Prioritizing risk genes for neurodevelopmental disorders using pathway information

Author List:
Hoang T. Nguyen[1*], Amanda Dobbyn[1,2], Alex Charney[1], Julien Bryois[3], Nathan G. Skene[4], Laura M. Huckins[1], Weiqing Wang[1], Douglas M Ruderfer[5], Xinyi Xu[6], Menachem Fromer[7], Shaun M Purcell[8], Matthijs Verhage[9], August B. Smit[10], Jens Hjerling-Leffler[4], Joseph D. Buxbaum[6], Dalila Pinto[6,11,12], Xin He[13], Patrick F Sullivan[14], Eli A. Stahl[1,15*]

Author Affiliations:

1. Division 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.
2. Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA.
3. Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden.
4. Laboratory of Molecular Neurobiology, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden.
5. Division of Genetic Medicine, Departments of Medicine, Psychiatry and Biomedical Informatics, Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, USA.
6. Seaver Autism Center, Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA.
7. Verily Life Sciences, 269 E Grand Ave, South San Francisco, CA.
8. Sleep Center, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA.
9. Department of Functional Genomics, The Center for Neurogenomics and Cognitive Research, VU University and VU Medical Center, Amsterdam, The Netherlands.
10. Department of Molecular and Cellular Neurobiology, The Center for Neurogenomics and Cognitive Research, VU University, Amsterdam, The Netherlands.
11. The Mindich Child Health & Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA.
12. Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA.
13. Department of Human Genetics, University of Chicago, Chicago, IL, USA.
14. Departments of Genetics and Psychiatry, University of North Carolina, Chapel Hill, North Carolina, USA.
15. Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA.

*Corresponding author: tan-hoang.nguyen@mssm.edu and eli.stahl@mssm.edu, Division 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.

ABSTRACT

Over the past decade, case-control studies of next-generation sequencing data have proven integral to understanding the contribution of rare inherited and *de novo* single-nucleotide variants to the genetic architecture of complex disease. Ideally, such studies would identify individual risk genes of moderate to large effect size to generate novel treatment hypotheses for further follow-up. However, due to insufficient power, gene set enrichment analyses have come to be relied upon for detecting differences between cases and controls, implicating sets of hundreds of genes rather than specific targets for further investigation. Here, we present a Bayesian statistical framework, termed gTADA, that integrates gene-set membership information with gene-level *de novo* (DN) and rare inherited case-control (rCC) counts to prioritize risk genes with excess rare variant burden. With this pipeline, arbitrary significance thresholds can be circumvented. Our method can leverage external gene-level information to identify additional risk genes. Applying gTADA to available whole-exome sequencing datasets for several neuropsychiatric conditions, we replicate previously reported gene set enrichment and identify novel risk genes. For epilepsy, gTADA prioritized 40 significant genes, of which 30 are not in the known gene list (posterior probabilities > 0.95) and 6 replicate in an independent whole-genome sequencing study. We found that epilepsy genes have high protein-protein interaction network connectivity, and their expression during human brain development. Finally, epilepsy risk genes are enriched for the targets of several drugs, including both known anticonvulsants and potentially novel repositioning opportunities.

INTRODUCTION

De novo mutations (DNMs) have been successfully used to identify genes associated with neurodevelopmental disorders (NDDs) [1-8]. Recently, additional risk genes have been reported by meta-analyzing DNMs and rare inherited/case-control (rCC) variants, an approach that has been particularly successful for autism spectrum disorders (ASD) [9, 10]. For epilepsy (EPI), multiple associated genes have been identified through DN based studies [4, 5, 11], and in recent years, a number of EPI significant genes have also been identified through CC studies [12, 13]. We hypothesized that, as in the case of ASD, additional significant EPI genes could be discovered through the integration of DN and CC data. EPI is a serious brain disorder which includes multiple subtypes. Studies of cases/controls and twins have shown that genetic components have played important roles in EPI [14-16]. Some of EPI's subtypes can be explained by single genes, but multiple subtypes might be caused by multiple genes [15]. It is still challenging to develop specific drugs for this disorder. There have been multiple antiepileptic drugs used for EPI treatments; however, 20-30% of EPI patients have not been successful in controlling their seizures by using current medications [17]. Identifying additional genes or gene sets might help better understand its etiology as well as better design drug targets for the disorder.

Due to the high polygenicity of these phenotypes, gene set (GS) based tests have also been used to identify specific pathways relevant to disease etiology [18-23]. A typical approach is that top significant genes are identified from the analyzed data, and then these genes are tested in established sets and pathways. Although GS enrichment can be tested directly from either DN data [20] or rCC data [21], testing for enrichment in meta-analysis of DN and rCC variants is challenging. One straightforward approach is to test for GS enrichment using the summary statistics resulting from meta-analysis; however, a drawback of this approach is that it does not

account for heterogeneity between DN and rCC data types [9]. We here propose an alternative method that circumvents this issue by jointly modeling CC/DN variant and gene set membership information. Through the integration of rare variant genotype data with gene set information, we show the increased ability in the identification of risk genes.

In this work, we introduce a method that tests gene-set enrichment directly from DN and rCC data, and leverages enriched gene sets (eGSs) to prioritize risk genes. This approach allows genes to be prioritized if they are in enriched GSs/tissues, for a given strength of genetic evidence. This pipeline can be used for discrete or continuous gene-set data, and therefore it can incorporate gene expression data to obtain additional significant genes based on tissue or cell-type expression information. It is a **g**eneralized framework of our extended **T**ransmission **A**nd **D**e novo **A**ssociation, gTADA. We demonstrate through multiple simulations that this pipeline is an improvement upon extTADA [18] which only integrates DN and rare CC information in identifying significant genes. We apply gTADA to large DN and rCC variant data sets, incorporating neuropsychiatric candidate gene sets, drug-target gene sets, and GTEx expression data in order to prioritize NDD and CHD genes. With recent large rare CC data sets of EPI, we further analyze results of this disorder. We identify multiple significant EPI genes, and validate top genes in an independent data set. We provide further support for our significant genes through the analysis of expression data and protein-protein interaction networks.

RESULTS

The gTADA pipeline

We have developed a pipeline to integrate DN mutations, rare CC variants, and pathway/gene-set information to prioritize significant genes (Figure 1). This pipeline could also be used for gene expressions in place of gene sets. To simplify, the term 'gene-set' (GS) rather than 'gene-set/pathway and gene expression' is used in this manuscript.

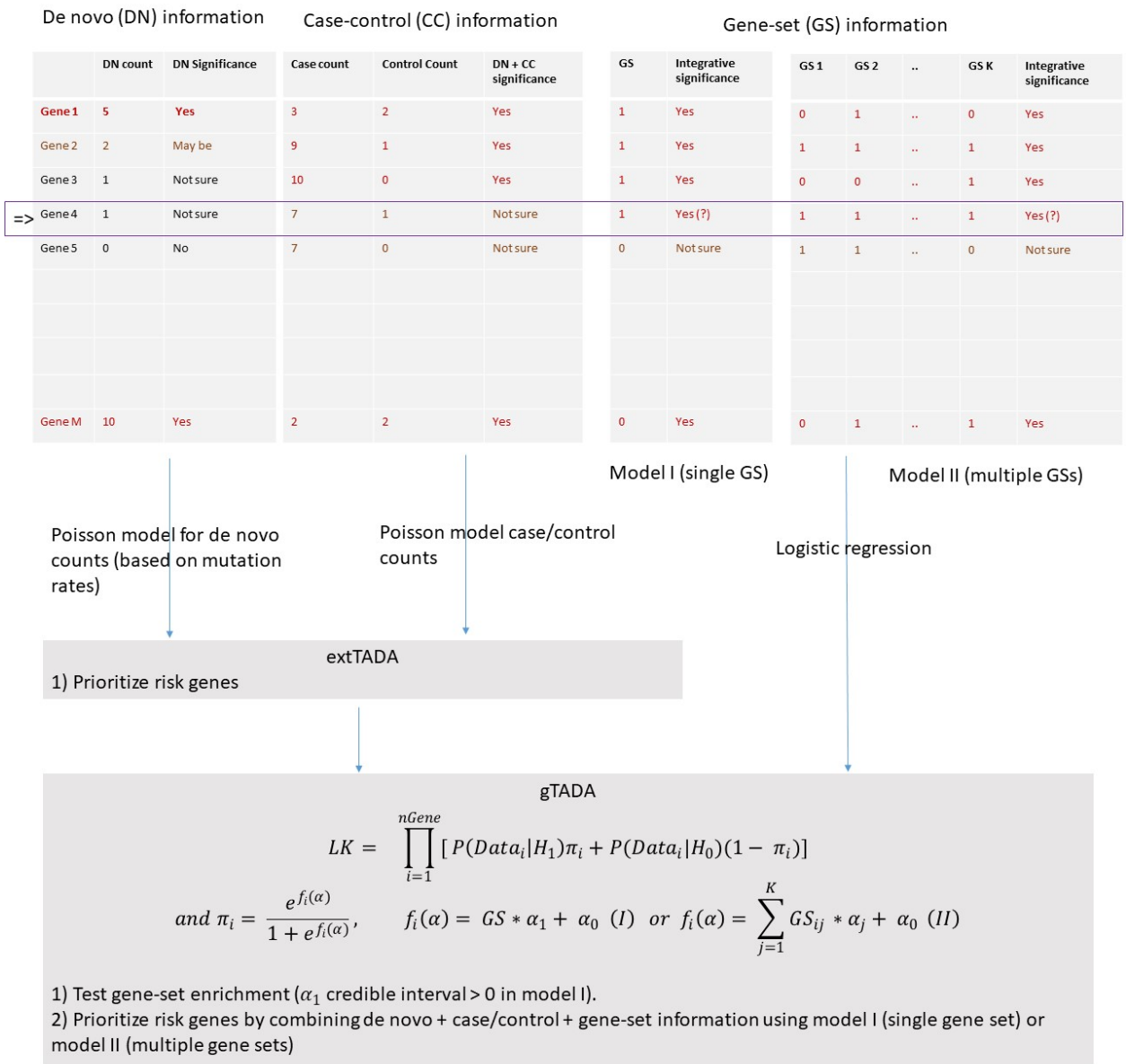
The pipeline employed the **T**ransmission **A**nd **D**enovo **A**ssociation test (TADA [9] and extTADA [18]) to model and integrate DN and CC data, and then combined gene-set information by using a logistic regression model (Figure 1). This means this pipeline could be considered a generalized framework of extTADA and TADA [9, 18]. To summarize, with each gene, for each variant category, all variants were collapsed and considered as a single count (x). Table S1 presents the details of statistical models of the counts, their parameters, and the hyper parameters of DN and CC data. For each gene, gTADA compared two hypotheses: risk-gene (H_1) and non-risk gene (H_0). Similar to TADA, our model assumes that rare variant counts in a risk gene are elevated by γ fold, comparing with chance expectation, and $\gamma \sim Gamma(\bar{\gamma} * \beta, \beta)$. For non-risk gene, γ = 1. We assumed that there was a probability π_i for the ith gene to be a risk gene. This π_i was connected to a GS by $\pi_i = e^{f_i(\alpha)}/(1 + e^{f_i(\alpha)})$, and $f_i(\alpha) = \alpha_0 + GS_i * \alpha_1$, GS_i in the function is the value of the GS at the ith gene which can be 0/1 or a continuous value. This was very different from TADA and extTADA in which π_i was assumed to be the same across genes.

The likelihood for the data was $P(x|parameters) = P(x|H_1)\pi_i + P(x|H_0)(1 - \pi_i)$. All

parameters, and hyper parameters of gene data as well as α_j (j = 0..1) were jointly estimated from the likelihood function across the whole genes. Similar to extTADA, if there were multiple categories or population samples, their parameters would be also jointly estimated inside this step. The main model for testing GS enrichment and prioritizing significant genes was the single-GS model.

We used a Markov Chain Monte Carlo (MCMC) method to sample parameters. Modes which were considered as the estimated values, and Bayesian credible intervals (CIs) of MCMC results were used in all the inferences. A GS was considered enriched if the lower boundary of its α CI was positive.

One of the advantages of gTADA is that after learning gene sets, it can use that knowledge to increase the power of finding risk genes, because genes in the enriched GS will have higher prior probabilities [24]. We used posterior probabilities (PPs) to prioritize risk genes with $PP_i = \frac{P(x|H_1)\pi_i}{P(x|H_1)\pi_i + P(x|H_0)(1-\pi_i)}$ for the ith gene. After testing multiple gene sets, we reported genes with PP>0.95 with any gene set as significant gTADA genes, and also conducted follow-up analyses on significant and suggestive genes with PP>0.8 with any gene set. We conducted simulation analyses based on real data, to show the effect of taking the union of results across gene sets.



	DN count	DN Significance	Case count	Control Count	DN + CC significance	GS	Integrative significance	GS 1	GS 2	..	GS K	Integrative significance
Gene 1	5	Yes	3	2	Yes	1	Yes	0	1	..	0	Yes
Gene 2	2	May be	9	1	Yes	1	Yes	1	1	..	1	Yes
Gene 3	1	Not sure	10	0	Yes	1	Yes	0	0	..	1	Yes
=> Gene 4	1	Not sure	7	1	Not sure	1	Yes (?)	1	1	..	1	Yes (?)
Gene 5	0	No	7	0	Not sure	0	Not sure	1	1	..	0	Not sure
Gene M	10	Yes	2	2	Yes	0	Yes	0	1	..	1	Yes

Figure 1: The framework of gTADA. The pipeline combines de novo (DN), case/control (CC) data (via variant counts of genes) and gene set (GS) information. It can test the enrichment of GS directly from the data (use α_1 information from single-GS model), and prioritize risk genes using model I (single GS) or model II (multiple GSs). For example, Gene 4 might have a small posterior probability (PP) to be a risk gene because it does not have strong genetic information; however, the gene's PP would be high when it is supported by GS information from eGSs.

Results of simulated data

We have simulated different gene sizes using genetic parameters from previous ASD studies [9, 18]. Enriched gene sets (eGSs) were simulated by using the results of known enriched gene sets

for ASD [18] (see Methods). Non-enriched gene sets were simulated by randomly choosing genes from the whole genes. Different trio numbers were used in the simulation process ranging from 1,000 to 50,000. Genetic parameters of simulated data are presented in Table S2 (See Methods). To compare the results between gTADA and extTADA, we calculated all gene counts and true gene counts for different PP thresholds 0.95 and 0.8.

We tested for both models: single gene sets and multiple gene sets (Supplementary Information). For single-GS model, the number of risk genes increased when eGSs were used (Figure S1). In addition, the Type I errors of calling non-eGSs as eGSs were also well calibrated (Table S3). For multiple-GS model, the number of risk genes increased when eGSs' numbers increased. However, with larger numbers of gene sets, we observed higher numbers of false positive results for identified genes in some simulations with small sample sizes (Figure S2), suggesting that FDR is not well controlled when multiple genes are modeled simultaneously. For this reason, we focused our analyses on single gene set models, and combined sets of significant or suggestive genes across single gene set models.

Application of gTADA to exome sequence variants in NDDs and CHD

We applied gTADA to available rare variant data of four NDDs and CHD to obtain top prioritized genes for these disorders (Figure 2). In summary, this data included 4293, 1012, 1213, 5122 and 356 trios of DD, ID, CHD, ASD and EPI; plus 4058 and 5704 cases/controls of ASD and EPI. These data were annotated and divided into different categories by using the approach of [25]. We used loss-of-function (LoF) and missense damaging (MiD) categories of these annotation. For EPI case/control, we only used count data from [13] which were annotated by the authors (Details in the Method).

To test GS enrichment and to prioritize genes, we only used the single-GS model in real data sets because FDRs could increase if multiple gene sets with different sizes were used (Figure S2). Similar to simulated data, a GS was called enriched if its α_1's lower boundary was positive. We also calculated p values for GSs, and a GS was called significantly enriched if its adjusted p value was < 0.05 and its α_1's lower boundary was positive. To identify significant genes for each eGS, we set a stringent PP threshold of 0.95. We also examined the properties of sets of prioritized genes having PPs > 0.8. Each gene had multiple PPs from multiple GSs; therefore, the PP of the gene was the maximum PP of all GSs.

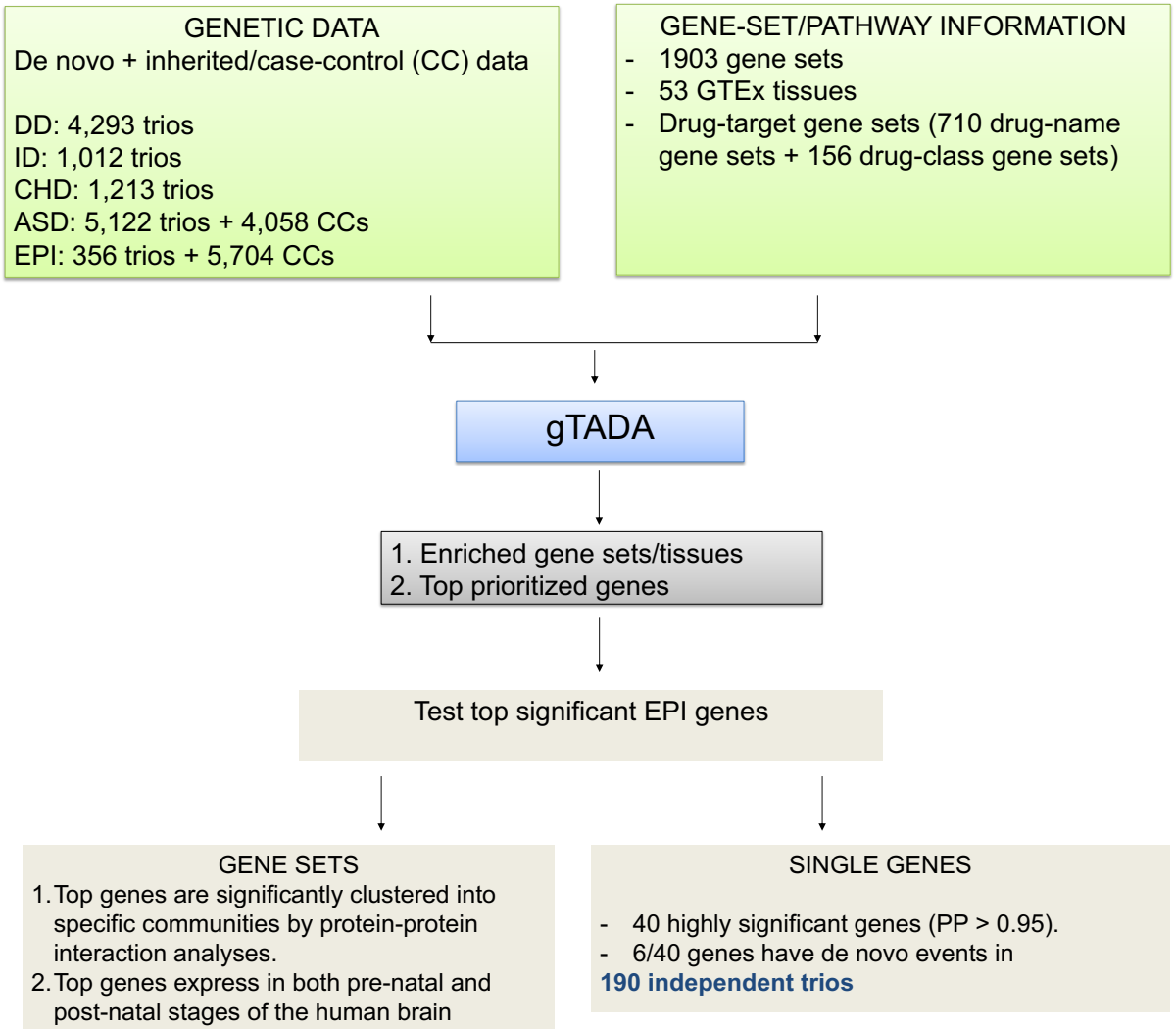


Figure 2: Whole data analysis in the study. Four neurodevelopmental disorders (NDDs) and congenital heart disease (CHD) are analyzed. Results of epilepsy (EPI) are validated by using different methods and an independent data set.

Results of gene set enrichments

We tested 1,903 GSs used in our previous study [25], including 186 candidate gene sets, and 1,717 gene sets with 100 to 4,955 genes from MSigDB [26] and the Gene Ontology data base [27] (Table S4). For each disorder, we did not use its known risk genes or its DNM gene set to avoid inflating identified genes; therefore, the number of tested GSs were slightly different between disorders as presented in Table 1.

gTADA identified multiple eGSs for all disorders (Table 1). All gTADA GS enrichment results are presented in Table S5. For ASD, DD and ID, we compared these results with gene set enrichments in extTADA results reported in our previous study [25]. To compare directly with the

previous results, we calculated and adjusted p values for the 186 GSs. We observed high correlations between gTADA adjusted p values and our previous methods (rho > 0.75, p ~ 0, Figure S3). gTADA was able to re-call 100% of significant gene sets identified by permutation, and > 89% seGS reported by the PP-based method (Figure S3).

We combined all eGSs to prioritize risk genes for each disorder. Based on PP>0.95, DD had the highest number of genes (167) followed by ASD, ID and EPI (64, 59 and 40 respectively) (Table 1). Both SCZ and CHD had only 12 prioritized genes. These combined gene counts were much higher than extTADA's results (Table 1). For PP > 0.8, the gene-count differences between gTADA and extTADA were >= 64 for all four NDDs. Full prioritized genes are presented in Table S6. One gene, STXBP1, was observed across four NDDs with PP > 0.95. In addition, 18 genes (PP > 0.95) were in at least three disorders (Table S6).

To better understand the performance of gTADA on each eGS individually, we chose top ten eGSs from each disorder based on α_1's estimated values and compared the gene-count results of gTADA and extTADA using a threshold PP>0.95. For each eGSs, the majority of gene-count differences were less than 5 (Figure 3). The largest differences were for the gene set of ID DNMs: 19 and 9 genes for ASD and DD respectively.

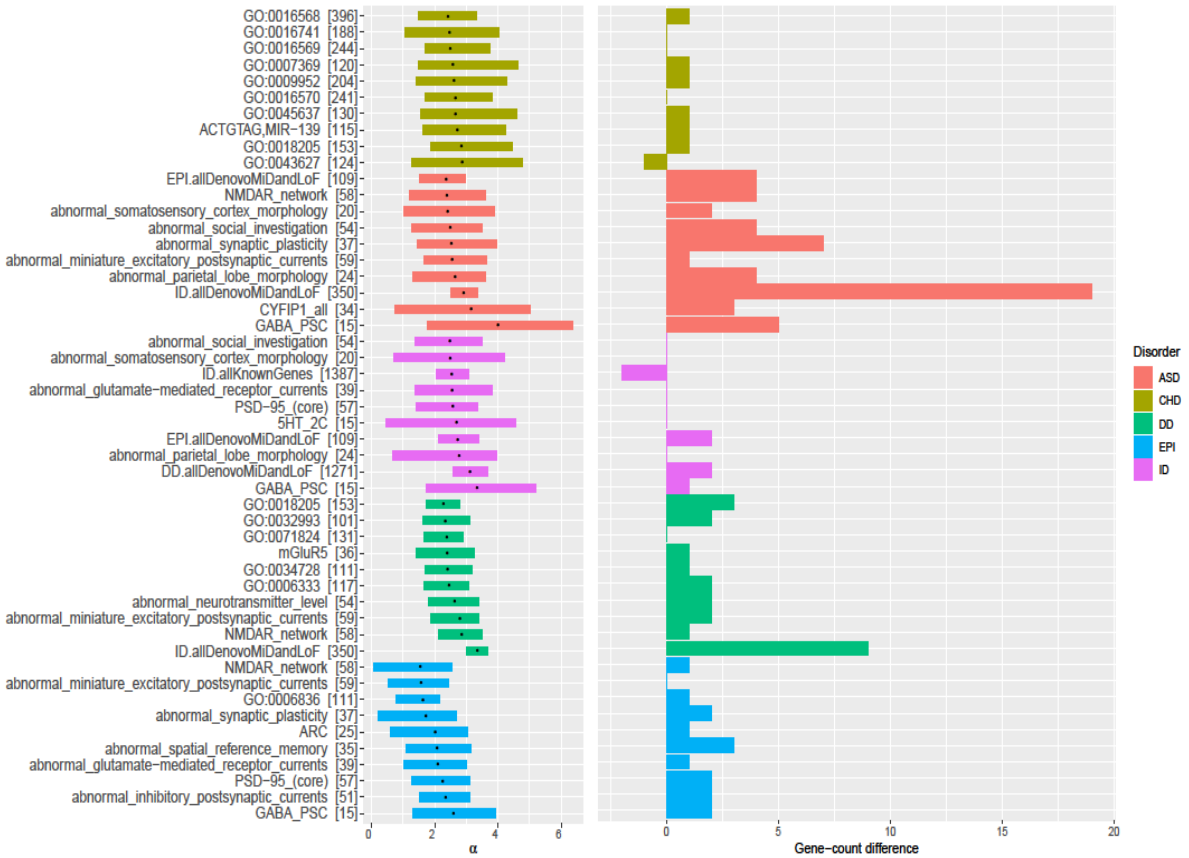


Figure 3: Top enriched gene sets (eGSs) of disorders. These are top ten eGSs of the analyzed disorders (based on α_1's estimated values). Y-axes are names of the eGSs and their sizes (e.g, GO:0016568 has 396 genes). The left picture shows α_1's credible intervals and modes of eGSs.

The right picture describes the differences in gene counts (posterior probabilities > 0.95) between using GSs (gTADA) and not using GSs (extTADA).

Results of tissue enrichments using GTEx data

We applied gTADA to expression data of 53 tissues from GTEx Consortium [28]. Only 6 tissues were enriched for EPI while 28 issues were enriched for ASD (Table 1). Interestingly, >= 50 tissues showed enrichment for ID, DD and CHD. Six brain tissues were significantly enriched across CHD and four NDDs ($p < 0.05$ and low CI > 0) (Brain-Amygdala, Brain-Anterior cingulate cortex BA24), Brain-Caudate (basal ganglia), Brain-Frontal Cortex (BA9), Brain-Hippocampus, Brain-Nucleus accumbens (basal ganglia)). All enrichment results are presented in Table S7 and in Figure S4.

Similar to the results of candidate gene sets above, gTADA prioritized more risk genes than extTADA with different thresholds (Table 2). However, the risk-gene numbers were not as high as results from candidate GSs. One possible reason was that the estimated α_1 values of GTEx were not high (less than 1, Table S7) because we used continuous values of the whole genes. Another reason was that there were only 53 tissues while there were nearly 2000 GSs analyzed above. The gene STXBP1 above was also in the list of genes with PP > 0.95 for ID, DD and EPI, but only in the list with PP > 0.8 for ASD. There were 11 genes with PP > 0.95 in at least three disorders. All these 11 genes were inside the 15 genes which were reported above (Table S8).

Results of drug-target GS enrichment

Two types of drug-target gene sets are used in this study. Details of these drug-target gene sets were described in [29]. Briefly, these drug targets were predicted by using the Similarity Ensemble Approach (SEA) [30]; data from DrugBank version 4.1 [31] and ChEMBL-14. These gene sets are based on the Anatomical Therapeutic Chemical (ATC) classification system. We used GSs from Level 3 and Level 5 of this system. To distinguish between two types of GSs, we used the drug-name and drug-class GSs for the Level 5-based and the Level 3-based types of drug-target GSs respectively. There were 710 and 156 drug-name and drug-class GSs respectively.

We first tested for GS enrichment. Regarding the drug-name GSs, gTADA identified multiple eGS (CIs > 0) for all disorders, except for CHD. EPI had the highest number of eGSs (88) followed by DD, ID and ASD with 39, 36 and 31 eGSs respectively (Table 1). There were four eGSs across four NDDs: cyproheptadine, enflurane, ketanserin, zuclopenthixol; and 21 eGSs across at least 3 NDDs. Regarding seGS numbers, EPI still had a high number of seGSs (67) while DD and ID had only 2 and 3 seGS respectively (Table S9). Next, we tested the enrichment of 156 drug-class GSs. EPI had 18 eGS while DD, ID, ASD and CHD had only 4, 8, 3 and 2 eGS respectively. However, after correcting p values, these numbers were only 13 and 4 for EPI and ID; and no enriched GS was for other disorders. Two common drug classes (ANTIEPILEPTICS and ANTIPROPULSIVES) were observed in enrichment results of both EPI and ID (Table S10).

We then prioritized EPI genes using enriched drug-target GSs. The majority of gene-count results were higher than those of GTEx based gTADA, but lower than those of candidate-GS based gTADA (Table 1). No gene was observed across four NDDs with PP > 0.95. The gene STXBP1 was also in EPI, DD, ID with PP > 0.95, but it was not in the list of ASD genes with PP > 0.8 as GTEx based results (Table S11). Regarding drug-class GSs, gene counts of gTADA were higher

than those of extTADA in all disorders (Table 1, Table S12). However, these gene-count results were not as high as results from drug-name eGSs or candidate eGSs. One possible reason was that the numbers of eGSs from this analysis were not as high as those from two other analyses (Table 1).

Disorder	Test gTADA on different gene sets/tissues												Results: number of prioritized genes									
													extTADA		gTADA							
	Candidate GS			GTEx tissue			Drug-target GS			Drug-target (class) GS					Candidate GS		GTEx tissue		Drug-name GS		Drug-class GS	
	nGS	nsGS	nsGS*	nGS	nsGS	nsGS*	nGS	nsGS	nsGS*	nGS	nsGS	nsGS*	PP > 0.95	PP > 0.8	PP > 0.95	PP > 0.8	PP > 0.95	PP > 0.8	PP > 0.95	PP > 0.8	PP > 0.95	PP > 0.8
ASD	1901	381	338	53	29	28	710	31	0	156	3	0	24	51	64	191	33	59	36	64	31	57
ID	1901	495	485	53	52	52	710	36	2	156	8	4	43	51	59	177	45	74	48	64	46	61
DD	1901	686	679	53	53	53	710	39	3	156	4	0	109	148	167	288	129	198	116	160	114	152
EPI	**1901**	**108**	**50**	**53**	**9**	**6**	710	**88**	**67**	156	**18**	**13**	**24**	**63**	**40**	**135**	**26**	**82**	**36**	**101**	**34**	**96**
CHD	1902	280	241	53	50	50	710	0	0	156	2	0	3	11	12	101	6	16	0	0	6	16

Table 1: The number of prioritized genes for all disorders. extTADA only uses de novo (DN) and rare case/control (CC) information while gTADA uses DN, rare CC and enriched gene-set information. nGS and nsGS are the number of tested GSs and the number of enriched/significant GSs (lower CIs > 0) respectively. nsGS is the number of tested GSs whose lower CIs are > 0 and adjusted p values are < 0.05. For each column with PP > a threshold, the number in each cell is the number of prioritized genes.*

Insight of the rare variant genetic architecture of EPI

We focus on EPI because this disorder had full DN+CC data, including recent rCC studies [12, 13]: three types of EPI rCC data: familial non-acquired focal epilepsy (familiar NAFE), familial genetic generalized epilepsy (familiar GGE), and sporadic non-acquired focal epilepsy (NAFE). In addition, multiple EPI genes were prioritized by gTADA. With a large data set of cases and controls, we used the gTADA results without GS (which is the same as extTADA) to better understand the genetic architecture of EPI. The proportion of risk genes was 4.9% (Table S13). This proportion was higher than the estimated proportion reported in [18]. However, Nguyen et al., 2017 only used DN data, therefore, this situation was the same as SCZ results in Nguyen et al., 2017: adding CC data into DN data might increase the number of risk genes. Based on this proportion, the mean DN RRs (estimated $\bar{\gamma}$) were >15 (16 and 18 for MiD and LoF mutations respectively). The mean RRs of the three CC samples were > 4 (Supplementary Information). Details of other analyses are in Section 1.2 of SI, and Table S14.

Validation of top prioritized genes for EPI

Because multiple genes were prioritized by gTADA, we sought to validate these results. Here, we focused on the results from 1903 GSs because higher numbers of significant genes were obtained from these GSs. In addition, top prioritized genes from GTEx were also inside the top prioritized genes from these 1903 GSs, and the majority of the top prioritized genes from the drug-target GSs were also inside these results (Figure S5). gTADA prioritized 40 genes with PP>0.95 from eGSs of 1903 GSs. Visually checking the top genes with PPs > 0.95, we saw that the majority of them were inside enriched GSs. Being inside enriched GSs helped increase the PPs of these genes (Figure 4). Table 2 describes details of eGSs of these 40 genes as well as the effect sizes of the eGSs.

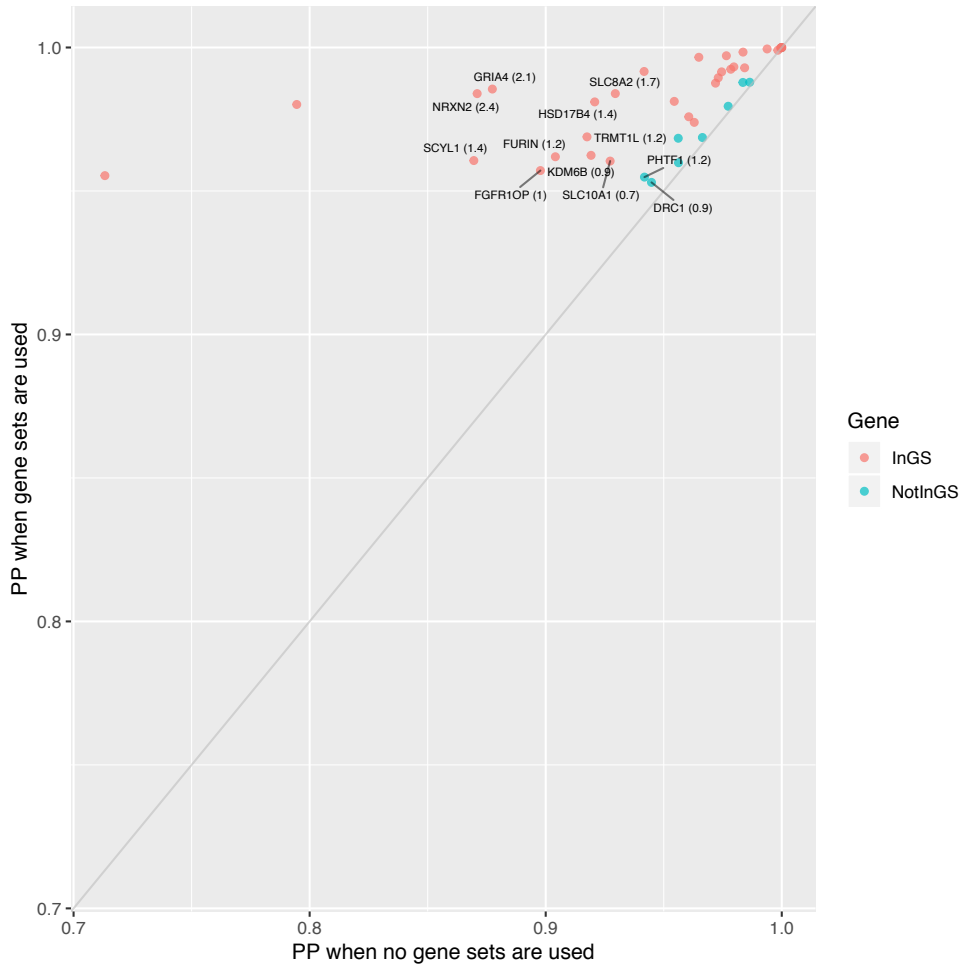


Figure 4: Comparing posterior probabilities (PPs) for top epilepsy (EPI) genes (PPs > 0.95). The x-axis shows the PPs when no gene sets are used while the y-axis shows the PPs when enriched GSs are used. Points with gene names describe novel risk genes whose PPs are less than 0.95 if no GSs are used. Genes in the red color are inside enriched GSs while genes in the blue color are not inside enriched GSs.

Gene	Gene Set	α_1	l_α_1	u_α_1	PP_gTADA	PP_extTADA
STX1B	PSD-95_(core)	2.25	1.29	3.11	0.96	0.71
GABRG2	GABA_PSC	2.60	1.32	3.93	0.98	0.79
NRXN2	abnormal_inhibitory_postsynaptic_currents	2.36	1.51	3.13	0.98	0.87
GRIA4	abnormal_spatial_reference_memory	2.08	1.11	3.15	0.99	0.88
SCYL1	abnormal_microglial_cell_morphology	1.45	0.25	2.47	0.96	0.87
HSD17B4	abnormal_microglial_cell_morphology	1.45	0.25	2.47	0.98	0.92
FGFR1OP	GO:0050730	1.01	0.17	1.69	0.96	0.90
FURIN	GO:1901215	1.21	0.04	2.01	0.96	0.90
SLC8A2	abnormal_synaptic_plasticity	1.73	0.23	2.67	0.98	0.93
TRMT1L	abnormal_response_to_new_environment	1.21	0.36	1.86	0.97	0.92

GNAO1	PSD-95_(core)	2.25	1.29	3.11	0.99	0.94
KDM6B	ID.allDenovoMiDandLoF	0.93	0.34	1.38	0.96	0.92
SLC10A1	GO:0006820	0.72	0.06	1.16	0.96	0.93
GABRA1	GABA_PSC	2.60	1.32	3.93	1.00	0.96
CSNK1E	Synaptic_vesicle	1.13	0.53	1.56	0.98	0.95
GABBR2	abnormal_inhibitory_postsynaptic_currents	2.36	1.51	3.13	1.00	0.98
COPB1	GO:0030135	0.89	0.07	1.54	0.99	0.97
EHD4	GO:0050730	1.01	0.17	1.69	0.99	0.97
PMPCA	REACTOME_METABOLISM_OF_PROTEINS	0.99	0.37	1.47	0.99	0.97
GPAM	GO:0031975	0.76	0.23	1.17	0.98	0.96
GABRB3	GABA_PSC	2.60	1.32	3.93	1.00	0.98
ATP8B1	ID.allDenovoMiDandLoF	0.93	0.34	1.38	0.99	0.98
TYRO3	GO:1901215	1.21	0.04	2.01	0.99	0.98
PHTF1	abnormal_astrocyte_morphology	1.18	0.01	1.93	0.95	0.94
C5orf42	abnormal_learning|memory|conditioning	0.96	0.49	1.45	0.97	0.96
GPR87	GO:0060089	0.44	0.05	0.82	0.97	0.96
NFATC3	CAGCAGG,MIR-370	0.89	0.09	1.67	0.99	0.98
DRC1	GO:0030135	0.89	0.07	1.54	0.95	0.94
CACNA1B	abnormal_spatial_reference_memory	2.08	1.11	3.15	1.00	0.99
KEAP1	abnormal_astrocyte_morphology	1.18	0.01	1.93	0.99	0.98
CEP89	abnormal_glutamate-mediated_receptor_currents	2.10	1.02	3.00	0.96	0.96
TBCK	abnormal_astrocyte_morphology	1.18	0.01	1.93	0.97	0.97
SAMD9L	GO:0006820	0.72	0.06	1.16	0.98	0.98
GIGYF1	abnormal_astrocyte_morphology	1.18	0.01	1.93	0.99	0.99
SLC9A2	GO:0015077	0.74	0.17	1.29	1.00	1.00
KCNQ2	abnormal_behavioral_response_to_xenobiotic	1.16	0.65	1.66	1.00	1.00
LGI1	abnormal_glutamate-mediated_receptor_currents	2.10	1.02	3.00	1.00	1.00
DEPDC5	REACTOME_NEUROTRANSMITTER_RECEPT BINDING_AND_DOWNSTREAM_TRANSMISSION_IN_THE_POSTSYNAPTIC_CELL	1.41	0.56	2.06	1.00	1.00
STXBP1	PSD-95_(core)	2.25	1.29	3.11	1.00	1.00
SCN1A	GO:0030424	1.11	0.51	1.68	1.00	1.00

Table 2: The information of the top identified EPI genes. These are 40 genes which had posterior probabilities (PPs) > 0.95. PP_extTADA is PPs of extTADA while PP_gTADA is the largest PP of PPs from gTADA's analyses. The second column describes gene sets' names which have largest PPs. The third, fourth and fifth columns are α_1's estimated values, lower and upper boundaries respectively.

We preformed simulation to assess observed false discovery rates (oFDRs) of top EPI genes (See Methods). We tested the oFDRs of genes when top prioritized genes from multiple gene sets were pooled together as in our current work. The genetic parameters of EPI and 1901 gene sets which were estimated above were used in the simulation process. oFDRs increased when GS numbers increased (Figure S6). For PP > 0.95, FDRs were always less than 0.1. Similarly, FDRs were less than 0.3 with PP > 0.8.

gTADA prioritized multiple novel significant EPI genes, and replicated previous results from the same data set

In the 40 most significant EPI genes (PP > 0.95, Table 2, S6) from gTADA, 10 genes were in the list of known EPI genes, and 30 genes are novel (ATP8B1, C5orf42, CACNA1B, CEP89, COPB1, CSNK1E, DRC1, EHD4, FGFR1OP, FURIN, GABBR2, GIGYF1, GPAM, GPR87, GRIA4, HSD17B4, KDM6B, KEAP1, NFATC3, NRXN2, PHTF1, PMPCA, SAMD9L, SCYL1, SLC10A1, SLC8A2, SLC9A2, TBCK, TRMT1L, TYRO3). One gene (GABBR2) in the 30 genes was in our recent result (PP > 0.95) in which only DN data were used [18].

In the 30 novel genes, pLI information was available for 29 genes. 12/29 genes (CACNA1B, COPB1, CSNK1E, FURIN, GABBR2, GIGYF1, GRIA4, KDM6B, NFATC3, NRXN2, TRMT1L, TYRO3) were highly intolerant genes (pLI > 0.9). Interestingly, 13/30 genes (ATP8B1, C5orf42, CEP89, DRC1, EHD4, GPAM, HSD17B4, PHTF1, PMPCA, SAMD9L, SCYL1, SLC10A1, TBCK) had pLI < 0.1. We investigated these genes and saw that the significant signal of these 11 genes was from CC data.

The current results replicated the results of the [13]. The [13] used the same CC data set as our current study and reported 6 significant autosomal genes (DEPDC5, GABRG2, GRIN2A, KCNQ2, LGI1, PCDH19, SCN1A). All of these six genes were in the list of genes having PP > 0.9. Interestingly, except for GRIN2A, 5/6 genes were in the list of genes having PP > 0.95.

Top EPI genes are also present in independent whole genome sequence data

Recently, [32] sequenced the whole genomes of 197 trios with developmental and epileptic encephalopathy (DEE). We used this data set as an independent data set. First, we tested for DN mutations. From the 40 genes identified by gTADA using candidate GSs, there were 6 genes (CSNK1E, GABBR2, GABRG2, GNAO1, KCNQ2, SCN1A) that had *de novo* mutations from the 197 trios (p value for this overlap < 5.9e-5). Interestingly, SCNA1 had 6 DN mutations and GNAO1 had 2 DN mutations. In the 30 novel genes, two genes GABBR2 and CSNK1E had one nonsynonymous *de novo* variant each. The gene GABRB2 was reported as a significant risk gene for DEE by [32]. Next, the gTADA results were tested for copy-number variant (CNV) data. From the 40 genes, GABRB3 was in a *de novo* duplication with 6 copies.

Significant EPI genes were connected to communities by PPI analysis

We analyzed the connection of the top EPI genes using GeNets [33]. First, to increase the number of significant gene counts for the analysis, we used a threshold PP > 0.8. There were 135 genes from this threshold. GeNets also found other candidate genes to make a total of 192 genes. 100/192 genes were well connected to five communities (all overall connectivity and community connectivity p values < 2e-3, Figure 5). These communities showed enrichment for multiple pathways: ion channel transport, neurotransmitter receptor binding, GABA receptor activation, ligand gated ion channel transport (Community 2, Figure 5); JAK-STAT signaling pathway, regulation of IFNA signaling, RIG-I-like receptor signaling pathway, interferon alpha/beta signaling, and autoimmune thyroid disease (Community 4, Figure 5, Table S15).

Community 4 were mainly enriched for immune genes; therefore, we tested all communities used single cell RNA sequencing (scRNAseq) data (Figure 5, Figure S7). As expected, microgloria cells were strongly enriched in Community 4 (p = 2.34e-5) but not in other communities. Endothelial mural, vascular leptomeningeal, astrocytes_ependymal cells were also enriched in this community but were not as strong as microgloria cells. We also saw that pyramidal CA1, SS cells were enriched in other communities. This was similar to results reported recently by our group for four NDDs [25]. We suspected that these results could be affected by candidate genes from the PPI network analysis. Therefore, we also tested the enrichment of scRNAseq data for these communities by using only genes from gTADA. Similar to the full results with candidate genes, microglia cell types were also enriched for the community 4 (p = 1.5e-3) followed by endothelial mural cell types (p = 0.02). Finally, we also saw that two drug-class gene sets: ANTIINFLAMMATORY_AGENTS, IMMUNOSTIMULANTS were enriched in EPI data. This showed convergent results between the PPI network analysis and enriched drug-target gene sets.

We also used the STRING database [34] to obtain the information of physical interactions of the communities of GeNets. Using only sources from experiments, the interactions between of the 135 genes were significant (13 observed edges versus 7 expected edges, p = 0.0248).

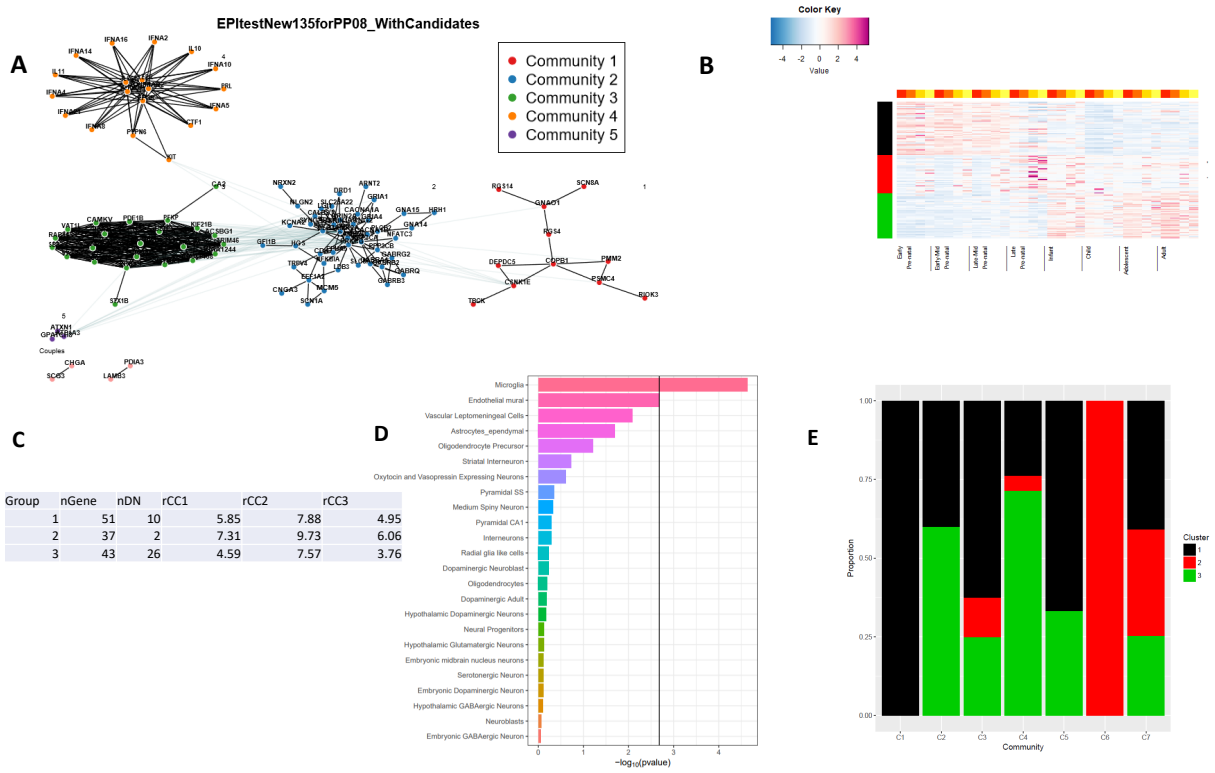


Figure 5: Results of the top prioritized EPI genes. A: PPI analysis for top prioritized EPI genes; B: spatiotemporal gene expressions across prioritized EPI genes; C: the number of de novo mutations (nDN), three ratios of case/control data (rCC) in clusters of the data in B; D: enrichment results of different mouse cell types using single-cell RNA data for Community 4; E: the proportion of genes in clusters from (B) in communities from (A). These are genes prioritized from gTADA using a threshold of PP > 0.8.

EPI prioritized genes showed differences in spatiotemporal expression

We used the data of spatiotemporal gene expressions to test the prioritized genes. The EPI prioritized genes showed expression during all developmental stages of the human brain (Figure 5). These results were much different from DD and ID, CHD genes which were strongly expressed in pre-natal stages (Figure S9).

We also tested clusters from this expressed data set in depth. We found that the data have some clear clusters for prenatal, late prenatal and postnatal genes; therefore, we counted all the DN mutations and calculated CC ratios for all the clusters to see whether the DN or CC signal was specific for these clusters. Surprisingly, 26 *de novo* mutations were observed in 43 postnatal genes (Figure 5). Next, we tested the results using gene expression and the results using PPI network. gTADA genes in the immune community (Community 4) from PPI-network analysis were in late prenatal genes (Figure 5).

Convergence of top genomic based genes and drug-target genes in EPI

The number of significant gene sets from drug-target data were very high for EPI, therefore, we also sought to understand these results. From drug-name-GS results, we chose genes with PPs>0.8 and tested the appearances of these genes in enriched drug-name GSs. There were multiple common genes across these seGSs, and they were mainly GABA receptor genes. Genes GABRA1, GABRA2, GABRA3, GABRA5 were in at least 45/67 seGSs (Figure 6).

We next tested whether these enriched drug-target GSs were specific for *de novo* or CC data. We ran gTADA for only *de novo* data, and only CC data. These GSs showed more enrichment for DN data than for CC data. Regarding the drug-name GSs, there were 143 and 4 GSs with lower CIs > 0 and adjusted p values < 0.05 for DN and CC data respectively. Two GSs (isoxsuprine and nitrazepam) were significantly observed in both DN and CC data. One gene (SCN1A) was present in these two GSs. For drug-class GSs, seGS numbers were 44 and 4 for DN and CC data. Three drug classes (anxiolytics, drugs for functional gastrointestinal disorders, other dermatological preparations) were enriched in both DN and CC data.

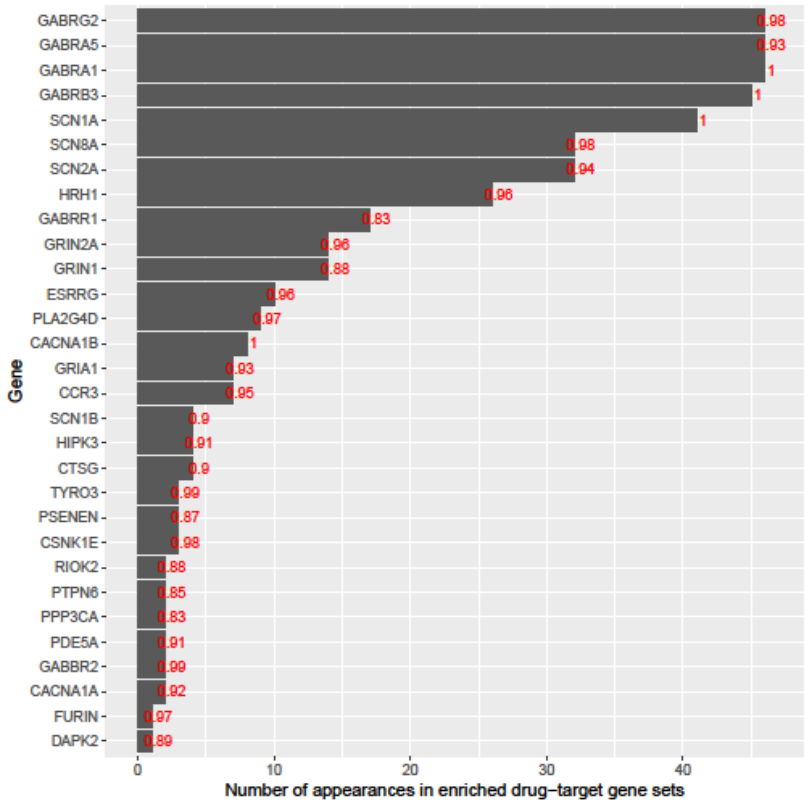


Figure 6: The number of appearance times of genes in enriched drug-target gene sets. For each gene, the number in red is the highest posterior probability of the gene from gene-set results.

Discussion

We have presented a pipeline (gTADA) that incorporates *de novo* mutations (DNMs), rare inherited/case-control (CC) variants and pathway/gene-set/expression information to prioritize disease genes. This pipeline is based on our previous work, extTADA [18], but gTADA is a generalized framework of extTADA. gTADA can be extTADA if the gene-set information is not used. Recently, methods which use annotation/expression information to impute more risk genes have been actively developed for common variants [35-37]. These methods have been successfully used to prioritize risk genes, and elucidate biological pathways for schizophrenia, bipolar [25, 38] and breast cancer [39]. gTADA might be the first tool using this approach for rare variants. There are many benefits of this approach [35]. First, it can identify significant genes which might be missed by using typical genetic-data based methods. Second, significant genes can assist in understanding the structure of enriched gene sets. Another advantage of gTADA is that the package can test gene set enrichment directly from data. This enrichment test has been shown more powerful than traditional ways in the analyses of ChiP-Seq data sets [40]. We hope that gTADA will be helpful in rare-variant based studies. The code is available online on https://github.com/hoangtn/gTADA.

We have used gTADA to identify enriched tissue/gene sets (GSs) (from candidate GSs, drug-target GSs and GTEx tissues); and to prioritize genes for neurodevelopmental disorders (NDDs) and congenital heart disease (CHD). We have seen that six human brain-region tissues and multiple candidate GS are enriched across NDDs and CHDs (Table 1). Interestingly, regarding drug-target GSs, multiple GSs are enriched in EPI, but just a few are enriched in DD and ID, and no enriched GS is for ASD and CHD. Based on enriched GSs, multiple significant genes are identified for all these disorders. gTADA identifies more significant genes than extTADA which only uses DN and CC data (Table 1).

We analyzed the results of EPI in depth because new rare CC data sets have been recently studied and multiple enriched drug-target GSs are identified by gTADA (Table 1). There are 40 genes with posterior probabilities PPs > 0.95. This number is much higher than the only DN and CC based approach. 30/40 genes are not in the list of EPI genes, and 2/30 genes have *de novo* events in a new trio data set. We also sought to learn the DN and CC genetic architecture of EPI. The number of predicted risk genes of these EPI combined data sets (~ 950) is higher than that of the DN-only based genetic architecture [18]. Interestingly, gTADA shows that mean relative risks (RRs) are nearly equal for three CC population samples in this study (Supplementary Information, Table S14), and top gTADA genes have higher differences between cases and controls than other well-known gene sets including known EPI genes, and FMRP gene sets (Table S14). This gives more information about EPI, especially for the sporadic non-acquired focal epilepsy in which its top prioritized genes converge to top prioritized genes of other types of EPI. Using gTADA for drug-target gene sets, we see multiple enriched gene sets (eGSs) while there are few eGSs for other NDDs and not for CHD. Top drug-target eGSs are still observed for *de novo* data and CC data if they are analyzed separately. This enrichment for EPI is because of some main genes: GABRA1, GABRA2, GABRA3, GABRA5, SCN1A, SCN8A; especially four genes GABRA1, GABRB3, GABRG2, SCN1A (Figure 6). These genes have been discussed by other studies in developing drug targets specific for EPI as well as neurodevelopmental disorders [41-43]. Further studies which focus on deeply understanding genetic variants in these genes would help better design drug targets for EPI. Finally, the top EPI genes (PP > 0.8) are well connected to communities by PPI network analysis, and these genes express in different development stages of the human brain. Interestingly, one community from the PPI network analysis is enriched in immune pathways. In addition, based on scRNAseq, we have seen that microglia cells were strongly enriched in this community but not enriched in other communities.

While this study uses a novel approach to integrate different types of genomic data, it does have some limitations. First, gTADA is partly an imputation tool, therefore some of its prioritized genes rely on reference data sets (e.g., gene sets, tissues, Figure 3, Table 1). This has been observed in tools developed for common variants [36, 37]. However, gTADA mainly uses DN and CC data inside the model; therefore, the top prioritized genes are usually from DN and CC data, not from reference data sets. When tested on simulation, as expected, when large GSs or multiple GSs are used, false discovery genes increase. One obvious reason is that if a GS size is larger than the number of risk genes, imputed genes outside the range of risk genes would be called false positive genes. However, for real data, large enriched gene sets might help in identifying more novel risk genes (Figure 3). For the current multiple-GS model, to control for FDR, users can increase a PP threshold when the number of gene sets increase (Figure S1, S2). For the single-GS model, false positive results can also happen if gTADA is applied to separate GSs and then top genes are obtained from each eGS. In this situation, we suggest that a stringent threshold should be used to

obtain more reliable results. For example, we used a threshold PP>0.95 to obtain top EPI genes. Based on simulation data from EPI genetic parameters, the prioritized genes should have FDRs < 0.1 with this PP threshold (Figure S6). Using the same PP threshold, we also saw that FDRs increased quickly when few gene sets were added; however, FDRs slightly increased when more gene sets were added (Figure S6). This might be because the enriched GSs overlap with each other. As a result, more significant genes are not identified when the number of GSs increases. Therefore, FDRs do not change much after adding a number of GSs. As earlier discussed by [18] and also in the simulation data of the current study (Figure S1), the weaknesses could be improved in future studies with larger sample sizes and more comprehensive variant categories. gTADA as well as its previous pipelines [9, 18] model variant-count data using statistical distributions (e.g, Poission distribution for rare variants); therefore, count data should follow these distributions to obtain optimal results. gTADA uses posterior probabilities (PPs) to prioritize genes, and these values should rely on the analyzed data sets. One possible solution is that users should set a high threshold of PPs to obtain more reliable significant genes. Finally, top prioritized EPI genes here are based on meta-analyzing multiple types/population samples of EPI; therefore, these results might not be totally similar to results from separate analyses.

Methods and data

Data

Gene-set data

We used 1903 gene sets curated by [25]. These included 186 known gene sets with prior evidence of involvement in ASD and SCZ, and 1717 gene sets (whose lengths were between 100 and 4995 genes) from different databases: the Gene Ontology database [44], KEGG, and REACTOME, and the C3 motif gene sets from the Molecular Signatures Database (MSigDB) [45]. The information of these gene sets was presented in detail in the Table S2 of Nguyen et al. 2017.

Drug-target gene sets were processed and classified as [46]. Briefly, drugs were classified according to the level of the Anatomical Therapeutic Chemical (AUC) classification system. The ATC system divides drugs into 5 levels from anatomical group (level 1) to chemical substance (level 5). Drug targets which were classified as level 3 (therapeutic subgroup) and level 5 (specific drug) were used in this study. We used 156 GSs from level 3, and 710 gene sets from level 5 whose lengths were ≥ 5 genes from the curated GSs of [29].

To compare the current results with previous results, known EPI genes were downloaded from two sources. The first was 76 genes from https://www.cureepilepsy.org/egi/genes.asp [11], and the second was 218 genes from https://www.omim.org/phenotypicSeries/PS308350 of the Online Mendelian Inheritance in Man, OMIM [47].

Transcriptomic data

Gene expression specific for tissues were downloaded from the GTEx project (V6p) [28] . Letting x_{ij} be the expression value of the ith gene at the jth tissue, we used $log2(x_{ij} + 1)$ in our analyses. Spatiotemporal transcriptomic data were downloaded from BRAINSPAN [48]. Using the approach of [49], the data were divided into eight developmental time points (four pre-natal and four post-natal).

Single-cell RNA sequencing (scRNAseq) data were obtained from [50]. Briefly, this data set included 9970 mouse cells. These cells were clustered into 24 Level 1 brain cell types and 149 Level 2 cell types [50]. 24 Level 1 cell types were used in this study.

Variant data

We used DN and rare CC data of NDDs from our previous publication [18], a recent EPI study [13] and CHD data from the denovo-db database [51]. The data of [18] were collected from multiple publications and were described in detail in Table S1 of [18]. In summary, the DN data included 5122, 4293, 1012 and 356 trios for ASD, DD, ID and EPI respectively), 404 cases for ASD, 3654 controls ASD respectively. We also used CHD data of 1213 trios from [52]. Variants were annotated and divided into different categories. There were categories which included loss of function (LoF) variants/mutations, missense damaging (MiD) variants/mutations. The data from [13] consisted of 5696 samples: 640 cases of familial genetic generalized epilepsy, 522 cases of familial non-acquired focal epilepsy, 662 cases of sporadic non-acquired focal epilepsy and 3877 controls. We used the ultra-rare variant counts of all genes from Table S10, S11, S12 of [13]. These variants had minor allele frequencies <= 0.05% and MAF = 0% in ExAC (http://exac.broadinstitute.org/about) and in EVS (http://evs.gs.washington.edu/EVS/) . They were annotated by SnpEff [53] as loss-of-function, inframe indels, or missense "probably damaging" predicted by PolyPhen-2 (HumDiv).

In addition, we also used an independent EPI data set of 197 trios [32] to validate our results. This is whole-genome-sequencing (WGS) data of individuals with EPI and DD and their parents.

Based on these data sets, DD, ID and CHD had only two DN categories (LoF and MiD); ASD had two DN categories (LoF, MiD) and one LoF+MiD CC population sample; EPI two DN categories (LoF, MiD), and three CC population samples.

Simulated data

To evaluate the new method, ASD genetic parameters were used to simulate DN and CC data. Simulation parameters were from previous ASD studies [10, 18] as described in Table S2. We first simulated exact parameters of ASD to compare gene counts between gTADA and extTADA and test type I errors of gTADA in the identification of eGSs. After that, we simulated different sample sizes to have a better understanding of gTADA. There were three sample sizes: case, control and family numbers. Therefore, to reduce the complexity of the simulation process, only family numbers were changed.

Method

The gTADA pipeline

gTADA was designed with two main aims. The first aim is to test the enrichment of a gene set directly from DN+CC data. The second is to use enriched gene sets as prior information to improve the identification of novel significant genes associated with the tested trait—this is considered a key feature of the pipeline.

The main pipeline of gTADA is described in Figure 1 and is presented in the Results section. In summary, gTADA combined *de novo* mutations, rare inherited/case-control variants and

pathway/gene-set (GS) information to jointly estimate genetic and enrichment parameters. GS information could be from gene sets or from expression data. For variant data of each gene, we used the statistical models of extTADA as described in Table S1. For GS data, there were two situations. If it was a gene set, we coded a gene as 1 or 0 corresponding with the presence or absence in all tested genes. If it was gene expression data, we used log2(1 + expression values). To incorporate GS information, we improved the main approach of extTADA. We assumed that for each ith gene, there was a probability π_i for the gene to be a risk gene. This was connected to a GS by $\pi_i = \frac{1}{1+e^{f_i(\alpha)}}$ with $f_i(\alpha) = \alpha_0 + GS * \alpha_1$ or to multiple GSs by $f(\alpha) = \alpha_0 + \sum_{j=1}^{K} \alpha_j GS_{ij}$. The likelihood (LK) function for all parameters:

$$LK = \prod_{i=1}^{nGene} [P(Data_i|H_1)\pi_i + P(Data_i|H_0)(1 - \pi_i)]$$

$$and\ \pi_i = \frac{e^{f_i(\alpha)}}{1 + e^{f_i(\alpha)}},\ \ f_i(\alpha) = GS * \alpha_1 + \alpha_0\ \ (I)\ \ or\ \ f_i(\alpha) = \sum_{j=1}^{K} GS_{ij} * \alpha_j + \alpha_0\ \ (II)$$

Based on the result of the equation above, GS was considered an enriched GS if the low boundary of its credible interval (CI) was positive. We did not adjust gene lengths inside the GS model because the statistical models of *de novo* data adjusted mutation rates (Table S1) and mutation rates were positively correlated with gene lengths.

From enriched gene sets, we chose a group of optimal gene sets that improved the model fit. We started with the model without any gene set (only α_0). Then, we looped over all gene sets, and a gene set was added into the model if it improved the value of the likelihood function by a given threshold and the 95% CI was positive. To reduce a computational burden, we used a reduced forward-selection strategy. All enriched GSs were sorted ascendingly according to their corresponding α values, and GSs were added into the combined model based on this order.

The final optimal gene sets were used in the identification process of risk genes. Their α values and genetic parameters were re-estimated to use for the calculation of posterior probabilities (PPs).

Generation of simulated data

To evaluate gTADA, we simulated the data as follows:

1. Simulate data without GSs (as extTADA)
   - Input α_0 to calculate $\pi_i = \frac{e^{\alpha_0}}{1+e^{\alpha_0}}$ for the ith gene.
   - Sample the characteristics of a gene (risk or not-risk genes) $z_i \sim Binomial(2, \pi_i)$:
     - $z_i = 1\ (risk\ gene): \gamma_i \sim Gamma(\overline{\gamma} * \beta, \beta)$
     - $z_i = 0\ (not - risk\ gene): \gamma_i = 1.$
   - Sample CC and DN counts for each gene from statistical models in Table S1.
2. Simulate gene sets

We simulated different GS sizes. To simulate not eGSs, random genes were chosen from the whole genes. To simulate enriched GSs, we used prior information from [18] as follows. Overlaps between eGSs and top significant genes from DN and CC data are not random. Therefore, to make the distribution of genes in gene sets more realistic in the simulation process, we used results from 186 candidate gene sets of our previous study for ASD. Briefly, eGSs from the 186 gene sets were chosen. We used extTADA [18] to obtain posterior probabilities (PPs) for genes from the simulation data, and then ranked the genes according to their PPs. After that, for each gene set, we made a table of overlapping-gene numbers between the gene set and genes in different groups (e.g., top 50 genes, 51st to 100th genes, ..). In the simulation process, we allocated genes into different groups using this table. The allocation was also based on the gene size of each simulated gene set.

Estimation of genetic and gene-set parameters

We used Markov Chain Monte Carlo (MCMC) methods implemented in the rstan package [54] to jointly estimate all genetic and gene-set parameters. The convergence of each parameter from MCMC results was diagnosed by the estimated potential scale reduction statistic ($\hat{R}$) inside the rstan package. The Locfit [55] was used to obtain credible intervals (CIs), modes of parameters.

To obtain eGSs for prioritizing risk genes, we only used GSs whose low boundaries of CIs were positive. To obtain comparable results with other studies, we used posterior sampling results. A one-tail p value for each GS was calculated as the probability of GS's alpha less than 0 if alpha's posterior mode was positive and larger than 0 if alpha's posterior mode was negative. All p values were adjusted by using the method of [56].

Validation of significant genes

GeNets [33] was used to test protein interactions from the identified genes. Connectivity p values were obtained by using default parameters from the GeNets server (http://apps.broadinstitute.org/genets#computations, January 26, 2018). We also used STRING database to further obtain the information of protein interactions of genes. To test the enrichment of scRNAseq data, we used the same method described in [25]. The information of the probabilities of LoF tolerance was downloaded from ftp://ftp.broadinstitute.org/pub/ExAC_release/release0.3.1/functional_gene_constraint/ [57].

Simulation of data to test the false discovery rates of top prioritized EPI genes

To check the observed false discovery rates (FDRs) of the top prioritized EPI genes, we simulated data similarly to the general simulation framework above. All genetic parameters which were estimated by gTADA for one trio population sample and three case/control population samples were used (Table S13). We used all 98 enriched GSs from the 1903 GSs.

Author's contributions

Conceived and designed the experiments: HTN, EAS. Designed the pipeline used in analysis, performed the experiments, analyzed the data and drafted the manuscript: HTN. Analyzed single-cell data: JB. Contributed reagents/materials/analysis tools: HTN, AD, AC, JB, NGS,

LMH, WW, DMR, XX, MF, SMP, MV, ABS, JH, JDB, DP, XH, PFS, EAS. Wrote the paper: HTN, AD, AC, XH, EAS.

Acknowledgements
This work is supported by NIH grant R01MH105554 to E.A.S, and by NIH grant R01MH110555 to D.P. 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.

1. Sanders, S.J., et al., *De novo mutations revealed by whole-exome sequencing are strongly associated with autism.* Nature, 2012. **485**(7397): p. 237-41.
2. Iossifov, I., et al., *The contribution of de novo coding mutations to autism spectrum disorder.* Nature, 2014. **515**(7526): p. 216-21.
3. Myers, C.T., et al., *De Novo Mutations in PPP3CA Cause Severe Neurodevelopmental Disease with Seizures.* Am J Hum Genet, 2017. **101**(4): p. 516-524.
4. Euro Epinomics- RES Consortium, Epilepsy Phenome/Genome Project, and E.K. Consortium, *De novo mutations in synaptic transmission genes including DNM1 cause epileptic encephalopathies.* Am J Hum Genet, 2014. **95**(4): p. 360-70.
5. Epi K. Consortium, et al., *De novo mutations in epileptic encephalopathies.* Nature, 2013. **501**(7466): p. 217-21.
6. Lelieveld, S.H., et al., *Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability.* Nat Neurosci, 2016. **19**(9): p. 1194-6.
7. Deciphering Developmental Disorders Study, *Prevalence and architecture of de novo mutations in developmental disorders.* Nature, 2017. **542**(7642): p. 433-438.
8. Deciphering Developmental Disorders Study, *Large-scale discovery of novel genetic causes of developmental disorders.* Nature, 2015. **519**(7542): p. 223-8.
9. He, X., et al., *Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes.* PLoS Genet, 2013. **9**(8): p. e1003671.
10. De Rubeis, S., et al., *Synaptic, transcriptional and chromatin genes disrupted in autism.* Nature, 2014. **515**(7526): p. 209-15.
11. Epi P. M. Consortium, *A roadmap for precision medicine in the epilepsies.* Lancet Neurol, 2015. **14**(12): p. 1219-28.
12. Zhu, X., et al., *A case-control collapsing analysis identifies epilepsy genes implicated in trio sequencing studies focused on de novo mutations.* PLoS Genet, 2017. **13**(11): p. e1007104.
13. Epi K. consortium and Epilepsy Phenome/Genome Project, *Ultra-rare genetic variation in common epilepsies: a case-control sequencing study.* Lancet Neurol, 2017. **16**(2): p. 135-143.
14. Speed, D., et al., *Describing the genetic architecture of epilepsy through heritability analysis.* Brain, 2014. **137**(10): p. 2680-2689.
15. Nabbout, R. and I.E. Scheffer, *Genetics of idiopathic epilepsies*, in *Handbook of clinical neurology*. 2013, Elsevier. p. 567-578.

16. Miller, L.L., et al., *Univariate genetic analyses of epilepsy and seizures in a population-based twin study: The Virginia twin registry.* Genetic epidemiology, 1998. **15**(1): p. 33-49.
17. Löscher, W., et al., *New avenues for anti-epileptic drug discovery and development.* Nature reviews drug discovery, 2013. **12**(10): p. 757.
18. Nguyen, H.T., et al., *Bayesian Integrated analysis of multiple types of rare variants to infer risk genes for schizophrenia and other neurodevelopmental disorders.* bioRxiv, 2017: p. 135293.
19. Purcell, S.M., et al., *A polygenic burden of rare disruptive mutations in schizophrenia.* Nature, 2014. **506**(7487): p. 185-90.
20. Fromer, M., et al., *De novo mutations in schizophrenia implicate synaptic networks.* Nature, 2014. **506**(7487): p. 179-84.
21. Genovese, G., et al., *Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia.* Nat Neurosci, 2016. **19**(11): p. 1433-1441.
22. Singh, T., et al., *The contribution of rare variants to risk of schizophrenia in individuals with and without intellectual disability.* Nat Genet, 2017. **49**(8): p. 1167-1173.
23. Jansen, A., et al., *Gene-set analysis shows association between FMRP targets and autism spectrum disorder.* Eur J Hum Genet, 2017. **25**(7): p. 863-868.
24. Carbonetto, P. and M. Stephens, *Integrated enrichment analysis of variants and pathways in genome-wide association studies indicates central role for IL-2 signaling genes in type 1 diabetes, and cytokine signaling genes in Crohn's disease.* PLoS Genet, 2013. **9**(10): p. e1003770.
25. Nguyen, H.T., et al., *Integrated Bayesian analysis of rare exonic variants to identify risk genes for schizophrenia and neurodevelopmental disorders.* Genome Med, 2017. **9**(1): p. 114.
26. Liberzon, A., et al., *Molecular signatures database (MSigDB) 3.0.* Bioinformatics, 2011. **27**(12): p. 1739-1740.
27. Gene Ontology Consortium, *Gene ontology consortium: going forward.* Nucleic acids research, 2014. **43**(D1): p. D1049-D1056.
28. GTEx Consortium, *Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans.* Science, 2015. **348**(6235): p. 648-60.
29. Ruderfer, D.M., et al., *Polygenic overlap between schizophrenia risk and antipsychotic response: a genomic medicine approach.* The Lancet Psychiatry, 2016. **3**(4): p. 350-357.
30. Keiser, M.J., et al., *Predicting new molecular targets for known drugs.* Nature, 2009. **462**(7270): p. 175.
31. Law, V., et al., *DrugBank 4.0: shedding new light on drug metabolism.* Nucleic acids research, 2013. **42**(D1): p. D1091-D1097.
32. Hamdan, F.F., et al., *High Rate of Recurrent De Novo Mutations in Developmental and Epileptic Encephalopathies.* Am J Hum Genet, 2017. **101**(5): p. 664-685.
33. Hu, Y., et al., *Joint modeling of genetically correlated diseases and functional annotations increases accuracy of polygenic risk prediction.* PLoS genetics, 2017. **13**(6): p. e1006836.

34. Szklarczyk, D., et al., *The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible.* Nucleic Acids Res, 2017. **45**(D1): p. D362-D368.
35. Zhu, X. and M. Stephens, *A large-scale genome-wide enrichment analysis identifies new trait-associated genes, pathways and tissues across 31 human phenotypes.* bioRxiv, 2017: p. 160770.
36. Gamazon, E.R., et al., *A gene-based association method for mapping traits using reference transcriptome data.* Nat Genet, 2015. **47**(9): p. 1091-8.
37. Mancuso, N., et al., *Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits.* Am J Hum Genet, 2017. **100**(3): p. 473-487.
38. Huckins, L.M., et al., *Gene expression imputation across multiple brain regions reveals schizophrenia risk throughout development.* bioRxiv, 2017.
39. Hoffman, J.D., et al., *Cis-eQTL-based trans-ethnic meta-analysis reveals novel genes associated with breast cancer risk.* PLoS Genet, 2017. **13**(3): p. e1006690.
40. Welch, R.P., et al., *ChIP-Enrich: gene set enrichment testing for ChIP-seq data.* Nucleic Acids Res, 2014. **42**(13): p. e105.
41. Kang, J.-Q. and R.L. Macdonald, *Making sense of nonsense GABA A receptor mutations associated with genetic epilepsies.* Trends in molecular medicine, 2009. **15**(9): p. 430-438.
42. Kang, J.-Q. and R.L. Macdonald, *Molecular pathogenic basis for GABRG2 mutations associated with a spectrum of epilepsy syndromes, from generalized absence epilepsy to dravet syndrome.* JAMA neurology, 2016. **73**(8): p. 1009-1016.
43. Braat, S. and R.F. Kooy, *The GABAA receptor as a therapeutic target for neurodevelopmental disorders.* Neuron, 2015. **86**(5): p. 1119-1130.
44. Euro Epinomics- R. E. S. Consortium, Epilepsy Phenome/Genome Project, and E.K. Consortium, *De novo mutations in synaptic transmission genes including DNM1 cause epileptic encephalopathies.* Am J Hum Genet, 2014. **95**(4): p. 360-70.
45. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proceedings of the National Academy of Sciences, 2005. **102**(43): p. 15545-15550.
46. Ruderfer, D.M., et al., *Polygenic overlap between schizophrenia risk and antipsychotic response: a genomic medicine approach.* Lancet Psychiatry, 2016. **3**(4): p. 350-7.
47. Amberger, J.S., et al., *OMIM. org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders.* Nucleic acids research, 2014. **43**(D1): p. D789-D798.
48. Miller, J.A., et al., *Transcriptional landscape of the prenatal human brain.* Nature, 2014. **508**(7495): p. 199-206.
49. 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, 2015. **85**(4): p. 742-54.
50. Skene, N.G., et al., *Genetic identification of brain cell types underlying schizophrenia.* bioRxiv, 2017: p. 145466.
51. Turner, T.N., et al., *denovo-db: a compendium of human de novo variants.* Nucleic Acids Res, 2017. **45**(D1): p. D804-D811.

52. Homsy, J., et al., *De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies.* Science, 2015. **350**(6265): p. 1262-1266.
53. Cingolani, P., et al., *A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3.* Fly (Austin), 2012. **6**(2): p. 80-92.
54. Carpenter, B., et al., *Stan: A probabilistic programming language.* Journal of Statistical Software, 2016. **20**: p. 1-37.
55. Loader, C., *Locfit: Local regression, likelihood and density estimation.* R package version, 2007. **1**.
56. Benjamini, Y. and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* Journal of the royal statistical society. Series B (Methodological), 1995: p. 289-300.
57. Karczewski, K.J., et al., *The ExAC browser: displaying reference data information from over 60 000 exomes.* Nucleic Acids Res, 2017. **45**(D1): p. D840-D845.