflecting four alternative hypotheses: the gene is associated with neither trait (H_0), the first trait only (H_1), the second trait only (H_2), or both traits (H_3). We assume prior probabilities π_j ($j = 0..3$) for the four models and these π_j are estimated from data and single-trait studies. DNMs are modeled using Poisson distributions with mean relative risks, mutation rates and sample sizes as main parameters¹⁰ (Methods). For each gene, four posterior probabilities (PP), which are abbreviated as PP0, PP1, PP2 and PP3 respectively, are used to infer the status of the gene for the four models. To summarize the evidence for association with a given trait, we use the sum of PPs of models including the risk gene hypotheses for that trait, i.e., PP1 + PP3 for trait one and PP2 + PP3 for trait two (Fig. 1, Table 1, Methods).

Results of mTADA on simulated data. To validate the new method, we conducted simulation studies by using genetic parameters from real-data analyses of previous studies (Methods).

Power for single-trait risk gene discovery. We compared gene numbers identified by mTADA and our previous single-trait method, extTADA, using the same threshold $PP > 0.8$. For $\pi_3 = 0$ (no overlapping information), mTADA and extTADA reported nearly the same positive gene numbers (Fig. 2). However, mTADA identified more genes than extTADA when π_3 increased.

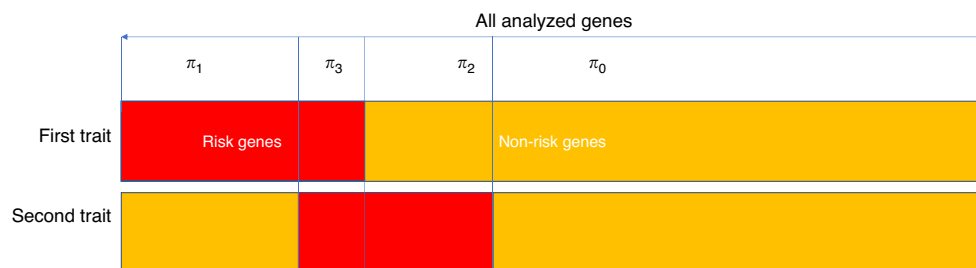


Fig. 1 The multiple trait transmission and de novo association test (mTADA). For each trait, mTADA divides the all tested genes into two sets: risk and non-risk genes. Therefore, there are four sets when two traits are combined: risk genes for neither of traits (H_0), for the first trait only (H_1), for the second trait only (H_2), and for both traits (H_3). Statistical details of four models for these four hypotheses are described in Table 1. π_j ($j = 0..3$) are prior probabilities for the four models. From mTADA's analysis results, each gene has four posterior probabilities (PPs) of the four models (e.g., PP0, PP1, PP2 and PP3 for Model 0, Model 1, Model 2 and Model 3 respectively).

Table 1 Statistical models of mTADA.

Hypothesis	Proportion	First trait	Second trait
H_0	π_0	$x_{i1} \sim \text{Poisson}(2N_1\mu_i)$	$x_{i2} \sim \text{Poisson}(2N_2\mu_i)$
H_1	π_1	$x_{i1} \sim \text{Poisson}(2N_1\mu_i\gamma_{i1})$ $\gamma_{i1} \sim \text{Gamma}(\bar{\gamma}_1\beta_1, \beta_1)$	$x_{i2} \sim \text{Poisson}(2N_2\mu_i)$
H_2	π_2	$x_{i1} \sim \text{Poisson}(2N_1\mu_i)$	$x_{i2} \sim \text{Poisson}(2N_2\mu_i\gamma_{i2})$ $\gamma_{i2} \sim \text{Gamma}(\bar{\gamma}_2\beta_2, \beta_2)$
H_3	π_3	$x_{i1} \sim \text{Poisson}(2N_1\mu_i\gamma_{i1})$ $\gamma_{i1} \sim \text{Gamma}(\bar{\gamma}_1\beta_1, \beta_1)$	$x_{i2} \sim \text{Poisson}(2N_2\mu_i\gamma_{i2})$ $\gamma_{i2} \sim \text{Gamma}(\bar{\gamma}_2\beta_2, \beta_2)$

Statistical models for four hypotheses in mTADA for one category of variants in each trait at the i^{th} gene. mTADA assumes that the gene can be in one of four models M_0, M_1, M_2, M_3 . π_j ($j = 0, 1, 2, 3$) is the prior probability of the j^{th} model. x_k and N_k ($k = 1, 2$) are the data and the sample size of the k^{th} trait. μ_i is the mutation rate of the gene. For each trait, the relative risks of shared and specific genes (γ_k) are from a Gamma distribution with two parameters: $\bar{\gamma}_k$ (mean relative risk) and β_k (to control the variance of relative risks).

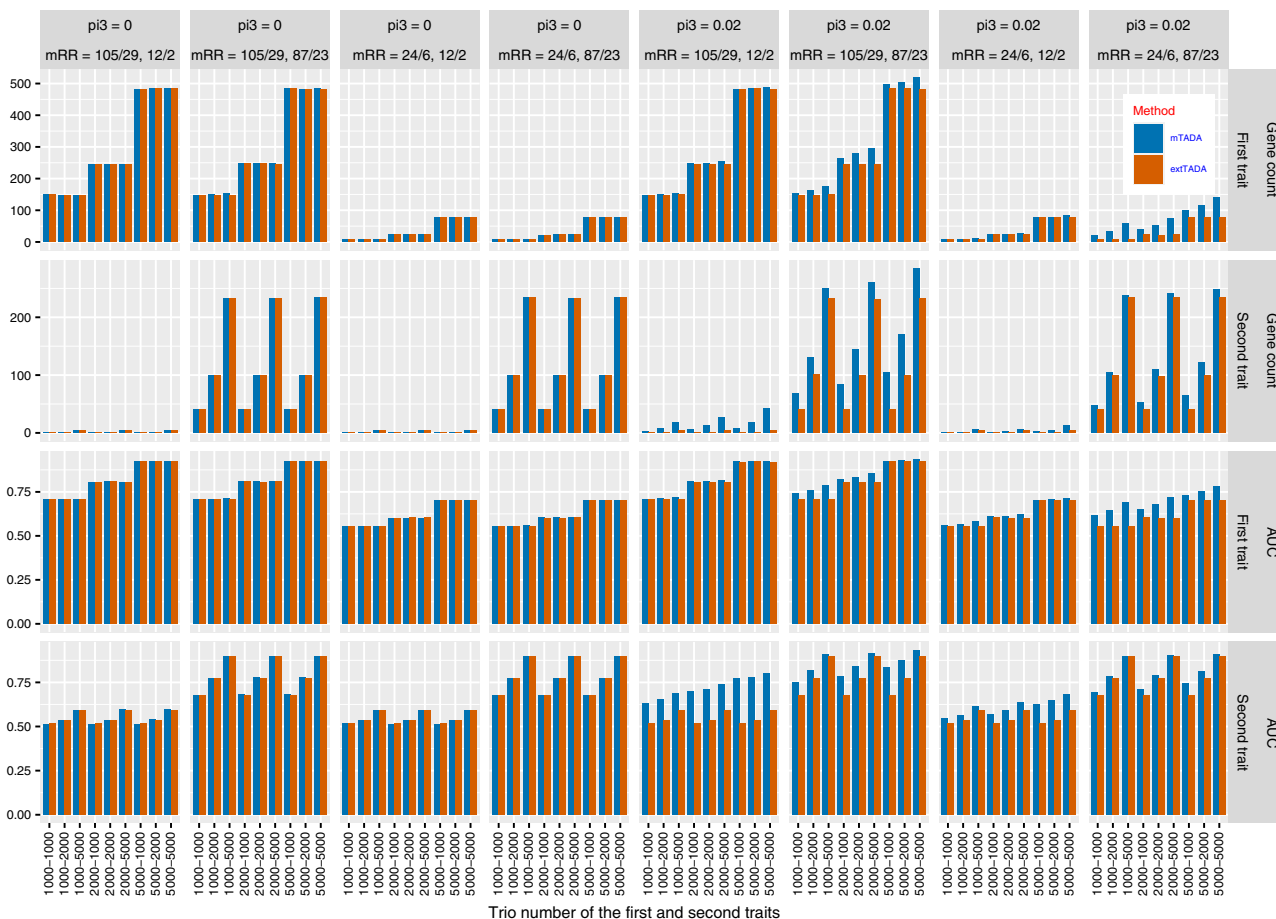


Fig. 2 Comparison results of simulated data for the current multi-trait approach (mTADA) and a previous single-trait approach (extTADA) in single-trait analyses. For each bar, its height shows the average value of 100 simulations. mTADA performs better than extTADA when the proportions of overlapping risk genes (π_3) are larger than zero. The top two lines describe gene counts (posterior probability >0.8 , while the two bottom lines show area under the Receiver Operating Characteristic (ROC) curves (AUCs). mRR describes mean relative risks and the trio number along the bottom describes the sample sizes. These results are for two variant categories. For example, “mRR = 105/29, 12/2” describes the mRRs of the first trait as 105 and 29, and the mRRs of the second trait as 12 and 2.

In addition, mTADA’s gene counts were also higher than those of extTADA when higher mean relative risks were used.

Comparison of risk-gene classification for single traits. We designed a simulation experiment to assess the performance in the classification of risk and non-risk genes. We applied extTADA to single-trait data from our simulated data. We then calculated areas under the Receiver Operating Characteristic curves (AUCs) for mTADA and extTADA using classification results from single-trait data. AUCs of both were equal when

$\pi_3 = 0$ (Fig. 2). However, AUCs of mTADA were higher than those of extTADA when π_3 ’s values were larger. In addition, mTADA also performed better than extTADA with larger mean relative risks.

The proportion of false positive shared risk genes for two traits with non-genetic overlaps. We estimated this information for identifying shared risk genes (i.e. associated with both traits). We simulated data with $\pi_3 = 0$ and calculated the proportion of shared risk genes (per 19,358 tested genes) using different PP

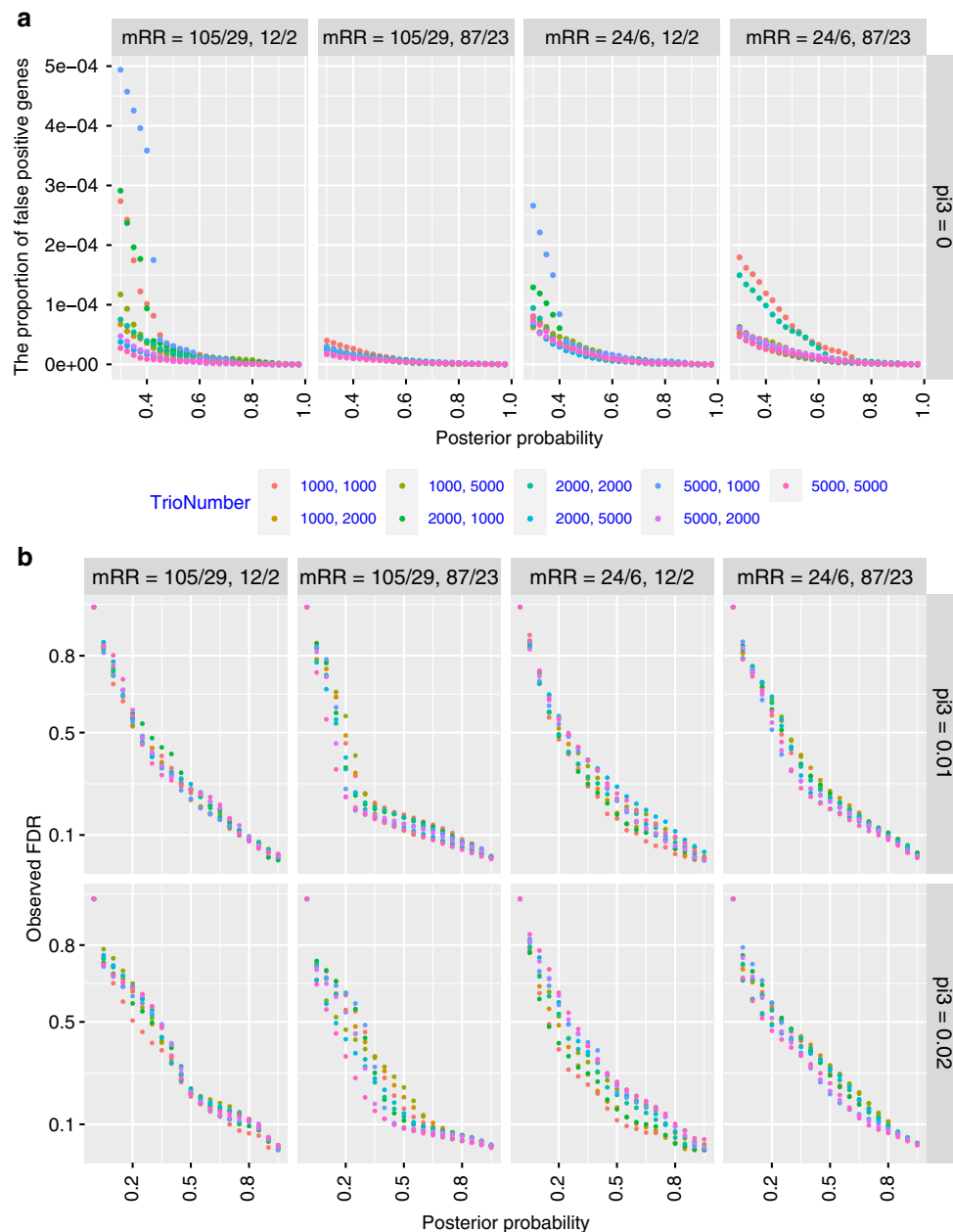


Fig. 3 Validation of shared risk gene identification using mTADA on simulated data. **a** The proportion of false positive genes (per 19,358 analyzed genes): X-axes are posterior probabilities of Model III while Y-axes are the proportions of false positive shared risk genes. **b** The correlation between posterior probabilities (x-axis) and observed false discovery rates (FDRs, y-axis). These are for the combination of different sample sizes (ntrio) and mean relative risks (mRR).

thresholds of Model III (PP3). Overall, these proportions were very small for different PP thresholds ($<4.5 \times 10^{-4}$, Fig. 3a).

Correlations between posterior probabilities and observed false discovery rates (FDRs). Since mTADA makes inference on risk genes using PPs, instead of commonly used FDR, we compared these two metrics. We calculated the correlation between PPs and observed false discovery rates (oFDRs) for all situations. For PP3, we found that PP = 0.8 and 0.5 approximately correspond to oFDR = 0.1 and 0.25, respectively. Small mean relative risks could lead to higher FDRs, but this inflation was modest (Fig. 3b). These results were similar for other situations: when genes were associated with only the first trait, only the second trait, single traits (e.g., Trait 1 or Trait 2 genes) (Supplementary Figs. 1, 2).

The correlation between simulated and estimated values of π_3 was also assessed. For large mean relative risks, high correlations

were observed for all sample sizes. For smaller mean relative risks (≤ 24), π_3 's values were over- or underestimated (Supplementary Fig. 3). However, these small differences did not affect the results of risk-gene identification (Fig. 2, Supplementary Figs. 1, 2).

The effects of misdiagnosis and ascertainment bias on the results. When sample phenotypes are misdiagnosed (a patient of one trait is mis-assigned to another), the estimated parameters of mTADA may be biased and this may affect the results. In another scenario, samples from one trait may contain a larger number of patients of the second trait than expected based on the comorbidity in the population. This ascertainment bias may also have an effect on mTADA's estimates. We tested the impact of these scenarios. Overall, π_3 and downstream results were not strongly affected when there was ascertainment bias. Similar results were also observed for misdiagnosis rates of 5–10% if the

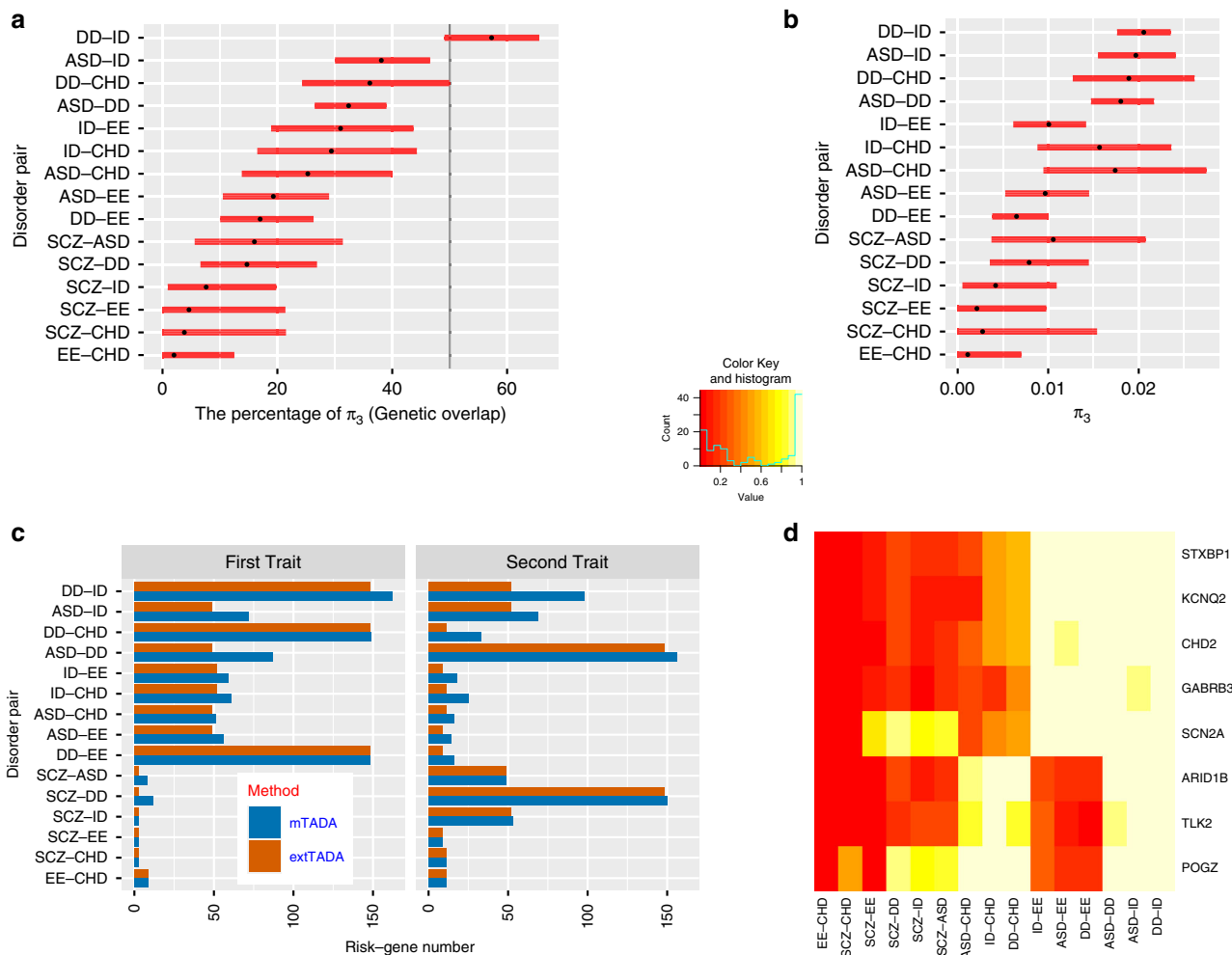


Fig. 4 Analysis results of mTADA for pairs of disorders. **a** The estimated gene-level genetic overlaps (gOs) of pairs of disorders from Markov Chain Monte Carlo sampling results. Each pair shows the credible interval and the black dot is the estimated value. The vertical black line describes $gO = 50\%$. **b** The estimated proportion of overlapping risk genes (π_3) in the mTADA model. **c** Comparison of mTADA and extTADA in the prioritization of top genes by using a threshold of posterior probability (PP) > 0.8. In mTADA, the column ‘First trait’ and ‘Second trait’ are inferred by summing the PPs of model 1 and 3 (PP1 + PP3), and model 2 and model 3 (PP2 + PP3) in Fig. 1 respectively. **d** These genes appear in at least 4 pairs of disorders (PP > 0.8). Cells show the PP values. Y-axis shows gene names and x-axis describes pairs of disorders.

mean relative risks of the tested traits were not highly imbalanced. If the mean relative risks of one trait were substantially higher than those of the other trait, overestimation of π_3 might arise for misdiagnosis rates of $\geq 5\%$. Detailed results are in the Supplementary Note 1.

Application of mTADA to neuropsychiatric disorders and CHD data. mTADA was applied to DNM datasets of 15 pairs of six disorders: five neuropsychiatric disorders including DD, ID, ASD, SCZ, EE; and CHD. These DNMs were classified into different categories using annotation tools. Based on previous results^{10,22}, we used loss of function (LoF), missense damaging (MiD) DNMs for all disorders and also added synonymous DNMs within DNase I hypersensitive sites for SCZ (Methods). We defined the gene-level genetic overlap (gO) of two disorders as: $gO = 100\% \times \pi_3 / (\pi_1 + \pi_2 + \pi_3)$. DD based results showed strong convergence with smaller credible intervals (CIs) because of its large sample size as well as high relative risks of DNMs (Fig. 4a/b, Supplementary Table 1). As expected, high gOs were observed for pairs of DD, ID, and ASD ($gO > 32\%$, $\pi_3 > 0.018$). CHD and EE had the lowest gO ($gO = 2\%$, $\pi_3 = 0.001$) followed

by SCZ-EE ($gO = 4.6\%$, $\pi_3 = 0.0023$). Supplementary Fig. 4 shows sampling results of the proportions of overlapping risk genes for pairs of these traits, and Supplementary Fig. 5 shows the percentage of genetic overlaps for traits. The gO of ASD and SCZ which was approximately 16% (CI = 5.6–31.4%) was similar to previous studies (Supplementary Table 2).

We also compared mTADA and extTADA in the identification of risk genes for single traits using a threshold of PP > 0.8. For DD and ID, mTADA always performed better than extTADA (Fig. 4c). Similar results were observed for ASD; except for the pair ASD-SCZ in which mTADA was slightly better than extTADA for SCZ but extTADA was better than mTADA for ASD. For CHD, EE and SCZ, mTADA was better than extTADA when CHD was combined with DD.

Insights into top genes prioritized by mTADA. To better understand the top genes prioritized by mTADA, we extracted genes with PP > 0.8 for further analyses.

Overlapping genes between two traits. The highest number of overlapping genes was observed for DD and ID (89 genes) followed by ASD-DD (65 genes) and ASD-ID (47 genes).

Four pairs of traits (CHD-EE, SCZ-EE, CHD-SCZ, SCZ-ID) had no overlapping genes. 152 genes were supported by the two-trait model in at least one pair ($PP_3 > 0.8$, Supplementary Data 1). Eight genes (*ARID1B*, *GABRB3*, *KCNQ2*, *STXBP1*, *CHD2*, *TLK2*, *POGZ*, *SCN2A*) were observed for at least six pairs of disorders (Fig. 4d). *POGZ* and *SCN2A* were present in eight pairs of disorders. *POGZ* was significant for pairs relating to ASD, DD, ID, CHD and SZ while *SCN2A* was significant for pairs relating to ASD, DD, EE, ID and SCZ. We checked DNMs of these two genes. As expected, *POGZ* had no DNMs for EE, and *SCN2A* had no DNMs for CHD. Interestingly, in the latest CHD study²³, *POGZ* was one of the top CHD genes while no DNMs were observed for *SCN2A*. In addition, in a recent study of 6,753 parent-offspring trios with neurodevelopmental disorders and epilepsy²⁴, 16 DNMs were in *POGZ*, but only one DNM was from a patient who has both ID and epilepsy.

Significant genes of single traits. To demonstrate the application of mTADA in the identification of additional risk genes, we tested three disorders (CHD, EE, and SCZ) whose DN-based genes have not been reported as often as the three other disorders. We used DD-based results because the number of risk genes for the three disorders highly increased when their datasets were jointly analyzed with the DD dataset in Fig. 4c.

CHD. 33 genes were prioritized. 20/31 were not in the list of known CHD genes and in the meta-analysis results of a recent CHD study of Jin et al.²³ (Table 2). We validated these results by using different approaches. First, we tested the protein-protein interactions (PPIs) of these 33 genes by using the STRING database²⁵. The number of edges was higher than expected between 33 protein nodes (PPI $p = 6e-12$, Fig. 5). Multiple protein products of novel and known genes interacted with each other. The number of interactions decreased when tested with only PPIs from experiments but was still significant (PPI $p = 0.0174$). Second, we tested these CHD genes from an independent data set which includes 1,241 trios and 226 cases²³. From the 1,241 trios, three genes (*CTNNB1*, *CUL3*, *LZTR1*) of the 20 novel genes had LoF or MiD DNMs (Poisson-test $p < 2.0e-4$, Table 2). Each of these three genes had only one DNM in the primary analysis. In addition, these genes were not called significant genes by extTADA. Finally, we compared these 33 genes with the top 25 genes meta-analyzed by Jin et al.²³. 8/33 were in the 25-gene list (permutation $p < 9.99e-05$; Table 2).

EE. There were 16 genes. Similar to top CHD genes, their protein products also had more interactions than expected by using the STRING database (PPI $p = 3e-11$, Supplementary Fig. 6). Three genes *HECW2*, *MLL*, *WDR19* were not in the list of EE genes on the Online Mendelian Inheritance in Man²⁶. These three genes only had $PP < 0.3$ in extTADA. Interestingly, *HECW2* had a DNM in a whole-genome-sequencing study recently²⁷.

SCZ. There were 12 genes including *AUTS2*, *BRPF1*, *CHD8*, *HIST1H1E*, *HIVEP3*, *MAP4K4*, *MKI67*, *POGZ*, *SCN2A*, *SETD1A*, *SYNGAP1*, *TAF13*. These genes' protein products were significantly connected by using the STRING database (PPI $p = 1.6e-03$, Supplementary Fig. 6). In these genes, only *TAF13* and *SETD1A* were suggested as top genes in previous studies^{10,28}. In addition, *AUTS2* was reported as a SCZ gene from a common variant based study²⁹. We tested these genes on an independent dataset of 4,877 cases and 6,203 controls, *HIST1H1E* showed nominally significant (Supplementary Table 3).

Biological insights into shared and specific genes from mTADA's analysis. To demonstrate the application of mTADA in helping to understand the shared and specific biological mechanism of two analyzed disorders. We extracted three gene lists (shared and

specific genes) for each pair of disorders using a threshold of $PP > 0.5$. To increase the sample size for CHD, we combined both tested and independent datasets (Methods). We then focused on ID and CHD in this analysis because this pair of disorders had high numbers of risk genes for the three lists (30 shared genes, 40 ID-specific genes and 30 CHD-specific genes, Supplementary Data 2). Different systems biology approaches were used to test these three gene lists. First, we conducted gene-set enrichment analyses³⁰ using gene-ontology (GO) gene sets³¹. The majority of top enriched GO gene sets were related to heart/cardiocyte-development for CHD-specific genes, to chromatin modification or DNA binding for shared or ID-specific genes (Fig. 6a). Next, we used gene sets from a human single-cell RNA sequencing (scRNAseq) dataset of ~4,000 cardiac cells from human embryos³². No overlaps were observed between shared or ID-specific genes with these gene sets, but interestingly the CHD-specific genes were enriched in multiple gene sets (Fig. 5b). We then tested the three gene lists by using mouse scRNAseq gene expression datasets from different brain regions³³. The three gene lists were not significantly enriched in these cell types; however, for pyramidal cells, ID-specific genes were nominally significant while CHD-specific genes did not have the same direction (Fig. 5c). Finally, we used BrainSpan RNAseq gene expression data to cluster these three gene lists into spatiotemporal groups. Using eight time points and four regions as in recent studies^{10,34}, shared and ID-specific genes were strongly expressed in the prenatal stages of the human brain while CHD-specific genes were expressed in both prenatal and postnatal stages for Region 3 including hippocampus, amygdala and striatum (Fig. 5d, Supplementary Fig. 7). The three other brain regions did not show strong differentiations between these three gene lists (Supplementary Fig. 7).

Discussion

In this paper, we propose a method to jointly analyze two traits (mTADA) using de novo exome sequencing data. The method is an extension of our previous work for single traits^{10,11}. mTADA estimates the proportion of overlapping risk genes (π_3) between two traits, and then uses this information to infer how many overlapping risk genes exist between two traits. The pipeline is also able to infer the number of risk genes for each trait by calculating posterior probabilities (PPs) of genes for each trait. On simulated data, mTADA performs better than a single-trait approach, extTADA, on the identification of risk genes (Fig. 2). We applied mTADA to more than 13,000 trios of five neuropsychiatric disorders and congenital heart disease, and reported overlapping genes between these disorders. We also saw that mTADA reported more risk genes for these disorders than extTADA (Fig. 4). This suggests that mTADA can help in the identification of additional risk genes, especially for disorders whose large sample sizes are challenging to obtain or whose mean relative risks are small. For such disorders, users can combine the data of the disorders with large public data sets (e.g., trio data of ASD, DD) to prioritize risk genes. Using one-trait information to leverage the information for other traits has been successful in fine-mapping³⁵ and common-variant³⁶ studies. Based on our best knowledge, mTADA is the first tool using this approach for de novo mutation data. We hope that mTADA (<https://github.com/hoangtn/mTADA>) will be generally useful for analyzing de novo mutation data across complex traits.

By using mTADA for prioritizing top genes, multiple overlapping genes were observed for CHD, DD, ID and ASD. This replicates a recent study³⁷ in which high overlapping genes were observed for CHD and neurodevelopmental disorders. Interestingly, CHD did not show any overlapping information with

Table 2 Information of genes prioritized for congenital heart disease (CHD).

Gene	dn_MID_DD	dn_LoF_DD	dn_MID_CHD	dn_LoF_CHD	PP	dn_MID_CHD2018	dn_LoF_CHD2018	pPoisson	Known_Gene	Top 25 genes from Jin et al., 2018
KDMSB	0	3	0	3	1	0	0	1.00E+00	N	2.90E-04
MLL2	0	2	0	4	1	0	0	5.50E-15	Y	8.50E-19
NAAT5	0	2	0	2	1	0	0	1.00E+00	N	-
CHD7	2	2	0	2	1	3	0	3.98E-24	Y	7.50E-19
RBFOX2	0	0	0	3	1	0	0	1.00E+00	N	1.10E-06
PTPN11	2	0	3	0	1	2	0	1.17E-04	Y	1.80E-15
POGZ	0	6	1	1	1	0	1	1.32E-02	N	2.90E-04
CTNNB1	0	11	0	1	0.97	0	0	1.90E-02	N	-
TCF12	1	2	0	1	0.97	0	0	1.00E+00	N	-
KANS1L1	0	8	0	1	0.97	0	0	1.00E+00	N	-
MEIS2	0	2	0	1	0.97	0	0	1.00E+00	Y	-
EIF4A2	1	1	0	1	0.97	0	0	1.00E+00	N	-
WHSCI	0	3	0	1	0.97	0	0	1.00E+00	N	-
KAT6B	0	8	0	1	0.97	0	0	1.00E+00	Y	-
MAP2	0	2	0	1	0.97	0	0	1.00E+00	N	-
CUL3	0	2	0	1	0.97	0	0	2.21E-02	N	-
ARID1B	0	30	0	1	0.97	0	0	1.00E+00	N	-
KAT6A	0	8	0	1	0.97	0	0	1.00E+00	N	-
NSD1	1	7	0	1	0.96	0	0	7.82E-04	Y	1.30E-04
EP300	3	12	0	1	0.96	0	0	1.00E+00	N	-
CACNA1A	5	0	0	1	0.95	0	0	1.00E+00	N	-
MEAT	0	1	0	1	0.95	0	0	1.00E+00	N	-
ZNF623	0	1	0	1	0.94	0	0	1.00E+00	N	-
GANAB	2	0	1	1	0.94	0	0	1.00E+00	N	-
COL4A3BP	4	0	1	0	0.92	0	0	1.00E+00	N	-
LZTR1	2	1	1	0	0.91	1	0	2.65E-02	N	-
RARGAP1L	0	1	1	1	0.9	0	0	1.00E+00	N	-
MED13L	5	13	1	0	0.89	0	0	1.00E+00	Y	-
TLK2	2	0	0	1	0.87	0	0	1.00E+00	N	-
ARID1A	1	2	1	0	0.87	0	0	1.00E+00	Y	-
SRRM2	0	2	0	1	0.87	0	0	1.00E+00	N	-
CHD4	5	1	1	0	0.86	1	0	7.30E-02	Y	-
SMAD2	0	0	1	1	0.8	0	0	1.00E+00	N	1.60E-04

These 33 genes are prioritized by mTADA using the information of undiagnosed developmental disorders (DD). Columns 'dn_LoF/MID' describe the number of loss-of-function/misense damaging de novo mutations, mTADA was applied to the DD and the CHD data in the 2nd, 3rd, 4th and 5th columns. The column 'PP' describes the posterior probabilities of these genes from mTADA's analyses. Columns 'dn_MID_CHD2018' and 'dn_LoF_CHD2018' are data from an independent dataset. Column 'pPoisson' describes p values of the Poisson test for the independent dataset. Column 'Known gene' shows whether a gene is in the list of known genes (Yes/Y) or not (No/N). The last column shows p-values calculated by Jin et al.²³ for their top significant genes.

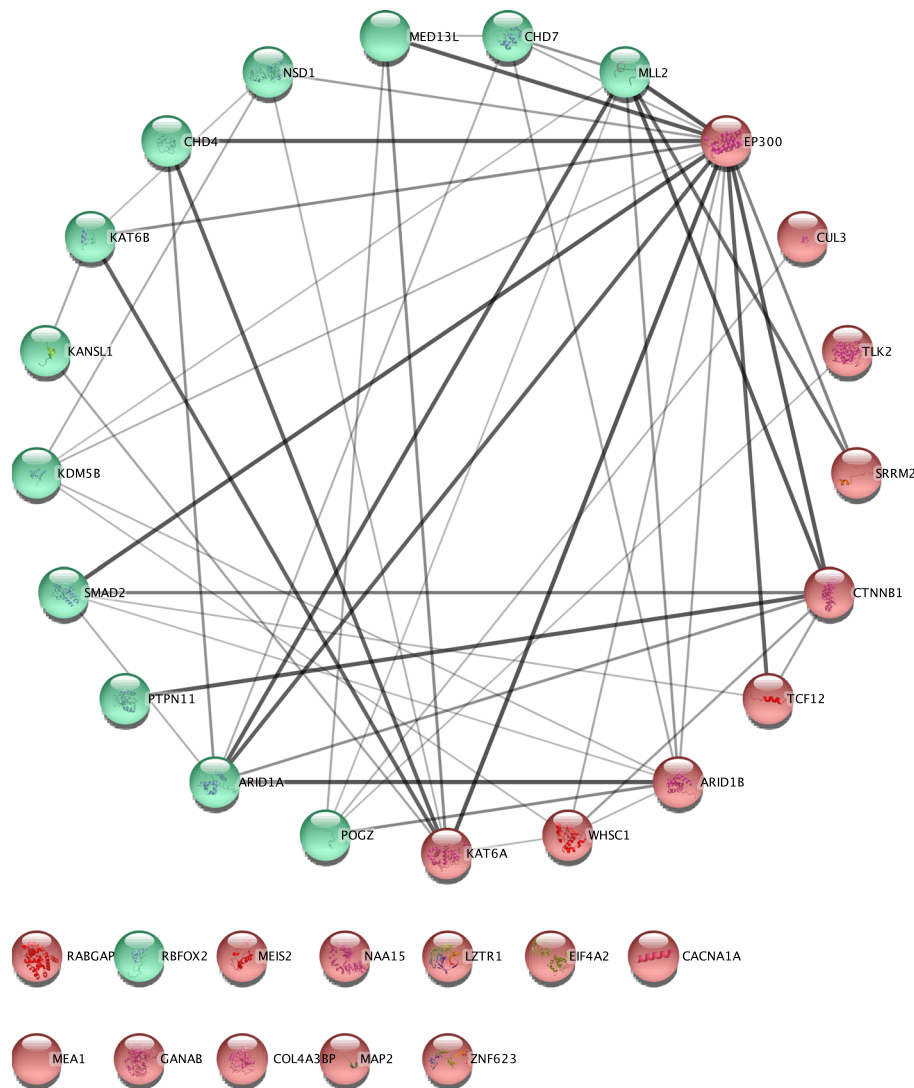


Fig. 5 Result of protein-protein interaction analysis for genes associated with congenital heart disease (CHD). These genes were prioritized by using undiagnosed developmental disorders (DD) information. This is the top 33 genes, posterior probabilities > 0.8 , identified by mTADA using the data set of Homsy et al.¹⁶. Novel genes have red background and known genes have green background. Additional information for these genes is in Table 2.

another neurodevelopmental disorder: EE. Two genes *SCN2A* and *POGZ* which have been reported as risk genes for some of these disorders^{23,38–40} are top overlapping genes from mTADA (Fig. 4d), but they show different trends. No *SCN2A* DNMs are in CHD data and no *POGZ* DNMs are in EE data. One possible reason is that the sample size of EE is small in this study (356 trios). Another hypothesis might be that they do not have strong overlapping biological pathways. We did not see any overlapping information between SCZ and CHD, or SCZ and EE. We analyzed in depth the top prioritized genes of CHD, EE and SCZ (Fig. 5, Supplementary Fig. 6). Some top risk CHD and EE genes from mTADA are also reported in recent studies^{23,27}. Multiple top CHD genes have only one DNM, but have DNMs in independent data sets (Table 2). This suggests that they might be real risk-genes for this disorder. Interestingly, we identify 20 CHD genes (posterior probabilities > 0.8) which are not in the list of 253 curated known human/mouse CHD genes. 3 of these 20 genes have DNMs in an independent data set. This shows the benefit of using mTADA in the prediction of risk genes for CHD by borrowing the information of DD (Fig. 5, Table 2). We used different systems biology approaches to understand the shared and specific risk gene lists of ID and CHD. Some specific information

emerged from these analyses. CHD-specific genes were enriched for heart/cardiocyte pathways, cell types while shared and ID-specific genes were strongly expressed in the prenatal stages of the human brain and enriched in regulatory and binding pathways (Fig. 6). This suggests that a model-based approach as mTADA can help shed light on the shared and specific biological mechanism between disorders with larger sample sizes.

Although mTADA performs better than the single-trait based extTADA, it does have some limitations. mTADA uses the parameters of single traits from extTADA to infer π_3 . Using parameters from extTADA makes mTADA much faster in its calculation, it means mTADA relies on the results of the single-trait pipeline extTADA that uses a full Bayesian approach. Also, mTADA as well as extTADA use de novo counts for each gene and divide these counts into different categories similar to other rare variant based studies^{12,41–43}. In this current pipeline, we estimated π_3 directly from data. However, common-variant based genetic co-heritabilities⁴⁴ and transcriptomic correlations⁴⁵ for multiple pairs of NPDs are available now. Other studies which are able to incorporate the annotation information of each mutation, prior information for π_3 from previous studies may increase the power of mTADA or similar tools. In the current version, users

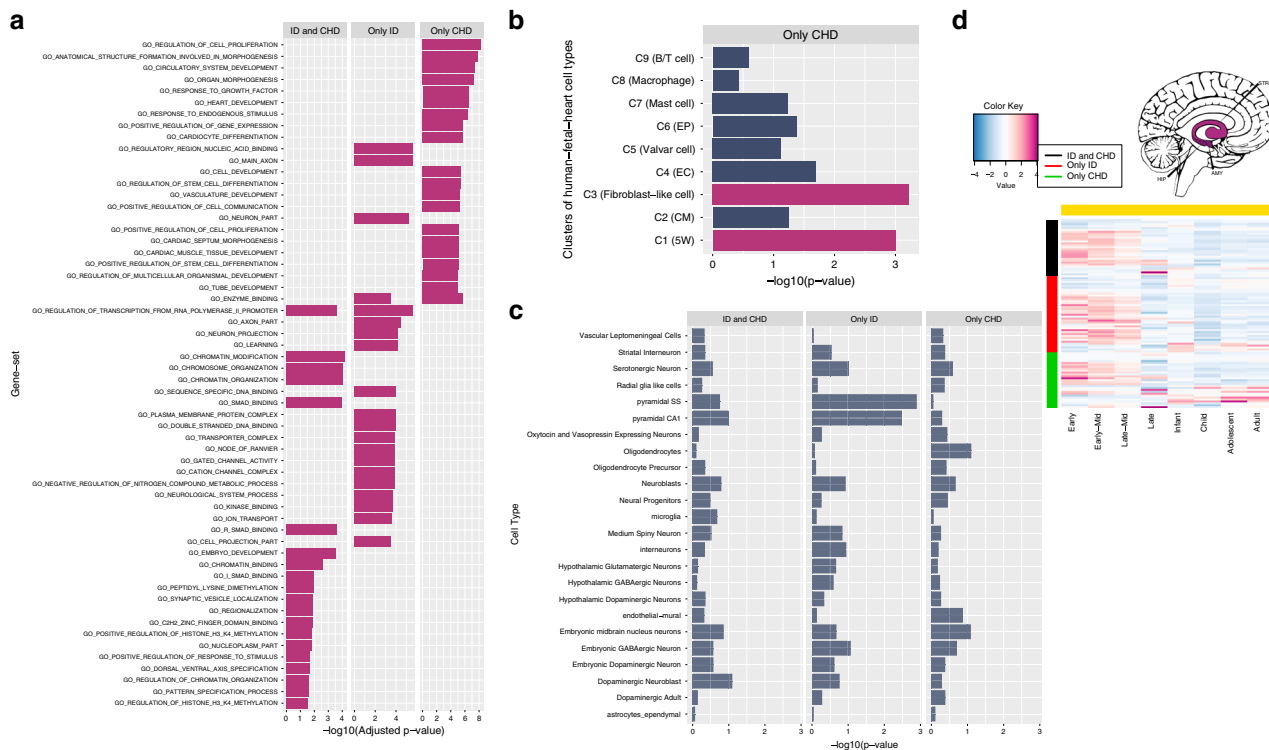


Fig. 6 The analysis results of shared and specific gene lists for ID and CHD (Only CHD: CHD-specific genes, Only ID: ID-specific genes, ID and CHD: shared genes). **a** Top enrichment results of gene-ontology (GO) gene sets. These are the top 20 enriched gene sets of each gene list. All these results have adjusted- p -value < 0.05 . **b** Enrichment results of human single-cell RNA sequencing (scRNAseq) datasets. These cells are from cardiac cells of the human fetal heart. They were clustered into 9 clusters (e.g., C1 to C9). The information of these clusters is in brackets (5W: 5-week hearts, ECs: endothelial cells, CMs: cardiomyocytes, epicardial cells: Eps). Magma-red bars are for results with adjusted p -value < 0.05 **c** Enrichment results of mouse scRNAseq expression data. **d** BrainSpan expression results for the three gene lists. This is for Region 3 as defined by Huckins, et al.³⁴ including hippocampus (HIP), amygdala (AMY), striatum (STR) regions. The package cerebroViz⁶⁷ was used to draw brain regions.

can set a prior or change a distribution for π_3 . Comorbid information might be used as prior information for π_3 in the analyses of mTADA. For example, our estimated gene-level genetic overlaps which are inferred from π_3 's estimations are very high for pairs of EE and ID (~31%, CI = 18.9–43.7%), ASD and ID (~38.1, CI = 30–46.6%). These three disorders are also highly comorbid^{46,47}; therefore, this information may be used as priors. However, genomic results and comorbid information might not always have the same trends. For example, the genetic overlapping information between ASD and SCZ are high in our study (~16%, CI = 5.1–31.4%), in previous common-variant based ($r_g = 0.16$, $se = 0.06$, $p = 0.0071$) and transcriptomic studies ($\rho \sim 0.5$, $p < 0.001$, Supplementary Table 2), but the comorbidity of the two disorders might not be strong or almost zero in some recent studies^{48,49}. In addition, ASD and SCZ can have overlapping copy number variable regions^{50,51}; however, duplications can be significantly seen for one disorder and deletions can highly present in the other disorder⁵¹. Finally, for all analyzed disorders, even though we observed multiple overlapping genes for pairs of disorders, the origin of these overlaps could be different. For example, for each pair, some overlapping genes could have more loss-of-function DNMs for one disorder and more missense damaging DNMs for the other disorder, and vice versa. Future rare-variant studies which are able to obtain comorbidity information from the overlapping samples and compare this with the genetic information will shed light on the genetic and clinical relationship of these disorders. Also, studies which are designed to understand in depth the information of variant categories for overlapping genes can elucidate the genomic mechanism of disorders.

Our analysis of de novo mutation data of neuropsychiatric disorders and CHD also has some limitations, in particular, overlapping phenotypes may lead to violation of mTADA assumptions. DD samples include people with different disorders⁵² and some of CHD samples may have other neuropsychiatric disorders²³. In a recent study, the DD dataset was combined with the ID dataset to create a larger ID dataset because of the high proportion of people with ID inside the DD cohort⁵³. In this study, even though we analyzed DD and ID separately to better understand the gene-level genetic overlaps between ID and other disorders, overlapping phenotypes may still affect the current results. We tested possible scenarios of overlapping phenotypes (Supplementary Note 1). The proportion of overlapping risk genes was modestly affected by ascertainment bias or by low percentages of misdiagnosed cases ($< 20\%$). However, this metric might be overestimated if misclassification rates were substantially high and the gene-level mean relative risks of one disorder were greatly different from those of the other disorder (Supplementary Note 1). The inflation could have an impact on analysis results, especially shared and specific risk genes. Nevertheless, mTADA always performed better than extTADA in the identification of risk genes for single-trait analyses in tested scenarios. It is possible that mTADA would benefit by jointly modeling these biases and this will be a future extension of the method.

With further development, the mTADA approach can be generalized further to consider more than two traits simultaneously, and the increased information could increase the number of identified risk genes but at a cost of increased computational time. Currently, the number of hypotheses

increases exponentially to 2^N with N being the number of traits. To reduce computational time, another approach which uses a small number of latent probability vectors⁵⁴ might be used for more than two-trait studies.

In conclusion, mTADA can be very useful for better understanding the genetic correlation across disorders (via the proportion of overlapping risk genes), and to prioritize additional risk genes for disorders. The approach of mTADA can be used to identify shared/specific risk genes for different categories of one trait (e.g., loss of function and missense de novo mutations). Genetic information of de novo mutations and rare case/control variants can be different⁵⁵, mTADA might be adopted to pipelines which are able to apply to DNMs and rare case/control variants as two traits.

Methods

mTADA: statistical models and parameter estimation. The mTADA is designed to jointly analyze two traits using DNMs. We use statistical models of extTADA, a single-trait method, to model DNM counts for each trait in mTADA as presented in Table 1. The likelihood of the data across all N genes can be computed as $L = \prod_{i=1}^N \sum_{j=0}^3 \pi_j P_{ij}^1 P_{ij}^2$ with $P_{ij}^k = P(x_{ki} | \phi_{kj})$, where x_{ki} and ϕ_{kj} are the i^{th} gene data and the j^{th} model's parameters for trait k ($k = 1, 2$). In addition, if the data include multiple categories of variants then $P_{ij}^k = \prod_{l=1}^{n_c} P_{ij}^{k,l}$ with n_c being the number of categories. For gene i , the statistical support for the j^{th} model is captured by its posterior probability ($PP_{ij} = \frac{\pi_j P_{ij}^1 P_{ij}^2}{\sum_{m=0}^3 \pi_m P_{im}^1 P_{im}^2}$, abbreviated as PP0, PP1, PP2 or PP3 for a gene).

We use our single-trait pipeline, extTADA, to estimate the proportions of risk genes (π_1^S and π_2^S), mean relative risks (\bar{y}_1^S and \bar{y}_2^S) and dispersion parameters (β_1^S and β_2^S) for each single trait (described as the superscript). We use these values inside mTADA: $\pi_1 = \pi_1^S - \pi_3$, $\pi_2 = \pi_2^S - \pi_3$, and $\pi_0 = 1 - (\pi_1^S + \pi_2^S - \pi_3)$ because of $\sum_{j=0}^3 \pi_j = 1$. We assume that $\bar{y}_1 = \bar{y}_1^S$, $\bar{y}_2 = \bar{y}_2^S$; $\beta_1 = \beta_1^S$ and $\beta_2 = \beta_2^S$. Therefore, we only estimate π_3 inside mTADA. Bayesian models are built using the *rstan* package⁵⁶. We use Markov Chain Monte Carlo (MCMC) within *rstan* to estimate π_3 . Convergence is diagnosed by the estimated potential scale reduction statistic (\hat{R}) and visualizing traces of results. The *Locfit* package⁵⁷ is used to obtain the mode, CI of π_3 . We use the mode as the estimated value of π_3 . We also tested a model with different mean relative risks for shared and specific risk genes. The model was more complex for the estimation process of parameters but did not improve the risk-gene identification. Therefore, this complex model was not used in our analysis.

Generation and analyses of simulated data. We simulated DNMs for genes under the mTADA model in Table 1. All 19,538 genes and their mutation rates from our current real dataset were used. A gene was assigned to one of the four groups (four models) by using the probability ($\pi_0, \pi_1, \pi_2, \pi_3$). We used $\pi_1^S = 0.05$ and $\pi_2^S = 0.03$ which are approximately equal to ASD, ID and DD results in our single-trait study¹⁰. π_3 was simulated with different values between 0 to min (π_1^S, π_2^S); and π_0, π_1 and π_2 were calculated as described in the section above. A range of mean relative risks were simulated for each of the two traits. Two mutation categories were simulated for each trait; therefore, there were four mean relative risks for the two traits. We used results from our previous studies^{10,11} and other studies^{58,59} for simulated values of mean relative risks. We simulated 100 values of each combination of π_3 and mean relative risks. We then calculated the mean of these 100 simulation results.

To calculate the proportion of false positive genes when there was not a genetic overlap between two tested disorders, we simulated different combinations of genetic parameters with $\pi_3 = 0$. For each PP threshold, we divided the number of identified overlapping genes by the total tested genes ($n = 19,358$ genes in our analysis).

We also used simulated data to assess the correlation between true and observed π_3 values and between PPs and oFDRs. An oFDR at a PP threshold was defined as the number of false positive genes divided by the number of identified genes. To use mTADA for single traits, for the i^{th} gene, we calculated $PP_{i1} + PP_{i3}$ and $PP_{i2} + PP_{i3}$ for the first and second trait respectively.

To compare risk gene classification performance between mTADA and extTADA on single traits, we used AUCs. We calculated true and false positive rates for extTADA and mTADA across PP thresholds, and calculated the areas under these ROC curves.

Real datasets of de novo mutations and variants. For primary analyses, we used the DNM data collected by Nguyen et al.¹⁰ and CHD data from Homsy et al.¹⁶. These data included 356 EE trios; 5,122 ASD trios; 4,293 DD trios; 1,012 ID trios; 1,077 SCZ trios; and 1,213 CHD trios. DNMs were annotated and classified into multiple categories as in our previous work¹⁰ as follows. For EE, ASD, DD, ID, and

CHD, we used two categories¹⁰: loss-of-function (LoF) and missense damaging (MiD) DNMs. The LoF category included nonsense, essential splice site, and frameshift DNMs defined by Plink/Seq⁶⁰ while the MiD category included DNMs annotated as missense by Plink/Seq and predicted damaging by each of seven methods⁴¹: SIFT, Polyphen2_HDIV, Polyphen2_HVAR, LRT, PROVEAN, MutationTaster, and MutationAssessor. For SCZ, we used LoF, MiD and synonymous mutations within DNase I hypersensitive sites because this category showed significant DNM enrichment in SCZ probands²² and non-null mean relative risk in extTADA¹⁰. Mutation rates were calculated as described by Fromer, et al.⁶⁰ and Nguyen, et al.¹⁰.

For the validation of mTADA's results and for better understanding the specific and shared risk genes between tested disorders, other datasets were used in the analysis. First, we used independent datasets to validate mTADA results. For CHD, we extracted variant data of 2,871 probands from Jin et al.²³. These samples include 2,445 trios (1,204 trios are inside the data set of extTADA and used in the primary analysis of this study) and 226 singletons²³. Only independent CHD samples were used in the validation process. For EE, we used the whole-genome-sequencing trio data of Hamdan et al.²⁷. This dataset includes 197 trios not included in our mTADA analyses. For SCZ, a case/control independent SCZ dataset from Genovese et al.⁴¹ was used. Disruptive and damaging ultra-rare variants from 4,877 cases and 6,203 controls were extracted from Table S3 of the study⁴¹.

Known risk-gene datasets. We extracted lists of known risk genes from two sources. 253 curated known human/mouse CHD genes were obtained from the supplementary data set 2 of Jin et al.²³. A list of EE genes from the Online Mendelian Inheritance in Man²⁶ was downloaded on September 02, 2019 using keywords "epileptic encephalopathy" and "epileptic encephalopathies".

Gene expression datasets. Human scRNAseq expression datasets of 4,000 cardiac cells were from 18 human embryos which ranged from 5 weeks (5W) to 25W of gestation. These were classified into four major cell types (cardiomyocytes (CMs), cardiac fibroblasts, endothelial cells (ECs), and valvar interstitial cells (VICs)), and also filtered and clustered into nice clusters. Gene lists of the nine clusters were extracted from Table S2 of Cui et al.³². scRNAseq transcriptome datasets were obtained from Skene et al.³³ via the link: http://www.hjerling-leffler-lab.org/data/scz_singlecell/ (Downloaded on August 01, 2018). These datasets included 9,970 single cells. These cells were clustered into 24 different cell types. Spatio-temporal transcriptome data were obtained from BrainSpan⁶¹, divided into eight developmental time points (four prenatal and four postnatal)⁶². The average expression at each spatiotemporal point was calculated across samples. For each gene, average expression values were standardized across spatiotemporal points to obtain z-scores^{10,34}. Z-scores were used for visualizing gene lists.

Analysis of de novo mutations using mTADA. extTADA was used to obtain the proportions of risk genes and the mean relative risks of each category for each disorder. These values were then used as input for mTADA to estimate π_3 and then to calculate PP_{ij} ($i = 1..N$, $j = 0..3$, $N = 19,358$ genes) for each pair of traits. The default algorithm, No-U-Turn Sampler (NUTS), in the *rstan* package was used to estimate π_3 . Two independent chains and 10,000 steps for each chain were used in the sampling process. Only 1,000 samples from each chain were chosen for further analyses.

For primary analysis, we applied mTADA to NPDs and 1,213 CHD trios. For understanding the specific and shared risk genes between tested disorders, we combined both tested and independent datasets of CHD (2,445 trios) in jointly analyzing with other disorders to increase power. Finally, we also applied mTADA to the two CHD datasets (tested and independent datasets) to test the performance of the method as described in Supplementary Note 2.

Other statistical methods for real data analyses. We used the EWCE package⁶³ to calculate the enrichment of our gene lists and the expression data from the 24 mouse cell types. To test the significance of the overlap of two gene sets, a permutation approach was used. We chose two random gene sets whose lengths are the same as the two tested gene sets from the background genes (19,358 genes from mTADA). This was carried out N times ($N = 10,000$ in this study) and the numbers of overlapping genes were recorded in a vector m . A p-value was calculated as $(\text{length}(m[m > m_0]) + 1) / (\text{length}(m) + 1)$ in which m_0 is the observed number of overlapping genes between the two tested gene sets. To conduct PPI analyses, we used the STRING database and the package STRINGdb²⁵ from the Bioconductor project⁶⁴, and p-values of protein-protein interactions were extracted from these analyses. To examine expression information of identified genes, we used the package mclust⁶⁵ to cluster BrainSpan gene expression data (z-scores) in heatmap analyses. To test the significance for individual genes from DNMs, we used a Poisson test. The R function `ppois(y - 1, lambda = 2 * Ntrio * mu, lower.tail = FALSE)` in which y and μ are the number of DNMs and the mutation rate of the tested gene; $Ntrio$ is the number of trios. All analyses were carried out using the R software⁶⁶.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All analyzed results are in Supplementary Data 1 and 2. These supplementary datasets are also available at: <https://github.com/hoangtn/mTADA>.

Code availability

mTADA package and examples for reproducible studies are available online (<https://github.com/hoangtn/mTADA>).

Received: 2 September 2018; Accepted: 6 May 2020;

Published online: 10 June 2020

References

- Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M. & Smoller, J. W. Pleiotropy in complex traits: challenges and strategies. *Nat. Rev. Genet.* **14**, 483–495 (2013).
- Zhernakova, A. et al. Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci. *PLoS Genet.* **7**, e1002004 (2011).
- Galesloot, T. E., van Steen, K., Kiemeny, L. A., Janss, L. L. & Vermeulen, S. H. A comparison of multivariate genome-wide association methods. *PLoS ONE* **9**, e95923 (2014).
- Allison, D. B. et al. Multiple phenotype modeling in gene-mapping studies of quantitative traits: power advantages. *Am. J. Hum. Genet.* **63**, 1190–1201 (1998).
- Pickrell, J. K. et al. Detection and interpretation of shared genetic influences on 42 human traits. *Nat. Genet.* **48**, 709–717 (2016).
- Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
- Turley, P. et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat. Genet.* **50**, 229–237 (2018).
- Lutz, S. M., Fingerlin, T. E., Hokanson, J. E. & Lange, C. A general approach to testing for pleiotropy with rare and common variants. *Genet. Epidemiol.* **41**, 163–170 (2017).
- Guo, B. & Wu, B. Integrate multiple traits to detect novel trait-gene association using GWAS summary data with an adaptive test approach. *Bioinformatics* **35**, 2251–2257 (2019).
- Nguyen, H. T. et al. Integrated Bayesian analysis of rare exonic variants to identify risk genes for schizophrenia and neurodevelopmental disorders. *Genome Med.* **9**, 114 (2017).
- He, X. et al. Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. *PLoS Genet.* **9**, e1003671 (2013).
- De Rubeis, S. et al. Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209–215 (2014).
- Iossifov, I. et al. The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216–221 (2014).
- Li, J. et al. Genes with de novo mutations are shared by four neuropsychiatric disorders discovered from NPdenovo database. *Mol. Psychiatry* **21**, 290 (2016).
- Hoischen, A., Krumm, N. & Eichler, E. E. Prioritization of neurodevelopmental disease genes by discovery of new mutations. *Nat. Neurosci.* **17**, 764 (2014).
- Homsy, J. et al. De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. *Science* **350**, 1262–1266 (2015).
- Willsey, A. J. et al. The psychiatric cell map initiative: a convergent systems biological approach to illuminating key molecular pathways in neuropsychiatric disorders. *Cell* **174**, 505–520 (2018).
- Lelieveld, S. H. et al. Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. *Nat. Neurosci.* **19**, 1194–1196 (2016).
- Deciphering Developmental Disorders Study. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**, 433–438 (2017).
- Wang, S. et al. De novo sequence and copy number variants are strongly associated with tourette disorder and implicate cell polarity in pathogenesis. *Cell Rep.* **24**, 3441–3454 e12 (2018).
- Cappi, C. et al. De novo damaging DNA coding mutations are associated with obsessive-compulsive disorder and overlap with tourette's disorder and autism. *Biol. Psychiatry* <https://doi.org/10.1016/j.biopsych.2019.09.029> (2019).
- Takata, A., Ionita-Laza, I., Gogos, J. A., Xu, B. & Karayiorgou, M. De novo synonymous mutations in regulatory elements contribute to the genetic etiology of autism and schizophrenia. *Neuron* **89**, 940–947 (2016).
- Jin, S. C. et al. Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. *Nat. Genet.* **49**, 1593–1601 (2017).
- Heyne, H. O. et al. De novo variants in neurodevelopmental disorders with epilepsy. *Nat. Genet.* **50**, 1048–1053 (2018).
- Szklarczyk, D. et al. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* **45**, D362–D368 (2017).
- Amberger, J. S., Bocchini, C. A., Scott, A. F. & Hamosh, A. OMIM.org: leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Res.* **47**, D1038–D1043 (2019).
- Hamdan, F. F. et al. High rate of recurrent de novo mutations in developmental and epileptic encephalopathies. *Am. J. Hum. Genet.* **101**, 664–685 (2017).
- Fromer, M. et al. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am. J. Hum. Genet.* **91**, 597–607 (2012).
- Zhang, B. et al. Association study identifying a new susceptibility gene (AUTS2) for schizophrenia. *Int. J. Mol. Sci.* **15**, 19406–19416 (2014).
- Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
- Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
- Cui, Y. et al. Single-cell transcriptome analysis maps the developmental track of the human heart. *Cell Rep.* **26**, 1934–1950 e5 (2019).
- Skene, N. G. et al. Genetic identification of brain cell types underlying schizophrenia. *Nat. Genet.* **50**, 825–833 (2018).
- Huckins, L. M. et al. Gene expression imputation across multiple brain regions provides insights into schizophrenia risk. *Nat. Genet.* **51**, 659–674 (2019).
- Kichaev, G. et al. Improved methods for multi-trait fine mapping of pleiotropic risk loci. *Bioinformatics* **33**, 248–255 (2017).
- Maier, R. M. et al. Improving genetic prediction by leveraging genetic correlations among human diseases and traits. *Nat. Commun.* **9**, 989 (2018).
- Ware, J. S., Samocha, K. E., Homsy, J. & Daly, M. J. Interpreting de novo variation in human disease using denovolyzeR. *Curr. Protoc. Hum. Genet.* **87**, 7.25.1–15 (2015).
- White, J. et al. POGZ truncating alleles cause syndromic intellectual disability. *Genome Med.* **8**, 3 (2016).
- Stessman, H. A. F. et al. Disruption of POGZ is associated with intellectual disability and autism spectrum disorders. *Am. J. Hum. Genet.* **98**, 541–552 (2016).
- Ben-Shalom, R. et al. Opposing effects on NaV1.2 function underlie differences between SCN2A variants observed in individuals with autism spectrum disorder or infantile seizures. *Biol. Psychiatry* **82**, 224–232 (2017).
- Genovese, G. et al. Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. *Nat. Neurosci.* **19**, 1433–1441 (2016).
- Allen, A. S. et al. Ultra-rare genetic variation in common epilepsies: a case-control sequencing study. *Lancet Neurol.* **16**, 135–143 (2017).
- Sanders, S. J. et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237 (2012).
- Cross-Disorder Group of the Psychiatric Genomics Consortium. et al. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat. Genet.* **45**, 984–994 (2013).
- Gandal, M. J. et al. Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. *Science* **359**, 693–697 (2018).
- Amiet, C. et al. Epilepsy in autism is associated with intellectual disability and gender: evidence from a meta-analysis. *Biol. Psychiatry* **64**, 577–582 (2008).
- Tuchman, R. & Cuccaro, M. Epilepsy and autism: neurodevelopmental perspective. *Curr. Neurol. Neurosci. Rep.* **11**, 428–434 (2011).
- Volkmar, F. R. & Cohen, D. J. Comorbid association of autism and schizophrenia. *Am. J. Psychiatry* **148**, 1705–1707 (1991).
- Leyfer, O. T. et al. Comorbid psychiatric disorders in children with autism: interview development and rates of disorders. *J. Autism Dev. Disord.* **36**, 849–861 (2006).
- Kushima, I. et al. Comparative analyses of copy-number variation in autism spectrum disorder and schizophrenia reveal etiological overlap and biological insights. *Cell Rep.* **24**, 2838–2856 (2018).
- Crespi, B., Stead, P. & Elliot, M. Evolution in health and medicine Sackler colloquium: comparative genomics of autism and schizophrenia. *Proc. Natl Acad. Sci. USA* **107** Suppl 1, 1736–1741 (2010).
- DDD Study. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**, 433–438 (2017).
- Taylor, J. L. et al. Paternal-age-related de novo mutations and risk for five disorders. *Nat. Commun.* **10**, 3043 (2019).
- Wei, Y., Tenzen, T. & Ji, H. Joint analysis of differential gene expression in multiple studies using correlation motifs. *Biostatistics* **16**, 31–46 (2015).

55. Sifrim, A. et al. Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing. *Nat. Genet.* **48**, 1060–1065 (2016).
56. Carpenter, B. et al. Stan: a probabilistic programming language. *J. Stat. Softw.* **20**, 1–37 (2016).
57. Loader, C. Locfit: local regression, likelihood and density estimation. *R package version 1* (2007).
58. Willsey, A. J. et al. De novo coding variants are strongly associated with tourette disorder. *Neuron* **94**, 486–499 e9 (2017).
59. Epi K. Consortium. et al. De novo mutations in epileptic encephalopathies. *Nature* **501**, 217–221 (2013).
60. Fromer, M. et al. De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179–184 (2014).
61. Miller, J. A. et al. Transcriptional landscape of the prenatal human brain. *Nature* **508**, 199–206 (2014).
62. Lin, G. N. et al. Spatiotemporal 16p11.2 protein network implicates cortical late mid-fetal brain development and KCTD13-Cul3-RhoA pathway in psychiatric diseases. *Neuron* **85**, 742–754 (2015).
63. Skene, N. G. & Grant, S. G. Identification of vulnerable cell types in major brain disorders using single cell transcriptomes and expression weighted cell type enrichment. *Front. Neurosci.* **10**, 16 (2016).
64. Gentleman, R. C. et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004).
65. Scrucca, L., Fop, M., Murphy, T. B. & Raftery, A. E. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *R. J.* **8**, 289–317 (2016).
66. R Core Team. R: A Language and Environment for Statistical Computing. (Vienna, Austria, 2018).
67. Bahl, E., Koomar, T. & Michaelson, J. J. cerebroViz: an R package for anatomical visualization of spatiotemporal brain data. *Bioinformatics* **33**, 762–763 (2017).

Acknowledgements

We thank Silviu-Alin Bacanu and Kenneth S. Kendler for useful discussions. This work is supported by NIH grant R01MH105554 to E.A.S., by NIH grant R01MH110555 to D.P., by NARSAD grant 28599 to T.H.N., by R01MH110531 to X.H., by K08 HD092610 to R.C.M., by U01MH111661 and R01MH097849 to J.D.B. The Sweden exome sequencing data generation and analysis are supported by the Stanley Center for Psychiatric Research and NIH grant R01 MH077139 to P.F.S. This work was supported in part through the computational resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai. We are deeply grateful for the participation of all subjects contributing to this research.

Author contributions

Designed the pipeline used in analysis; performed the experiments, analyzed the data and drafted the manuscript: T.H.N.; conceived and designed the experiments: T.H.N., X.H., E.A.S.; contributed reagents/materials/analysis tools: T.H.N., A.D., B.P.R., D.P., J.D.B., X.H., P.F.S., E.A.S.; wrote the paper: T.H.N., A.D., P.F.S., X.H., E.A.S., R.C.B.

Competing interests

P.F.S. reports the following potentially competing financial interests: Lundbeck (advisory committee, grant recipient), Pfizer (Scientific Advisory Board), Element Genomics (consultation fee), and Roche (speaker reimbursement). The remaining authors declare that they have no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41467-020-16487-z>.

Correspondence and requests for materials should be addressed to T.-H.N., X.H. or E.A.S.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contributions to the peer review of this work.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020